Fourier Analysis of the Preconditioned Helmholtz Equation for Scattering Problems

ERASMUS MUNDUS MASTER PROGRAM
COMPUTER SIMULATIONS IN SCIENCE AND ENGINEERING (COSSE)

Master Thesis By: Luis A. García Ramos
Student Number: 4126394

TU Delft Coordinator: Prof. Dr. Kees Vuik
TU Berlin Supervisor: Prof. Dr. Reinhard Nabben

June 2017

Contents

Acknowledgements

I would like to thank in the first place my supervisor Prof. Reinhard Nabben, for his advice and constant encouragement throughout all the stages of this work. I am also very grateful to Prof. Kees Vuik, for his mentoring during my year in Delft and for suggesting this thesis topic to me.

I feel privileged for having shared my two years of studies with the other students of the COSSE program. I am thankful to all of them for their friendship, very especially to Carlos Echeverría and Matthew The in Berlin, and Agus Darmawan, Arna skarsdottir, and Carolyn Langen in Delft.

My friends and family have always given me unconditional support, and I am infinitely thankful to them. I specially thank Hilda Botero for her guidance and many words of wisdom. To my mother Rosalba, who taught me the most valuable lessons in life and gave me all her love and care throughout the years, I am grateful beyond any words. This thesis is dedicated to her.

Bogotá, January 2013

Chapter 1

Introduction

Scattering theory is the study of the effect that a medium or object has on an incident wave. This phenomenon has attracted the attention of physicists and mathematicians, and their investigations have led to a range of techniques that are used in applications to seismics, acoustics, medical imaging and optics, among other disciplines (see [8], [11]).

In these applications it is often required to determine the features of an object or medium, such as its shape or material parameters. Such features are difficult to measure directly, but they can be reconstructed from the observation of waves scattered by the object or medium, if the incident wave is known. Problems of this form are called *inverse scattering problems*. An intermediate step to the solution of these problems are *direct scattering problems*, where the incident wave and the properties of the scatterer are given (or can be estimated) and one is interested in finding the scattered wave.

The fundamental equation describing scattering problems of acoustic waves is the *Helmholtz equation*:

$$-\Delta u(x) - k^2(x)u(x) = f(x), \tag{1.1}$$

where u and f are complex-valued functions and the wavenumber $k = \omega/c$ is a function of the spatial variable. In this thesis we will study the numerical solution of the linear system of equations arising from the discretization of Helmholtz' equation, motivated by its applications to scattering problems in seismics.

1.1 Seismic Imaging and Migration

In exploration geophysics it is important to have an accurate image of the inner layers of the earth. For this purpose, a technique known as *migration* is usually employed. Acoustic waves generated from a source placed on the surface of the earth are shot through the subsurface,

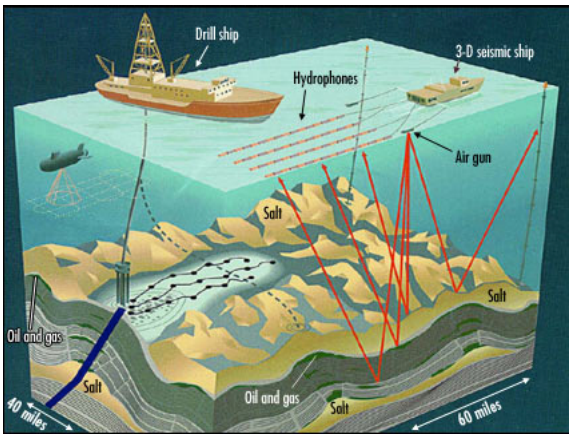


Figure 1.1: Obtaining seismic information for subsurface imaging. Hydrophones streaming from a 3-D seismic ship record the reflection of sound waves as they bounce back from the earth's subsurface. (Hutchins et al. (Eds), World Oil's 4-D Seismic Handbook, Gulf Publishing, 1997.)

and the difference between reflection times is recorded. This method allows to obtain information of the earth's layers up to 6000m below the surface. The data is processed afterwards using a technique known as *migration*. For an extensive presentation of the method, see the book by Bleistein, Cohen and Stockwell [4].

1.2 Scattering of Acoustic Waves

Let us consider acoustic waves propagating in $\mathbb{R}^3$ through an inhomogeneous medium, such as a fluid. If $\rho_0(x)$ and $c(x)$ are the density of the medium and the speed of sound respectively, depending only on space, and γ is a damping coefficient, this propagation is modelled by the *wave equation*

$$\frac{\partial^2 p(x,t)}{\partial t^2} + \gamma \frac{\partial p(x,t)}{\partial t} = c(x)\rho_0(x)\,\mathrm{div}\left[\frac{1}{\rho_0(x)}\nabla p(x,t)\right], \tag{1.2}$$

where $p(x,t)$ is the pressure disturbance on the medium (representing the wave) at point $x \in \mathbb{R}^3$ and time $t \in \mathbb{R}$. If the waves are time-periodic, i.e., have the form

$$p(x,t) = \mathrm{Re}\,[u(x)e^{-i\omega t}],$$

with frequency $\omega > 0$ and $u(x)$ a complex-valued function only depending on the space variable, equation (1.2) can be reduced to the *Helmholtz equation*

$$-\Delta u(x) - \frac{\omega^2}{c(x)^2}\left(1 + i\frac{\gamma}{\omega}\right)u(x) = 0. \tag{1.3}$$

In free space, $c = c_0$ is constant and there is no damping, so $\gamma = 0$. We define the *wavenumber* k and the *index of refraction* n as

$$k = \frac{\omega}{c_0} \quad \text{and} \quad n(x) = \frac{c_0^2}{c(x)^2}\left(1 + i\frac{\gamma}{\omega}\right),$$

equation (1.3) then takes the form

$$-\Delta u - k^2 n u = 0, \tag{1.4}$$

where n is a complex-valued function with $\mathrm{Re}\,[n(x)], \mathrm{Im}\,[n(x)] \geq 0$. Also, we assume that the inhomogeneous medium is bounded, that is, there is $a \in \mathbb{R}$ such that $n(x) = 1$ for all $x \in \mathbb{R}^3$ with $\|x\| \geq a$. From this assumptions, we have that

$$\Omega = \{x \in \mathbb{R}^3 : \|x\| < a\}$$

is our domain of interest. Further, we assume that there are no wave sources in Ω. We will treat here the cases of a point source and a plane wave. In the first case, we have an incident spherical wave p^I of the form

$$p^I(x,t) = \frac{1}{\|x-z\|}\,\mathrm{Re}\,[e^{ik\|x-z\|\omega t}], \text{ corresponding to } u^I(x) = \frac{e^{ik\|x-z\|}}{\|x-z\|}$$

where z is a point source outside Ω. In the second case, the incident plane wave is

$$p^I(x,t) = \mathrm{Re}\,[e^{ik\hat{\theta}x - i\omega t}], \text{ corresponding to } u^I(x) = e^{ik\hat{\theta}x},$$

where $\hat{\theta}$ is a unit vector representing the direction of the wave. Each type of wave is a solution to the Helmholtz equation

$$-\Delta u - k^2 u = 0$$

in $\overline{\Omega}$. The perturbation introduced by the medium is represented by the index of refraction n, and it produces a scattered wave u^S. The total field $u = u^I + u^S$ satisfies the Helmholtz equation (1.4) outside the sources. We require an additional condition to ensure a unique solution to the problem. As we expect that the scattered field behaves as a spherical wave propagating far away from the medium, we require that u^S satisfies the *Sommerfeld radiation condition*

$$\frac{\partial u^S(x)}{\partial r} - iku^S(x) = O(1/r^2) \text{ as } r = \|x\| \to \infty, \tag{1.5}$$

uniformly in all directions. We can now formulate the direct scattering problem.

Direct Scattering Problem for Acoustic Waves: Let the wavenumber $k > 0$, the index of refrac-

tion $n = n(x)$ with $n(x) = 1$ for $\|x\| > a$, and the incident field u^I be given. Determine the total field u that satisfies the Helmholtz equation (1.4) in $\mathbb{R}^3$ outside the source region, such that the scattered field $u^S = u - u^I$ satisfies the radiation condition (1.5).

In most applications, only information about the wave at far distances from the source can be measured. The behavior of the wave at far distances is known as the *far-field pattern*, that we now introduce. Let $u : \mathbb{R}^3 \to \mathbb{C}$ be a solution of the direct scattering problem for some incident field u^I, wavenumber k and refraction index n. Let

$$S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}.$$

The far-field pattern $u_\infty : S^2 \to \mathbb{C}$ is defined as

$$u_\infty(\hat{x}) = \frac{k^2}{4\pi} \int_{y \in \Omega} [n(y) - 1] e^{-ik\hat{x}\cdot y} u(y)\, dy.$$

The inverse scattering problem can now be (informally) stated.

Inverse Scattering Problem of Acoustic Waves: Determine the index of refraction n from knowledge of the far-field patterns u_∞ corresponding to various known incident fields u^I and different wavenumbers k.

Details on the conditions for the existence and uniqueness of solutions to both problems can be found in [24]. We remark that in order to solve the inverse problem it is required to solve several instances of the direct problem, often for a range of frequencies ω (consequently, for a range of wavenumbers k). In the rest of this thesis, we will focus on solving the direct problem numerically, keeping in mind that this is an intermediate step for many applications in which the solution of the inverse scattering problem is sought.

1.3 The Numerical Solution of Helmholtz' Equation

Helmholtz' equation (1.1) can be discretized using finite-differences leading to a linear system of the form

$$Au = f.$$

For a high degree of accuracy, it is required that the number of grid points grows at least quadratically in the wavenumber, leading to a very large system of equations. For 2-dimensional problems, direct methods such as the LU factorization can be used, but for 3-dimensional problems the size of the matrix is too large, and direct methods lead to excessive fill in due to the large bandwidth of the matrix [11]. Alternatively, the use of iterative methods for Helmholtz problems has been proposed.

The iterative solution of Helmholtz' equation has been a subject of research since the early 1980's. Because of the large size of the matrix A and its sparsity, methods based on *Krylov subspaces* are convenient. The article [2] by Bayliss, Goldstein, and Turkel is the first work in this direction. There, the conjugate-gradient-like method known as CGNR is proposed to solve the Helmholtz equation. Later on, iterative methods based on ADI (alternating-direction-implicit) schemes, domain decomposition, and multigrid, among others, were introduced. The reader is referred to the article [11] by Erlangga for a full survey. All these methods, however, have a limited range of application and no standard method for the Helmholtz equation has been developed. Ideally, the number of iterations of such a method would be independent on the wavenumber, or only mildly dependent.

The main difficulty in obtaining an efficient solver for the Helmholtz equation is the eigenvalue distribution of the matrix A, which is not favorable for Krylov subspace methods. As a remedy for this, various preconditioning methods have been investigated, since a well-designed preconditioner can convert the system into an equivalent one with better spectral properties. In this area, the use of a *shifted Laplace* operator as a preconditioner for the Helmholtz equation has attracted much interest, beginning with the work of Giles and Laird [25] and more recently with the *complex shifted Laplace* (CSL) preconditioner introduced by Erlangga, Osterlee and Vuik in [13].

In [14] Erlangga and Nabben propose a *multilevel Krylov* (MK) method to solve boundary value problems, and extend it further to incorporate multigrid preconditioners in a *Multilevel Krylov Multigrid* (MKMG) method that can be applied to Helmholtz' problems preconditioned by the CSL [15]. A variant of this method has been further studied theoretically and experimentally in [31] by Sheikh, Lahaye and Vuik. These methods are the starting point of our work.

1.4 Scope of the Thesis and Overview

In this thesis the analysis of the methods proposed in [15] and [31] is extended. We intend to compare the theoretical aspects of both methods, which have shown good numerical results, with a nearly constant number of iterations for a range of wavenumbers. For this, we perform rigorous Fourier analysis on a Helmholtz model problem, and study the distribution of the eigenvalues with respect to the various parameters. The analysis proposed here clarifies some aspects of the two-level method that have been observed before only through numerical experiments, and gives theoretical evidence for the suitability of the MKMG method to solve Helmholtz problems with large wavenumbers.

In chapter 2 we shortly review the subject of iterative and multigrid methods. In chapter 3 we formulate the 1-dimensional discrete Helmholtz problem that we will analyze, and discuss

some of the preconditioning methods that have been used for general Helmholtz problems. In chapter 4 two versions of a two-level preconditioner are studied using Fourier analysis. In the last chapter we discuss some conclusions and propose further directions of research.

Chapter 2

Iterative Methods for Linear Systems

In this chapter we review the subject of iterative methods for solving linear systems, with an emphasis on Krylov subspace methods and preconditioning. We also present shortly the multigrid method. We refer the reader to the books by Saad [29] and Van der Vorst [35] for a complete treatment of iterative methods, and the books by Trottenberg et al. [34] and Briggs et al. [7] for a full presentation of the multigrid method. The survey article [3] gives a historical overview of these methods.

2.1 Basic Iterative Methods

Given a linear system of equations

$$Ax = b \tag{2.1}$$

where $A \in \mathbb{C}^{N\times N}$ and $b, x \in \mathbb{C}^N$, we can find an approximate solution by constructing a sequence $x_0, x_1, \ldots$ beginning from an initial guess x_0 and with each x_n obtained by solving a system related to (2.1). Writing A in the form $A = B - C$ for a pair of matrices B, C, we have that (2.1) is equivalent to

$$Bx = b + Cx. \tag{2.2}$$

This suggests an iterative scheme of the form

$$Bx_{n+1} = b + Cx_n, \;\; n = 0, 1, \ldots \tag{2.3}$$

If the matrix B is nonsingular we have

$$\begin{aligned} x_{n+1} &= B^{-1}(b + Cx_n) \\ &= B^{-1}b + (I - B^{-1}A)x_n \\ &= x_n + B^{-1}r_n, \end{aligned} \tag{2.4}$$

where $r_n = b - Ax_n$ is the *residual* at the n-th iteration. For such a scheme to be efficient it is necessary that the matrix B is easily invertible.

The simplest iterative scheme is *Richardson's iteration*, where $B = I$ and $C = I - A$. If A is splitted in the form $A = D - L - U$, where D is the diagonal part of A, and the matrices $-L$ and $-U$ are the strict upper and strict lower diagonal parts of A, we obtain the *Jacobi* method when $B = D$ and $C = -(L+U)$, and the *Gauss-Seidel* method when $B = D - L$ and $C = -U$.

For a parameter $\omega \in (0, 1)$ we can form the splitting

$$\omega A = (D - \omega L) - (\omega U + (1 - \omega)D),$$

from which the *Succesive Over Relaxation* (SOR) method is obtained. Equation (2.4) shows that for all these basic methods the iterates can be obtained by a fixed point iteration of the form

$$x_{n+1} = Gx_n + f, \tag{2.5}$$

for a matrix G and a vector f. If the method is convergent, equations (2.2) and (2.3) show that $\lim_{n\to\infty} x_n = x$, so the sequence of iterates converges to the solution of the system.

Recall that for a matrix $G \in \mathbb{C}^{N\times N}$ the *spectral radius* $\rho(G)$ of G is defined as

$$\rho(G) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } G\}.$$

The following theorem ([29], theorem 4.1) gives a necessary and sufficient condition for the convergence of a basic iterative method.

Theorem 2.1. *Let G be a square matrix with $\rho(G) < 1$. Then $I - G$ is nonsingular and the iteration* (2.5) *converges for every f and x_0. Conversely, if for every f and x_0 the iteration* (2.5) *converges, then $\rho(G) < 1$.*

2.2 Krylov Subspace Methods

An alternative approach to solve the linear system

$$Ax = b$$

is based on *projections*. In a projection method, we start with an initial guess $x_0 \in \mathbb{R}^N$ and then construct a sequence of approximations $x_1, x_2, \ldots$ requiring that each x_n lies in an affine subspace

$$x_n \in x_0 + \mathcal{S}_n. \tag{2.6}$$

A unique x_n can be determined if n linear constraints are fixed. These can be set by requiring that the residual $r_n = b - Ax_n$ is orthogonal to some n-dimensional subspace $\mathcal{C}_n$:

$$b - Ax_n \perp \mathcal{C}_n. \tag{2.7}$$

The subspaces $\mathcal{S}_n$ and $\mathcal{C}_n$ are called the *search* and *constraint* subspaces respectively. Let $S = [s_1, \ldots, s_n]$ and $C_n = [c_1, \ldots, c_n]$ be matrices of size $N \times n$ whose columns form bases for the search and constraint subspaces $\mathcal{S}_n$ and $\mathcal{C}_n$. Condition (2.6) can be reformulated as

$$x_n = x_0 + S_n y_n, \text{ for some } y_n \in \mathbb{C}^N \text{ to be determined,}$$

and (2.7) leads to

$$\begin{aligned} 0 &= C_n^*(b - Ax_n) \\ &= C_n^* b - C_n^* A x_0 - C_n^* A S_n y_n \\ &= C_n^* r^0 - C_n^* A S_k y_n. \end{aligned}$$

Hence, in the n-th step of the method, the projected system

$$C_n^* A S_n y_n = C_n^* r_0$$

is solved. We say that the method is *well-defined* at step n if the matrix $C_n^* A S_n$ is nonsingular. The following result provides conditions under which the method is well-defined ([29], proposition 5.1).

Theorem 2.2. *A projection method is well-defined at step n if A, $\mathcal{S}_n$, and $\mathcal{C}_n$ satisfy one of the two following conditions:*

(i) A is positive definite and $\mathcal{S}_n = \mathcal{C}_n$.

(ii) A is invertible and $\mathcal{C}_n = A\mathcal{S}_n$.

The family of projection methods known as *Krylov subspace methods* uses search subspaces of the form

$$\mathcal{K}_n(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{n-1} r_0\},$$

these are called *Krylov subspaces*. A key property of these methods is that the n-th iterant can be written in the form

$$x_n = x_0 + p(A) r_0,$$

where p is a polynomial of degree at most $n - 1$. This is a crucial property that is used for the convergence analysis of Krylov subspace methods, that we will review later.

