Interim Thesis

- Iterative Helmholtz Solver -

A scalable version of the deflation based ADEF-solver

V. Dwarka, 4262964
MSc. Applied Mathematics - Computational Science and Engineering

December 4, 2016

Abstract

In this interim thesis report we will discuss the building blocks of the ADEF-solver applied to the Helmholtz problem. The main focus will be on the occurrence of the clustering eigenvalues around zero, which seem to appear as soon as the wave number k becomes very large. We start by discussing the literature, summarizing the main results. Moreover, the main findings from pivoting papers will also be reconstructed in order to gain thorough insights into the behavior of the eigenvalues as this will aid us in future research purposes as regards the potential scalability of the ADEF-solver.

In Chapter 1 we start by exploring the properties of both the continuous and discrete Helmholtz operator. Chapter 2 focuses on numerical techniques, suitable for the Helmholtz problem. Chapter 3 involves around appropriate preconditioning techniques, in particular the CSLP-preconditioner, whereas Chapter 4 and discusses the deflation-based ADEF-preconditioner. Finally, Chapter 5 concludes with a proposal for future research.

Contents

List of Figures

List of Tables

Chapter 1

The Helmholtz equation

In this chapter we will start by exploring the Helmholtz equation. The Helmholtz equation, named after its creator, Hermann von Helmholtz, German physician and physicist, is a second order partial differential equation widely used in engineering practices. It models wave phenomena and is thus suitable for applications in various areas of physics and mathematics such as electromagnetic radiation, seismology, acoustics and optics.

After giving a brief introduction into the derivation of the Helmholtz equation, we will proceed by construing an analytical and exact solution to the defined problem. This exact solution will serve as a reference for future analysis in the subsequent chapters of this literature study.

Figure 1.1: *Herman von Helmholtz*

1.1 Derivation

The Helmholtz equation can be derived from the time-dependent wave equation after applying the method of separation of variables. The resulting equation models harmonic wave propagation through a homogeneous medium. We can thus start by considering the propagation of time harmonic waves, which is governed by the following equation

$$(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})u(\mathbf{x},t) = \mathbf{0}. \tag{1.1}$$

In equation(1.1) the vector $\mathbf{x}$ denotes the spatial variable in some subspace Ω of $\mathbb{R}^n$, which represents the physical domain. The real constant c and the real variable t represent the wave speed and time variable respectively.

A solution to equation (1.1) can be obtained by separating the variables into a spatial and time component, which can be represented as follows

$$u(\mathbf{x}, t) = \varphi(\mathbf{x})\, T(t). \tag{1.2}$$

Letting equation (1.2) represent a potential solution to equation(1.1), we substitute the previous equation into the former to obtain

$$(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})(\varphi(\mathbf{x})\, T(t)) = \mathbf{0}, \tag{1.3}$$

$$\frac{\partial^2 \varphi}{\varphi} = \frac{1}{Tc^2}\frac{\partial^2 T}{\partial t^2} = -k^2. \tag{1.4}$$

Note that in order for the solution to satisfy equation (1.1), we had to equate both sides of equation to a constant $-k^2$. Rearraging the left hand side of equation (1.1), which now is completely separated from the time component, leads to the homogeneous Helmholtz equation

$$(-\nabla^2 - k^2)\, \varphi(\mathbf{x}) = \mathbf{0} \tag{1.5}$$

Intuitively $\varphi(\mathbf{x})$ can best be interpreted as the wave function, whereas k stands for the wavenumber, which relates the wavelength λ and the angular frequency. General expressions for the before mentioned are

$$k = \frac{2\pi}{\lambda} \tag{1.6}$$

Practical applications of the Helmholtz equation often involve the non-homogeneous Helmholtz equation. In this case the right hand side of (1.5) consist of a source function $f(\mathbf{x})$

$$f(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x_s}) \tag{1.7}$$

Additionally, in some applications, such as modeling phenomena through an inhomogeneous medium, a non-constant wavenumber $k(\mathbf{x})$ is enforced to capture different velocity profiles.

1.2 Boundary Conditions

Solving the Helmholtz equation on a bounded physical domain Ω requires the reinforcement of boundary conditions. In the absence of such conditions the problem becomes ill-posed; the equation in its current form models the indefinite propagation of waves. Therefore, the following boundary conditions are often implemented when solving the Helmholtz equation at the boundary of Ω, which we will denote by $\partial\Omega$:

- Vanishing boundary conditions: vanishing boundary conditions can be modelled by imposing homogeneous Dirichlet conditions

$$\varphi(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in \partial\Omega. \tag{1.8}$$

- Reflecting boundary conditions: reflecting boundary conditions can be modelled by imposing homogeneous Neumann conditions, where $\mathbf{n}$ denotes the outward normal unit vector with respect to the boundary $\partial\Omega$

$$\left(\frac{\partial}{\partial \mathbf{n}}\right) \varphi(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in \partial\Omega. \tag{1.9}$$

- Mixed boundary conditions: mixed boundary conditions can be modelled by imposing both homogeneous Dirichlet and Neumann conditions instantaneously. Within the context of the Helmholtz equation, one important mixed boundary condition is often referred to as rhe *Sommerfeld Radiation condition* (also known as an absorbing boundary condition), where i represents the imaginary unit and $\partial \mathbf{n}$ denotes the outward normal unit vector with respect to the boundary $\partial \Omega$ and represents an artificial boundary $\partial \tilde{\Omega}$ such that

$$\left(\frac{\partial}{\partial \mathbf{n}} + ik\right) \varphi(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in \partial \Omega. \tag{1.10}$$

- Perfectly Matched Layer (PML) boundary conditions [1]: in the literature this boundary condition is implemented to prevent the outgoing waves at the boundary to be reflected back into the domain. Similarly like Sommerfeld Radiation conditions, an artifical boundary $\partial \tilde{\Omega}$ is appended to the original domain Ω. Consequently, a thin layer around the artificial boundary $\partial \tilde{\Omega}$ is modeled. Inside the layer, the PDE is altered such that 1) waves are damped rapidly and 2) no reflections are introduced at $\partial \tilde{\Omega}$. The alteration to the PDE must be done cautiously in order to prevent artificial reflections. At the outer end of the layer any boundary condition can be used, for instant homogeneous Dirichlet.

1.3 Analytical Model

For the sake of theoretical thoroughness, we focus on a simple one-dimensional mathematical model, allowing for extensive and robust testing. Consequently, for the purpose of this interim research report, we will proceed with the following simple mathematical model

$$\begin{aligned}
&-\frac{d^2 u}{dx^2} - k^2\, u = \delta(x - \frac{L}{2}),\\
&u(0) = 0, u(L) = 0,\\
&x \in \Omega = [0, L] \subseteq \mathbb{R},\\
&k \in \mathbb{N} \setminus \{0\}.
\end{aligned} \tag{1.11}$$

Note that we have placed the harmonic point source at the center of the analytical domain and keep k constant and integer valued. The stated model problem (Helmholtz boundary value problem) allows for simplicity while assuring that results obtained from the one-dimensional case can easily be extended to multi-dimensional problems analogously. Moreover, it has been suggested in Sheikh (2014) that the homogeneous Dirichlet conditions produce the most adverse spectral properties due to the absence of damping, staggering the convergence of iterative numerical methods, and thus being particularly suitable for theoretical research purposes when k becomes very large.

1.3.1 Analytical Solution

We can express the analytical solution to our model problem in terms of the Green's function $G(x, x')$. Let $\mathcal{L}(x)$ be the general Sturm-Liouville operator

$$\mathcal{L}(x) = \frac{d}{dx}\left[p\left(x\right)\frac{d}{dx}\right] + q(x) \tag{1.12}$$

Setting $p(x) = -1$ and $q(x) = -k^2$, we obtain the Sturm-Liouville operator for the Helmholtz boundary value problem, which we will continue to denote by $\mathcal{L}(x)$. Consequently, let $G(x, x')$ be the Green's function satisfying

$$\begin{aligned}
&\mathcal{L}(x)G(x, x') = \delta(x - x')\\
&G(0, x') = G(L, x') = 0,\ x \in \partial\Omega\\
&x \in \Omega = [0, L] \subset \mathbb{R}
\end{aligned} \tag{1.13}$$

[1] This point has been summarized from Runborg (2012), p. 13

The eigenfunctions and eigenvalues of the Sturm-Liouville problem in equation (1.13) are

$$\begin{aligned}\phi_j(x) &= \sqrt{\frac{2}{L}}\sin(j\pi x/L) \\ \lambda_j(x) &= \frac{j^2\pi^2 - k^2}{L^2} \\ j &= 1, 2, 3, \ldots \end{aligned} \tag{1.14}$$

Due to the eigenfunctions being sines, we can postulate a Fourier series solution using sine functions. Thus, we define $G(x, x')$ in terms of the following series representation

$$\begin{aligned}G &= \sum_{j=1}^{\infty} \alpha_j(x')\sin(j\pi x/L) \\ \delta(x - x') &= \sum_{j=1}^{\infty} \frac{2}{L}\sin(j\pi x/L)\sin(j\pi x'/L)\end{aligned} \tag{1.15}$$

Equation coefficients of the $\sin(j\pi x/L)$ term will allow us to solve for $\alpha_j(x')$

$$\begin{aligned}\left(\frac{j^2\pi^2}{L^2} - k^2\right)\alpha_j(x') &= \sin(j\pi x'/L) \\ \alpha_j(x') &= \frac{2}{L}\left(\frac{\sin(j\pi x'/L)}{j^2\pi^2 - k^2/L}\right)\end{aligned} \tag{1.16}$$

Substituting these expressions into equation 1.15 and letting $x' = L/2$ gives us the solution the model problem as defined in equation (1.11) [2]

$$\begin{aligned}G(x, L/2) &= \frac{2}{L}\sum_{j=1}^{\infty} \frac{\sin(j\pi x/L)\sin((j\pi L/2/L))}{j^2\pi^2 - k^2/L} \\ k^2 &\neq \frac{j^2\pi^2}{L} \\ j &= 1, 2, 3, \ldots\end{aligned} \tag{1.17}$$

In the event that $k^2 = \frac{j^2\pi^2}{L}$, the eigenfunction expansion would become defective as this would imply resonance and unbounded oscillations in the absence of dissipation. Therefore, we explicitly need to warrant for the latter case and impose the extra condition $k^2 \neq \frac{j^2\pi^2}{L}$ asserting that our analytical solution remains valid.

1.3.2 Continuous Spectrum

Equation 1.14 immediately provides us with an expression for the analytical eigenvalues. It is apparent that within the bounded domain $[0, L]$ there are an infinite number of eigenpairs. We will employ this expression for the eigenvalues in upcoming sections, where we will compare them with the discrete eigenvalues for the linear system of equations.

1.3.3 Dimensionless Analytical Model

Equation 1.11 has been defined over an arbitrary domain $[0, L]$. We can apply a linear transformation to map the problem onto the unit domain $[0, 1]$. We introduce the following change of variable

$$\widehat{x} = \frac{x}{L} \tag{1.18}$$

[2] We have not used the exponential form of the Green's function, given that it satisfies the inhomogeneous Helmholtz equation with equipped with both Dirichlet and Sommerfeld radiation conditions.

Substituting $\widehat{x}$ into equation 1.11 and noting that

$$\frac{d\widehat{x}}{dx} = \frac{1}{L} \tag{1.19}$$

Equation 1.11 can be written as

$$\begin{aligned}
&(-\nabla^2 - \widehat{k}^2)\, u(\widehat{x}) = \widehat{f}(\widehat{x}) \\
&u(0) = 0, u(1) = 0 \\
&\widehat{x} \in \Omega = [0,1] \subset \mathbb{R} \\
&\widehat{k} = Lk \\
&\widehat{f}(x) = L^2 f(x)
\end{aligned} \tag{1.20}$$

Equation 1.20 provides a convenient way to extrapolate results from a physical domain onto the unit domain, without affecting the solution and spectral properties. Undifferentiated transformations can be applied to higher dimensions in order to map the problem onto the squared and cubed unit domain respectively.

Unless stated otherwise, we will resort to a coherent notation in the upcoming sections, where $\widehat{k} = k$ and $\widehat{f}(\widehat{x}) = f(x)$.

1.4 Numerical Model

Solving the Helmholtz Equation analytically in higher dimensions is unpractical. Especially at high frequencies, numerical schemes are necessary to provide functional solutions. In principle, both finite differences and finite elements methods are considered as preferable schemes.
In this study we will focus on the finite difference method and accordingly will elaborate on the discretization of the Helmholtz equation. [3]

1.4.1 Finite Differences

We will discretisize the model problem on the finite domain $\Omega = [0,1]$, using a second-order accurate central difference scheme. Starting with the one-dimensional case, we can naturally extend the discretization to the two-dimensional case.

Discretization of the Geometry

For the discretization of the model problem we let n denote the number of elements on a uniform grid $G_{h_{1D}}$ consisting of $n+1$ nodes, including the boundary $\partial\Omega$. Given the unit domain, we get the following numerical domain, with step size $h = \frac{1}{n}$

$$G_{h_{1D}} = \{(x_i) \,|\, x_i = ih, h = \frac{1}{n},\, 0 \leq i \leq n,\, n \in \mathbb{N} \setminus \{0\}\}$$

In the two-dimensional case, our finite domain becomes the unit square domain $\Omega = [0,1] \times [0,1]$. We remain a stepsize of $h = \frac{1}{n}$, where now n represents the number of mesh elements, which induces $(n+1)^2$ nodes on an uniform grid $G_{h_{2D}}$

$$G_{h_{2D}} = \{(x_i, y_j) \,|\, x_i = ih,\, y_j = jh, h = \frac{1}{n},\, 0 \leq i, j \leq n,\, n \in \mathbb{N} \setminus \{0\}\}$$

[3] Most of this section is a summary of §3.5 − 3.7 (Vuik and Lahaye, 2012) applied to the Helmholtz problem.