2.2.1 The Conjugate Gradient Method

The Conjugate Gradient (CG) method, introduced by Hestenes and Stiefel in 1952 [21], is the method based on the choice of search and constraint subspaces

$$\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(A, r_0),$$

with the additional requirement that the matrix A should be Hermitian positive definite (HPD). This choice leads to the following optimality property for the n-th iterant:

$$\|x - x_n\|_A = \min_{y \in \mathcal{K}_n(A, r_0)} \|x - y\|_A, \tag{2.8}$$

where $\|\cdot\|_A$ is the norm induced by the inner product

$$(v, w)_A = v^T A w.$$

The CG algorithm is given below.

Data: A HPD matrix $A \in \mathbb{C}^{N \times N}$, and vectors $b, x_0 \in \mathbb{C}^N$.
begin
  $r_0 := b - Ax_0$, $p_0 := r_0$;
  for $k = 0, 1, \ldots$ *until convergence* **do**
    $\alpha_k := (r_k, r_k)/(Ap_k, p_k)$;
    $x_{k+1} := x_k + \alpha_k p_k$;
    $r_{k+1} := r^k - \alpha_k A p_k$;
    $\beta_k := (r^{k+1}, r^{k+1})/(r_k, r_k)$;
    $p_{k+1} := r^{k+1} + \beta_k p_k$
  end
end

Algorithm 1: CG Method

2.2.2 The GMRES Method

In the GMRES method, the choices of search and constraint subspaces are $\mathcal{S}_n = \mathcal{K}_n$ and $\mathcal{C}_n = A\mathcal{K}_n$. The method was proposed in 1986 by Saad and Schultz [30], and does not require any additional properties of the linear system. Moreover, an optimality property for the residual of the n-th iterant holds:

$$\|r_n\| = \|b - Ax_n\| = \min_{y \in \mathcal{K}_n(A, r_0)} \|b - Ay\|.$$

The GMRES algorithm is given below.

Data: A matrix $A \in \mathbb{C}^{N\times N}$, and vectors $b, x_0 \in \mathbb{C}^N$.
begin
  $r_0 := b - Ax_0$, $v_1 := r_0/\|r_0\|$;
  Let $H_m \in \mathbb{C}^{(m+1)\times m}$, $H_m := 0$;
  for $j = 1, 2, \ldots m, \ldots$ **do**
    for $i = 1, 2, \ldots, j$ **do**
      $h_{ij} := (Av_j, v_i)$;
    end
    $\hat{v}_{j+1} := Av_j - \sum_{i=1}^{j} h_{ij}v_j$;
    $h_{j+1j} := \|\hat{v}_{j+1}\|$;
    $v_{j+1} := \hat{v}_{j+1}/h_{j+1j}$;
  end
  $y_m := \mathrm{argmin}_y \|\beta e_1 - H_m y\|$;
  $x_m := x_0 + V_m y_m$;
end

Algorithm 2: GMRES Method

2.2.3 The BiCG Method

The Biconjugate Gradient (BiCG) method was first proposed by Lanczos in [26] and in an alternate form by Fletcher in [18]. The method works for general nonsigular matrices, and results from projecting in each iteration onto the search subspaces $\mathcal{S}_n = K_n(A, r_0)$, orthogonally to the constraint subspaces $\mathcal{C}_n = K_n(A^*, r_0)$. The choice of the adjoint matrix A^* for the constraint subspaces leads to the solution of a dual linear system $A^*\tilde{x} = \tilde{b}$ which is usually ignored. The BiCG method is implemented in algorithm 3.

Data: A matrix $A \in \mathbb{C}^{N\times N}$, and vectors $b, x_0 \in \mathbb{C}^N$.
begin
  $r_0 := b - Ax_0$, $p_0 := r_0$;
  Choose r_0^* with $(r_0^*, r_0) \neq 0$, $p_0^* := r^0$;
  for $k = 0, 1, \ldots$ *until convergence* **do**
    $\alpha_k := (r_k^*, r_k^*)/(Ap_k, p_k^*)$;
    $x_{k+1} := x_k + \alpha_k p_k$;
    $r_{k+1} := r_k - \alpha_k A p_k$;
    $r_{k+1}^* := r_k - \alpha_k A^* p_k$;
    $\beta_j := (r_{k+1}, r_{k+1}^*)/(r_k, r_k^*)$;
    $p_{k+1} := r_{k+1} + \beta_k p_k$;
    $p_{k+1}^* := r_{k+1}^* + \beta_k p_k^*$
  end
end

Algorithm 3: BiCG Method

2.2.4 Convergence of Krylov Subspace Methods

In this section we review the convergence estimates for the CG and GMRES methods. Since the iterants of the BiCG method do not satisfy an optimality property, the convergence of this method is more difficult to analyze. Recall that given a nonsingular matrix A with maximum and minimum eigenvalues (in modulus) $\lambda^{\max}$ and $\lambda^{\min}$, the spectral condition number of A ([29], p. 45) is given by

$$\kappa(A) = \frac{|\lambda^{\max}|}{|\lambda^{\min}|} \tag{2.9}$$

We have previously noted that the n-th iterate in a Krylov subspace method has the form

$$x_n = x_0 + p(A)r_0,$$

for a polynomial $p(A)$. Combining this with the minimization property of the CG method

$$\|x - x_n\|_A = \min_{y \in \mathcal{K}_n(A, r_0)} \|x - y\|_A,$$

and using the optimality properties of the Chebyshev polynomials, the classical convergence estimate for CG is obtained. The following bound for the error of the n-th iterate of the CG method in the A-norm holds:

$$\|x - x_n\|_A \leq 2 \left[\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right] \|x - x_0\|_A. \tag{2.10}$$

One should keep in mind that this bound only gives a worst-case estimate for the error for all possible eigenvalue distributions on the interval $[\lambda^{\min}, \lambda^{\max}]$. Moreover, this bound for the error shows that a small condition number leads to a faster convergence of the CG method, but it does not show that a large condition number leads to slow convergence, as explained in [27]. The next theorem ([30], proposition 4) gives a convergence estimate for the GMRES method.

Theorem 2.3. *Let A be a diagonalizable matrix with $A = X\Lambda X^{-1}$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues of A, and let*

$$\epsilon^{(n)} = \min_{p \in \Pi_n} \max_k |p(\lambda_k)|,$$

where Π_n is the space of all polynomials of degree less or equal to n with value 1 at the origin. Then, the norm of the residual of the n-th GMRES iterate satisfies the inequality

$$\|r_n\| \leq \epsilon^{(n)} \kappa(X) \|r_0\|.$$

If the spectrum of A is contained in a disk in the complex plane not containing the origin, centered at C with radius R, one has the following bound for the quantity $\epsilon^{(m)}$ from the previous theorem (see [27])

$$\epsilon^{(n)} \leq \left(\frac{R}{C}\right)^n.$$

This implies that a fast convergence can be expected when the spectrum is contained in a small circle far from the origin. However, we remark that the convergence of GMRES applied to a linear system cannot be predicted in general by the spectrum of the matrix, in view of the result in [19], where it is shown that any nonincreasing curve of relative GMRES residual norms is attainable for a matrix A having any prescribed eigenvalues.

2.3 Preconditioning

We have seen that the eigenvalues of a matrix play an important role in the convergence of Krylov subspace methods. The idea behind preconditioning is that a linear system can be transformed into an equivalent one which has a better eigenvalue distribution and is easier to solve. If M is a nonsingular matrix, the linear system

$$Ax = b$$

is equivalent to

$$M^{-1}Ax = M^{-1}b. \tag{2.11}$$

The matrix M is said to be a (left) *preconditioner*, and it should be easily invertible and approximate A in some sense. The system (2.11) can now be solved using a Krylov subspace method. In practice the matrix M^{-1} is not computed explicitly, and products of the form $u = M^{-1}Av$ that need to be formed during the iterative process are obtained by computing

$$t = Av, \;\; u = M^{-1}t. \tag{2.12}$$

Alternatively, a right preconditioner can be used, leading to the equivalent system

$$\begin{aligned} AM^{-1}\tilde{u} &= b, \\ \hat{u} &= Mu, \end{aligned}$$

The convergence of a Krylov subspace method applied to (2.11) now only depends on the eigenvalues of $M^{-1}A$. In view of the discussion on the preceding section, an improved convergence can be obtained if the eigenvalues of this system are more clustered and bounded away from zero. Hence, for a preconditioner to be useful we should at least require that

$\kappa(M^{-1}A) < \kappa(A)$ holds.

2.4 The Multigrid Method

The multigrid method goes back to the work by Fedorenko [17], and was further developed by Brandt [5], and Hackbush [20], among others, to solve linear systems arising from the discretization of partial differential equations. We outline here its main ideas.

Suppose that we look for a solution to the system

$$A_h u_h = f_h,$$

where $A_h \in \mathbb{R}^{N\times N}$ corresponding to the discretization of a PDE on a uniform grid Ω_h with grid size h. Let u_0 be an initial approximation to the exact solution $\tilde{u}$, and $e_0 = \tilde{u} - u_0$ the initial error. If a basic iterative method (such as the methods from section 2.1) is applied to compute a sequence of approximations $u_1, u_2, \ldots$, only the high frequency components of the error $e_n = \tilde{u} - u_n$ will decrease after a few iterations. More precisely, if $\{\phi_k\}_{k=1}^N$ is the orthonormal basis of Fourier modes in $\mathbb{R}^N$, the error can be written in the form

$$e_n = \sum_{k=1}^{N} a_k^{(n)}\phi_k = \sum_{\text{low}} a_k^{(n)}\phi_k + \sum_{\text{high}} a_k^{(n)}\phi_k,$$

for some coefficients $a_k^{(n)}$, where we have splitted the sum into 'high' and 'low' frequencies corresponding to the frequencies of the Fourier modes. The high frequency error decreases rapidly, and after a few iterations we have

$$\left\| \sum_{\text{low}} a_k^{(n)}\phi_k \right\| \ll \left\| \sum_{\text{low}} a_k^{(0)}\phi_k \right\|.$$

This is known as the *smoothing property* of basic iterative methods, and for this reason these methods are called *smoothers* in the context of multigrid methods. Since the low frequency components of the error $\tilde{u} - u_n$ remain and they cannot be decreased by the smoother, the convergence stagnates. In order to overcome this problem, the smooth error can be transferred to a coarser grid $\Omega_H \subset \Omega_h$, where the low frequency components of the error are now 'visible' as high frequency ones, and a basic iterative method can be used effectively to solve the error equation. A correction is then computed on the coarse grid, and interpolated back to the fine grid. A recursive application of this idea on increasingly coarse grids leads to the multigrid method.

2.4.1 Basic Components of Multigrid

Setting up a multigrid method requires several basic components: A choice of smoothers, a sequence of grids, and the corresponding transfer operators between them. We begin with a simple two-grid scheme for a 1-D model problem. Consider two equidistant grids Ω_h, Ω_H on the domain $\Omega = (0, 1)$, where Ω_h has gridsize $h = 1/n$ and $\Omega_H \subseteq \Omega_h$ is obtained by standard grid coarsening, so $H = 2h$. We also need intergrid transfer operators I_h^H, I_H^h (restriction and prolongation), and the discretization matrices A_h and A_H for a PDE problem on the fine and coarse grid. We will write $\tilde{x} = SM^\nu(A, b, x_0)$ for the result of applying ν iterations of a basic iterative method (smoother) to solve the linear system $Ax = b$ with the initial guess x_0. Algorithm 4 is a simple implementation of the two-grid method with ν_1 presmoothing steps and ν_2 postsmoothing steps.

$v_h \leftarrow TGM^{\nu_1,\nu_2}(A_h, f_h, v_h)$.

begin

Presmoothing
$v_h = SM^{\nu_1}(A_h, f_h, v_h)$.

Coarse grid correction:
Set $r_h := f_h - A_h v_h$ and restrict it to coarse grid $r_H := I_h^H r_h$.
Solve exactly the error equation $e_H := A_H^{-1} r_H$
Prolongate the coarse error to the fine grid $e_h := I_H^h e_H$ and correct $v_h \leftarrow v_h + e_h$.

Postsmoothing
$v_h = SM^{\nu_2}(A_h, f_h, v_h)$

end

Algorithm 4: Two-grid Method

The two-grid method can be extended by applying the same idea recursively, on a sequence of grids $\Omega_h \supseteq \Omega_{2h} \supseteq \cdots \supseteq \Omega_{2^m h}$. This leads to a multigrid V-cycle, which is an intermediate step in the construction of more complex methods. The following two algorithms implement the V-cycle and the full multigrid method.

$v_h \leftarrow V_h^{\nu_1,\nu_2}(A_h, f_h, v_h)$.
begin
  Presmoothing
  $v_h = SM^{\nu_1}(A_h, f_h, v_h)$.
  if *on coarsest grid* **then**
    goto *Postsmoothing*
  end
  else
    Coarse grid correction
    Set $r_h := f_h - A_h v_h$, $f_{2h} := I_h^{2h} r_h$, $v_{2h} := 0$.
    $v_{2h} \leftarrow V_{2h}^{\nu}(A_{2h}, f_{2h}, v_{2h})$
    Correct $v_h \leftarrow v_h + I_{2h}^h v_{2h}$
  end
  Postmoothing
  $v_h = SM^{\nu_2}(A_h, f_h, v_h)$
end

Algorithm 5: Multigrid V-cycle

$v_h \leftarrow FMG_h^{\nu}(A_h, f_h, v_h)$.
begin
  if *on coarsest grid* **then**
    Set $v_h := 0$.
    goto *V-cycle*
  end
  else
    Coarse grid correction
    Set $f_{2h} := I_h^{2h} r_h$, $v_{2h} = I_h^{2h} v_h$,
    $v_{2h} \leftarrow FMG_{2h}^{\nu}(A_{2h}, f_{2h}, v_{2h})$.
    Correct $v_h \leftarrow v_h + I_{2h}^h v_{2h}$.
  end
  V-cycle $v_h = V_h^{\nu}(A_h, f_h, v_h)$
end

Algorithm 6: Full Multigrid Cycle

Chapter 3

Numerical Methods for the Helmholtz Equation

In this chapter we formulate our Helmholtz model problem and discuss some of the methods that are used for solving linear systems arising from Helmholtz model problems.

3.1 Discretization of the Helmholtz Equation

We introduce now the discretization of the Helmholtz problem in a bounded domain with a source term

$$-\Delta u - k^2 u = f \quad \text{in } \Omega \subseteq \mathbb{R}^d. \tag{3.1}$$

The boundary conditions to be imposed should mimic Sommerfeld's radiation condition (1.5), adapted to the domain Ω, to ensure that waves are traveling in the outwards direction and there is no artificial reflection of waves. The following first-order *absorbing boundary conditions* have been proposed in [10]:

$$\partial_n u - iku = 0 \text{ on } \partial\Omega. \tag{3.2}$$

The discretization of problem [(3.1), (3.2)] leads to a complex-valued matrix, which is *complex-symmetric* (A=A^T but $A \neq A^*$) and *indefinite*, i.e., has positive and negative eigenvalues. We will only treat the case of Dirichlet boundary conditions

$$u = 0 \text{ on } \partial\Omega, \tag{3.3}$$

as the matrix arising from the discretization of [(3.1), (3.2)] is non-normal, and cannot be analyzed with the techniques we use here. However, the results from our analysis will also

be useful for the complex Helmholtz system, since the preconditioning methods that we study have shown in general better results for the complex Helmholtz systems [13], and the properties and location of the spectrum of the preconditioned systems are similar. Hence, in terms of convergence analysis, the Dirichlet problem can be considered as a worst-case version of the Sommerfeld problem

We will focus on the following Helmholtz model problem:

1-Dimensional Helmholtz Model Problem. Given a source function $f(x)$ and a wavenumber k, find the wavefield u such that

$$\begin{aligned} -\frac{d^2u}{dx^2}(x) - k^2u(x) &= f(x) \text{ for } x \in (0,1), \\ u(0) &= u(1) = 0, \end{aligned} \tag{3.4}$$

Equation (3.4) can be discretized on a uniform grid Ω_h with meshwidth $h = 1/n$ and gridpoints $x_0, \ldots, x_n$ where $x_i = ih$. A finite difference approximation of second order on the grid is given by the stencil

$$[A_h] = \frac{1}{h^2}[\ -1 \quad 2-\kappa^2 \quad -1\ ],$$

where $\kappa = kh$. After elimination of the unknowns corresponding to the boundary points, the discrete problem is reduced to a linear system of equations

$$A_h u_h = f_h, \tag{3.5}$$

where

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2-\kappa^2 & -1 & 0 & \cdots & 0 \\ -1 & 2-\kappa^2 & -1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 2-\kappa^2 \end{pmatrix} \tag{3.6}$$

has size $(n-1)\times(n-1)$. To obtain an accurate representation of the wavefield u on the finite grid, at least 10 gridpoints per wavelength should be used [1]. A time-harmonic wave with wavenumber k has a wavelength $\lambda = 2\pi/k$, so the number p of gridpoints per wavelength equals

$$p = \frac{\lambda}{h} = \frac{2\pi}{kh},$$

and we have the restriction

$$p \geq 10, \text{ or equivalently, } kh \leq \pi/5. \tag{3.7}$$

A more restrictive bound on the quantity h^2k^3 is also required to avoid instability [22], how-

ever, we will only enforce (3.7) since we are mostly interested in the iterative solution of the system.