Discretization of the Physics

On both $G_{h_{1d}}$ and $G_{h_{2D}}$ respectively, we introduce spatial grid vectors in order to approximate the source function $f(\mathbf{x})$ and the wave function $u(\mathbf{x})$. Due to the vanishing boundary conditions, the numerical wave function $u(\mathbf{x})$ will solely be defined at the internal grid nodes of $G_{h_{1d}}$, which gives

$$\begin{aligned} &f(x) \approx f(x_i) = f_{ih}, \\ &u(x) \approx u(x_i) = u_{ih}, \\ &x \in G_{h_{1D}}. \end{aligned}$$

For the two-dimensional case, we have

$$\begin{aligned} &f(x,y) \approx f(x_i,y_j) = f_{i,jh} \\ &u(x,y) \approx u(x_i,y_j) = u_{i,jh} \\ &(x,y) \in G_{h_{2D}} \end{aligned}$$

Linear System Formulation

We arrive at a linear system formulation after approximating the continuous second order derivatives by central finite difference approximations. For the one-dimensional case we have

$$\frac{-u_{i-1h} + 2u_{ih} - u_{i+1h}}{h^2} - k^2 u_{ih} = f_{ih},\ 1 \leq i \leq n-1$$

Working in two dimensions, the discretization results in:

$$\frac{-u_{i,j-1h} - u_{i-1,jh} + 4u_{i,jh} - u_{i,j+1h} - u_{i+1,jh}}{h^2} - k^2 u_{i,jh} = f_{i,jh},\ 1 \leq i,j \leq n-1$$

Note that the equations for the nodes corresponding to the homogeneous Dirichlet boundary are redundant. Consequently, the linear system can be formulated exclusively on the basis of the internal grid points. This approach will be referred to as *with elimination of the boundary conditions.* In the case that the boundary conditions are eliminated, we can implement an x-line lexicographic ordering of the internal nodes, allowing us to assemble the unknown grid values u_{ih} and f_{ih} into column vectors of dimension $(n-1)$ for the one dimensional case. Consequently we can compose a linear system of equations

$$A_h = \frac{1}{h^2}\text{tridiag}[-1 \ \ 2-k^2 \ \ -1]$$

$$= \frac{1}{h^2}\begin{bmatrix} 2-k^2h^2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2-k^2h^2 & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & -1 & 2-k^2h^2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2-k^2h^2 \end{bmatrix}$$

$$u_h = \begin{bmatrix} u_{1h} \\ \vdots \\ u_{ih} \\ \vdots \\ u_{nh} \end{bmatrix},\ f_h = \begin{bmatrix} f_{1h} \\ \vdots \\ f_{ih} \\ \vdots \\ f_{nh} \end{bmatrix}$$

$$1 \leq i \leq n-1$$

Regarding the two-dimensional problem, we obtain an equivalent linear system of equations

$$A_h = \frac{1}{h^2}\text{tridiag}[-1 \;\; 2-k^2 \;\; -1] \otimes \frac{1}{h^2}\text{tridiag}[-1 \;\; 2-k^2 \;\; -1]$$

$$= \frac{1}{h^2}\begin{bmatrix} 4-k^2 & -1 & 0 & \dots & -1 & 0 & \dots & 0 \\ -1 & 4-k^2 & -1 & 0 & \dots & -1 & 0 & \dots \\ 0 & -1 & 4-k^2 & -1 & 0 & \dots & -1 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & \dots & 0 & -1 & 4-k^2 & -1 & 0 \\ 0 & -1 & \dots & \dots & \dots & -1 & 4-k^2 & -1 \\ \dots & \dots & -1 & 0 & \dots & \dots & -1 & 4-k^2 \end{bmatrix}$$

$$u_h = \begin{bmatrix} u_{i,1h} \\ u_{i,2h} \\ \vdots \\ u_{i,n-1h} \end{bmatrix}, \; f_h = \begin{bmatrix} f_{i,1h} \\ f_{i,2h} \\ \vdots \\ f_{i,n-1h} \end{bmatrix}$$

$$1 \leq i \leq n-1$$

We have transformed the continuous partial differential Helmholtz equation into a linear system of equations. Resultantly, solving the Helmholtz boundary value problem now constitutes to solving the system

$$\begin{aligned} A_h u_h &= f_h \\ A_h &\in \mathbb{R}^{(n-1)\times(n-1)} \\ u_h, f_h &\in \mathbb{R}^{(n-1)} \end{aligned} \tag{1.21}$$

The upper system formulation comprises the one-dimensional case. For the two-dimensional case, we obtain

$$\begin{aligned} A_h u_h &= f_h \\ A_h &\in \mathbb{R}^{(n-1)^2\times(n-1)^2} \\ u_h, f_h &\in \mathbb{R}^{(n-1)^2} \end{aligned} \tag{1.22}$$

In (Sheikh, 2014) it is mentioned that as a rule of thumb for second order accurate finite differences discretizations at least 10 grid points per wavelength λ should be efmployed

$$\begin{aligned} \kappa = kh = \frac{2\pi}{10} \approx 0.625 \\ \Rightarrow n = \lfloor \frac{k}{\kappa} \rfloor \end{aligned} \tag{1.23}$$

We therefore use the restriction from equation 1.23 , where κ should be interpreted as the grid refinement and h will be determined accordingly for a given wave number k. Thus, unless stated otherwise we employ a grid resolution of $kh = 0.625$.

For large values of k, solutions become oscillatory, requiring more refined grids and higher resolutions due to the wavelength λ decreasing. If the grid resolution is not adapted to reflect these concerns, a pollution of the numerical solution is reported in various sources (Ihlenburg and Babuska, 1997), (Deraemaeker et al., 1999), (Gerdes and Ihlenburg, 1999), and (Wang and Wong, 2014). The pollution error will be discussed in Chapter 2, section 2.6.2.

1.4.2 Discrete Spectrum

In section 1.3.2 we expressed the analytical eigenvalues to be as in equation 1.14. Given that our numerical model is exerted on the unit domain, the eigenvalues for the one-dimensional problem become

$$\lambda_j = j^2\pi^2 - k^2,\ j = 1, 2, 3, \dots \tag{1.24}$$

A comparable result holds in the discrete case [4].

$$\begin{aligned} \lambda_{l_h} &= \frac{1}{h^2}\left[2 - 2\cos(l\pi h) - k^2h^2\right] \\ l &= 1, 2, 3, \dots, n-1 \end{aligned} \tag{1.25}$$

Similarly, the two-dimensional eigenvalues are given by

$$\begin{aligned} \lambda_{l,mh} &= \frac{1}{h^2}\left[4 - 2\cos(l\pi h) - 2\cos(m\pi h) - k^2h^2\right] \\ l, m &= 1, 2, 3, \dots, n-1 \end{aligned} \tag{1.26}$$

Moreover, we obtain the following set of eigenvectors for the one-dimensional model problem

$$\phi_{lh} = \sin(l\pi\mathbf{x}),\ 1 \leq l \leq n-1, \tag{1.27}$$

where $\mathbf{x} = [x_i]$, $1 \leq i \leq n-1$ represents the gridvector on $G_{h_{1d}}$.

Similarly, in the two-dimensional case we obtain

$$\phi_{l,mh} = \sin(l\pi\mathbf{x})\sin(m\pi\mathbf{y}),\ 1 \leq l, m \leq n-1, \tag{1.28}$$

where again $\mathbf{x}$ and $\mathbf{y}$ represent the grid vectors in x and y direction respectively.

1.4.3 Linear System Properties

The coefficient matrices of the linear systems obtained in section 1.4.1 are real symmetric matrices. Due to an orthonormal basis of eigenvectors, see equation 1.27 and 1.28, the matrices are also normal. The one-dimensional coefficient matrix is a tridiagonal matrix, leading to a sparse matrix A_h. The two-dimensional matrix is a sparse penta-diagonal matrix. However, unlike in the one-dimensional case, the appearance of non-zero diagonals within the bandwidth of A_h is common. Note that in case of Sommerfeld Radiation conditions, the coefficient matrix becomes complex non-Hermitian symmetric with complex eigenvalues.

One can immediately notice from the one-dimensional case that the coeffient matrix A_h is in fact the discretisized Laplacian including a term involving k^2

$$A_h = \Delta_h - k^2 I_h \tag{1.29}$$

where I_h represents the $(n-1)\times(n-1)$ identity matrix. For large enough k^2, the matrix becomes highly indefinite due to the increasing number of negative eigenvalues.

1.5 Negative Eigenvalues

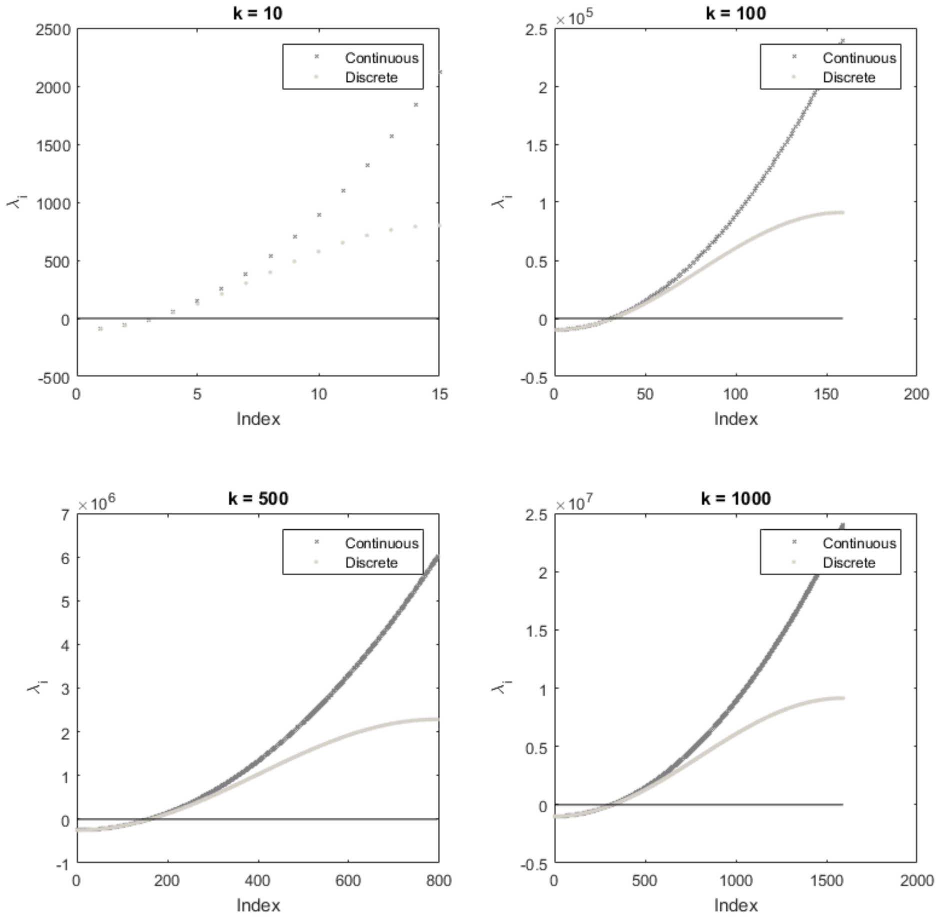
As the emphasis in this study is on instances where k is very large, we start by exploring the eigenvalues of the coefficient matrix A_h starting from $k = 10$ up to $k = 1000$. For the sake of brevity, we will drop the notation A_h in this section and resort to denoting the coefficient matrix by A. A similar notation

[4] These eigenvalues can obtained by using the expression for the Toeplitz matrix, see section 2.2.1 of Vuik and Lahaye (2012)

will be adapted for the eigenvalues.

In figure 1.2 the eigenvalues for the discrete and continuous Helmholtz operator are illustrated. Already for small k, a discrepancy between the analytical and discrete eigenvalues exists. However, this is not characteristcally inherent to the Helmholtz problem, as the same discrepancy exists when solving the Poisson problem using finite differences. Interestingly, the smaller the discrete eigenvalues, the better they approximate the analytical eigenvalues.
As k increases, the number of negative eigenvalues increases as well. However, increasing the number of grid points per wave length does not administer in resolving this issue as the ratio between the number of negative eigenvalues and total eigenvalues remains constant for wave numbers $k > 100$ and fixed κ (Sheikh, 2014).

Figure 1.2: *Eigenvalues of the continuous and discrete Helmholtz operator*



Generally, the effect of the eigenvalues on the convergence behavior is anticipated by considering the condition number of a matrix. However, due to the negative eigenvalues of the discrete Helmholtz operator, the condition number as a metric for convergence becomes meaningless. Therefore, we solely look at the behavior of the negative eigenvalues in relation to the total number of eigenvalues. Table 1.1 provides detailed aspects of the various cases of large k. The number of negative eigenvalues reflects the indefiniteness of the coefficient matrix and seems inherently dependent on k.