When k^2 is larger than the smallest eigenvalue of the Laplace operator $-\Delta_h$, the discrete Helmholtz operator has both positive and negative eigenvalues, making it *indefinite*. We can calculate these eigenvalues explicitly. Recall the basis of *Fourier modes* in $\mathbb{R}^{n-1}$

$$V_h = \{v_l : 1 \leq l \leq n-1\},$$

where

$$v^l = \sin(l\pi\underline{x}) = [\sin(l\pi h), \sin(2l\pi h), \ldots, \sin((n-1)l\pi h)]^T,$$

these are the eigenvectors of the discrete Laplace operator $-\Delta_h$ with corresponding eigenvalues

$$\lambda^l(-\Delta_h) = \frac{1}{h^2}[2 - 2\cos(l\pi h)] \text{ , for } l = 1, \ldots n-1. \tag{3.8}$$

To check this we compute the entries of the matrix-vector product $-\Delta_h v^l$, using the trigonometric identity $\sin(\alpha+\beta) - \sin(\alpha-\beta) = 2\sin(\alpha)\cos(\beta)$:

$$\begin{aligned}
[-\Delta_h v^l]_j &= \sum_{k=1}^{n-1} [-\Delta_h]_{jk} v_k^l \\
&= \frac{1}{h^2}[-\sin(l\pi(j-1)h) + 2\sin(l\pi jh) - \sin(l\pi(j+1)h)] \\
&= \frac{1}{h^2}[2\sin(l\pi jh) - 2\sin(l\pi jh)\cos(l\pi h)] \\
&= \frac{2}{h^2}[1 - \cos(l\pi h)]\sin(l\pi jh) \\
&= \frac{2}{h^2}[1 - \cos(l\pi h)][v^l]_j
\end{aligned}$$

Also note that

$$\lambda^{n-l}(-\Delta_h) = \frac{1}{h^2}[2 - 2\cos((n-l)\pi h)] = 2 + 2\cos(l\pi h), \text{ for } l = 1, \ldots, n/2 - 1,$$

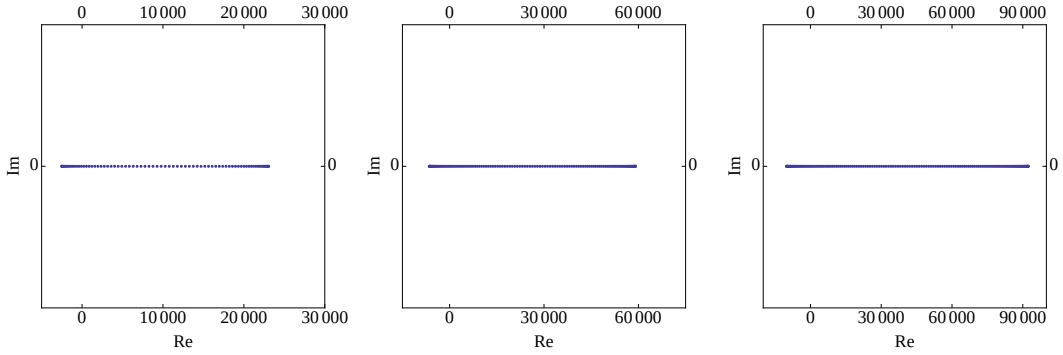
and

$$\lambda^{n/2} = \frac{2}{h^2}[1 - \cos((n/2)\pi h)] = \frac{2}{h^2}[1 - \cos(\pi/2)] = \frac{2}{h^2}.$$

We have then that the eigenvalues of the discrete Helmholtz operator A_h are

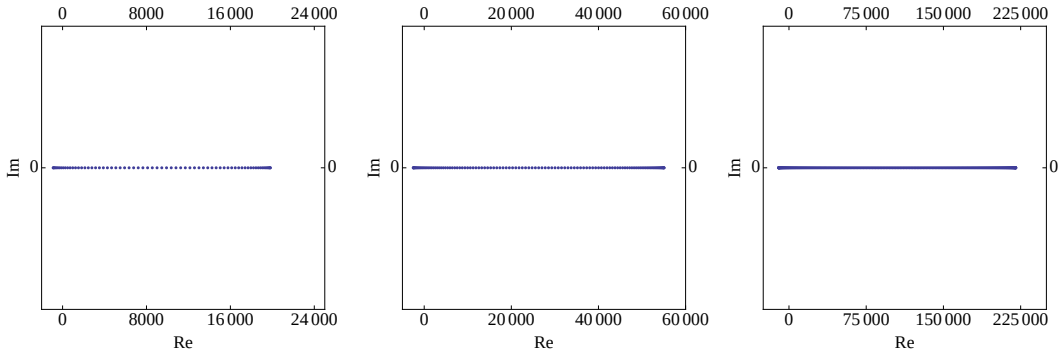
$$\lambda^l(A_h) = \frac{1}{h^2}(2 - 2\cos(l\pi h) - \kappa^2), \text{ for } l = 1, \ldots n-1. \tag{3.9}$$

The plots from figures 3.1 and 3.2 show the spectrum of the Helmholtz system for several values of the wavenumber k, the number of gridpoints per wavelength p and $\kappa = kh$.



(a) $k = 50$ (b) $k = 80$ (c) $k = 100$

Figure 3.1: Spectrum of A_h for $p = 10$ and $\kappa = 0.628$



(a) $k = 50$ (b) $k = 50$ (c) $k = 100$

Figure 3.2: Spectrum of A_h for $p = 15$ and $\kappa = 0.418$

From the formula for the eigenvalues of the 1-D Helmholtz operator of our model problem one can obtain some insight that applies also for higher dimensional Helmholtz problems and their corresponding discretization matrices, since the spectrum of such matrices is similar. Note that the sign of the eigenvalues equals the sign of

$$2 - 2\cos(l\pi h) - k^2h^2.$$

For a fixed wavenumber k, increasing the grid size h may change the sign of an eigenvalue. This poses a difficulty for the error correction steps in multigrid methods: Even if high frequency components of the error are well resolved on coarse grids and reduced by smoothing steps, the correction may be added with the wrong sign. The same phenomenon also occurs when multigrid is used for other indefinite and nearly singular problems [6]. Another drawback for the use of multigrid methods in Helmholtz problems, discussed in [16], is that using standard smoothers to dampen oscillatory modes of the error leads to amplification of the

smooth modes. A possible solution which is proposed in [9] is to incorporate Krylov subspace methods as smoothers on the intermediate grids.

3.2 The Shifted Laplace Preconditioner

The use of a shifted Laplace operator of the form

$$\mathcal{M}_{(\beta_1,\beta_2)} = -\Delta - (\beta_1 - i\beta_2)k^2, \;\; \beta_1, \beta_2 \in \mathbb{R},$$

as a preconditioner for the Helmholtz equation was first initiated by Giles and Laird, who proposed using a Laplace preconditioner with a real shift in [25]. More recently, Erlangga, Osterlee, and Vuik considered a more efficient complex shifted Laplace (CSL) preconditioner in which the imaginary part of the shift is nonzero. The main advantage of using such an operator is that the imaginary term adds damping and allows multigrid methods to be effective for the inversion of the preconditioner, while the spectrum of the preconditioned system becomes suitable for the use of Krylov subspace methods.

For our 1-D model problem, the eigenvalues of the preconditioned system can be computed explicitly and the effect of the choice of the shift parameters can be analyzed. Let A_h be the Helmholtz matrix for our model problem, and $M_{h,(\beta_1,\beta_2)}$ the finite-difference discretization matrix of the 1-D shifted Laplace operator with homogeneous boundary conditions. The preconditioned matrix is

$$\hat{A}_{h,(\beta_1,\beta_2)} = M^{-1}_{h,(\beta_1,\beta_2)} A_h.$$

The set of eigenvectors of this matrix is the set V_h of Fourier modes, and their corresponding eigenvalues are

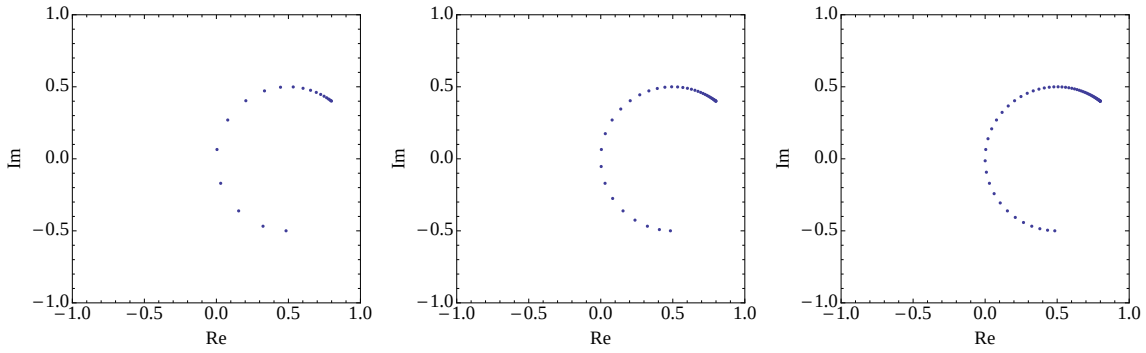
$$\lambda^l(\hat{A}_{h,(\beta_1,\beta_2)}) = \frac{2 - 2c_l - \kappa^2}{2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)}, \;\; \text{for } l = 1, 2, \ldots, n-1. \tag{3.10}$$

Figures 3.3 to 3.5 show the spectrum of $\hat{A}_{h,(\beta_1,\beta_2)}$ for several values of the parameters k, κ, β_1 and β_2. Note that most of the eigenvalues are clustered around 1, but some of them have small magnitude and are close to zero. Also, the number of small eigenvalues increases for larger k, and the clustering improves when more gridpoints per wavelength are used. Moreover, the eigenvalues are located on a circle, which is independent of the wavenumber k. The following theorem is a special case of theorem 3.2 in [36].

Theorem 3.1. *For every κ, β_1 and $\beta_2 \neq 0$, the eigenvalues of the preconditioned matrix $\hat{A}_{h,\beta_1,\beta_2}$ given by the expression* (3.10) *are located on the circle*

$$|z - c|^2 = r^2,$$

with center $c = \left(\frac{1}{2}, \frac{1-\beta_1}{2\beta_2}\right)$ and radius $r = \frac{1}{2}\sqrt{1 + \frac{(1-\beta_1)^2}{\beta_2^2}}$.



(a) $k = 50$ (b) $k = 100$ (c) $k = 150$

Figure 3.3: Spectrum of $\hat{A}_{h,(1,0.5)}$ for $\kappa = 0.628$

Proof. From the expresion for $\lambda^l(\hat{A}_{h,\beta_1,\beta_2}) = \operatorname{Re}\lambda^l + i\operatorname{Im}\lambda^l$ we get

$$[\operatorname{Re}\lambda^l + i\operatorname{Im}\lambda^l][2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)] = 2 - 2c_l - \kappa^2$$

Multiplying the left hand side, and equating the real and imaginary parts of both sides of the equation:

$$\begin{aligned}\operatorname{Re}\lambda^l(2 - 2c_l - \kappa^2\beta_1) - \operatorname{Im}\lambda^l(\kappa^2\beta_2) &= 2 - 2c_l - \kappa^2\\ \operatorname{Im}\lambda^l(2 - 2c_l - \kappa^2\beta_1) + \operatorname{Re}\lambda_l(\kappa^2\beta_2) &= 0.\end{aligned}$$

From the second equation we obtain

$$2 - 2c_l = \kappa^2\beta_1 - \kappa^2\beta_2\frac{\operatorname{Re}\lambda^l}{\operatorname{Im}\lambda^l},$$

substituting in the first equation and multiplying by $-\operatorname{Im}\lambda_l$ we get

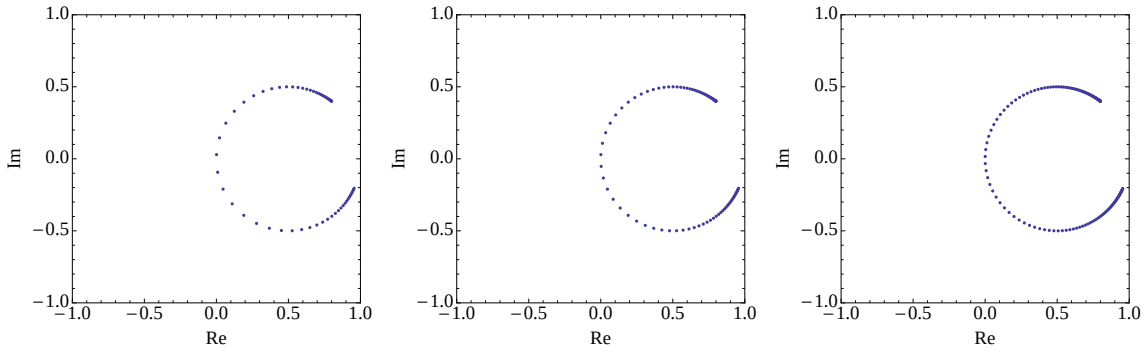
$$\kappa^2\beta_2(\operatorname{Re}\lambda^l)^2 + \kappa^2\beta_2(\operatorname{Im}\lambda^l)^2 - \kappa^2\beta_2\operatorname{Re}\lambda_l + \kappa^2(\beta_1 - 1)\operatorname{Im}\lambda^l = 0.$$

Dividing by $\kappa^2\beta_2$ and completing the squares gives the equation of the circle

$$\left(\operatorname{Re}\lambda^l - \frac{1}{2}\right)^2 + \left(\operatorname{Im}\lambda^l - \frac{(1-\beta_1)}{2\beta_2}\right)^2 = \frac{1}{4} + \frac{(1-\beta_1)^2}{4\beta_2^2}.$$

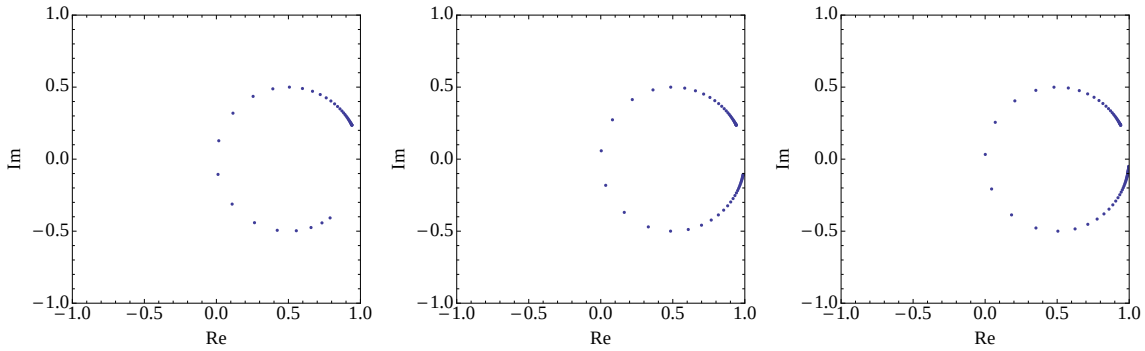
□

The previous theorem and the convergence bounds from section 2.2.4 can be used to find near optimal values of the shift with respect to the number of Krylov iterations, as is done in [36] for GMRES, and validated experimentally. The choice of parameters $(\beta_1, \beta_2) = (1, 0.5)$



(a) $k = 100$ (b) $k = 150$ (c) $k = 250$

Figure 3.4: Spectrum of $\hat{A}_{h,(1,0.5)}$ for $\kappa = 0.418$



(a) $\kappa = 0.628$ (b) $\kappa = 0.418$ (c) $\kappa = 0.314$

Figure 3.5: Spectrum of $\hat{A}_{h,(1,0.25)}$ for $k = 100$

is considered to be near optimal in terms of number of Krylov iterations and is the standard choice in the literature.

One should also take into account the difficulties explained in section 3.1. In particular, if the imaginary shift is too small, one obtains a Helmholtz-like operator which cannot be inverted using multigrid. The trade-off between the constraint on the shift for a small number of Krylov iterations and the requirement for multigrid convergence can be understood from a closer inspection to formula (3.10). The following analysis is proposed in [16]. If $\beta_1 = 1$ and β_2 is small, a Taylor expansion of (3.10) in the variable β_2 gives

$$\lambda^l(\hat{A}_{h,(1,\beta_2)}) = 1 - i\frac{\kappa^2\beta_2}{2 - 2c_l - \kappa^2} + O(\beta_2^2),$$

so in order to have the spectrum clustered in an arc around $(1, 0)$ and bounded away from the origin we must require

$$\beta_2 \ll \min_{l=1,\ldots,n-1} |2 - 2c_l - \kappa^2|.$$

If $\kappa = kh$ is fixed, given a wavenumber k let

$$d(k) = \min_{l=1,\ldots,n-1} |2 - 2c_l - \kappa^2|.$$

We show the dependence of this quantity with respect to the wavenumber k in figure 3.6.

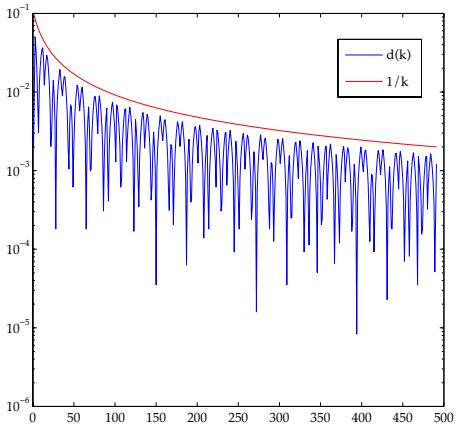


Figure 3.6: Illustration of how small the shift β_2 should be chosen in the CSL to obtain a preconditioner that bounds the spectrum away from the origin.

We can see from the plot that the condition $\beta_2 < 1/k$ must be met to have a good preconditioner. If this condition is met, for increasing values of k the multigrid convergence will deteriorate, indicating that that a different preconditioning technique is needed to overcome this problem.

3.3 Deflation and Multilevel Methods

3.3.1 Deflation

Deflation is a technique to deal with the deterioration in the speed of convergence of Krylov subpace methods due to the presence of small eigenvalues. To describe this technique we consider a linear system of the form

$$Ax = b \tag{3.11}$$

Our goal is to remove the components of the residual corresponding to the smallest r eigenvalues, that are responsible for the slow convergence. Let $Z, Y \in \mathbb{R}^{n\times r}$ be full rank matrices, where $r \ll n$. We define the (left) *projection matrix* P_D

$$P_D = I - AZE^{-1}Y^T, \text{ where } E = Y^T AZ. \tag{3.12}$$

and the corresponding right variant

$$Q_D = I - ZE^{-1}Y^T A. \tag{3.13}$$

The matrix Z is called the *deflation matrix*, and, since E is similar to the coarse grid matrix of multigrid or domain decomposition methods, we call it the *Galerkin matrix*. If we apply the projection P_D on the left to (3.11) we obtain the deflated system

$$P_D A = P_D b.$$

The following theorem explains the effect of the projection matrix on (3.11).

Theorem 3.2. *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix and $Z, Y \in \mathbb{R}^{n\times r}$ full rank matrices, with $r \ll n$. If P_D and Q_D are defined as in* (3.12) *and* (3.13)*, then*

1. $P_D A = AQ_D$.
2. $P_D AZ = 0$.
3. *If $\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subset \mathbb{C}$ are the eigenvalues of A ordered increasingly in magnitude, and the columns of Z are the eigenvectors of A corresponding to the smallest r eigenvalues, we have*

$$\sigma(P_D A) = \{0, \lambda_{r+1}, \lambda_n\},$$

And for arbitrary Z and Y,

$$\sigma(P_D A) = \{0, \mu_{r+1}, \ldots, \mu_n\}$$

Proof. The first and second parts follow from a direct computation. For the rest, see lemma 3.1 from [14] and theorem 3.1 from [15]. □

Part 4 of the previous theorem shows that the small eigenvalues are removed from the spectrum by the deflation preconditioner, and the remaining eigenvalues may be shifted, unless eigenvectors are used in the deflation matrix Z. To solve the original system we use the following procedure. Let x^* be the solution to (3.11), and write it in the form

$$x^* = (I - Q_D)x^* + Q_D x^*.$$

We have

$$(I - Q_D)x^* = ZE^{-1}Y^T Ax = ZE^{-1}Y^T b,$$

as E is a smaller system that can be inverted with less work, it only remains to compute $Q_D x^*$. Moreover, in view of the equality $AQ_D = P_D A$, if $\tilde{x}$ is a solution of the deflated

system

$$P_D A x = P_D b \tag{3.14}$$

we have

$$AQ_D\tilde{x} = P_D A\tilde{x} = P_D b = P_D A x^* = AQ_D x^*$$

hence $Q_D\tilde{x} = Q_D x^*$ and we have reduced our problem to computing $\tilde{x}$. The deflated system (3.14) is singular but consistent (i.e., there exists at least one solution), and can be solved using a Krylov subspace method.

If A is symmetric positive definite one typically sets $Z = Y$. The resulting deflated system is positive semidefinite and can be solved by CG [23]. For arbitrary A, Z and Y the deflated system has to be solved using a general method, for instance, GMRES. In any case, the original system has been replaced by the deflated system (3.14) which is expected to converge faster, because the small eigenvalues have been removed.

Note that if a large deflation subspace is used the matrix E should be inverted using iterative methods, hence for the method to be effective the computation of P_D should be insensitive to an inaccurate computation of E^{-1}. This is not always the case, as shown in [28] for SPD systems. A full analysis of this and other aspects of deflation is given in [33].

3.3.2 Multilevel Krylov Methods

To begin our discussion of multilevel methods, we consider a more general version of the deflation preconditioner in which a standard preconditioner is incorporated. Suppose we want to solve (3.11), and let $\hat{A} = M^{-1}A$ and $\hat{b} = M^{-1}b$. For full rank matrices $Z, Y \in \mathbb{R}^{n\times r}$, the corresponding projection operator is

$$\hat{P}_D = I - \hat{A}Z\hat{E}^{-1}Y^T, \tag{3.15}$$

where $\hat{E} = Y^T\hat{A}Z$ is the coarse grid matrix. In order to solve the deflated system

$$\hat{P}_D\hat{A} = \hat{P}_D\hat{b},$$

using an iterative Krylov method, the Galerkin coarse matrix $\hat{E}$ should be inverted. Naturally, one can apply the same method and deflate the coarse system. If this technique is used recursively, a multilevel Krylov (MK) method arises. In principle, there are no restrictions for Z and Y, other than the requirement that they should be full rank matrices. Possible choices of the deflation matrices Z and Y that have been studied are transfer operators from geometric multigrid methods (see [14] and [15]) and matrices based on algebraic multigrid considerations (see [12]).

It was remarked earlier that the computation of P_D may be sensitive to an inaccurate computation of E^{-1}. Adding a shift term to (3.12) leads to an alternative projection preconditioner that is more robust in this sense. This preconditioner has the form

$$\hat{P}_N = I - \hat{A}Z\hat{E}^{-1}Y^T + \tilde{\lambda}_n Z\hat{E}^{-1}Y^T = \hat{P}_D + \tilde{\lambda}_n Z\hat{E}^{-1}Y^T, \tag{3.16}$$

where $E = Y^T AZ$ and $\tilde{\lambda}_n$ is the largest eigenvalue of $\hat{A}$ (or an estimate for it). Strictly speaking, the operator $\hat{P}_N$ is not a projection ($\hat{P}_N^2 \neq \hat{P}_N$), but it is classified as such because it projects some of the eigenvalues of $\hat{A}$ to an appropriate value for acceleration of convergence. For right preconditioning, the operator

$$\hat{Q}_N = I - Z\hat{E}^{-1}Y^T\hat{A} + \tilde{\lambda}_n Z\hat{E}^{-1}Y^T. \tag{3.17}$$

is used. The following theorem explains the effect of these projection-type preconditioners.

Theorem 3.3. *For a nonsingular matrix $\hat{A} \in \mathbb{R}^{n\times n}$ and full rank matrices $Z, Y \in \mathbb{R}^{n\times r}$, the following properties hold:*

1. *If $\hat{P}_D$ and $\hat{P}_N$ are given by* (3.15) *and* (3.16)*, and*

$$\sigma(\hat{P}_D\hat{A}) = \{0, \mu_{r+1}, \ldots, \mu_n\},$$

   then

$$\sigma(\hat{P}_N\hat{A}) = \{\tilde{\lambda}_n, \mu_{r+1}, \ldots, \mu_n\}.$$

2. *If $\hat{P}_N$ and $\hat{Q}_N$ are given by* (3.16) *and* (3.17)*, we have:*

$$\sigma(\hat{P}_N\hat{A}) = \sigma(\hat{A}\hat{Q}_N).$$

We have then that the smallest r eigenvalues are not shifted to 0 but to an estimate of the largest eigenvalue $\tilde{\lambda_n}$. In the general case (Z and Y arbitrary) one needs to take into account that the remaining $n-r$ eigenvalues of $\hat{P}_D\hat{A}$ can also be shifted, and $\tilde{\lambda}_n$ should be scaled by a factor ω. If GMRES is used as a Krylov solver, it is preferable to use the right preconditioner (3.17) since it computes the actual residual during the iteration, unlike the left preconditioner (3.16) which calculates a preconditioned residual that is not useful to terminate the iteration. The right preconditioned GMRES solves the system

$$\hat{A}\hat{Q}_N\hat{x} = \hat{b},$$

or, with $\hat{A} = AM^{-1}$,

$$AM^{-1}\hat{Q}_N\hat{x} = b, \text{ where } u = M^{-1}Q_N\hat{u} \tag{3.18}$$

Since in every iteration of GMRES a different preconditioner is used, a modified version of algorithm 2 known as flexible GMRES (FGMRES) is used, which allows variable preconditioning [32]. Algorithm 7 is an implementation of the FGMRES method with right preconditioning to solve (3.18).

Data: A matrix $A \in \mathbb{C}^{N\times N}$, and vectors $b, x_0 \in \mathbb{C}^N$, an estimate for the largest eigenvalue $\tilde{\lambda}_n \in \mathbb{C}$, and a scaling factor $\omega \in \mathbb{R}$.
begin
  Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2, v_1 = r_0/\beta$.;
  for $j = 1, 2, \ldots k$ **do**
    $x_j := \hat{Q}_N v_j$;
    $w := AM^{-1}x_j$;
    for $i = 1, 2, \ldots, j$ **do**
      $h_{ij} := (w, v_i)$;;
      $w := w - h_{ij}v_i$.;
    end
    $h_{j+1j} := \|w\|$. ;
    $v_{j+1} := w/h_{j+1j}$.;
  end
  Set $X_k := [x_1 \ldots x_k]$ and $\hat{H}_k = \{h_{ij}\}_{1\le i\le j+1; 1\le j\le k}$.;
  Compute $y_k := \text{argmin}_y \|\beta e_1 - H_k y\|$ and $x_k := x_0 + M^{-1}X_k y_k$;
end

Algorithm 7: FGMRES preconditioned by M and $\hat{Q}_N$

The first line of the outer iteration of algorithm 7 requires that we compute $x_j := \hat{Q}_N v_j$. We have

$$\begin{aligned} x_j &= (I - Z\hat{E}^{-1}\hat{A} + \omega\tilde{\lambda}_n Z\hat{E}^{-1}Y^T)v_j \\ &= v_j - Z\hat{E}^{-1}Y^T v' \end{aligned}$$

where $v' = (\hat{A} - \omega\tilde{\lambda}_n I)v_j$. Let $s = \hat{A}v_j$ and $v'_R = Y^T(s - \omega\tilde{\lambda}_n v_j)$. Then

$$x_j = v_j - Z\hat{E}^{-1}v'_R.$$

To compute $v_R = \hat{E}^{-1}v'_R$, we solve the Galerkin system

$$\hat{E}v_R = v'_R, \tag{3.19}$$

and prolongate the solution to obtain $t = Zv_R$. Finally, we compute $x_j = s - t$. Algorithm 8 is a more detailed version of the two-level method.

To accelerate the convergence, the Galerkin system corresponding to the matrix $\hat{E}$ can be solved using a projection preconditioner. Using this technique recursively leads to a multi-

Data: A matrix $A \in \mathbb{C}^{N\times N}$, and vectors $b, x_0 \in \mathbb{C}^N$, an estimate for the largest eigenvalue $\tilde{\lambda}_n \in \mathbb{C}$, and a scaling factor $\omega \in \mathbb{R}$.

begin

Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2, v_1 = r_0/\beta$.;

for $j = 1, 2, \ldots k$ **do**

$s := AM^{-1}v_j$.;

Restriction: $v'_R := Y^T(s - \omega\tilde{\lambda}_n v_j)$.;

Solve for v_R: $\hat{E}v_R = v'_R$.;

Prolongation: $t := Zv_R$. ;

$x_j := v_j - t$.;

$w := AM^{-1}x_j$.;

for $i = 1, 2, \ldots, j$ **do**

$h_{ij} := (w, v_j)$;;

end

$w := w - \sum_{i=1}^{j} h_{ij}v_j$.;

$h_{j+1j} := \|w\|$. ;

$v_{j+1} := w/h_{j+1j}$.;

end

Set $X_k := [x_1 \ldots x_k]$ and $\hat{H}_k = \{h_{ij}\}_{1\le i\le j+1; 1\le j\le k}$.;

Compute $y_k := \operatorname{argmin}_y \|\beta e_1 - H_k y\|$ and $x_k := x_0 + M^{-1}X_k y_k$;

end

Algorithm 8: FGMRES preconditioned by M and $\hat{Q}_N$

level Krylov (MK) method.

We will now focus on the MK method with approximate Galerkin matrices, following [15]. When applied to Helmholtz problems, the preconditioner M is the shifted Laplace preconditioner, which is inverted using one multigrid iteration. Therefore, M^{-1} is not explicitly available for building the Galerkin matrix $\hat{E} = Y^T A M^{-1} Z$. Using the approximation $M^{-1} \approx Z(Y^T M Z)^{-1} Y^T$ we obtain the approximate Galerkin matrix

$$\begin{aligned}\hat{E} &= Y^T \hat{A} Z = Y^T A M^{-1} Z \\ &\approx Y^T A Z (Y^T M Z)^{-1} Y^T Z = A_H M_H^{-1} B_H\end{aligned}$$

where $A_H = Y^T A Z$, $M_H = Y^T M Z$, $B_H = Y^T Z$ are the Galerkin matrices associated to A, M and I. The Galerkin system (3.19) now reads as

$$A_H M_H^{-1} B_H v_R = v_R'. \tag{3.20}$$

Further, a projection preconditioner can be applied to this system. To describe the full multilevel method, we update our notation and let $A^{(1)} = A$, $M^{(1)} = M$, $Z^{(1,2)} = Z$ and $Y^{(1,2)} = Y$. With this notation, we have

$$\hat{A}^{(2)} = A^{(2)} {M^{(2)}}^{-1} B^{(2)},$$

where

$$\begin{aligned}A^{(2)} &= {Y^{(1,2)}}^T A^{(1)} Z^{(1,2)} \\ M^{(2)} &= {Y^{(1,2)}}^T M^{(1)} Z^{(1,2)} \\ B^{(2)} &= {Y^{(1,2)}}^T I^{(1)} Z^{(1,2)}.\end{aligned}$$

The matrix $\hat{A}^{(2)}$ is the second level Galerkin matrix associated with $\hat{A}^{(1)} = A^{(1)} {M^{(1)}}^{-1}$. If $A^{(2)}$ is sufficiently small, the system

$$\hat{A}^{(2)} v_R^{(2)} = (v_R')^{(2)}$$

can be solved exactly, otherwise it is solved using a Krylov subspace method. In the latter case a shift operator can be applied to $\hat{A}^{(2)}$, namely

$$\hat{Q}_N^{(2)} = I - Z^{(2,3)} {\hat{A}^{(3)}}^{-1} {Y^{(2,3)}}^T \hat{A}^{(2)} + \omega^{(2)} \tilde{\lambda}_n^{(2)} Z^{(2,3)} {\hat{A}^{(3)}}^{-1} {Y^{(2,3)}}^T,$$

where $\hat{A}^{(3)} = {Y^{(2,3)}}^T \hat{A}^{(2)} Z^{(2,3)}$, and the resulting linear system

$$A^{(2)} {M^{(2)}}^{-1} B^{(2)} \hat{Q}_N^{(2)} \tilde{v}_R^{(2)} = \hat{A}^{(2)} \hat{Q}_N^{(2)} \tilde{v}_R^{(2)} = (v_R')^{(2)},$$

where $v_R^{(2)} = \hat{Q}_N^2 \tilde{v}_R^{(2)}$, is solved by a Krylov method. If the same procedure is used for the system corresponding to $\hat{A}^{(3)}$ we arrive at the multilevel Krylov method.

Suppose that we use m levels, so that at level $m-1$ the Galerkin problem is sufficiently small to be solved exactly. Algorithm 9 (algorithm 1 in ([15]) implements the multilevel Krylov method.

3.3.3 The Multilevel Krylov Multigrid Method

We close this chapter with a discussion of the multilevel Krylov multigrid method. The MK method presented before requires several preconditioner solves that can be done using multigrid, as it is the case when the method is combined with the CSL to solve Helmholtz problems. Consider a sequence of grids $\Omega_1 \supset \Omega_2 \cdots \supset \Omega_m$, with corresponding intergrid transfer operators

$$
\begin{aligned}
I_j^{j+1} &: \mathcal{G}(\Omega_j) \to \mathcal{G}(\Omega_{j+1}) \quad \text{(restriction)} \\
I_{j+1}^{j} &: \mathcal{G}(\Omega_{j+1}) \to \mathcal{G}(\Omega_j) \quad \text{(prolongation)}
\end{aligned}
$$

In this setting, if a linear system corresponding to $M = M_{MG}^{(1)}$ is to be solved on the finest grid Ω^1 using multigrid, the Galerkin matrices associated with the coarse grid correction steps are

$$
M_{MG}^{(j)} = I_{(j)}^{(j+1)} M_{MG}^{(j+1)} I_{(j)}^{(j+1)}
$$

For the multigrid method, the matrices I_j^{j+1}, I_{j+1}^{j} should be chosen to represent accurate processes of restriction and interpolation of smooth functions. On the other hand, the only requirement for transfer operators $(Z^{(j,j+1)}, Y^{(j,j+1)})$ in the MK method is that they have full rank. Since the multigrid transfer operators satisfy this requirement, they can be incorporated into the MK method, leading to a multilevel Krylov multigrid (MKMG) method. Hence, in the MKMG method we have

$$
M^{(j)} = M_{MG}^{(j)}.
$$

If m levels are used for the MK method, and in level j we use a multigrid V-cycle (algorithm 5, section 2.4) with $m-j$ levels for the preconditioner solve, we obtain a method that can be schematically represented as in the figure.