Table 1.1: *Number of negative eigenvalues relative to the problem size. The last column contains the ratio of the negative eigenvalues to the total number of eigenvalues*

k	n	h	No. Neg. Eig.	Ratio
100	160	6.211180e-03	32	2.000000e-01
500	800	1.248439e-03	161	2.012500e-01
1000	1600	6.246096e-04	323	2.018750e-01
1500	2400	4.164931e-04	485	2.020833e-01
2000	3200	3.124024e-04	647	2.021875e-01
2500	4000	2.499375e-04	809	2.022500e-01
3000	4800	2.082899e-04	970	2.020833e-01
3500	5600	1.785395e-04	1132	2.021429e-01
4000	6400	1.562256e-04	1294	2.021875e-01
4500	7200	1.388696e-04	1456	2.022222e-01
5000	8000	1.249844e-04	1618	2.022500e-01
7500	12000	8.332639e-05	2427	2.022500e-01
10000	16000	6.249609e-05	3237	2.023125e-01
15000	24000	4.166493e-05	4855	2.022917e-01
20000	32000	3.124902e-05	6474	2.023125e-01

We find that the number of negative eigenvalues increases along with k. As previously mentioned by Sheikh (2014), the ratio between the number of negative eigenvalues and the total number of eigenvalues remains constant as the problem size increases.

1.5.1 Near-null Eigenvalues

Equation 1.24 provides us with an expression for the analytical solution. In section 1.3.2, a particular mentioning had been granted to the case where $\lambda_j = j^2\pi^2 = k^2$, as this would imply singularity, causing the system to become severely insolvable. Effectively this means that the eigenvalues of the continuous *Laplacian* operator are moving towards k^2, causing the eigenvalues of the continuous Helmholtz operator to move closer to zero.

From Figure 1.2 we can deduce that the close to zero eigenvalues of the matrix A are located at the intersection with the origin. In order to locate the intersection point, we use the expressions obtained for the analytical and discrete eigenvalues respectively. For the continuous counterpart we have obtained

$$\begin{aligned}\lambda_j = 0 &\Rightarrow j^2\pi^2 = k^2\pi^2, \\ &\Rightarrow j = \lfloor \frac{k}{\pi} \rfloor \text{ or } \lceil \frac{k}{\pi} \rceil. \end{aligned} \tag{1.30}$$

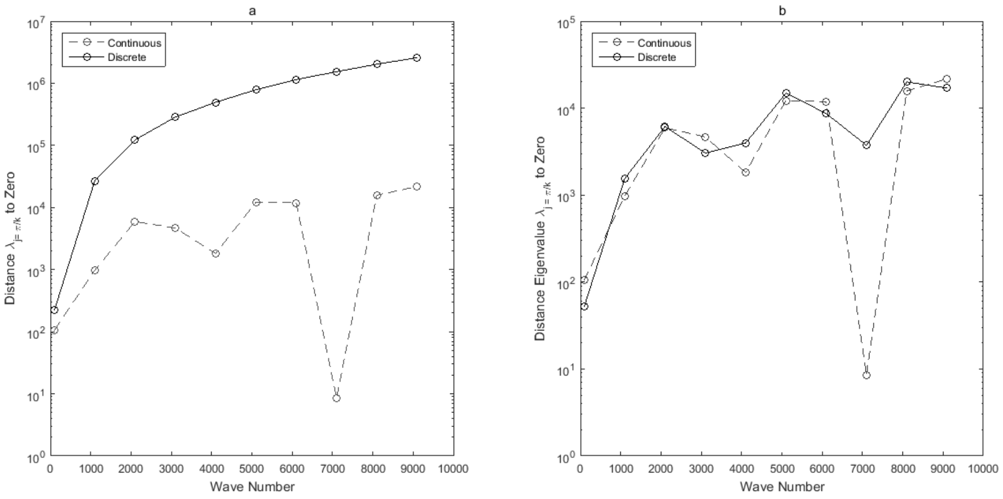
Whereas an equivalent condition for the discrete case led to

$$\lambda_l = 0 \Rightarrow \frac{1}{h^2}\left[2 - 2\cos(l\pi h)\right] \approx k^2 \tag{1.31}$$

Letting $j = \widehat{l}$ according to equation 1.30, we let $\widehat{l}$ represent the index where the discrete eigenvalue should be closest to zero in case the discrete eigenvalue is a satisfactory approximation of the continuous eigenvalue at that point.

We initially plot the eigenvalues using $\kappa = 0.625$ and $\kappa = 0.0625$, where we take $\widehat{l} = \lfloor \frac{k}{\pi} \rfloor$ for both the analytical and discrete case. Figure 1.3 (a) immediately affirms a pattern pointing to the conclusion that the index $\widehat{l}$ does not point to the near-null eigenvalues of the discrete operator when using 10 grid points per wave length. While the index indeed points to the reference where the negative eigenvalue turns positive in the continuous case, a similar conclusion can not be extended to the discrete case. Despite the latter, after resorting to a finer grid in Figure 1.3 (b), both the continuous and discrete eigenvalues seem to be approaching each other. Thus, while refining the grid leads to a better approximation of the analytical eigenvalues, it does not prevent the negative eigenvalues of appearing. As a result, the number of negative eigenvalues seems independent of the number of grid points per wave length, which determines the step size h.

Figure 1.3: *Absolute distance of the closest eigenvalue to zero. In Figure (a) a grid resolution of $kh = 0.625$ was used, while Figure (b) a resolution of $k^3h^2 = 0.625$ was used.*



In light of the previous, it has been mentioned by Sheikh et al. (2016) that initially it suffices to take $\widehat{l} = \lfloor \frac{k}{\pi} \rfloor$. However, as k increases, the gap between $\widehat{l}$ and $\lfloor \frac{k}{\pi} \rfloor$ increases resulting in the intersection point moving further away. Consequently, we can use the approximation as mentioned in (Sheikh et al., 2016) to allocate the index to the point where the eigenvalues of A intersect the origin

$$\widehat{l} = \text{round}[\frac{\arccos 1 - \kappa^2}{\pi h}] \tag{1.32}$$

We now plot the eigenvalues of A without restricting $\widehat{l}$ to be equal to $\lfloor \frac{k}{\pi} \rfloor$. We let the index $\widehat{l}$ be determined by equation 1.32.

Figure 1.4: *Eigenvalues of the continuous and discrete Helmholtz operator. Results are plotted near the index corresponding to the close to zero eigenvalue.*

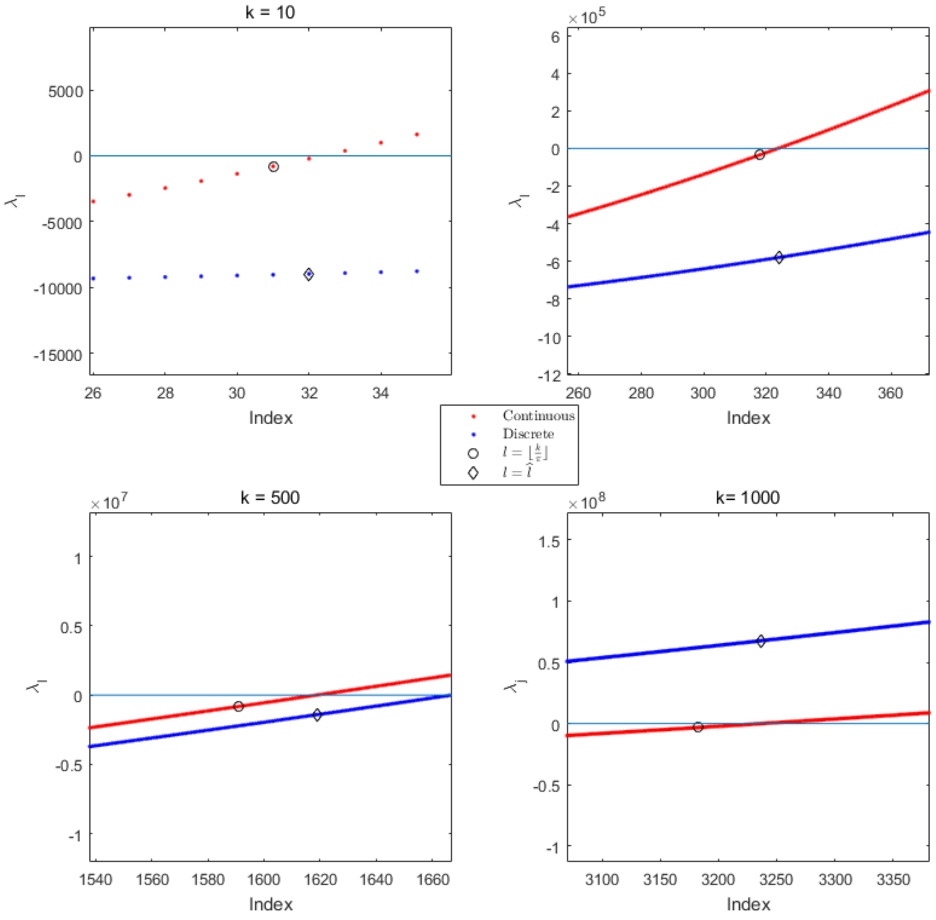
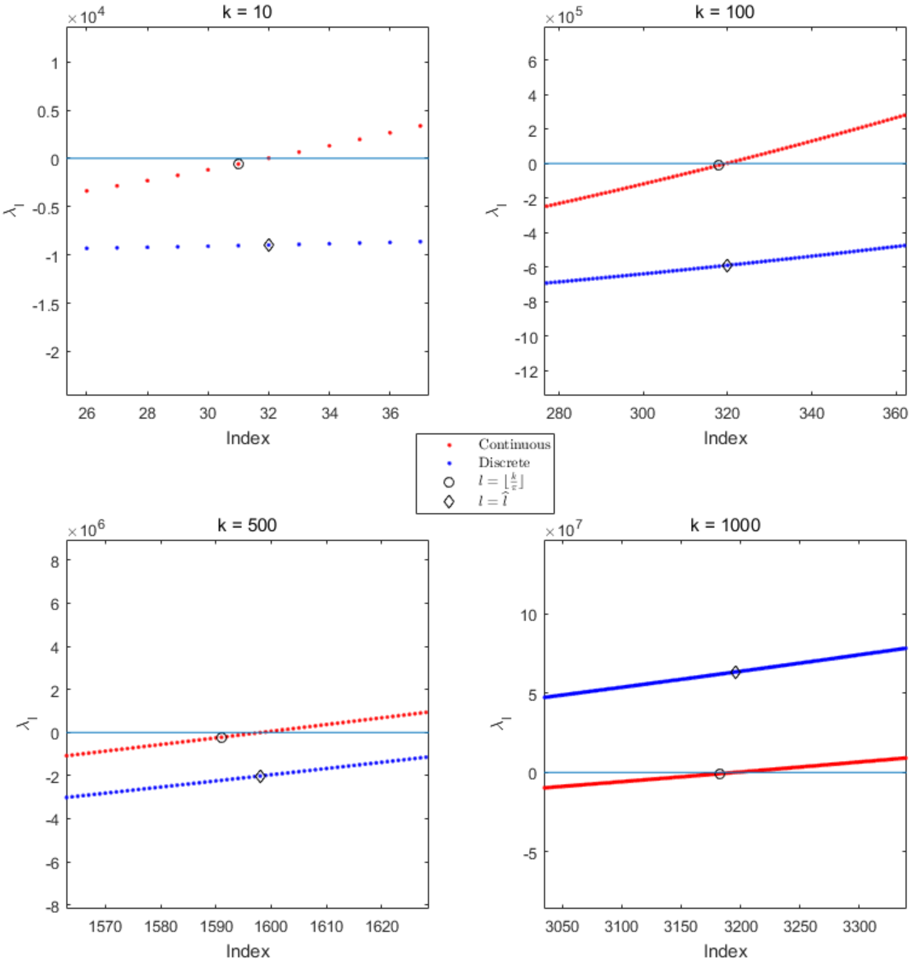


Figure 1.5 confirms the notion by Sheikh et al. (2016) that the index moves further down the axis as k increases. We now repeat the same analysis for $\kappa = 0.3125$, which is equivalent to using approximately 20 grid points per wave length.

Figure 1.5: *Eigenvalues of the continuous and discrete Helmholtz operator using 20 grid points per wavelength ($\kappa = 0.3125$). Results are plotted near the index corresponding to the close to zero eigenvalue.*



Refining the grid in terms of the resolution κ reduces the difference between the indices $\widehat{l} = \lfloor \frac{k}{\pi} \rfloor$ and $\widehat{l}$ according to equation 1.32. As a result, the intersection point with the origin is more closely located near $\lfloor \frac{k}{\pi} \rfloor$, as the two marked indices start to overlap. We therefore expect $\widehat{l} \longrightarrow \lfloor \frac{k}{\pi} \rfloor$ as the number of grid points per wave length is increased. This notion is supported by a similar observation using Figure 1.3 (b).

1.6 Concluding Remarks and Summary

In this chapter we discuss the literature and some basic results regarding the Helmholtz boundary value problem and its spectral properties. We can summarize our findings in the following main points:

- Discretisizing the inhomogeneous Helmholtz equation using Dirichlet boundary conditions, leads to an indefinite normal but symmetric coefficient matrix with real but partially negative eigenvalues
- The number of negative continuous and discrete eigenvalues increases along with k
- The ratio between the number of negative discrete eigenvalues and the total number of discrete eigenvalues scales accordingly relative to the problem size and remains constant

- Refining the grid leads to a better approximation of the analytical eigenvalues, yet has no influence on the grade of indefiniteness

Chapter 2

Numerical Solution Methods

In Chapter 1 we saw that the problem size of the system to be solved depends on the wave number k. As a consequence, large values of k not only cause the indefiniteness of the matrix to develop further and further, but also result in large linear systems which need to be solved. In general, direct numerical solution methods are suitable for medium sized problems. Though still workable for the one-dimensional case, solving large problems with direct numerical solution methods becomes computationally expensive due to increasing memory requirements, especially in higher dimensions. Despite these drawbacks, direct numerical solution methods serve as subdomain solvers in domain decomposition methods and multigrid methods, see section 2.5.