Initialization:;
For $j = 1$, set $A^{(1)} = A$, $M^{(1)} = M$, $B^{(1)} = 1$, construct $Z^{(1,2)}$ and choose $\tilde{\lambda}_n^{(1)}$ and $\omega^{(1)}$. With this information, $\hat{A}^{(1)} = A^{(1)}{M^{(1)}}^{-1}$ and $\hat{Q}_N$ are in principle determined.;
For $j = 2, \ldots, m$, choose $Z^{(j-1,j)}$ and $Y^{(j-1,j)}$ and compute

$$A^{(j)} = {Y^{(j-1,j)}}^T A^{(j-1)} Z^{(j-1,j)},$$
$$M^{(j)} = {Y^{(j-1,j)}}^T M^{(1)} Z^{(j-1,j)},$$
$$B^{(j)} = {Y^{(j-1,j)}}^T I^{(1)} Z^{(j-1,j)},$$

which define

$$\hat{A}^{(j)} = A^{(j)}{M^{(j)}}^{-1} B^{(j)}$$

For $j = 2, \ldots, m-1$, set $\omega^{(j)}$ and $\tilde{\lambda}_n^{(j)}$, and define

$$\hat{Q}_N^{(j)} = I - Z^{(j-1,j)}{\hat{A}^{(j)}}^{-1}{Y^{(j-1,j)}}^T(\hat{A}^{(j-1)} - \omega^{(j)}\tilde{\lambda}_n^{(j)} I)$$

Iteration phase:;
begin;
$j = 1$;
Solve $A^{(1)}{M^{(1)}}^{-1}\tilde{u}^{(1)} = b$, $u^{(1)} = {M^{(1)}}^{-1}\tilde{u}^{(1)}$ with Krylov iterations by computing;
$v_M^{(1)} = {M^{(1)}}^{-1} v^{(1)}$;
$s^{(1)} = A^{(1)} v_M^{(1)}$;
$t^{(1)} = s^{(1)} - \omega^{(1)}\tilde{\lambda}_n^{(1)} v^{(1)}$;
Restriction: $(v_R')^{(2)} = {Y^{(1,2)}}^T t^{(1)}$;
if $j = m$ **then**;
$\tilde{v}^{(m)} = {\hat{A}^{(m)}}^{-1}\hat{v}^{(m)}$;
else;
$j = 2$;
Solve $A^{(2)}{M^{(2)}}^{-1}B^{(2)}v_R^{(2)} = (v_R')^{(2)}$ with Krylov iterations by computing;
$v_M^{(2)} = {M^{(1)}}^{-1}B^{(2)}v^{(2)}$;
$s^{(2)} = A^{(2)}v_M^{(2)}$;
$t^{(2)} = s^{(2)} - \omega^{(2)}\tilde{\lambda}_n^{(2)}v^{(2)}$;
Restriction: $(v_R')^{(3)} = Y^{(2,3)T}t^{(2)}$;
if $j = m$ **then**;
$v_R^{(m)} = {\hat{A}^{(m)}}^{-1}(v_R')^{(m)}$;
else;
$j = 3$;
Solve $A^{(3)}{M^{(3)}}^{-1}B^{(3)}v_R^{(3)} = (v_R')^{(3)}$;
$\cdots$;
Interpolation: $v_I^{(2)} = Z^{(2,3)}v_R^{(3)}$;
$q^{(2)} = v^{(2)} - v_I^{(2)}$;
$w^{(2)} = {M^{(2)}}^{-1}B^{(2)}q^{(2)}$;
$p^{(2)} = A^{(2)}w^{(2)}$;
Interpolation: $v_I^{(1)} = Z^{(1,2)}v_R^{(2)}$;
$q^{(1)} = v^{(1)} - v_I^{(1)}$;
$w^{(1)} = {M^{(1)}}^{-1}q^{(1)}$;
$p^{(2)} = A^{(1)}w^{(1)}$;
end;

Algorithm 9: Multilevel Krylov method with approximate Galerkin matrices

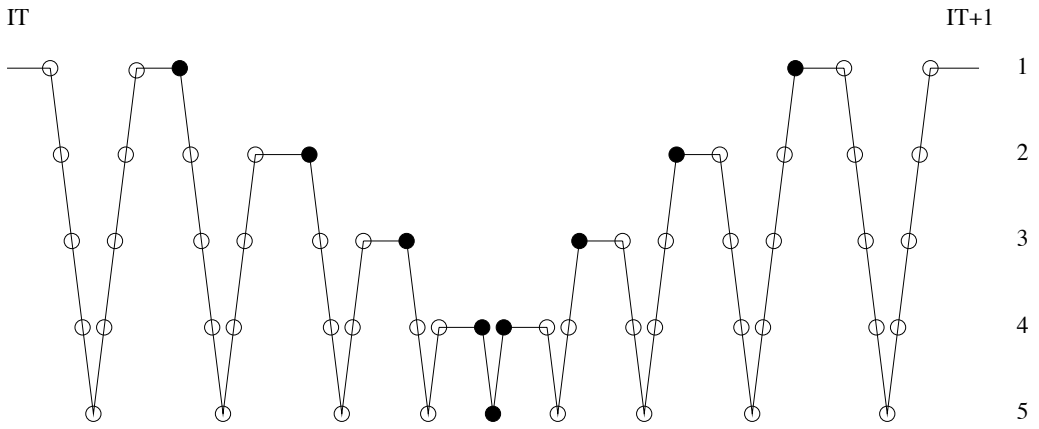


Figure 3.7: Multilevel Krylov-multigrid cycle with $m = 5$. • Multilevel Krylov step. ◦: Multigrid Step.

Chapter 4

Fourier Analysis of the Preconditioned 1-D Helmholtz Equation

In this chapter we return to the Helmholtz problem on $(0,1)$ with homogenous Dirichlet boundary conditions

$$-\frac{d^2u}{dx^2} - k^2u = f(x), \;\; x \in (0,1) \tag{4.1}$$

$$u(0) = u(1) = 0, \tag{4.2}$$

where $k > 0$, f and u are the wavenumber, source function and unknown field respectively. We will study two variants of the two-level version of the MK method based on multigrid. Our goal is to obtain a block diagonal decomposition of the two preconditioned systems, together with explicit formulas for its eigenvalues, to be able to establish the effect of the distinct choices of the parameters.

Let us recall the discretized version of our model problem from section (3.1). A finite-difference approximation of equation (4.1) on a uniform mesh Ω_h with gridsize $h = 1/n$, leads, after elimination of the boundary conditions, to a linear system of equations of the form

$$A_h u_h = f_h, \tag{4.3}$$

where

$$A_h = -\Delta_h - k^2 I.$$

The discrete complex shifted Laplace (CSL) operator is given by

$$M_{h,(\beta_1,\beta_2)} = \Delta_h - k^2(\beta_1 - i\beta_2)I, \tag{4.4}$$

so the Helmholtz system preconditioned by the CSL is

$$\hat{A}_{h,(\beta_1,\beta_2)} = M^{-1}_{h,(\beta_1,\beta_2)} A_h. \tag{4.5}$$

We assume that $n = 2m$ for some integer $m > 1$ and consider the coarse grid $\Omega_H \subset \Omega_h$, where $H = 2h$. Let I_h^H and I_H^h be two-grid transfer operators (we will give and an explicit choice for these operators later). We will analyze two different preconditioned Helmholtz systems based on two-grid deflation. The first variant is proposed in [31]:

$$M^{-1}_{h,(\beta_1,\beta_2)} P_{h,H} A_h \tag{4.6}$$

where $P_{h,H}$ the deflation preconditioner from (3.12) corresponding to A_h, $Z = I_H^h$, $Y = I_h^H$. The second variant is the two-level version of MKMG from [15]:

$$\hat{P}_{h,H,(\beta_1,\beta_2)} \hat{A}_{h,(\beta_1,\beta_2)}, \tag{4.7}$$

with $\hat{P}_{h,H}$ the deflation preconditioner from (3.15), corresponding to $\hat{A}_{h,(\beta_1,\beta_2)}$, $Z = I_H^h$, and $Y = I_h^H$.

4.1 First Variant of the Two-level Preconditioner

We choose I_h^H and I_H^h to be the restriction and prolongation operators given by full weighting and bilinear interpolation. Since the dimension of the space of grid functions on Ω_H is $n/2-1$, we have that

$$I_h^H : \mathbb{R}^{n-1} \to \mathbb{R}^{n/2-1}$$
$$I_H^h : \mathbb{R}^{n/2-1} \to \mathbb{R}^{n-1},$$

In stencil notation, the restriction operator can be represented as

$$[I_h^H] = \frac{1}{4}[\ 1 \quad 2 \quad 1\ ], \tag{4.8}$$

with this choice we have $I_H^h = (I_h^H)^T$. The deflation operator $P_{h,H} : \mathbb{R}^{n-1} \to \mathbb{R}^{n/2-1}$ is defined as

$$P_{h,H} = I - A_h Q_{h,H}, \text{ where } Q_{h,H} = I_h^H A_H^{-1} I_h^H \text{ and } A_H = I_h^H A_h I_H^h. \tag{4.9}$$

Since A_h and $Q_{h,H}$ are symmetric

$$P_{h,H}^T = I_h - Q_{h,H}A_h. \tag{4.10}$$

The operator resulting from applying the CSL preconditioner to the multigrid deflated system is

$$M_{h,(\beta_1,\beta_2)}^{-1})P_{h,H}A_h. \tag{4.11}$$

We will, however, focus our analysis on the operator

$$B_{h,H,(\beta_1,\beta_2)} = P_{h,H}^T M_{h,(\beta_1,\beta_2)}^{-1} A_h = P_{h,H}^T \hat{A}_{h,(\beta_1,\beta_2)}, \tag{4.12}$$

as a consequence of our analysis we will see that the operators in (4.11) and (4.12) share the same spectrum

$$\sigma(M_{h,(\beta_1,\beta_2)}^{-1}P_{h,H}A_h) = \sigma(P_{h,H}^T \hat{A}_{h,(\beta_1,\beta_2)}),$$

since both can be block-diagonalized in the basis of Fourier modes. We will obtain a block diagonalization of $B_{h,H,(\beta_1,\beta_2)}$ and an explicit formula for its eigenvalues. We recall the eigenpairs $(\lambda^l(-\Delta_h), v^l)$ of the discrete Laplace operator $-\Delta_h$ from section 3.1:

$$\lambda^l(-\Delta_h) = (2 - 2\cos(l\pi h)), \tag{4.13}$$

$$v^l = \sin(l\pi\underline{x}) = [\sin(l\pi hj)]_{j=1}^{n-1}, \quad \text{for } l = 1, \ldots, n-1. \tag{4.14}$$

The $n-1$ eigenvalues of the discrete Helmholtz operator A_h and the CSL preconditioner, corresponding to the same eigenvectors, are given by

$$\begin{aligned} \lambda^l(A_h) &= \frac{1}{h^2}(2 - 2c_l - \kappa^2), \\ \lambda^l(M_h) &= \frac{1}{h^2}(2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)), \end{aligned} \tag{4.15}$$

for $l = 1, \ldots, n-1$, using the notation $c_l = \cos(l\pi h)$. In consequence, the eigenvalues of the operator $\hat{A}_{h,(\beta_1,\beta_2)}$ are

$$\lambda^l(\hat{A}_{h,(\beta_1,\beta_2)}) = \frac{2 - 2c_l - \kappa^2}{2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)} \ \ (l = 1, 2, \ldots, n-1). \tag{4.16}$$

A block decomposition of the deflation operator $P_{h,H}^T$ can be obtained by reordering the set of eigenvectors in the form

$$V_h = \{(v^l, v^{n-l}) \,|\, l = 1, \ldots, n/2 - 1\} \cup \{v^{n/2}\}.$$

The deflation operator has a block representation in the basis V_h that we write as

$$P_{h,H}^T = [(P_{h,H}^T)^l]_{1\leq l\leq n/2}$$

Where the $l-$th block is given by

$$(P_{h,H}^T)^l = I - (I_H^h)^l (A_H^l)^{-1} (I_h^H)^l A_h^l.$$

The blocks of the prolongation operator are given by

$$(I_H^h)^l = \frac{1}{2}\begin{bmatrix} (1+c_l) \\ -(1-c_l) \end{bmatrix}, \tag{4.17}$$

for $1 \leq l \leq n/2-1$, and

$$(I_H^h)^{n/2} = 0.$$

We can now obtain the 1×1 blocks of the Galerkin coarsening matrix A_H:

$$A_H^l = (I_h^H)^l (A_h)^l (I_h^H)^l = \frac{1}{2h^2}[2(1-c_l^2) - \kappa^2(1+c_l^2)],$$

for $1 \leq l \leq n/2$ and $(A_H)^{n/2} = 0$. A block decomposition of the operator $Q_{h,H}$ is given by

$$(Q_{h,H})^l = (I_{h,H})^l (A_H^{-1})^l)(I_h^H)^l.$$

More explicitly:

$$\begin{aligned}(Q_{h,H})^l &= \left[\frac{2h^2}{2(1-c_l^2)-\kappa^2(1+c_l^2)}\right]\frac{1}{4}\begin{bmatrix} 1+c_l \\ -(1-c_l)\end{bmatrix}\begin{bmatrix} 1+c_l & -(1-c_l)\end{bmatrix} \\ &= \left[\frac{h^2}{4(1-c_l^2)-2k^2(1+c_l^2)}\right]\begin{bmatrix} (1+c_l)^2 & -(1-c_l^2), \\ -(1-c_l^2) & (1-c_l)^2\end{bmatrix}.\end{aligned}$$

From which we get

$$\begin{aligned}(Q_{h,H})^l(A_h)^l &= \left[\frac{h^2}{4(1-c_l^2)-2k^2(1+c_l^2)}\right]\begin{bmatrix} (1+c_l)^2 & -(1-c_l^2) \\ -(1-c_l^2) & (1-c_l)^2\end{bmatrix}\frac{1}{h^2}\begin{bmatrix} 2(1-c_l)-\kappa^2 & 0 \\ 0 & 2(1+c_l)-\kappa^2\end{bmatrix} \\ &= \left[\frac{1}{4(1-c_l^2)-2k^2(1+c_l^2)}\right]\begin{bmatrix} (1+c_l)^2(2-2c_l-\kappa^2) & (c_l^2-1)(2+2c_l-\kappa^2) \\ (c_l^2-1)(2-2c_l-\kappa^2) & (c_l-1)^2(2+2c_l-\kappa^2)\end{bmatrix}.\end{aligned}$$

The blocks of $P_{h,H}^T$ can now be computed:

$$\begin{aligned}(P_{h,H}^l)^T &= I - Q_{h,H}^l A_h^l \\ &= \frac{1}{C}\left[CI - CQ_{h,H}^l A_h^l\right] \\ &= \frac{1}{C}\begin{bmatrix} (c_l-1)^2(2+2c_l-\kappa^2)) & (1-c_l^2)(2+2c_l-\kappa^2) \\ (c_l^2-1)(-2+2c_l-\kappa^2) & (1+c_l)^2(-2+2c_l-\kappa^2)\end{bmatrix}.\end{aligned}$$

where $C = 4(1-c_l^2) - 2k^2(1+c_l^2)$. The l-th block of the operator $B_{h,H,(\beta_1,\beta_2)}$ is given by

$$B_{h,H,(\beta_1,\beta_2)}^l = \frac{1}{C}\begin{bmatrix} (c_l-1)^2(2+2c_l-\kappa^2)) & (1-c_l^2)(2+2c_l-\kappa^2) \\ (c_l^2-1)(-2+2c_l-\kappa^2) & (1+c_l)^2(-2+2c_l-\kappa^2)\end{bmatrix}\begin{bmatrix} \frac{2-2c_l-\kappa^2}{2-2c_l-\kappa^2(\beta_1-i\beta_2)} & 0 \\ 0 & \frac{2-2c_l-\kappa^2}{2+2c_l+\kappa^2(\beta_1-i\beta_2)})\end{bmatrix}$$

Each block $CB_{h,H,(\beta_1,\beta_2)}^l$ has a zero eigenvalue and another eigenvalue given by the formula

$$\lambda^l(CB_{h,H,(\beta_1,\beta_2)}) = \frac{2(2(c^2-1)+\kappa^2(\beta_1-i\beta_2)(1+c^2))(4c^2-(\kappa^2-2)^2)}{4c^2-(\kappa^2(\beta_1-i\beta_2)-2)^2}, \quad l = 1,\ldots,n/2-1.$$

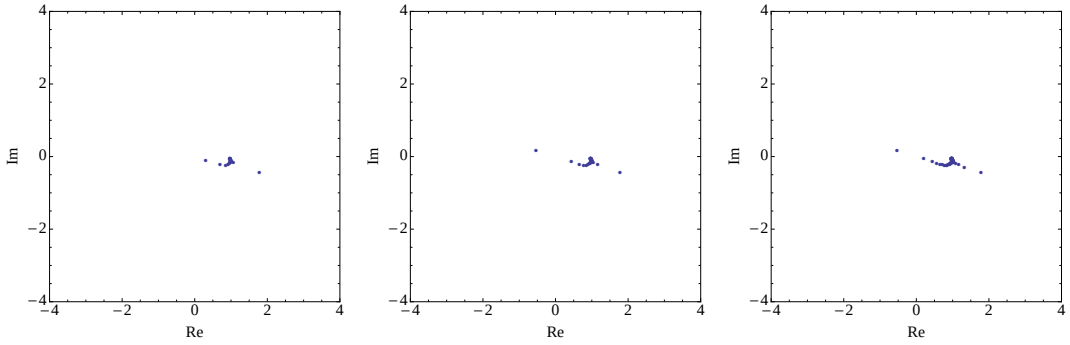
Recalling that $C = 4(1-c_l^2) - 2k^2(1+c_l^2)$, we have the following expression for the nonzero eigenvalues of $B_{h,H,(\beta_1,\beta_2)}$

$$\lambda^l(B_{h,H,(\beta_1,\beta_2)}) = \frac{D(c_l,\kappa,\beta_1,\beta_2)+iE(c_l,\kappa,\beta_1,\beta_2)}{F(c_l,\kappa,\beta_1,\beta_2)+iG(c_l,\kappa,\beta_1,\beta_2)}. \tag{4.18}$$

where

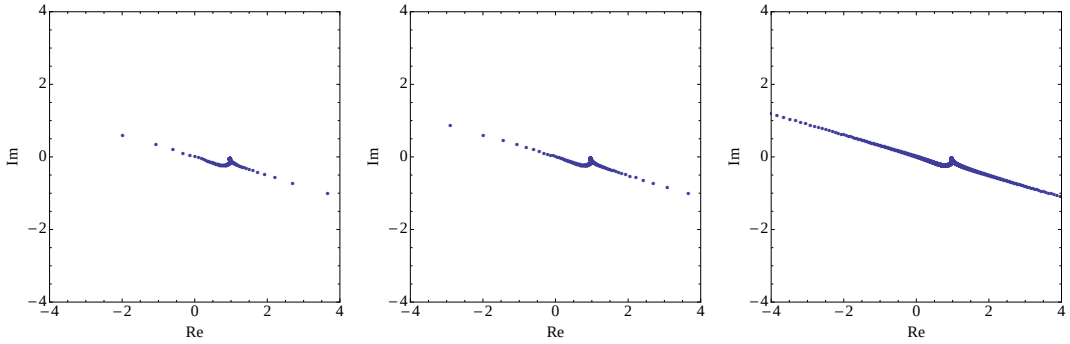
$$\begin{aligned}D(c_l,\kappa,\beta_1,\beta_2) &= [2+2c_l-\kappa^2][2-2c_l-\kappa^2][2-\beta_1\kappa^2-c_l^2(2+\beta_1k^2)], \\ E(c_l,\kappa,\beta_1,\beta_2) &= -\beta_2\kappa^2[2+2c_l-\kappa^2][2-2c_l-\kappa^2][1+c_l^2], \\ F(c_l,\kappa,\beta_1,\beta_2) &= [c_l^2(2+\kappa^2)+\kappa^2-2][4c_l^2+2\beta_1(\kappa^2+\beta_2\kappa^4)+\beta_1^2+\beta_2^2-4], \\ G(c_l,\kappa,\beta_1,\beta_2) &= 2\beta_2\kappa^2[\beta_1\kappa^2-2][2-\kappa^2-c^2(2+\kappa^2)].\end{aligned} \tag{4.19}$$

The spectrum of the system $B_{h,H(\beta_1,\beta_2)}$ is shown in figures 4.1 and 4.2 for several values of the shift parameters β_1, β_2 and κ. One can see that most of the eigenvalues are clustered around $(1,0)$, with some small eigenvalues and some negative eigenvalues as well. For high wavenumbers, the eigenvalues move closer to zero and to the left half of the complex plane, and large eigenvalues appear. Also, the spectrum is more clustered for smaller κ, hence for a larger number of gridpoints per wavelength. This is shown in figure 4.3.



(a) $k = 30$ (b) $k = 50$ (c) $k = 100$

Figure 4.1: Nonzero part of the spectrum of $B_{h,H,(1,0.5)}$ for $\kappa = 0.628$



(a) $k = 500$ (b) $k = 1000$ (c) $k = 10000$

Figure 4.2: Nonzero part of the spectrum of $B_{h,H,(1,0.5)}$ for $\kappa = 0.628$

The formula for the eigenvalues in (4.18) can also be written in the form

$$\lambda^l(B_{h,H}) = \frac{1}{F^2+G^2}\left[(DF+EG)+i(EF-DG)\right], \tag{4.20}$$

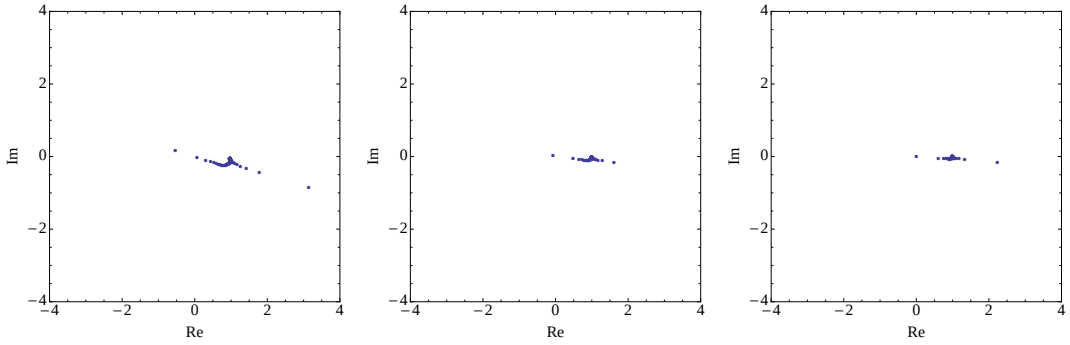
from which we can obtain the real and imaginary parts of the nonzero eigenvalues

$$\mathrm{Re}[\lambda^l(B_{h,H})] = \frac{(DF+EG)}{F^2+G^2}, \tag{4.21}$$

$$\mathrm{Im}[\lambda^l(B_{h,H})] = \frac{EF-DG}{F^2+G^2}. \tag{4.22}$$

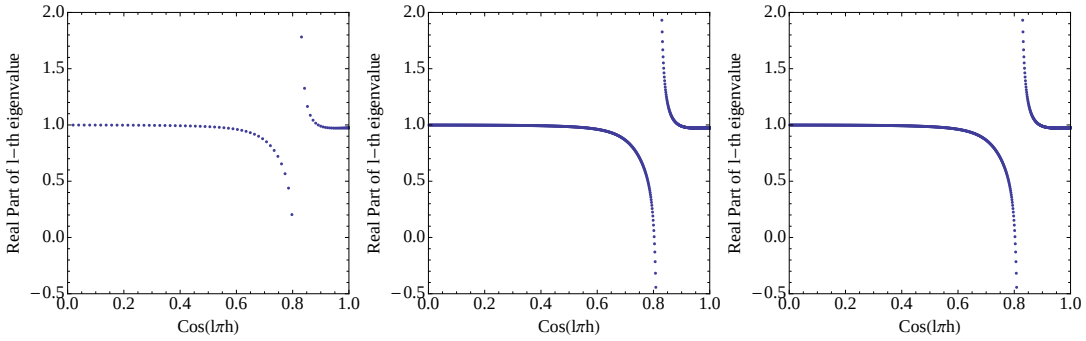
Figure 4.4 and 4.5 show some plots of the real and imaginary parts of the eigenvalues for the preconditioned system $B_{h,H,(1,0.5)}$ and different wavenumbers.

For fixed parameters κ, β_1, β_2, the modulus of the eigenvalues of the preconditioned system



(a) $\kappa = 0.628$ (b) $\kappa = 0.418$ (c) $\kappa = 0.314$

Figure 4.3: Nonzero part of the spectrum of $B_{h,H,(1,0.5)}$ for $k = 150$



(a) $k = 100$ (b) $k = 1000$ (c) $k = 1000$

Figure 4.4: Real part of the eigenvalues of $B_{h,H,(1,0.5)}$ for $\kappa = 0.628$

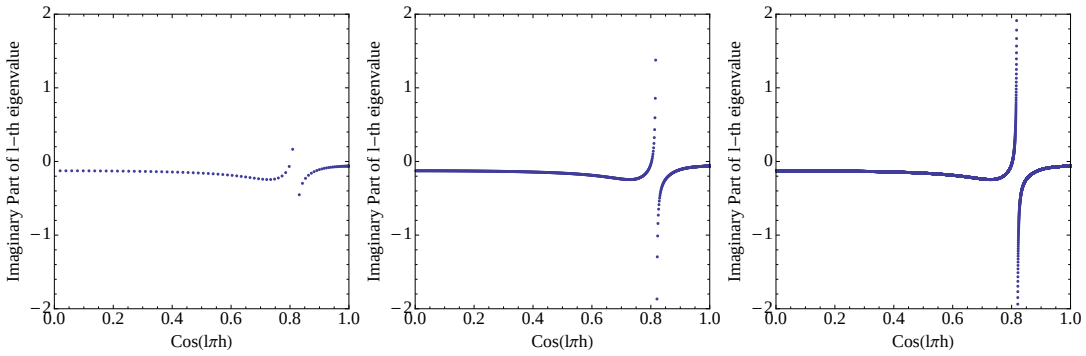
$B_{h,H,\beta}$ is given by $N(c_l)$, where N is the function

$$N_B(c) = \sqrt{\frac{D(c,\kappa,\beta_1,\beta_2)^2 + E(c,\kappa,\beta_1,\beta_2)^2}{F(c,\kappa,\beta_1,\beta_2)^2 + G(c,\kappa,\beta_1,\beta_2)^2}}.$$

This function has a unique zero at $C_0 = 1 - \frac{\kappa^2}{2}$, which is the unique common zero of the functions D and E, and an infinite discontinuity at $C_\infty = \sqrt{\frac{2-\kappa^2}{2+\kappa^2}}$ corresponding to the common zero of F and G. We have then: $|\lambda^l(B_{h,H,(\beta_1,\beta_2)})| \approx 0$ for $c_l \approx C_0$ and $|\lambda^l(B_{h,H,(\beta_1,\beta_2)})| \to \infty$ for $c_l \approx C_\infty$. Following these observations, we can study the dependence of the magnitude of the eigenvalues with respect to the wavenumber. Given fixed parameters κ, β_1 and β_2, for a wavenumber $k \in \mathbb{R}$ we denote the maximum and minimum (in magnitude) nonzero eigenvalues of the preconditioned system $B_{h,H,(\beta_1,\beta_2)}$ by

$$\lambda_B^{\max}(k) = \max\{|\lambda^l(B_{h,H,\beta})| : 1 \le l \le n/2 - 1\}, \tag{4.23}$$

$$\lambda_B^{\min}(k) = \min\{|\lambda^l(B_{h,H,\beta})| : 1 \le l \le n/2 - 1\}, \tag{4.24}$$



(a) $k = 100$ (b) $k = 1000$ (c) $k = 10000$

Figure 4.5: Imaginary part of the eigenvalues of $B_{h,H,(1,0.5)}$ for $\kappa = 0.628$

where the number of points $n = n(k)$ of the discretization satisfies the restriction imposed by κ, that is, $\kappa = kh = k/n$, and we have omitted that the system B depends of β_1, β_2 and κ for clarity. We have the following result.

Theorem 4.1. *Let κ, β_1, β_2 be fixed. Then the extreme eigenvalues* (4.23), (4.24) *of the preconditioned system $B_{h,H,(\beta_1,\beta_2)}$ satisfy:*

$$\lim_{k\to\infty} \lambda_B^{\max}(k) = \infty, \tag{4.25}$$

$$\lim_{k\to\infty} \lambda_B^{\min}(k) = 0. \tag{4.26}$$

Proof. We first show that that (4.25) holds. Consider the function N_B for the parameters κ, β_1, β_2. Clearly $N_B(c_l) = |\lambda^l(B_{h,H,(\beta_1,\beta_2)})|$ for $h = 1/n$ satisfying the restriction of κ and for every l with $1 \leq l \leq n/2 - 1$. From the expressions for F and G we can see that the function N_B has an infinite discontinuity at $C_\infty = \sqrt{\frac{2-\kappa^2}{2+\kappa^2}}$, and it is clear that $0 < C_\infty < 1$. Choose k sufficiently large so that $c_{n/2-1} < C_\infty < c_1$, with $n = n(k)$. This is posible since $c_1 = \cos(\pi/n) \to 1$ and $c_{n/2-1} = \cos(\pi/2 - \pi/n) \to 0$ as $n \to \infty$ (hence as $k \to \infty$). Since N_B is a rational function, we can choose n (and k accordingly) such that N_B is monotonically increasing on $[C_\infty - \frac{\pi}{n}, C_\infty)$ and monotonically decreasing on $(C_\infty, C_\infty + \frac{\pi}{n}]$. Now let l^* be such that $c_{l^*+1} < C_\infty < c_{l^*}$, with $1 < l^* < \frac{n}{2} - 1$. Since the function $f(x) = \cos(x)$ is Lipschitz continuous with Lipschitz constant $C = 1$ it follows that

$$\begin{aligned} |c_{l^*} - c_{l^*+1}| &= |\cos(l^*\,\pi\,h) - \cos((l^*+1)\,\pi\,h)| \\ &\leq |\pi\,h| = \pi/n. \end{aligned}$$

It is clear that $c_{l^*}, c_{l^*+1} \in [C_\infty - \frac{\pi}{n}, C_\infty + \frac{\pi}{n}]$ and we have

$$\begin{aligned}\lambda_B^{\max}(k) &= \max\{|\lambda^l(B_{h,H,(\beta_1,\beta_2)})| : 1 \leq l \leq n/2-1\} \\ &= \max\{N_B(c_l) : 1 \leq l \leq n/2-1\} = \max\{N_B(c_{l^*}), N_B(c_{l^*+1})\},\end{aligned}$$

that is, the eigenvalues $\lambda^l(B_{h,H,(\beta_1,\beta_2)})$ corresponding to the c_l closest to the infinite discontinuity of N_B have the largest modulus. Since

$$\max\{N_B(c_{l^*}), N_B(c_{l^*+1})\} \geq \max\{N_B(C_\infty - \pi/n), N_B(C_\infty + \pi/n)\},$$

and the right hand side can be made arbitrarily large by choosing n large enough (and k large enough), the conclusion follows. For the limit (4.26), a similar analysis can be done for C_0 and the function $1/N$. □

An immediate consequence is that for fixed parameters κ and $\beta = \beta_1 - i\beta_2$ the effective condition number

$$\mathrm{Cond}_B(k) = \frac{\lambda_B^{\max}(k)}{\lambda_B^{\min}(k)}$$

of the preconditioned Helmholtz system $B_{h,H,(\beta_1,\beta_2)}$ grows unbounded asymptotically with respect to the wavenumber. In comparison to the CSL system, in which the largest eigenvalue is bounded and the smallest eigenvalue goes to zero as the wavenumber increases, one should expect that this effect is more pronounced for the deflated system (see corollary 4.2).

Regarding the smallest nonzero eigenvalue of the preconditioned system $B_{h,H,(\beta_1,\beta_2)}$, in the proof of theorem (4.1) we have shown that, for sufficiently large k, we can approximate $\lambda_B^{\min}(k)$ by $N_B(c)$ with $c \approx C_0$. This also holds for the minimum eigenvalue $\lambda_{\hat{A}}^{\min}(k)$ of the CSL preconditioned system $\hat{A}$ and the corresponding function $N_{\hat{A}}(c)$, defined analogously from formula (4.16). Therefore, we can compare the minimum nonzero eigenvalues of each system by calculating the asymptotic ratio

$$L = \lim_{k\to\infty} \frac{\lambda_{\hat{A}}^{\min}(k)}{\lambda_B^{\min}(k)} = \lim_{c\to C_0} \frac{N_{\hat{A}}(c)}{N_B(c)}, \tag{4.27}$$

and obtain an estimate of the form

$$\lambda_{\hat{A}}^{\min}(k) \approx L\, \lambda_B^{\min}(k)$$

for large wavenumbers k. We would like to have $L < 1$ for at least some choice of the parameters β_1, β_2, κ, as this would imply that the deflation operator removes the small eigenvalues of the spectrum and does not shift the remaining eigenvalues closer to zero. A computation

shows that

$$L = \sqrt{\frac{M(\kappa, \beta_1, \beta_2)}{N(\kappa, \beta_1, \beta_2)}},$$

where

$$\begin{aligned} M(\kappa, \beta_1, \beta_2) &= 4\beta_2\kappa^4(\beta_1\kappa^2 - 2)^2 + [\beta_1^2 + \beta_2^2 + 2(\beta_1 - 1)\kappa^2 + (2\beta_1\beta_2)\kappa^4]^2 \\ N(\kappa, \beta_1, \beta_2) &= 4[(\beta_1 - 1)^2 + \beta_2^2)(4(\kappa^2 - 4)^2 + 4\beta_1(k - 2)(\kappa^4 - 4\kappa^2 + 8) \\ &\quad + (\beta_1^2 + \beta_2)^2(\kappa^4 - 4\kappa^2 + 8))]. \end{aligned}$$

Figure 4.6 shows that fixing $\beta_1 = 1$ we have that $L > 1$ only for very small values of β_2. For the standard choice of $\beta_2 = 0.5$ the the smallest eigenvalue of the deflated Helmholtz system is approximately 3 times larger (in modulus) than the smallest eigenvalue of the CSL preconditioned system. Choosing a larger β_2 leads to an improvement since it shifts the smallest eigenvalue further from the origin ($L \approx 0.15$). Note that a large shift β_2 is more convenient for the inversion of the preconditioner. The small values for which $L > 1$ are not used in practice because they are not suitable for multigrid convergence.

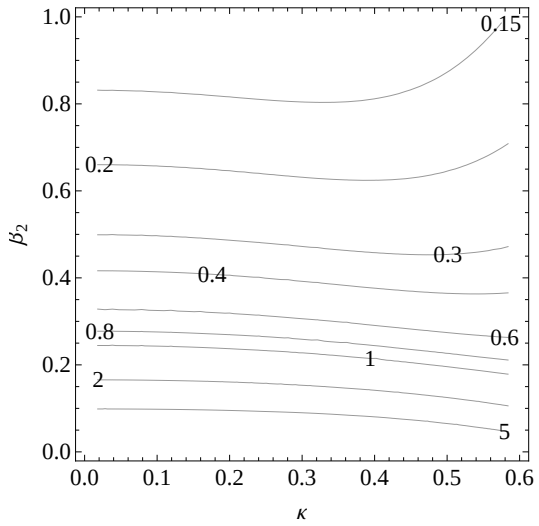


Figure 4.6: Asymptotic Ratio of Minimum Eigenvalues L for $\beta_1 = 1$

Given κ, β_1, β_2 fixed parameters, we denote by $\mathrm{Cond}_{\hat{A}}(k)$ the condition number of $\hat{A}_{h,\beta}$. We have the following corollary.

Corollary 4.2. *For all* κ, β_1, β_2

$$\mathrm{Cond}_{\hat{A}}(k) \leq \mathrm{Cond}_B(k) \text{ as } k \to \infty.$$

Proof. From the limits (4.25) and (4.27) we have, for k sufficiently large:

$$\lambda_{\hat{A}}^{\min}(k) \leq 1 \leq \lambda_B^{\max}(k) \left[\frac{\lambda_{\hat{A}}^{\min}(k)}{\lambda_B^{\min}(k)} \right].$$

Hence, as $k \to \infty$

$$\frac{\lambda_{\hat{A}}^{\max}(k)}{\lambda_{\hat{A}}^{\min}(k)} \leq \frac{\lambda_B^{\max}(k)}{\lambda_B^{\min}(k)}.$$

□

One can see from the plots that, for fixed parameters κ, β_1, β_2 and wavenumber k, the spectrum of B_{h,H,β_1,β_2} may contain negative eigenvalues. From formula (4.21) for the real part of the eigenvalues we have that

$$\lambda^l(B_{h,H,(\beta_1,\beta_2)}) < 0 \text{ if and only if } c_l \in (C_0, C_\infty),$$

where $C_0 = 1 - \frac{\kappa^2}{2}$ and $C_\infty = \sqrt{\frac{2-\kappa^2}{2+\kappa^2}}$, this implies that the number of negative eigenvalues is independent of the parameters β_1, β_2 and only depends on h and κ. For a fixed wavenumber k, we can study the dependence of the number of eigenvalues on the parameter κ, hence, on the number of points per wavelength p, since $\kappa = 2\pi/p$. Figure 4.7 shows that the fraction of negative eigenvalues increases for larger values of p, with approximately one third of the nonzero eigenvalues being negative for large values of p.