The problem size of the Helmholtz boundary value problem for large k requires the use of iterative solution methods. Basic iterative methods (BIMs) suffer from slow convergence behavior. As k increases, these methods become more dependent on the grid size and wave number, see Erlangga (2005). As our study serves the inspection of potential scalability of a deflation based Krylov subspace solver, we will primarily focus on the literature regarding Krylov subspace methods. Additionally, we will look into multigrid methods for the Helmholtz problem. Standard multigrid methods are not suitable for the Helmholtz problem, see Chapter 4, section 4.1.1. Therefore, we will only briefly describe the standard smoothing properties of BIMs in relation to multigrid methods.

2.1 Krylov Subspace Methods

Consider a general linear system [1]

$$\begin{aligned} &Au = f, \\ &A \in \mathbb{C}^{n\times n},\ u, f \in \mathbb{C}^n. \end{aligned} \tag{2.1}$$

Definition .1 *(Petrov-Galerkin Method) Given a linear system $Au = f$, let A be a matrix in $\mathbb{C}^{n\times n}$, u, f vectors in $\mathbb{C}^n$. Then a solution of equation 2.2.1 can be approximated by*

$$y = u_0 + s, s \in S \subset \mathbb{C}^n, \tag{2.2}$$

where u_0 is a predefined initial approximation and S is denoted as the search space. Let $r \in \mathbb{C}^n$ be defined as the residual vector such that we can define a constraint space C satisfying

$$r := f - Ay \perp C \subset \mathbb{C}^n. \tag{2.3}$$

Then a Petrov-Galerkin method is well defined if $\langle C, AS\rangle$ is nonsingular for any C and Y, where $C, Y \subseteq \mathbb{C}^n$.

[1] Most of this section contains summarizing parts from section 2.4, 2.6 and 2.7 of Chapter 2 from (Gaul, 2014). The theorems, propositions and corollaries in this section have been taken from beforementioned sections of Gaul (2014). A page-reference to the proofs will be made for each subsequently.

Here $\langle \bullet \rangle$ denotes the standard inner product defined on the complex space. If the latter condition is satisfied, we can get an approximate solution using the following theorem

Theorem .1 *(Petrov-Galerkin Method) Let A be a matrix in $\mathbb{C}^{n\times n}$, u, f vectors in $\mathbb{C}^n$ such that the Petrov-Galerkin method with search space S and constraint space C is well-defined. Then the approximate solution y and the corresponding residual r that satisfy Definition .1 are given by*

$$y = u_0 + S\langle C, AS\rangle^{-1}\langle C, r_0\rangle, \tag{2.4}$$

$$r = f - Ay = P_{C^\perp, AS} r_0, \tag{2.5}$$

where $r_0 = f - Au_0$ is the initial residual. Furthermore, the linear system from Definition .1 is solved if and only if $r_0 \in AS$

Proof *For a proof of this theorem, see Gaul (2014) corollary 2.26, p. 23.*

Using Theorem .1, we now have a practical way to find an approximate solution $y \approx x$ which solves the linear system $Au \approx Ay = f$. However, we would like to find an approximate solution which is not only optimal, but also unique. For this purpose, we can use the following theorem

Theorem .2 *(Well-definedness and Optimality) Consider a linear system $Au = f$ with A a matrix in $\mathbb{C}^{n\times n}$, u, f vectors in $\mathbb{C}^n$. Furthermore, let $u_0 \in \mathbb{C}^n$ be the inital guess vector and let S be an $n-$dimensional subspace. Then Petrov-Galerkin method with search space S and constraint space C is well defined and defines a unique approximate solution $y + u_0 \in S$ if one of the following conditions holds:*

1. *$C = S, S \cup \mathcal{N}(A) = \{0\}$, A is self-adjoint and positive semi-definite. Then*

$$\|u - y\|_A = \inf_{z \in u_0 + S} \|u - z\|_A\,,$$

where $\|\bullet\|_A$ is the norm defined by $\|z\|_A = \sqrt{\langle z, Az\rangle}$.

2. *$C = AS, S \cup \mathcal{N}(A) = \{0\}$. Then*

$$\|f - Ay\| = \inf_{z \in u_0 + S} \|f - Az\|\,.$$

Proof *For a proof of this theorem, see Gaul (2014) lemma 2.28, p. 23.*

Note that either the residual or the difference between the true and approximate solution is minimized, and thus we obtain an optimality certificate for constructing an approximate solution to the original linear system 2.2.1.

We now proceed by giving the definition of a general Krylov subspace, using an arbitrary vector $v \in \mathbb{C}^n$:

Definition .2 *(Krylov Subspace) Given a linear system $Au = f$, with u, f, v vectors in $\mathbb{C}^n$ Then the m-th Krylov subspace is defined by*

$$\begin{aligned} K_m(A, v) &= \operatorname{span}\{v, Av, ...A^{m-1}v\}, \\ K_0(A, v) &= \{0\}, m \geq 1. \end{aligned} \tag{2.6}$$

If the vectors from Definition .2, i.e. $v, Av, \ldots, A^{m-1}v$ are linearly independent, they form a basis for the Krylov subspace $\mathcal{K}_m(A, v)$. Furthermore, it has been shown in Gaul (2014) that there exists a minimal index d at which the Krylov subspace becomes invariant, i.e., $A\mathcal{K}_d(A, v) \subseteq \mathcal{K}_d(A, v)$. As a result, applying A to v will not result in an additional vector which can span the Krylov subspace any further. Using this index d, it has also been shown that a Krylov subspace, for a nonsingular matrix A, has the following properties:

1. *Dimension*: $\dim \mathcal{K}_m(A, v) = m$ for $m \leq d \leq n$.
2. *Nested sequence of subspaces*: $\mathcal{K}_{m-1}(A, v) \subseteq \mathcal{K}_m(A, v)$ for $m \geq 1$.
3. *The following statements are equivalent*:

- $A\mathcal{K}_d(A,v) = \mathcal{K}_d(A,v)$
- $\mathcal{K}_d(A,v) \cap \mathcal{N}(A) = \{0\}$
- $v \in A\mathcal{K}_d(A,v)$

A Krylov subspace method is essentially an iterative implementation of the Petrov-Galerkin method over the Krylov subpace from Definition .2 using $v = r_0$. If we take the Krylov subspace K_m as a basis for the search space as defined in Definition .1 and apply Theorem .2 up to the point where the subspace becomes invariant, we arrive at the heart of all Krylov subspace methods.

Corollary 2.1.1 *(Krylov Subspace Method) Consider a consistent linear system $Au = f$ with $A \in \mathbb{C}^{n\times n}$ and $f \in \mathbb{C}^n$. Let $u_0 \in \mathbb{C}^n$ be an initial guess corresponding to the initial residual $r_0 = f - Au_0$. Let $d < \infty$ be the minimal index at which $A\mathcal{K}_d(A,r_0) \subseteq \mathcal{K}_d(A,r_0)$ and let $\mathcal{K}_d(A,r_0) \cap \mathcal{N}(A) = \{0\}$. The sequence of iterates $\{u_m\}_{m\in 1,..,d}$ that satisfy*

$$u_m = u_0 + s_m,\ s_m \in S = \mathcal{K}_m(A,r_0),$$

$$r_m := f - Au_m \perp C_m,$$

is well defined and u_d is a solution of the linear system $Au_d = f$ if one of the following conditions holds:

1. *$C_m = \mathcal{K}_m(A,v)$, A is self-adjoint and positive semidefinite. Then the iterates u_m satisfy the optimality property*

$$\|u - u_m\|_A = \inf_{z\in u_0+\mathcal{K}_m(A,r_0)} \|u - z\|_A\,. \tag{2.7}$$

2. *$C_m = A\mathcal{K}_m(A,r_0)$. Then the iterates u_m satisfy the optimality property*

$$\|f - Au_m\| = \inf_{z\in u_0+\mathcal{K}_m(A,r_0)} \|f - Az\|\,. \tag{2.8}$$

Proof In both cases the well-definedness and optimality of the approximate solutions follow from Theorem .2. $Au_d = f$ follows from Theorem .1 and using the second property of the Krylov subspaces. For more details, see Gaul (2014) corollary 2.41, p. 31.

Theoretically, for $m \leq d \leq n$, the vectors $r_0, Ar_0, ..A^{m-1}r_0$ are linearly independent. They also form a basis for the Krylov subspace $\mathcal{K}_m(A,r_0)$. However, numerically this basis becomes indistinguishable from linear independence as the computation of the vector $A^i r_0$ using the power method usually points in the direction of the dominant eigenvector as i increases. As a result, if n is large, most of the vectors in $\mathcal{K}_m(A,r_0)$ will point to the same direction, rendering an ill-conditioned basis. Consequently, a Krylov subspace method is always constructed by implementing an basis orthonormalization process, such as the Arnoldi or Lanczos method (modified Gram-Schmidt), see Arnoldi (1951) and Lanczos (1952).

As a result of Corollary 2.1.1, different iterative Krylov subspace methods can be obtained by varying the constraint space C to be equal to either $\mathcal{K}_m$ or $A\mathcal{K}_m$. Indefiniteness of the coefficient matrix A restricts the applicability of several Krylov subspace methods for the Helmholtz equation, which are based on equation 2.7 from Corollary 2.1.1. For example, the well known CG-method[2] requires the input of a symmetric and positive-definite coefficient matrix A. Fortunately, the GMRES-method and the Bi-CGSTAB-method are considered suitable alternatives for this system, and are intrinsically based on equation 2.9 from Corollary 2.1.1.

2.2 GMRES-Method

The GMRES-method is based on the MINRES-method. The MINRES method was particularly developed as an extension of the Lanczos method to solve a linear system with a self-adjoint but indefinite

[2] The *Conjugate Gradient* method falls into the first category of Corollary 2.1.1, i.e. equation 2.7 and minimizes the error in terms of the A-norm. Where the GMRES-method uses an Arnoldi procedure for orthonormalizing the Krylov basis vectors, the CG-method uses the Lanczos method. The CG-method is widely used for large sparse SPD systems due to its superlinear convergence behavior. For more information, please refer to Vuik and Lahaye (2012) section 7.1.3 and Gaul (2014) section 2.8.

coefficient matrix A. The GMRES method was proposed for general matrices, interchanging the Lanczos method for the Arnoldi method. Both methods are characterized by minimizing the residual norm over the Krylov subspace. In essence, this translates into the minimization problem from Corollary 2.1.1, equation 2.9, which we can now reformulate specifically as

Theorem 2.2.1 *Consider a consistent linear system $Au = f$ with $A \in \mathbb{C}^{n\times n}$ and $f \in \mathbb{C}^n$. Let $u_0 \in \mathbb{C}^n$ and $r_0 = f - Au_0$ be such that $d < \infty$ and $\mathcal{K}_d(A, r_0) \cap \mathcal{N}(A) = \{0\}$ are fulfilled. Then, for $S_m = \mathcal{K}_m(A, r_0)$ and $C_m = A\mathcal{K}_m(A, r_0)$, the iterates $u_m = u_0 + s_m$, $s_m \in \mathcal{K}_m(A, r_0)$ minimize the residual norm, i.e.*

$$\|f - Au_m\| = \inf_{z \in u_0 + \mathcal{K}_m(A, r_0)} \|f - Az\|, \tag{2.9}$$

and u_d is a solution for the linear system .

Proof Applying Theorem .2 and Corollary 2.1.1 leads to the GMRES-method. For more details, please refer to Gaul (2014), section 2.9.1.

Note that Theorem 2.1.1 only holds for u_0 and r_0 satisfying $d < \infty$ and $\mathcal{K}_d(A, r_0) \cap \mathcal{N}(A) = \{0\}$. However, as long as A is non-singular, these conditions are automatically satisfied and the GMRES-method is well-defined for any initial choice u_0 (Gaul, 2014). Consequently, in upcoming sections we will present the results assuming that the coefficient matrix A is non-singular.

2.2.1 Arnoldi's-Method

We have previously mentioned that the application of Krylov subspace methods goes hand in hand with an orthonormalization procedure in order to obtain a well-conditioned basis for the Krylov subspace. As the GMRES-method is applicable to general and thus non-symmetric matrices, the Arnoldi procedure is used to construct a set of orthonormal basis vectors, which in algorithmic form is given below

Algorithm 1 Arnoldi's Orthonormalization Algorithm

```
Choose $v_1$ with $\|v_1\| = 1$
for $j = 1, 2, ...m$ do
    $w_j := Av_j$
    for $i = 1, 2, \ldots, j$ do
        $h_{i,j} := \langle Av_j, v_i \rangle$
        $w_j = w_j - \sum_{i=1}^{j} h_{i,j} v_i$
        $h_{j+1,j} := \|w_j\|$
        $v_{j+1} := \frac{w_j}{h_{j+1,j}}$
    end for
end for
```

Each step in the algorithm multiplies v_j by A and orthonormalizes the vector w_j with respect to all previous Arnolid vectors v_i from $i = 1$ to j. Using the Arnoldi method, we arrive at two widely used propositions, see Saad (2011), p. 129.

Proposition 2.2.2 *Assume that Arnoldi's algorithm does not stop before the $m-$th step. Then the vectors $v_1, v_2, \ldots, v_m$ form an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A, v_1)$.*

Proposition 2.2.3 *Let V_m be the $m \times m$ matrix with column vectors $v_1, v_2, \ldots, v_m$. Let $\widehat{H_m}$ be the $((m+1) \times m)$ Hessenberg matrix whose nonzero entries $h_{i,j}$ are defined by Arnoldi's method and let $e_m = \{0, 0, \ldots, 1\}^T$. If we let H_m be the matrix obtained from $\widehat{H_m}$ by deleting its last row, then the following relation holds*

$$AV_m = V_m H_m + w_m {e_m}^T, \tag{2.10}$$

$$= V_{m+1} \widehat{H_m}, \tag{2.11}$$

$${V_m}^T A V_m = H_m. \tag{2.12}$$

We can implement Arnoldi's method into Theorem 2.2.1, by noting that iterate vectors u_m can be written as $u_m = u_0 + V_m s_m$, where s_m is a vector in $\mathbb{C}^m$ and V_m is an orthonormal basis for the Krylov subspace.