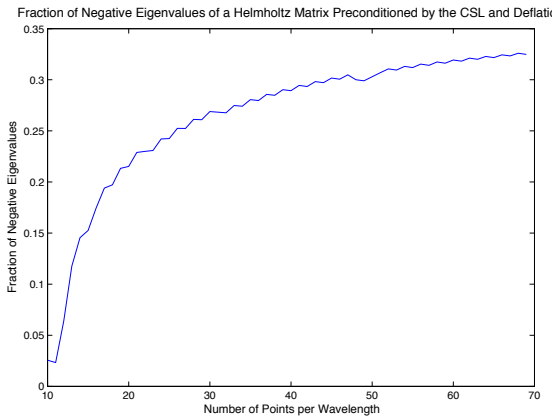


Figure 4.7: Fraction of negative eigenvalues of the preconditioned matrix $B_{h,H,(1,0.5)}$ for $k = 100$

4.2 Second Variant of the Two-Level Preconditioner

We now turn now our attention to the second variant of the two-grid preconditioner. Given a discrete Helmholtz system A_h and a discrete CSL operator $M_{h,(\beta_1,\beta_2)}$ corresponding to the shift parameters (β_1, β_2), let $\hat{A}_{h,(\beta_1,\beta_2)}$ be the Helmholtz operator preconditioned by the CSL, as in the previous section. We study here the two-grid deflation preconditioner built from $\hat{A}_{h,(\beta_1,\beta_2)}$, that is, the preconditioner $\hat{P}_{h,H,(\beta_1,\beta_2})$ given by

$$\hat{P}_{h,H,(\beta_1,\beta_2)} = I - \hat{A}_{h,(\beta 1,\beta_2)}\hat{Q}_{h,H,(\beta_1,\beta_2)}$$

where $\hat{Q}_{h,H,(\beta_1,\beta_2)} = I_H^h \hat{A}^{-1}_{H,(\beta_1,\beta_2)} I_h^H$ and $\hat{A}_{H,(\beta_1,\beta_2)} = I_h^H \hat{A}_{h,(\beta_1,\beta_2)} I_H^h$. The preconditioned system has the form

$$\hat{B}_{h,H,(\beta_1,\beta_2)} = \hat{P}_{h,H,(\beta_1,\beta_2)}\hat{A}_{h,(\beta_1,\beta_2)}.$$

We will obtain a block decomposition of $\hat{B}_{h,H,(\beta_1,\beta_2)}$ as in the previous section, from which we will be able to compute the eigenvalues of the preconditioned operator. Recall the reordered basis of Fourier modes

$$V_h = \{(v^l, v^{n-l}) \,|\, l = 1, \ldots, n/2 - 1\} \cup \{v^{n/2}\},$$

The preconditioning matrix has a block representation in this basis

$$\hat{P}_{h,H,(\beta_1,\beta_2)} = [\hat{P}^l_{h,H,(\beta_1,\beta_2)}]_{1\le l\le n/2}$$

where each block has the form

$$\hat{P}^l_{h,H,(\beta_1,\beta_2)} = I - \hat{A}^l_{h,(\beta_1,\beta_2)}(I_H^h)^l(\hat{A}^l_{H,(\beta_1,\beta_2)})^{-1}(I_h^H)^l.$$

Using the block representation of the restriction and prolongation operators I_h^H and I_H^h from (4.17) we compute the 1×1 blocks of the coarse-grid operator A_H:

$$\hat{A}^l_{H,(\beta_1,\beta_2)} = \frac{(c_l+1)^2(2-2c_l-\kappa^2)}{4(2-2c_l-(\beta_1-i\beta_2)\kappa^2)} + \frac{(c_l-1)^2(2-2c_l-\kappa^2)}{4(2+2c_l-(\beta_1-i\beta_2)\kappa^2)} \qquad (l = 1, \ldots, n/2-1)$$

and $A^{n/2}_{H,\beta} = 0$. After some computations, we get that the 2×2 blocks of the two-grid preconditioner are

$$\hat{P}^l_{h,H,(\beta_1,\beta_2)} = \frac{1}{Z}\begin{bmatrix} \frac{(c_l+1)^2(2+2c_l-\kappa^2)}{4(2+2c_l-(\beta_1-i\beta_2)\kappa^2)} & \frac{(1-c_l^2)(2-2c_l-\kappa^2)}{4(2-2c_l-(\beta_1-i\beta_2)\kappa^2)} \\ \frac{(1-c_l^2)(2+2c_l-\kappa^2)}{4(2+2c_l-(\beta_1-i\beta_2)\kappa^2)} & \frac{(1+c_l)^2(2-2c_l-\kappa^2)}{4(2-2c_l-(\beta_1-i\beta_2)\kappa^2)} \end{bmatrix},$$

where

$$Z = \frac{(c_l+1)^2(2-2c_l-\kappa^2)}{4(2-2c_l-(\beta_1-i\beta_2)\kappa^2)} + \frac{(c_l-1)^2(2-2c_l-\kappa^2)}{4(2+2c_l-(\beta_1-i\beta_2)\kappa^2)},$$

for $l = 1, \ldots, n-1$ and $\hat{P}^{n/2}_{h,H,(\beta_1,\beta_2)} = 1$. From this we have that the 2×2 blocks of the preconditioned system are given by

$$\hat{B}^l_{h,H,(\beta_1,\beta_2)} = \hat{P}^l_{h,H,(\beta_1,\beta_2)}\hat{A}^l_{h,(\beta_1,\beta_2)} = \frac{W}{YZ}\begin{bmatrix} (c_l-1)^2 & (c_l^2-1) \\ (c^2-1) & (1+c^2) \end{bmatrix},$$

where

$$\begin{aligned} W &= 4c_l^2 - (\kappa^2-2)^2, \\ Y &= 16c_l^2 + 4(2-(\beta_1-i\beta_2)\kappa^2)^2. \end{aligned}$$

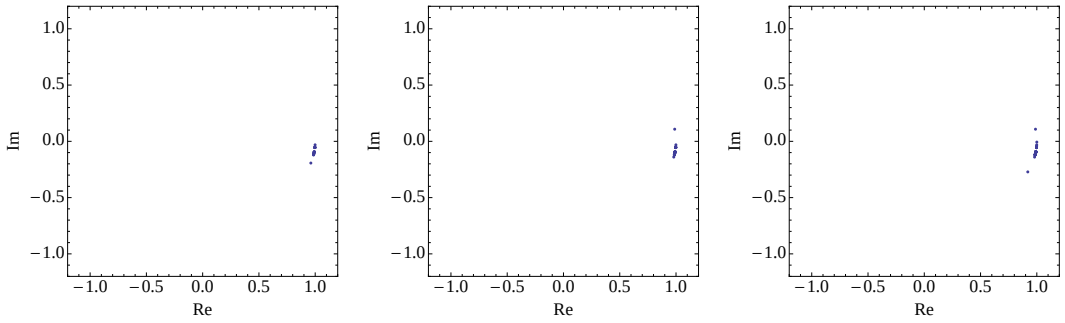
The eigenvalues of $\hat{B}^l_{h,H(\beta_1,\beta_2)}$ can now be computed. Each block has a zero eigenvalue and an eigenvalue given by the formula

$$\lambda^l(\hat{B}_{h,H,(\beta_1,\beta_2)}) = \frac{\hat{D}(c_l,\kappa)+i\hat{E}(c_l,\kappa)}{\hat{F}(c_l,\kappa)+i\hat{G}(c_l,\kappa)}. \;\; l = 1, \ldots, n/2-1.$$

where

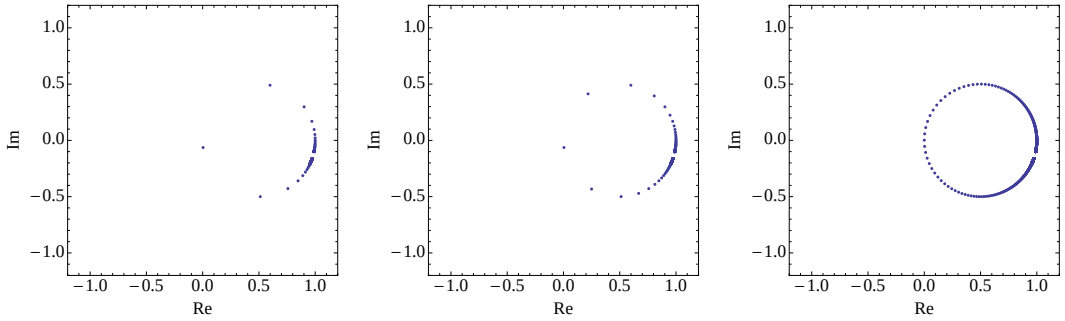
$$\begin{aligned} \hat{D}(c_l,\kappa) &= 4[1+c_l^2][2+2c_l-\kappa^2][2-2c_l-\kappa^2][4-4c_l^2-4\beta 1\kappa^2+\beta 1^2\kappa^4-\beta_2^2\kappa^4], \\ \hat{E}(c_l,\kappa) &= 8\beta_2\kappa^2[1+c_l^2][2+2c_l-\kappa^2][2-2c_l-\kappa^2][2-\beta_1\kappa^2], \\ \hat{F}(c_l,\kappa) &= \hat{F}_1(c_l,\kappa)+\hat{F}_2(c_l,\kappa), \\ \hat{F}_1(c_l,\kappa) &= 4[16c_l^6+(\kappa^2-2)(\beta_1\kappa^2-2)(4-4\beta_1\kappa^2+(\beta_1^2-3\beta_2^2)\kappa^4)] \\ &\quad -16c_l^4[4-2(3+\beta_1)\kappa^2+(\beta_1^2+\beta_1-\beta_2^2)\kappa^4], \\ \hat{F}_2(c_l,\kappa) &= 4c_l^2[16(\beta_1-1)\kappa^2+8(\beta_2^2+3\beta_1-\beta_1^2)\kappa^4 \\ &\quad +2[(\beta_1-5)\beta_1^2+(5-3\beta_1)\beta_2^2]\kappa^6+(\beta_1^3-3\beta_1\beta_2^2)\kappa^8-16], \\ \hat{G}(c_l,\kappa) &= \hat{G}_1(c_l,\kappa)+\hat{G}_2(c_l,\kappa), \\ \hat{G}_1(c_l,\kappa) &= 4\beta_2\kappa^2[4c_l^4[(2\beta_1+1)\kappa^2-2]+[2-\kappa^2][12-12\beta_1\kappa^2+(3\beta_1^2-\beta_2^2)\kappa^4]], \\ \hat{G}_2(c_l,\kappa) &= 4\beta_2\kappa^2c_l^2[8(2\beta_1-3)\kappa^2+2[(10-3\beta_1)\beta_1+\beta_2^2]\kappa^4+(\beta_2^2-3\beta_1^2)\kappa^6-16]. \end{aligned}$$

Figures 4.8 and 4.9 show the spectrum of $\hat{B}_{h,H,(\beta_1,\beta_2)}$ for several values of the parameters. Most of the eigenvalues are clustered around $(1,0)$ and there are only few eigenvalues close to the origin. This holds even for large wavenumbers ($k = 500$) and the minimum number of gridpoints per wavelength $p = 10$, corresponding to $\kappa = 0.628$.



(a) $k = 30$ (b) $k = 50$ (c) $k = 100$

Figure 4.8: Nonzero part of the spectrum of $\hat{B}_{h,H,(1,0.5)}$ for $\kappa = 0.628$



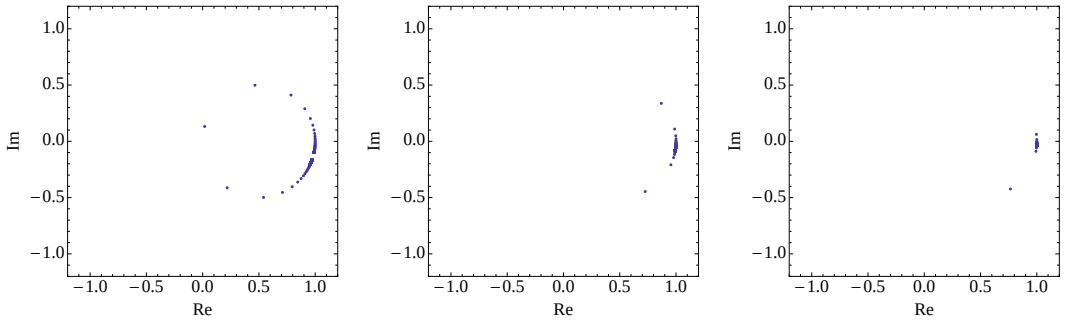
(a) $k = 500$ (b) $k = 1000$ (c) $k = 10000$

Figure 4.9: Nonzero part of the spectrum of $\hat{B}_{h,H,(1,0.5)}$ for $\kappa = 0.628$

Remarkably, the eigenvalues appear to lie on the same circles as the eigenvalues of the matrix $\hat{A}_{h,(\beta_1,\beta_2)}$. We were not able to establish this property analytically from the formulas, but varying the shift (β_1, β_2) strongly indicates that this holds. In consequence, there are no negative eigenvalues and the maximum eigenvalue is bounded in modulus by 1 for all wavenumbers.

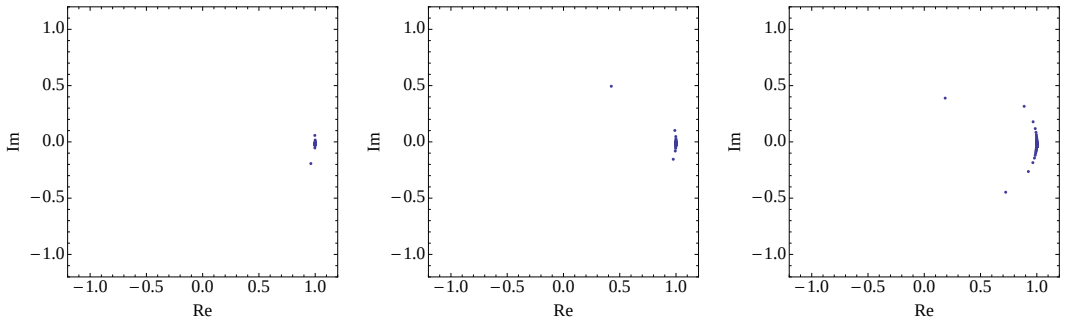
The plots in figure 4.10 show the dependence of the spectrum with respect to the parameter κ. They clearly show that a smaller κ leads to a more clustered spectrum. In figure 4.11 we show that using 20 gridpoints per wavelength suffices to obtain a clustered spectrum up to very large wavenumbers ($k = 5000$).

In figure 4.12 we show the dependence of the spectrum with respect to the imaginary shift β_2. The eigenvalues are more clustered and move away from zero for smaller values of β_2. This, however, cannot be exploited for Krylov acceleration, since it would affect the inversion of the CSL preconditioner using multigrid.



(a) $\kappa = 0.628$ (b) $\kappa = 0.418$ (c) $\kappa = 0.314$

Figure 4.10: Nonzero part of the spectrum of $\hat{B}_{h,H,(1,0.5)}$ for $k = 800$



(a) $k = 1000$ (b) $k = 2500$ (c) $k = 5000$

Figure 4.11: Nonzero part of the spectrum of $\hat{B}_{h,H,(1,0.5)}$ for $\kappa = 0.314$

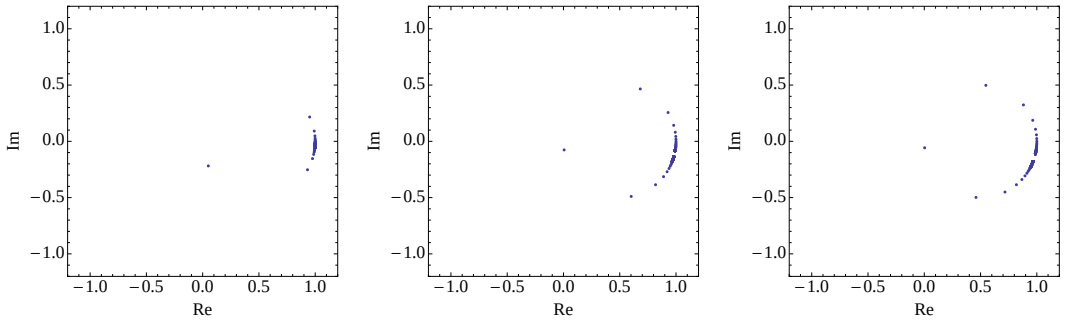
We compare now the minimum and maximum eigenvalues of the preconditioned matrix $\hat{A}_{h,(\beta_1,\beta_2)}$ and the deflated system $\hat{B}_{h,H,(\beta_1,\beta_2)}$. For fixed parameters κ and (β_1, β_2), let

$$\lambda_{\hat{B}}^{\min}(k) = \min\{|\lambda^l(\hat{B}_{h,H,(\beta_1,\beta_2)})| : 1 \leq l \leq n/2 - 1\},$$
$$\lambda_{\hat{B}}^{\max}(k) = \min\{|\lambda^l(\hat{B}_{h,H,(\beta_1,\beta_2)})| : 1 \leq l \leq n/2 - 1\}$$

Where the number of points $n = n(k)$ of the discretization satisfies the restriction imposed by κ, that is, $\kappa = kh = k/n$. From our previous discussion, assuming that the eigenvalues of $\hat{B}_{h,H,(\beta_1,\beta_2)}$ lie on the same circles we have $\lambda_{\hat{B}}^{\max}(k) < 1$. Now, consider the modulus function for the eigenvalues of $\hat{B}_{h,H,(\beta_1,\beta_2)}$ given by

$$N_{\hat{B}}(c) = \sqrt{\frac{\hat{D}(c,\kappa,\beta_1,\beta_2)^2 + \hat{E}(c,\kappa,\beta_1,\beta_2)^2}{\hat{F}(c,\kappa,\beta_1,\beta_2)^2 + \hat{G}(c,\kappa,\beta_1,\beta_2)^2}}.$$

We have that $N_{\hat{B}}(c_l) = |\lambda^l(\hat{B}_{h,H,(\beta_1,\beta_2)})|$ for $l = 1, \ldots, n/2 - 1$. This function has a unique zero at $C_0 = 1 - \frac{\kappa^2}{2}$, which is the unique common zero of the functions $\hat{D}$ and $\hat{E}$, hence



(a) $\beta_2 = 0.25$ (b) $\beta_2 = 0.75$ (c) $\beta_2 = 1$

Figure 4.12: Nonzero part of the spectrum of $\hat{B}_{h,H,(1,\beta_2)}$ for $k = 500$

$|\lambda^l(\hat{B}_{h,H,(\beta_1,\beta_2)})| \approx 0$ for $c_l \approx C_0$. Let $N_{\hat{A}}$ be the corresponding modulus function for the system $\hat{A}_{h,H,(\beta_1,\beta_2)}$, defined from (4.16). To compare the minimum eigenvalues of $\hat{A}_{h,H,\beta}$ and $\hat{B}_{h,H,\beta}$ we calculate, as in the previous section, the asymptotic ratio

$$\hat{L} = \lim_{k\to\infty} \frac{\lambda^{\min}_{\hat{A}}(k)}{\lambda^{\min}_{\hat{B}}(k)} = \lim_{c\to C_0} \frac{N_{\hat{A}}(c)}{N_{\hat{B}}(c)}$$

We obtain after some computations that

$$\hat{L} = \hat{L}(\kappa) = \frac{\kappa^4}{2(8 - 4\kappa^2 + \kappa^4)}. \tag{4.28}$$

Note that this asymptotic ratio is independent from the choice of the shift parameters (β_1, β_2), unlike the corresponding ratio for the first variant of the two-level preconditioner that we computed before. We show a plot of this quantity in figure 4.13.

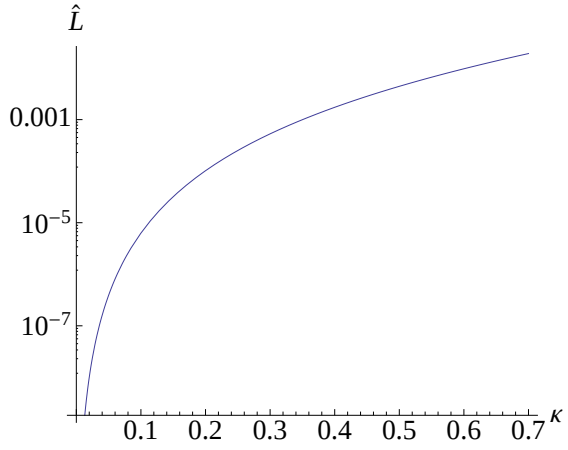


Figure 4.13: Asymptotic Ratio of Minimum Eigenvalues $\hat{L}$

We see that for 10 points per wavelength we have that $\hat{L}(0.628) = 0.011$, and so $\lambda^{\min}_{\hat{B}}(k)$ is approximately $1/0.011 \approx 100$ times larger than $\lambda^{\min}_{\hat{A}}(k)$ for large wavenumbers k. Since $\hat{L}$ decreases very rapidly, setting the number of gridpoints per wavelength to $p = 15$ leads to $\hat{L}(0.314) = 0.0006$. This explains why the clustering of the spectrum improves so dramatically for high wavenumbers when the number of gridpoints is increased (see figure 4.10).

Since $\hat{L} \ll 1$, we have that

$$\frac{1}{\lambda^{\min}_{\hat{B}}(k)} \ll \frac{1}{\lambda^{\min}_{\hat{A}}(k)}, \text{ for } k \text{ sufficiently large.}$$

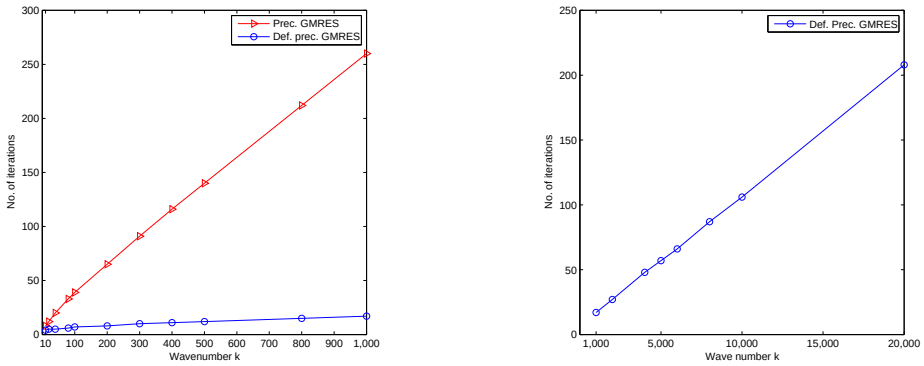
Using the approximation $\lambda^{\max}_{\hat{B}}(k) = \lambda^{\max}_{\hat{A}}(k) \approx 1$ for large wavenumbers, we obtain the property

$$\mathrm{Cond}_{\hat{B}}(k) \ll \mathrm{Cond}_{\hat{A}}(k) \text{ as } k \to \infty.$$

4.3 Numerical Results

In this section we review the numerical tests from [31] and [15] to validate our analysis.

We focus first on the results of the first variant of the two-grid preconditioner from [31]. Figure 4.14 shows the number of GMRES iterations depending on the wavenumber, for a 1-dimensional Helmholtz problem with Sommerfeld boundary conditions, preconditioned by the first variant of two-level deflation $P_{(h,H,(1,1))}$ (the shift is $(1,1)$) and 20 gridpoints per wavelength. The number of iterations is almost constant for the range of wavenumbers $10 \leq k \leq 800$ but grows linearly for high wavenumbers ($1000 \leq k \leq 20000$). These results are expected from our analysis on the asymptotic behavior of the extreme eigenvalues.



(a) $10 \leq k \leq 1000$ (b) $1000 \leq k \leq 20000$

Figure 4.14: Number of GMRES iterations to solve the one-dimensional Sommerfeld problem

Table **??** shows the number of iterations for a 2-D problem with constant wavenumber and Dirichlet boundary conditions, with and without the first variant of two-grid deflation. The lower diagonal part of the table shows grid resolutions of 20 gridpoints per wavelength or more. We can see that for a fixed wavenumber the number of iterations decreases as the resolution increases. This behaviour is also expected from our analysis, since we have shown that increasing the number of points per wavelength (hence decreasing κ) leads to a more clustered spectrum of the preconditioned system.

In table 4.1 we show the number of GMREs iterations for a 2-D Sommerfeld problem, with the standard shift $(1, 0.5)$ and the first variant of two-grid deflation. Similarly to the Dirichlet case, increasing the grid resolution leads to a decrease in the number of iterations. Note that, as we remarked earlier, the Sommerfeld problem requires fewer iterations to be solved.

We turn now our attention to the second variant of the two-grid deflation method from [15]. In this article, the results were obtained using a left preconditioner that shifts the small

Table 4.1: Number of GMRES iterations for a 2-D problem with Sommerfeld boundary conditions, for distinct wavenumbers and grid resolutions, using the CSL preconditioner $M_{h,1,0.5}$ with/without deflation

Size of the problem	$k = 10$	$k = 20$	$k = 30$	$k = 40$	$k = 50$	$k = 100$
32	**3/10**	8/17	17/31	35/50	52/80	13/14
64	3/10	**6/17**	10/30	17/47	24/63	221/552
96	3/10	5/17	**7/30**	11/46	15/62	209/220
28	3/10	5/17	6/30	**10/45**	11/62	90/196
160	3/10	4/17	5/30	8/45	**9/62**	65/194
320	2/10	3/17	4/30	5/45	6/61	**24/193**

eigenvalues to 1 (recall section 3.3.3). This two-level preconditioner has the form

$$\hat{Q}_{h,H,(\beta_1,\beta_2)} = I - I_H^h \hat{A}^{-1}_{H,(\beta_1,\beta_2)} I_h^H \hat{A}_{h,(\beta_1,\beta_2)} + I_H^h \hat{A}^{-1}_{H,(\beta_1,\beta_2)} I_h^H,$$

where $\hat{A}_{h,(\beta_1,\beta_2)} = A_h M^{-1}_{h,(\beta_1,\beta_2)}$. However, the results of our analysis carry over to this case in view of the spectral properties discussed in section 3.3. Table 4.2 shows that the number of

Table 4.2: Number of preconditioned GMRES iterations for a 1D Helmholtz problem. Equidistant grids equivalent to 30/15/8 gridpoints per wavelength are used.

	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
Standard	14/15/15	24/25/26	39/40/42	65/68/78	142/146/157
$\hat{Q}_{h,H,(\beta_1,\beta_2)}$, piece-wise constant	4/5/7	4/6/10	5/7/14	6/10/20	7/15/37
$\hat{Q}_{h,H,(\beta_1,\beta_2)}$, linear interpolation	3/4/5	3/4/7	3/4/8	3/5/10	3/5/12

GMRES iterations is nearly constant for the range of wavenumbers $\leq k \leq 500$. Here, the preconditioner M and the second level matrix are inverted exactly. Also, the use of piece-wise constant interpolation leads to a poorer performance compared to the linear interpolation that we have considered for the theoretical analysis. Note that the number of iterations decreases with an increasing number of gridpoints per wavelength, as expected.

Table 4.3 shows the result for the full multilevel MKMG method. Here the notation MKMG(6,6,2) indicates that 6 iterations of FGMRES are used at the second level , 6 at the third level and 2 for subsequent levels. Note that since it is important to solve accurately at the second level, the number of iterations there should be bigger. Compared to the results of the two-level version, more iterations are required, since at the second level the solution is no longer computed exactly. Still, the number of iterations grows only very mildly as the wavenumber increases. Last, table 4.4 shows the number of iterations of MKMG(8,2,2) and MKMG(8,2,1). The results are only slightly better in some of the cases than for MKMG(6,2,2).

Table 4.3: Number of GMRES iterations for 1D Helmholtz problems with constant wave number. p is the number of gridpoints per wavelength. Multilevel Krylov method with MKMG(6,2,2). MG is shown in parentheses.

p	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
15	11 (19)	11 (29)	11 (43)	15 (66)	25 (138)
30	9 (18)	11 (28)	12 (42)	14 (68)	22 (136)
60	9 (18)	9 (28)	12 (43)	12 (68)	19 (141)

Table 4.4: Number of GMRES iterations for 1D Helmholtz problems with constant wave number. p is the number of grid points per wavelength'. Multilevel Krylov method with MKMG(8,2,2) and MKMG(8,2,1) (in parentheses).

p	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
15	11 (11)	15 (16)	19 (18)	22 (21)	33 (33)
30	10 (10)	13 (13)	13 (13)	15 (15)	20 (20)
60	9 (9)	13 (13)	10 (12)	14 (14)	17 (18)

Chapter 5

Conclusions and Remarks

In this thesis we have studied the numerical solution of Helmholtz equation. Helmholtz' problems appear in many engineering applications, and lead to very large sparse linear systems that are difficult to solve due to the unfavorable distribution of the eigenvalues of the matrix.

We discussed some of the main difficulties in applying different iterative methods to Helmholtz problems. These methods have a limited range of application, and for increasing wavenumbers the number of iterations grows too large. Krylov and multigrid methods were reviewed, and we presented the shifted Laplace preconditioner and the technique known as deflation.

Using rigorous Fourier analysis we have analyzed two variants of a two-level method that combines the shifted Laplace preconditioner with multigrid deflation. The analysis was limited to a one-dimensional Helmholtz problem for Dirichlet boundary conditions. This allowed us to obtain exact formulas for the eigenvalues of the preconditioned Helmholtz systems, and study the dependence on the spectrum of the preconditioned systems with respect to distinct parameters: The wavenumber k, the number of gridpoints per wavelength p, and the shift (β_1, β_2).

The first variant of the two-level preconditioner works well for small wavenumbers (< 500), since the spectrum of the preconditioned system that one obtains is clustered, on the right side of the complex plane and bounded away from the origin with a few outliers. For larger wavenumbers, negative eigenvalues appear, as well as very small and very large eigenvalues. The clustering of the spectrum is controlled by the parameter p (more points per wavelength leads to better clustering), but no choice of parameters removes the large and small eigenvalues for large wavenumbers. Our analysis suggests that the method is no longer effective for large wavenumbers, and therefore not scalable, confirming previously reported numerical tests.

The second variant of the two-level preconditioner works very well for wavenumbers up to

approximately $k = 5000$, using a (reasonable) number of gridpoints per wavelength of $p = 20$. With this method, the smallest eigenvalues of the Helmholtz system preconditioned by the CSL are removed and the rest of them are shifted towards the largest one. A nice property that we could not verify analytically is that the eigenvalues of the system remain on a circle. As in the first variant, the clustering of the spectrum can be improved by increasing the number of gridpoints, we have found that $p = 20$ is enough to obtain a clustered spectrum with the standard choice of parameters for the CSL preconditioner $(1, 0.5)$. If the imaginary shift β_2 is increased, p should be increased as well. Our analysis serves to validate previously reported results for wavenumbers up to $k = 500$. Since this method leads to very similar clustered spectra for Helmholtz systems up to very large wavenumbers, we can expect that even in the multilevel extension the number of iterations is only mildly dependent on the wavenumber.

As of further directions of research, we think that an analysis of the 2-D Helmholtz problem with Dirichlet is feasible with the methods of rigorous Fourier analysis. Also, more numerical experiments should be made for Helmholtz problems with high wavenumbers and more realistic 2D problems.

References

[1] A. BAYLISS, C. GOLDSTEIN, AND E. TURKEL, *On accuracy conditions for the numerical computation of waves*, Journal of Computational Physics, 59 (1985), pp. 396 – 404.

[2] A. BAYLISS, C. I. GOLDSTEIN, AND E. TURKEL, *An iterative method for the Helmholtz equation*, Journal of Computational Physics, 49 (1983), pp. 443 – 457.

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

[4] N. BLEISTEIN, J. COHEN, AND J. STOCKWELL, *Mathematics of Multidimensional Seismic Imaging, Migration, and Inversion*, Interdisciplinary Applied Mathematics: Geophysics and Planetary Sciences, Springer, 2001.

[5] A. BRANDT, *Multi-level adaptive solutions to boundary-value problems*, Mathematics of Computation, 31 (1977), pp. 333–390.

[6] A. BRANDT AND S. TA'ASAN, *Multigrid method for nearly singular and slightly indefinite problems*, no. v. 178026 in NASA contractor report, Langley Research Center, NASA, 1985.

[7] W. BRIGGS, V. HENSON, AND S. MCCORMICK, *A Multigrid Tutorial*, Society for Industrial and Applied Mathematics, 2000.

[8] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Applied Mathematical Sciences, Springer, 1998.

[9] H. ELMAN, O. ERNST, AND D. O'LEARY, *A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations*, SIAM Journal on Scientific Computing, 23 (2001), pp. 1291–1315.

[10] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Mathematics of Computation, 31, p. 629.

[11] Y. ERLANGGA, *Advances in iterative methods and preconditioners for the Helmholtz equation*, Archives of Computational Methods in Engineering, 15 (2008), pp. 37–66.

[12] Y. ERLANGGA AND R. NABBEN, *Algebraic multilevel Krylov methods*, SIAM Journal on Scientific Computing, 31 (2009), pp. 3417–3437.

[13] Y. ERLANGGA, C. OOSTERLEE, AND C. VUIK, *A novel multigrid based preconditioner for heterogeneous Helmholtz problems*, SIAM Journal on Scientific Computing, 27 (2006), pp. 1471–1492.

[14] Y. A. ERLANGGA AND R. NABBEN, *Multilevel projection-based nested Krylov iteration for boundary value problems.*, SIAM Journal on Scientific Computing, 30 (2008), pp. 1572–1595.

[15] ——, *On a multilevel Krylov method for the Helmholtz equation preconditioned by the shifted Laplacian*, ETNA Electronic Transactions on Numerical Analysis, 31 (2008), pp. 403–424, (electronic only).

[16] O. ERNST AND M. GANDER, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, in Numerical Analysis of Multiscale Problems, I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, eds., vol. 83 of Lecture Notes in Computational Science and Engineering, Springer Berlin Heidelberg, 2012, pp. 325–363.

[17] R. FEDORENKO, *A relaxation method for solving elliptic difference equations*, USSR Computational Mathematics and Mathematical Physics, 1 (1962), pp. 1092 – 1096.

[18] R. FLETCHER, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis, G. Watson, ed., vol. 506 of Lecture Notes in Mathematics, Springer Berlin / Heidelberg, 1976, ch. 7, pp. 73–89.

[19] A. GREENBAUM AND Z. STRAKOS, *Any nonincreasing convergence curve is possible for GMRES*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 465–469.

[20] W. HACKBUSCH, *Multi-Grid Methods and Applications*, Springer Series in Computational Mathematics, Springer, 2003.

[21] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–436.

[22] F. IHLENBURG AND I. BABUŠKA, *Finite element solution of the Helmholtz equation with high wave number part i: The h-version of the FEM*, Computers & Mathematics with Applications, 30 (1995), pp. 9–37.

[23] E. KAASSCHIETER, *Preconditioned conjugate gradients for solving singular systems*, Journal of Computational and Applied Mathematics, 24 (1988), pp. 265 – 275.

[24] A. KIRSCH, *An Introduction to the Mathematical Theory of Inverse Problems*, Applied Mathematical Sciences, Springer, 2011.

[25] A. L. LAIRD AND M. B. GILES, *Preconditioned iterative solution of the 2-D Helmholtz equation*, Tech. Rep. 02/12, Oxford Computer Laboratory, Oxford, UK, 2002.

[26] C. LANCZOS, *Solution of systems of linear equations by minimized iterations*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 33–53.

[27] J. LIESEN AND P. TICHÝ, *Convergence analysis of Krylov subspace methods*, GAMM Mitteilungen der Gesellschaft für Angewandte Mathematik und Mechanik, 27 (2005), pp. 153–173.

[28] R. NABBEN AND C. VUIK, *A comparison of deflation and the balancing preconditioner*, SIAM Journal on Scientific Computing, 27 (2006), pp. 1742–1759.

[29] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd ed., 2003.

[30] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[31] A. SHEIKH, D. LAHAYE, AND C. VUIK, *A scalable Helmholtz solver combining the shifted Laplace preconditioner with multigrid deflation*, Report 11-01, Delft Institute of Applied Mathematics, Delft, The Netherlands, 2011.

[32] V. SIMONCINI AND D. B. SZYLD, *Recent computational developments in Krylov subspace methods for linear systems*, Numerical Linear Algebra with Applications, 14 (2007), pp. 1–59.

[33] J. TANG, R. NABBEN, C. VUIK, AND Y. ERLANGGA, *Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods*, Journal of Scientific Computing, 39 (2009), pp. 340–370.

[34] U. TROTTENBERG, C. OOSTERLEE, AND A. SCHÜLLER, *Multigrid*, Academic Press, 2001.

[35] H. VAN DER VORST, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2003.

[36] M. VAN GIJZEN, Y. ERLANGGA, AND C. VUIK, *Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian*, SIAM Journal on Scientific Computing, 29 (2007), pp. 1942–1958.