If we let $\beta = \|r_0\|$ and $v_1 = r_0 / \|r_0\|$, we can use use equation 2.11 to obtain

$$\begin{aligned} \|f - Au_m\| &= \|f - A(u_0 + V_m s_m)\| \, , \\ &= \|r_0 - AV_m s_m\| \, , \\ &= \left\| \beta v_1 - V_{m+1} \widehat{H_m} s_m \right\| , \\ &= \left\| V_{m+1} (\beta e_1 - \widehat{H_m} s_m) \right\| . \end{aligned} \tag{2.13}$$

By definition, the columns of V_{m+1} are orthonormal and we can rewrite equation 2.13 as follows

$$\left\| V_{m+1} (\beta e_1 - \widehat{H_m} s_m) \right\| = \left\| \beta e_1 - \widehat{H_m} s_m \right\| . \tag{2.14}$$

The optimality property from equation 2.9, Theorem 2.2.1 becomes

$$\begin{aligned} \|f - Au_m\| &= \left\| \beta e_1 - \widehat{H_m} s_m \right\|_m , \\ &= \min_{z \in \mathbb{C}^n} \left\| \beta e_1 - \widehat{H_m} z \right\| . \end{aligned} \tag{2.15}$$

As a result, the approximate solution is the unique z vector which minimizes $F(z) = \min_{z \in \mathbb{C}^n} \left\| \beta e_1 - \widehat{H_m} z \right\|$ over $\mathcal{K}_m(A, r_0)$ which iteratively reduces to finding

$$s_m = \arg \min_{z \in \mathbb{C}^n} \left\| \beta e_1 - \widehat{H_m} z \right\| .$$

2.2.2 GMRES-Algorithm

The GMRES-method can be implemented using the following algorithm:

Algorithm 2 GMRES-method $Au = f$

```
 1: Choose $u_0$ and compute $r_0 = f - Au_0$, $b_0 = ||r_0||$ and $v_1 = r_0/b_0$
 2: for $j = 1, 2, \dots n$ or until convergence do
 3:     $w_j := Av_j$
 4:     for $i := 1, 2, \dots, j$ do
 5:         $h_{i,j} := (w_j, v_i)$
 6:         $w_j := w_j - h_{i,j} v_i$
 7:     end for
 8:     $h_{j+1,j} := \|w\|$
 9:     $v_{j+1} := \frac{w}{h_{j+1,j}}$
10: end for
```

Note that steps 2 to 10 are in fact the Arnoldi orthonormalization algorithm. The GMRES-method is stable and only breaks down if $h_{j+1,j} = 0$. However, if $h_{j+1,j} = 0$ then $u_j = u$ and we retrieve the exact solution Vuik and Lahaye (2012).
The GMRES-method is considered inefficient in case a large number of iterations are needed. Due to its long recurrences, it requires increasing memory storage and computational force for the orthonormalization process. Several remedies have been opted to circumvent this drawback. For example, the GMRES-method can be restarted, see (Vuik and Lahaye, 2012), section 7.3.4.

In order to accelerate the convergence, preconditioning techniques are available for Krylov subspace methods. For the GMRES-method in particular, a preconditioned variant can be obtained by applying the GMRES-method to the following linear system

$$\begin{aligned} M^{-1}Au = M^{-1}f \Leftrightarrow AMy = f, u = My, \\ A \in \mathbb{C}^{n \times n}, \, u, x, f \in \mathbb{C}^n, \end{aligned}$$

where M is an invertible matrix in $\mathbb{C}^{n\times n}$. In general, a matrix M is eligible as a preconditioner if the eigenvalues of $M^{-1}A$ are clustered around 1 and $M^{-1}y$ can be obtained at low cost. We will treat the subject of preconditioning to more detail in Chapter 2.

2.2.3 Convergence

In this section we briefly describe the convergence properties of the GMRES-method, which is based on the following theorem

Theorem 2.2.4 *Let P_m be the space of all polynomials of degree less than m and let $\sigma = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$ represent the spectrum of A. Moreover, we define*

$$\varepsilon^m = \min_{p\in P_m, p(0)=1} \max_{\lambda_i \in \sigma} |p(\lambda_i)| .$$

Suppose that A is diagonalizable so that $A = XDX^{-1}$ where D is a diagonal matrix containing $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Then the residual norm of the $m-$th iterate satisfies

$$\frac{\|r_m\|_2}{\|r_0\|_2} = \min_{p\in P_m, p(0)=1} \frac{\left\|XP(D)X^{-1}r_0\right\|_2}{\|r_0\|_2} \leq K(X)\varepsilon_m \|r_0\|_2 , \tag{2.16}$$

where $K(X) = \|X\|_2 \left\|X^{-1}\right\|_2$.

Proof For a proof see Vuik and Lahaye (2012), Theorem 7.3.1. and Liesen and Tichý (2004), section 3.1.

Eiermann and Ernst (2001) state that it would be impossible to predict the convergence behavior of the GMRES-method *solely* in terms of the eigenvalues of A. In fact, the author argues that in case convergence is monitored through the spectrum, additional assumptions on *departure* from normality are a necessity. Liesen and Tichý (2004) have presented an extensive overview of the convergence properties of Krylov subspace methods. The problem with non-normality seems to be related to ill-conditioned eigenvectors resulting in very large $K(X)$ due to $\left\|X^{-1}r_0\right\| > \|r_0\|$. As a result, the bound in equation 2.16 may not be sharp and information regarding the convergence may be disconnected from spectral properties. However, Hannukainen (2015), Liesen and Tichý (2004) and Meurant and Tebbens (2015) all argue that for a large class of matrices, such as general normal and Hermitian matrices, the convergence results in terms of the spectral distribution properties hold. Liesen and Tichý (2004) even emphasize that theoretically, non-normality of a matrix does *not* lead to slower convergence, as for each non-normal matrix A there exist a normal matrix B with the same convergence behavior.

For normal matrices A in general, the eigenvectors form an orthonormal set making X in equation 2.16 well-conditioned. Due to the orthonormality, the eigenvalues have a predominant influence on the rate of convergence. Consequently, clustering and favorably distributed eigenvalues stimulate convergence, while eigenvalues close to the origin impede convergence. The resulting slow convergence can often be alleviated by eliminating these convergence hampering eigenvalues through the application of deflation techniques. We will treat this extensively in Chapter 4.

2.3 Bi-CGSTAB-Method

We have previously mentioned that the Bi-CGSTAB-method is another Krylov subspace method that can be applied to general matrices. It can be viewed as a combined version of the Bi-CG-method[3] and the GMRES-method. In essence, the Bi-CGSTAB-method uses two mutually bi-orthogonal bases to construct the Krylov subspace. The residual is kept bi-orthogonal to both bases by solving minimum residual polynomials. Despite being a suitable iterative method for general matrices, the convergence behavior is noted to be irregular for ill-conditioned matrices (Vuik and Lahaye, 2012). For the sake of completeness yet brevity, we present the algorithm in the next subsection.

[3] The *Bi-Conjugate Gradient* method builds on the CG-method, by using a Bi-Lanczos procedure to orthonormalize the Krylov basis vectors. It is generally used for non-symmetric linear systems. For more information, please refer to Vuik and Lahaye (2012) section 7.3.3.

2.3.1 Bi-CGSTAB-Algorithm

By preconditioning M, the linear system $Au = f$ can be solved by implementing the following preconditioned Bi-CGSTAB algorithm:

Algorithm 3 Bi-CGSTAB Method

```
Pick $u^0$ as an initial estimate and compute $r^0 = f - Au^0$;
Choosing $\bar{r}^0$ as an arbitrary vector such that $(\bar{r}^0, r^0) \neq 0$, e.g. for $\bar{r}^0 = r^0$;
$\rho_{-1} = \alpha_{-1} = \omega_{-1} = 1$;
$v^{-1} = p^{-1} = 0$;
for j=0, 1,2, ... do
    $\rho_i = (\bar{r}^0, r^i)\,; \beta_{i-1} = (\rho_i/\rho_{i-1})(\alpha_{i-1}/\omega_{i-1})\,;$
    $p^i = r^i + \beta_{i-1}(p^{i-1} - \omega_{i-1}v^{i-1})$;
    $\hat{p} = M^{-1}p^i$;
    $v^i = A\hat{p}$;
    $\alpha_i = \rho_i / (\bar{r}^0, v^i)$;
    $s = r^i - \alpha_i v^i$;
    if $\|s\| < \mathcal{O}^n$ where $n > 0$ is small, then;
        $u^{i+1} = u^i + \alpha_i \hat{p}$ ;
    end if
    $z = M^{-1}s$ ;
    $t = Az$ ;
    $\omega_i = (t, s) / (t, t)$;
    $u^{i+1} = u^i + \alpha_i \hat{p} + \omega_i z$;
    if $u^{i+1} < \mathcal{O}^n$ where $n > 0$ is small, then ;
    end if
    $r^{i+1} = s - \omega_i t$;
end for
```

2.4 Starting, Monitoring and Stopping

Implementing iterative solution methods requires an initial guess as primary input. Moreover, a process for monitoring the convergence is needed by choosing a stopping criterion. [4]

Starting Guess

Generally, the initial guess is determined by the context of the problem. In some cases the discretized linear system is solved on a coarser grid and interpolated back to the fine grid to serve as an initial guess. For the GMRES-method in particular, it has been noted in section 2.2 that the method converges for any initial guess.

Stopping Criterium

A stopping criterion aims at balancing the quality imposed on the solution relative to the computational costs. In practice, a solution method is set to iterate until the residual norm $\|r_n\| \leq \epsilon$ becomes smaller than some small number ϵ, where ϵ denotes the tolerance level. Three stopping criteria are available:

1. $\|r_n\| \leq \epsilon$: this stopping criterion is not scaling invariant and is not preferred.
2. $\frac{\|r_n\|}{\|r_0\|} \leq \epsilon$: this stopping criterion is dependent on the initial guess. As a result, the accuracy increases with the accuracy of the initial guess.
3. $\frac{\|r_n\|}{\|f\|} \leq \epsilon$: this stopping criteria is widely used in practice and preferred.

[4] This section contains summarizing parts from section 5.6, Vuik and Lahaye (2012).

Note that in case the residual vector can be constructed using a number of dominant eigenvector components, a random initial guess may be more suitable. However, for a right hand side f defined as a source function, the likelihood of such an event occurring is small. Therefore, in upcoming sections, we will implement a zero initial guess, unless stated otherwise.
With respect to the stopping criterium, we will employ a tolerance level of $\epsilon = 10^{-7}$ in upcoming sections, unless stated otherwise.

2.5 Multigrid Methods

In this section we will describe the basic idea behind multigrid methods using coarse grids, applied to our one-dimensional model problem, see Chapter 1, section 1.3. We will limit ourselves to describing the building blocks from multigrid methods which are of importance for the ADEF-preconditioner. As mentioned in the introductory part of this chapter, standard multigrid methods where a BIM acts as a smoother, are not suitable for the Helmholtz problem, unless the wave number is small enough relative to the step size. It has been shown by Ernst and Gander that damped Jacobi relaxation breaks down at high wave numbers. Therefore, we will only briefly discuss these smoothing properties in section 2.5.2. Most of this section is contained in section 6.2 of Vuik and Lahaye (2012), section 3.4.2 of Ernst and Gander (2012) and section 3.2 of Ernst and Gander.

The main idea behind the use of different grid refinement levels in multigrid methods, was the notion that the low frequency modes of the iteration error from solving a linear system using BIMs was not being reduced sufficiently. These low frequency modes are related to the eigenvectors corresponding to the small eigenvalues of the linear system. Suppose we would like to solve the linear system obtained from discretisizing a simple one-dimensional Poisson problem

$$\begin{aligned} &Au = f, \\ &A \in \mathbb{R}^{n\times n},\ u, f \in \mathbb{R}^n. \end{aligned} \tag{2.17}$$

The eigenmodes can be divided into low and high frequency modes. The low frequency modes are slowly varying grid vectors that correspond to the small eigenvalues of A. The eigenvectors of the matrix A are

$$\begin{aligned} v_{hl} &= \begin{pmatrix} \sin \pi l h \\ \sin \pi l 2h \\ \vdots \\ \sin \pi l (n-1) h \end{pmatrix}, \\ &1 \leq l \leq n-1. \end{aligned} \tag{2.18}$$

For now we assume $n-1$ to be even. Recall from Chapter 1, section 1.4.3 that the eigenvectors are sine-functions applied to the grid vectors $\mathbf{x} = [x_i] = ih$, with $i = 1, 2, \ldots, n-1$. For increasing l, the eigenvectors become more oscillatory. The indices $l = 1$ to $\frac{n}{2} - 1$ therefore relate to the low frequency modes, whereas the remaining eigenmodes represent high frequency modes. By transferring these low frequency eigenvectors onto a coarse grid, their smooth components become oscillatory and can be reduced. We will treat this in more detail in the next section, where we apply these ideas to our one-dimensional model problem from Chapter 1, section 1.3.

2.5.1 Coarse-Grid Correction

The key ingredient of multigrid methods is the use of coarser grids, where smooth components become oscillatory. This section contains some excerpts of Ernst and Gander, section 3.4.2, p.17 - 25. It essentially describes the application of a two-grid multigrid method applied to the discretisized one-dimensional Helmholtz problem. Recall that the eigenvectors corresponding to the linear system of our model problems coincide with equation 2.18.

We start by defining the transfer grid functions $u_H = [u_{H_1}, \ldots, u_{H_n}]$ from Ω_H to the fine grid Ω_h using a standard linear interpolation mapping

$$I_h^H : \Omega_H \to \Omega_h, \quad u_H \to I_h^H u_H \tag{2.19}$$

such that

$$\begin{cases} [u_H]_{i/2} & \text{if } i \text{ is even,} \\ \frac{1}{2}\left([u_H]_{(i-1)/2} + [u_H]_{(i-1)/2}\right) & \text{if } i \text{ is odd,} \end{cases} \quad i = 1, \ldots, n-1 \tag{2.20}$$

with matrix representation

$$I_H^h = \frac{1}{2}\begin{bmatrix} 1 & & & \\ 2 & & & \\ 1 & 1 & & \\ & 2 & & \\ & 1 & & \\ & & \ddots & 1 \\ & & & 2 \\ & & & 1 \end{bmatrix} \in \mathbb{R}^{n\times(n)/2-1} \tag{2.21}$$

Using the eigenvectors given in equation 2.18, we obtain the following proposition, see Ernst and Gander, p. 19.

Proposition 2.5.1 *The coarse-grid eigenvectors are mapped by the interpolation operator I_H^h according to*

$$I_H^h v_{H_l} = c_l^2 v_{h_l} - s_l^2 v_{h_{n-1-l}}, \quad l = 1, \ldots, \frac{n}{2} - l, \tag{2.22}$$

where we define

$$c_l := \cos\frac{l\pi h}{2}, \quad s_l := \sin\frac{l\pi h}{2}, \quad l = 1, \ldots, \frac{n}{2} - 1. \tag{2.23}$$

As a result, the coarse-grid modes v_{H_l} are mapped to a linear combination of their fine grid counterparts v_{h_l} and a complementary mode $v_{h_{l'}}$, where $l' := n - 1 - l$. Moreover, we have

$$c_{l'} = s_l \quad s_{l'} = c_l, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.24}$$

In order to transfer fine-grid functions to a coarse grid, we define the restriction operator

$$I_H^h : \Omega_h \to \Omega_H, \quad u_h \to I_H^h u_h \tag{2.25}$$

by

$$\left[I_H^h u_h\right]_i = \frac{1}{4}\left([u_h]_{2i-1} + 2[u_h]_{2i} + [u_h]_{2i+1}\right), \quad i = 1, \ldots, \frac{n}{2} - 1. \tag{2.26}$$

The associated matrix representation is given by $I_H^h = \frac{1}{2}\left[I_h^H\right]^T$. Note that the current standard restriction weights form a convex linear combination. Changing the weight coefficients, will result in a different interpolation and restriction operator respectively. The following proposition can be proven for I_H^h, see Ernst and Gander, p. 20.

Proposition 2.5.2 *The fine-grid eigenvectors are mapped by the restriction operator I_H^h according to*

$$I_H^h v_{h_l} = c_l^2 v_{H_l}, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.27}$$

$$I_H^h v_{h_{N+1-l}} = -s_l^2 v_{H_l}, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.28}$$

$$I_H^h v_{h_{n+1}} = 0. \tag{2.29}$$

Let u_h be an approximate solution to our model problem. Then the coarse-grid correction of u_h can be obtained by solving the error equation $A_h e_h = f - A_h u_h = r_h$ on the coarse grid. We start by defining a coarse-grid representation A_H of A_h and solve for $A_H^{-1} I_h^H r_h$, where r_h is first restricted to the coarse grid.

A_H is more commonly referred to as the *Galerkin Coarsening Matrix.* Note that $A_H^{-1} I_h^H r_h$ approximates the error $e_H = A_h^{-1} r_h$ on Ω_H. As a last step, e_h is interpolated to the fine grid by

$$u_h \leftarrow u_h + I_H^h A_H^{-1} I_h^H \left(b - A_h u_h\right), \tag{2.30}$$

with the associated error propagation operator

$$C := I - I_H^h A_H^{-1} I_h^H A_h. \tag{2.31}$$

We thus get the following recursive relation for the error

$$e_{h\,l+1} = C_{h\,l} e_0 \tag{2.32}$$

It has been noted that C spans two invariant subspaces corresponding to the index set $l = 1, \ldots, \frac{n}{2} - 1$ and $l' = n - 1 - l$. Recall that the eigenvalues of the discrete one-dimensional Laplacian operator on Ω_h and Ω_H are given by

$$\lambda_{h_l} = \frac{4}{h^2} \sin^2 \frac{l \pi h}{2} - k^2, \quad l = 1, \ldots, n - 1 \tag{2.33}$$

and

$$\lambda_{H_l} = \frac{4}{H^2} \sin^2 \frac{l \pi H}{2} - k^2, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.34}$$

Letting span $\left\{v_{h_l}, v_{h_{l'}}\right\}$ denote an invariant subspace, i.e.

$$C\left[v_{h_l}, v_{h_{l'}}\right] = \left[v_{h_l}, v_{h_{l'}}\right] C_l, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.35}$$

$$C v_{h_{n/2}} = v_{h_{n/2}}, \tag{2.36}$$

we can write C from equation 2.31 as follows

$$C_l = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} c_l^2 \\ -s_l^2 \end{bmatrix} \frac{1}{\lambda_{H_l}} \left[c_l^2 - s_l^2\right] \begin{bmatrix} \lambda_{h_l} & 0 \\ 0 & \lambda_{h_{l'}} \end{bmatrix} = \begin{bmatrix} 1 - c_l^4 \frac{\lambda_{h_l}}{\lambda_{H_l}} & c_l^2 s_l^2 \frac{\lambda_{h_{l'}}}{\lambda_{H_l}} \\ c_l^2 s_l^2 \frac{\lambda_{h_l}}{\lambda_{H_l}} & 1 - s_l^4 \frac{\lambda_{h_{l'}}}{\lambda_{H_l}} \end{bmatrix}. \tag{2.37}$$

Moreover, the following proposition can be proven, see Ernst and Gander, p. 23.

Proposition 2.5.3 *(Spectrum of C) The eigenvalues of the 2×2 blocks from equation 2.37 representing the coarse grid correction operator are given by*

$$\Lambda(C_l) = \left\{1 - \tfrac{c_l^4 \lambda_{h\,l} + s_l^4 \lambda_{h\,l'}}{\lambda_{H\,l}}, 1\right\}, l = 1, \ldots, \frac{n}{2} - 1 \tag{2.38}$$

with eigenvectors

$$w_l^1 = \begin{bmatrix} c_l^2 \\ -s_l^2 \end{bmatrix} \tag{2.39}$$

$$w_l^2 = \frac{4}{h^2} \begin{bmatrix} s_l^2 \left(c_l^2 - \tfrac{hk}{2}^2\right) \\ c_l^2 \left(s_l^2 - \tfrac{hk}{2}^2\right) \end{bmatrix}. \tag{2.40}$$

The non-unit eigenvalues reduce to

$$\nu_l = \nu_l(kh) = \frac{\frac{(kh)^2}{2}}{\frac{(kh)^2}{(2 s_l c_l)} - 1}, l = 1, \ldots, \frac{n}{2} - 1, \tag{2.41}$$

If k is zero, we obtain the discrete one-dimensional Laplacian operator and the expressions from equation 2.33, 2.34 and 2.37 simplify to

$$\frac{\lambda_{h_l}}{\lambda_{H_l}} = \frac{4 s_l^2}{\left(2 s_l c_l\right)^2} = \frac{1}{c_l^2} \quad \text{as well as} \quad \frac{\lambda_{h_{l'}}}{\lambda_{H_l}} = \frac{4 c_l^2}{\left(2 s_l c_l\right)^2} = \frac{1}{s_l^2}, \quad l = 1, \ldots, \frac{n}{2} - 1, \tag{2.42}$$

and therefore

$$C_l = \begin{bmatrix} 1 - c_l^2 & c_l^2 \\ s_l^2 & 1 - s_l^2 \end{bmatrix} = \begin{bmatrix} s_l^2 & c_l^2 \\ s_l^2 & c_l^2 \end{bmatrix}, \quad l = 1, \ldots, \frac{n}{2} - 1. \tag{2.43}$$

For $k = 0$, the operator C is an orthogonal projection and has only two eigenvalues 0 and 1. Also, the eigenvectors corresponding to 0 and 1 respectively are

$$w_l^1 = \begin{bmatrix} c_l^2 \\ -s_l^2 \end{bmatrix} \tag{2.44}$$

$$\tag{2.45}$$

$$w_l^2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2.46}$$

For small l, w_l^1 reduces to approximately $[0,1]^T$ since $c_l^2 \approx 1$ and $s_l^2 \approx 0$. Thus, as Ernst and Gander put it, the eigenmode w_l^1 eliminated by the coarse grid correction is closely aligned with the low-frequency eigenmode v_{h_l}. This alignment becomes less as l increases.

In case the wave number k is positive, Ernst and Gander (2012) state that for $k = 6.3\pi$ the unit eigenvalues of C remain but the zero eigenvalue starts to shift. As a result, low frequency modes corresponding to small eigenvalues of the Helmholtz operator may be partially unaffected.
We will show this by means of an example. Suppose the small eigenvalues of C corresponding to the eigenvectors in the low frequency range are of order ε, with $0 < \varepsilon << 1$. The eigenvectors corresponding to these eigenvalues for small l are the same for the one-dimensional Laplacian operator. Thus, for small l up to some index d, where $l \leq d \leq \frac{n}{2} - 1$, we assume that the eigenmodes w_l^1 corresponding to the zero eigenvalue of the operator C_l and the low frequency modes v_{h_l} are closely aligned.
From equation 2.32 we know that the error propagates as follows

$$e_{h_{l+1}} = C_l e_{h_0}. \tag{2.47}$$

If we decompose the initial error e_0 in terms of the eigenvectors of A we obtain

$$e_{h_0} = \begin{bmatrix} \gamma_l v_{h_l} \\ \gamma_l' v_{h_{l'}} \end{bmatrix}^T$$
$$l = 1, \ldots, d,$$

where γ_l corresponds to suitable coefficients for the low frequency range and $\gamma_{l'}$ represents the coefficients with respect to the high frequency range. Similarly, for $e_{h_{l+1}}$ we can write

$$e_{h_{l+1}} = \begin{bmatrix} \widehat{\gamma_l} v_{h_l} \\ \widehat{\gamma_l'} v_{h_{l'}} \end{bmatrix}^T,$$
$$l = 1, \ldots, d.$$

Applying the coarse-grid error propagation matrix C_l according to equation 2.47 thus gives

$$\begin{bmatrix} \widehat{\gamma_l} v_{h_l} \\ \widehat{\gamma_l'} v_{h_{l'}} \end{bmatrix}^T = C_l \begin{bmatrix} \gamma_l v_{h_l} \\ \gamma_l' v_{h_{l'}} \end{bmatrix}^T \tag{2.48}$$

Multiplying by $w_l^1 = [1, 0^T]$ on both sides gives

$$e_{h_{l+1}} w_l^1 = \begin{bmatrix} \widehat{\gamma_l} v_{h_l} \\ \widehat{\gamma_l'} v_{h_{l'}} \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \widehat{\gamma_l} v_{h_l}.$$

The left hand side of equation 2.48 becomes

$$C_j \begin{bmatrix} \gamma_l v_{h_l} \\ \gamma_l' v_{h_{l'}} \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

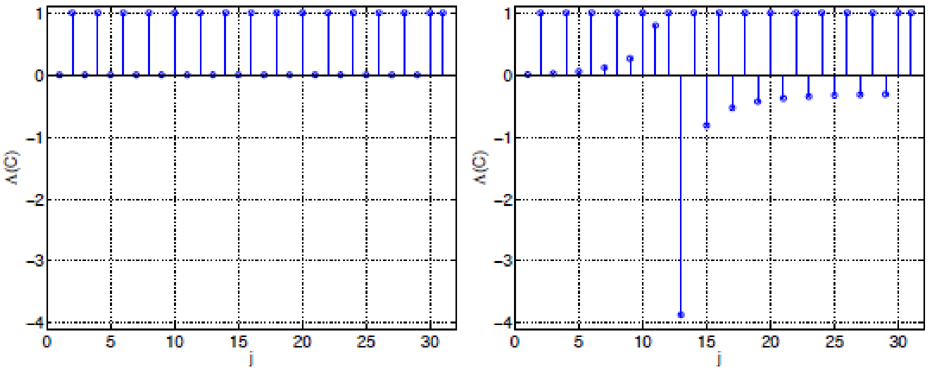
Using that $w_l^1 = [1, 0]^T$ is an eigenvector of C_l corresponding to $l = 1$ up to $l = d$ we can rewrite the expressions into

$$e_{h_{l+1}} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \widehat{\gamma_l} v_{h_l} = C_l \begin{bmatrix} \gamma_l v_{h_l} \\ \gamma_l' v_{h_{l'}} \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma_l v_{h_l} \\ \gamma_l' v_{h_{l'}} \end{bmatrix}^T \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix} = \gamma_l \varepsilon v_{h_l}$$

Thus, if ε is not equal to zero, the low frequency modes for $l = 1$ to $l = d$, do not get removed and propagate further as the error develops. A similar conclusion can be drawn by looking at Figure 3.1 [5]. Ernst and Gander have plotted the spectrum of C for $k = 0$ and $k = 6.3\pi$ in Figure 3.1.

[5] Figure 3.1 has been taken from Ernst and Gander, section 3.2, page 24.

Figure 2.1: *Eigenvalues of C for $n = 32$. Left $k = 0$ and right $k = 6.3\pi$.*



For $k > 0$ we indeed see some eigenvalues deviating from zero. On another note, some of these low frequency modes, instead of being projected onto zero, become amplified and do not partake in the error smoothing process. An effective remedy to obtain a better coarse-grid correction operator for the Helmholtz problem is by incorporating the dispersion properties of the discretization scheme, which will be treated in section 2.6.

2.5.2 Smoothing

In the introductory section of this chapter we briefly mentioned basic iterative methods. While inefficient for solving large scale Helmholtz problem, they are generally accepted as reducing high frequency components of the iteration error in a general multigrid setting. In order to reduce the low frequency components basic iterative methods, such as the damped Jacobi, are implemented as *smoothers*. By choosing an optimal smoothing parameter, the high frequency components of the iteration error are reduced as the damped Jacobi act as a low-pass filter. In section 2.5.1 we described the general two-grid method for the discrete Helmholtz and Poisson operator. The two-grid method can be combined with a basic iterative method as a smoother by solving the coarse-grid correction scheme from equation 2.30 using a basic iterative method, such as damped Jacobi or Gauss-Seidel. As a result, the error propagation matrix acting on the iteration error decomposes into pre and post-smoothing steps, see Vuik and Lahaye (2012) for more details.

2.6 Pollution

In section 2.5.1, we briefly pointed to the pollution effect due to differences between the exact and numerical wave number. The pollution effect was first mentioned by Deraemaeker et al. (1999) for Finite Element Solutions (FEM) of the Helmholtz equation. The accuracy of the numerical solution deteriorates due to the wave number for the numerical solution being different from the wave number of the true solution. These differing wave numbers appear to be responsible for numerical dispersion. Effectively this means that the numerical solution is traveling at a different wave speed than the exact solution introducing a phase error. This effect accumulates as k increases and results in non-accurate error estimates.

2.6.1 Error Bounds

[6] To understand how the dispersion error depends on $\kappa = kh$, recall from Chapter 1, section 1.3.3 that the dimensionless wave number k is represented by

$$k = \frac{2\pi fl}{c} = \frac{\omega}{c} = \omega m,$$

where ω denotes the angular frequency and $c = \frac{1}{m}$ is the wave speed [7]. Discretisizing the one-dimension Helmholtz equation based on the grid G_{h_1D} from Chapter ??, section 1.4.1 leads to

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} - m^2\omega^2 u_j = 0. \tag{2.49}$$

Moreover, a general continuous solution is given by

$$u\left(x\right) = e^{i\omega m x}. \tag{2.50}$$

Evaluation of expression 2.50 in the discrete points gives

$$u_j = e^{i\omega\tilde{m}x_j}. \tag{2.51}$$

Here i denotes the imaginary unit and $\tilde{m}$ contains a perturbation such that $\tilde{m} = m + \mathcal{O}((h\omega)^2)$. The perturbed wave speed $\tilde{c}$ becomes $\tilde{c} = 1/\tilde{m} = c + \mathcal{O}((h\omega)^2)$, which leads to an overal perturbed wave number $k = \omega\tilde{m}$. Substituting equation 2.51 into the equation 2.49 results in

$$u_{j+1} - 2u_j + u_{j-1} = e^{i\omega\tilde{m}x_j}\left(e^{i\omega\tilde{m}h} - 2 + e^{-i\omega\tilde{m}h}\right) = 2\left[\cos\left(\omega\tilde{m}h\right) - 1\right]e^{i\omega\tilde{m}x_j}. \tag{2.52}$$

Equation 2.52 holds if $\tilde{m}$ solves

$$\frac{2\,\cos\left(\omega\tilde{m}h\right) - 1}{h^2} - m^2\omega^2 = 2(\cos\left(\omega\tilde{m}h\right) - 1) - m^2\omega^2h^2 = 0.$$

Applying Taylor's expansion on the cosine term leads to

$$-\left(\omega\tilde{m}h\right)^2 + \mathcal{O}\left(h^4\tilde{m}^2\omega^4\right) - m^2\omega^2h^2 = -\tilde{k}^2h^2 - k^2h^2 + \mathcal{O}(\tilde{k}^4h^4) = 0, \tag{2.53}$$

which delivers the following bound on $\tilde{k}$ and k

$$\left|\tilde{k} - k\right| = \mathcal{O}(k^2h^2), \text{ for } \tilde{k} \approx k.$$

The error due to $\left|\tilde{k} - k\right| \neq 0$ becomes

$$\text{error}_{pollution} = \left|e^{ikx_j} - e^{i\tilde{k}x_j}\right| = \left|1 - e^{i\left(\tilde{k}-k\right)x_j}\right| \leq Ck\left|\tilde{k} - k\right| \leq Ck^3h^2.(2.54)$$

The factor Ck^3h^2 can be decomposed as follows. $\mathcal{O}(k^2h^2)$ provides the error in the numerical wave speed for a wave traveling one period. The extra factor k is called the *pollution error* and corrects the total pollution error by scaling the error over one wave length by the total number of wave lengths traveled over the entire numerical domain.

Deraemaeker et al. (1999) note that the error given in equation 2.54 solely relates to the dispersion caused by the differing wave numbers. The total error for the discretized one-dimension Helmholtz operator is given by

$$\text{error}_{total} = \tfrac{\|u-\hat{u}\|}{\|u\|} \leq C_1kh + C_2k^3h^2,\ kh < 1.(2.55)$$

While applying the rule of thumb $\kappa = kh \leq 0.625$ is sufficient for keeping the first term under control, it does not harbour properly against the propagation of the pollution error which grows rapidly with k, even if κ is kept small enough. Thus, it has been advocated to set the grid resolution to $k^3h^2 \leq \epsilon$ instead of $\kappa = kh \leq 0.625$ (Sheikh, 2014). The pollution effect is illustrated in Figure 2.2 for $k = 50$, where the real part of the exact and numerical solution is plotted. The numerical solution has been obtained by preconditioned GMRES-iterations [8]

[6] This section has been adjusted from Runborg (2012), p.9. and contains most results mentioned in the cited section.

[7] Runborg (2012) refers to m as the constant index of refraction, i.e. the inverse of speed propagation.

[8] The exact solution has been discussed in section ?? of Chapter ??.

Figure 2.2: *Exact and numerical solution for $k = 50$. The point-source has been placed at $x = 0.5$.*

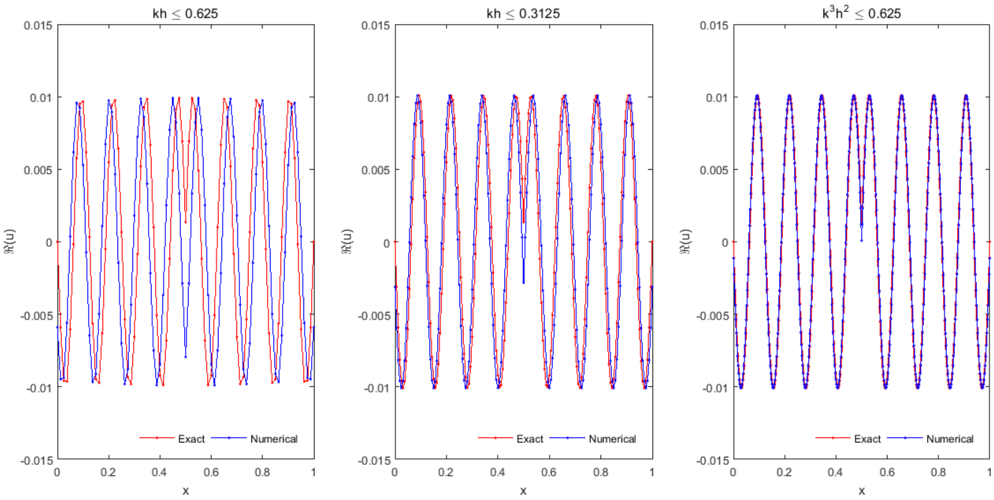


Figure 2.2 confirms that refining the grid leads to a more accurate numerical solution. When keeping $k^3h^2 \leq \epsilon$, the numerical dispersion error is indeed minimized. Minimizing the pollution error inevitably leads to large linear systems and becomes unpractical in higher dimensions.

2.6.2 Pollution and Multigrid

Ernst and Gander state that a more effective coarse-grid correction operator can be obtained by using a modified wave number k_H in order to bring the coarse-grid approximations in phase with the fine-grid approximations. This can be achieved by taking H instead of h in the following expressions:

$$k_H = \frac{\arccos\left(1 - \frac{\tilde{k}^2H^2}{2}\right)}{H} = k \Rightarrow \tilde{k} = \sqrt{\frac{2(1-\cos(kH))}{H^2}}. \tag{2.56}$$

An alternative would be to equate the *discrete* coarse-grid modified wave number k_H to the fine-grid *discrete* wave number $\tilde{k}$, i.e.

$$\tilde{k}_H = \tilde{k} \tag{2.57}$$

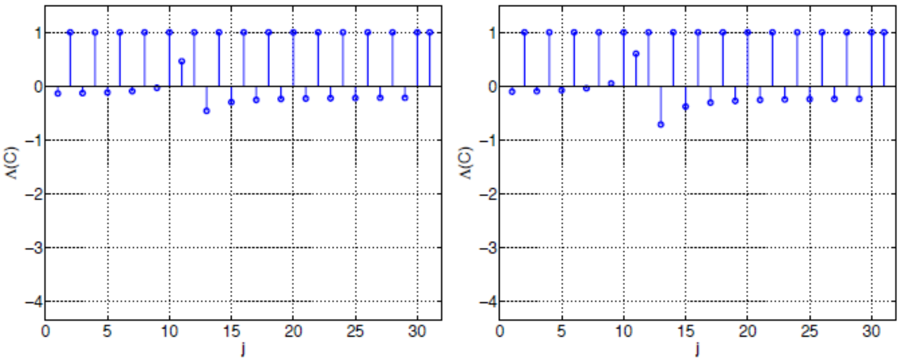
$$\frac{\arccos\left(1 - \frac{\tilde{k}^2H^2}{2}\right)}{H} = \frac{\arccos\left(1 - \frac{k^2h^2}{2}\right)}{h}. \tag{2.58}$$

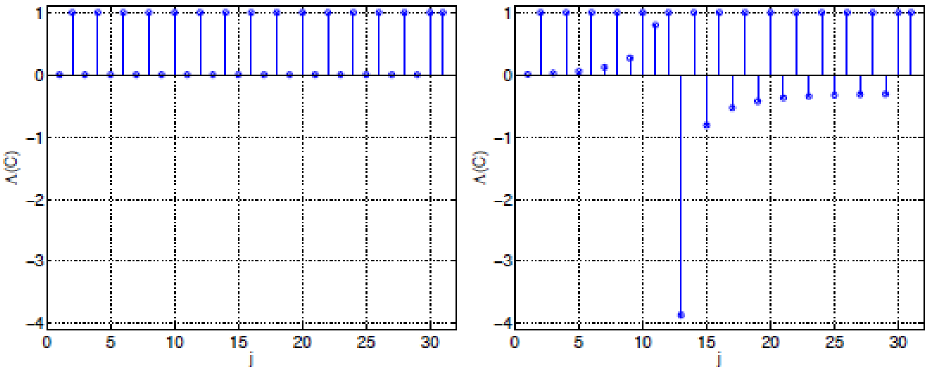
Using $\cos(2x) = 2\cos(x)^2 - 1$ and the fact that $H/h = 2$, equation 2.58 becomes

$$\tilde{k_H} = k\sqrt{1 - \frac{k^2h^2}{4}}. \tag{2.59}$$

Ernst and Gander have studied the effect of implementing the modified wave number $\tilde{k}$ according to equation 2.56 and 2.58 into the coarse-grid discrete operator C_l from section 2.5.1. Their results are presented in Figure 2.3 (upper left and right) and compared to the case where C is constructed on the basis of the one-dimensional Laplacian (lower left) and the unmodified wavenumber k (lower right).

Figure 2.3: *Eigenvalues of C for $k = 6.3\pi$ and $n = 32$ using equation 2.56 left and using equation 2.58 right.*



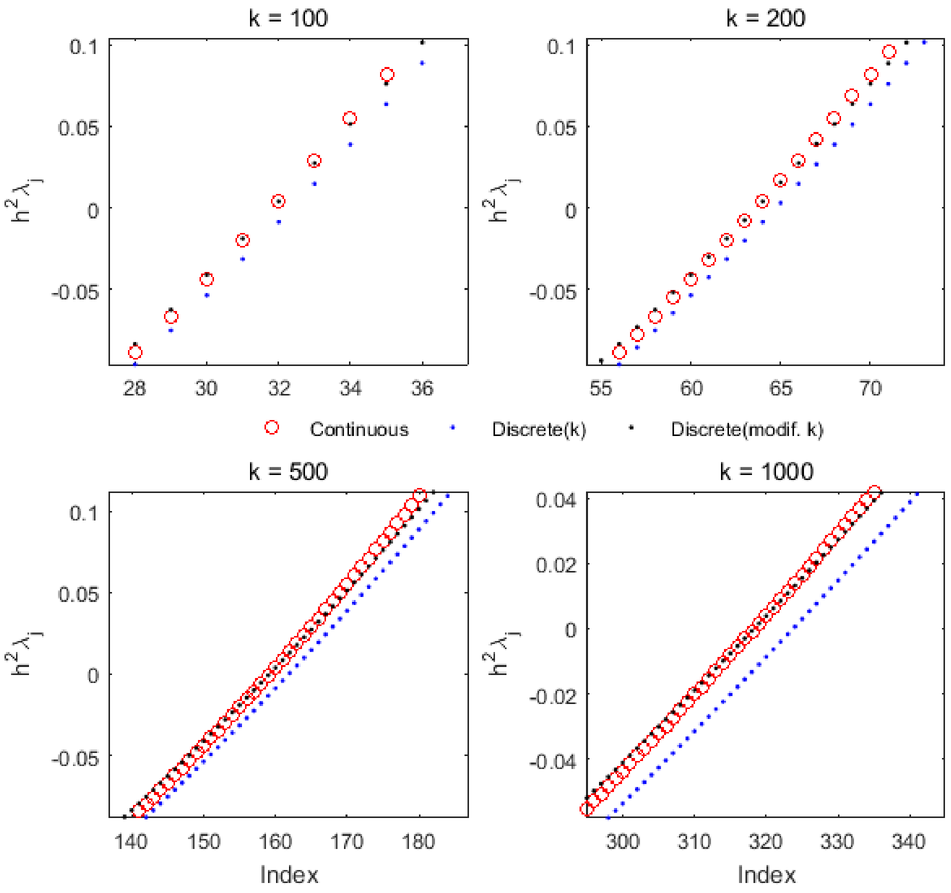


Imposing a modified wave number on the coarse-grid, keeps the modulus of the non-unit eigenvalues well below 1. Moreover, the amplification of the low frequency mode has been lifted. As a result, there seems to be a close correlation between the wave speed on the coarse-grid level and the accuracy of the non-unit eigenvalues of the coarse-grid projection operator C.

2.6.3 Pollution and Eigenvalues

When it comes to the influence of the pollution error on the spectrum of the discretized operator A, not much literature is available. However, the previous section illustrated that implementing a modified wave number $\tilde{k}$ or $\tilde{k_H}$ on the coarse grid Ω_H does influence the eigenvalues of the coarse-grid projection operator C.
In this section, we proceed with a similar analysis from section 2.6.1 and investigate the spectrum of the coefficient matrix $\tilde{A}$. Figure 2.4 plots the continuous eigenvalues of the one-dimensional Helmholtz operator against their discrete counterparts using k and $\tilde{k}$ respectively. We see that for the eigenvalues near zero, the discrete eigenvalues obtained by using $\tilde{k}$ are better approximations of the continuous eigenvalues. The more we move to the left of the origin, the better the continuous eigenvalues are approximated by k (blue dots) instead of $\tilde{k}$ (black dots). For both k and $\tilde{k}$, moving further to the right of the origin, the discrete eigenvalues become worse approximations of their continuous counterparts. Taking these results into consideration, there seems to exist some support for the preliminary notion that the accuracy of the near-zero eigenvalues and the accuracy of the overall numerical solution are potentially related.

Figure 2.4: *Eigenvalues of* h^2A *and* $h^2\tilde{A}$ *for* $k = 100, 200, 500$ *and* 1000 *using* $\kappa = 0.625$.



2.7 Concluding Remarks and Summary

This chapter deals with the general numerical solution methods available for solving the Helmholtz problem. We summarize our main findings in the following points

- Krylov subspace methods that based on minimizing the residual are suitable iterative methods for the Helmhotlz problem.
- The main method used in this interim thesis to solve the ADEF+CSLP-preconditioned system is the GMRES-method.
- Standard multigrid methods combined with BIMs for smoothing unwanted error components are not suitable for the Helmholtz problem at high wave numbers.
- A coarse-correction scheme is effective in reducing the unwanted error components by projecting these components onto zero. This projection operator serves as a basis for the ADEF-deflation projection operator.
- At high wave numbers, the numerical solution of the Helmholtz equation suffers from pollution; a phase error between the numerical and exact solution caused by deviating wave numbers.

Chapter 3

Preconditioning

In general, the performance of numerical iterative methods are commonly assisted by the use of precondtioners. The latter class of matrices are implemented to cluster the spectrum of A into a more favorable counterpart in order to speed up convergence. Instead of solving $Ax = b$, one resorts to solving $M^{-1}Au = M^{-1}f$, where the matrix M serves as the preconditioner matrix.

3.1 Preconditioning for the Helmholtz Problem

The use of preconditioners for the Helmholtz problem has been studied widely throughout the years. The suitable Krylov subspace methods generally do not perform well without incorporating a preconditioner. Several preconditioners have been tailored for the Helmholtz problem.
An important class is mentioned in (Sheikh et al., 2009) and (Erlangga, 2005), where an incomplete LU factorization of the coefficient matrix A_h serves as a preconditioner. However, ILU preconditioners are notoriously known to cause fill-in, destroying the original sparsity of the coefficient matrix and can especially become problematic for large wavenumbers.
An alternative has been opted by Gander and Nataf (2000), M. Gander (2005) and Gander and Nataf (2001), where an analytical ILU factorization has been proposed. A drawback of the AILU preconditioner is its applicability to constant-wave number problems as it diverges for non-constant wave number problems.
Finally, a class of preconditioners has been constructed which focuses on the operator in question and shows promising performance gains for medium sized wave numbers. In (Bayliss et al., 1983) the preconditioner matrix M_h is equal to the discretizized Laplacian operator $-\Delta_h$. Laird and Giles (2002) have further developed this class by including a positive real shift.

For large wavenumbers it seems that the most effective and robust results can be achieved by combining a real and complex shift in the Laplacian operator based preconditioner. Erlangga et al. (2006a) and Erlangga et al. (2006b) have first examined the behavior of the CSLP-preconditioned system for the Helmholtz equation. Despite achieving a substantial speed-up, the small eigenvalues of the preconditioned system rush to zero for the Helmholtz problem as the wave number increases. The upcoming sections will be dedicated to the cause and behavior of these eigenvalues.

3.2 CSLP Preconditioner

Let A_h be the resulting coefficient matrix after discretization. Recall that we can write A_h in terms of the discrete Laplacian operator $-\Delta_h$ and the $n \times n$ identity matrix I: as:

$$A_h := -\Delta_h - k^2 I, A_h, \in \mathbb{C}^{n\times n}. \tag{3.1}$$

The CSLP preconditioner is accordingly defined as

$$M_h := -\Delta_h - (\beta_1 + i\beta_2)k^2 I_h,\ A_h, \in \mathbb{C}^{n\times n},\ \beta_1, \beta_2 \in [0, 1], \tag{3.2}$$

where i denotes the imaginary unit and β_1 and β_2 represent the real and complex shift respectively. Initially, the coefficient matrix A_h is an indefinite real symmetric matrix in the absence of Sommerfeld radiation conditions. Incorporating a complex shift transforms the preconditioned coefficient matrix $M_h^{-1}A_h$ into a normal, complex, symmetric yet non-Hermitian matrix.
We will proceed by dropping the subscript h in the notation for A_h, M_h and proceed with A, M respectively. Moreover, we introduce a notation for the preconditioned linear system $\widehat{A}u = M^{-1}Au = \widehat{f} = M^{-1}f$.

The preconditioned system has a convenient way of relating the eigenvalues of the matrix A to the eigenvalues of the transformed system $\widehat{A}$ given that A and M^{-1} commute, which is shown below

$$\begin{aligned}\widehat{A} &= M^{-1}A,\\ &= M^{-1}(M + (\beta_1 + i\beta_2 - 1)k^2 I),\\ &= I + (\beta_1 + i\beta_2 - 1)k^2 M^{-1},\\ &= (M + (\beta_1 + i\beta_2 - 1)k^2 I)M^{-1} = AM^{-1}.\end{aligned} \tag{3.3}$$

As a result, the preconditioned system shares the same orthonormal eigenvectors as the original coefficient matrix A and we obtain an elegant expression for the eigenvalues of the preconditioned system

$$\begin{aligned}\lambda_j(\widehat{A}) &= \lambda_j(M^{-1}A),\\ &= \lambda_j(M^{-1})\lambda_j(A),\\ &= \frac{\lambda_j(A)}{\lambda_j(M)}.\end{aligned} \tag{3.4}$$

Using equation 3.4, it is easy to see that the eigenvalues of the continuous operator defining the preconditioner are

$$\lambda_j(\widehat{A}) = \frac{j^2\pi^2 - k^2}{j^2\pi^2 - (\beta_1 + i\beta_2)k^2}, \ \beta_1, \beta_2 \in [0, 1]. \tag{3.5}$$

The eigenvalues for the discretized Helmholtz operator are given in section 1.4.3, equation 1.25. Using equation 1.25, we obtain the following expression for the discrete eigenvalues of the preconditioned system, where ω denotes the eigenvalues of the discretized Laplacian operator

$$\begin{aligned}\omega &= \frac{1}{h^2}(2 - 2\cos(l\pi h)), \ l = 1, 2, \dots\\ \Rightarrow \lambda_l(M^{-1}A) &= \frac{\omega - k^2}{\omega - (\beta_1 + i\beta_2)k^2}, \ \beta_1, \beta_2 \in [0, 1].\end{aligned} \tag{3.6}$$

3.2.1 Optimal Shift

Various options for the shift parameters β_1 and β_2 have been considered, while respecting the condition that $\beta_1, \beta_2 \in [0, 1]$. In (Erlangga et al.) and (van Gijzen et al., 2007b) the spectral properties of the CSLP preconditioned system have been inspected. When the real shift parameter β_1 is set to 1 the condition number of the preconditioned coefficient matrix $\widehat{A}$ is minimized (Erlangga et al.). Letting $\beta_1 = 1$ leads to a tight circular distribution of the eigenvalues, remedying the high indefiniteness of the original coefficient matrix A and eventually positively affecting rate of convergence iterative Krylov subspace methods [1].

Thus far, the literature seems to suggest that the most optimal configuration would be to set β_1 to 1 and β_2 as small as possible. van Gijzen et al. (2007a) have studied the optimal complex shift parameter β_2, affirming that the complex shift parameter can be interpreted as the radius of the circular eigenvalue distribution when β_1 is fixed at 1. However, a word of caution is in place as decreasing the magnitude of β_2 leads to the matrix M resembling the original coefficient matrix A, making the inversion

[1] Choosing β_1 any larger than 1 would lead to a more indefinite preconditioner matrix M than the original matrix A.

and implementation of the preconditioner computationally expensive. van Gijzen et al. (2007a) postulate that the optimal shift (β_1, β_2) is obtained by letting $\beta_1 = 1$ and $\beta_2 = 0.5$, causing the real part of the eigenvalues to be bounded below by 0 and above by 1, while allowing the complex part to vary between $-0.5i$ and $0.5i$. In the figures below, we plot the eigenvalues for various k using the optimal shift in accordance with (van Gijzen et al., 2007a). Unless stated otherwise, we therefore use shifts $(\beta_1, \beta_2) = (1, 0.5)$ and $\kappa = 0.625$.

Figure 3.1: *Eigenvalues of the preconditioned system* $\widehat{A}$

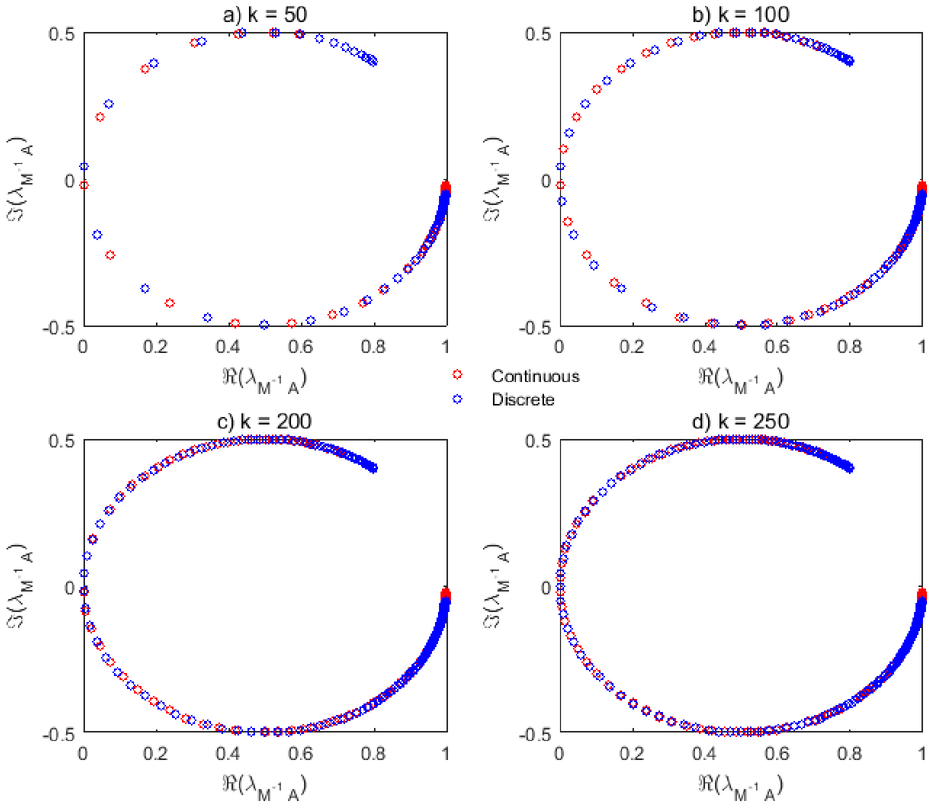


Figure 3.1 is illustrative of the problem at hand. As k increases, we indeed see small eigenvalues gathering around the origin for both the continuous and discrete systems. For $k = 250$, the clustering already seems dominantly prominent. When k grows very large, this effect accumulates. Section 1.5, Figure 1.2 was affirmative in demonstrating the discrepancy between the continuous and discrete eigenvalues. A similar effect echoes through when examining the eigenvalues of the preconditioned system $\widehat{A}$, where increasing the wave number k leads to more deviations between the analytical and discrete eigenvalues.

We repeat the procedure in Figure 3.2 where we plot the same eigenvalues using a finer grid resolution.

Figure 3.2: *Eigenvalues of the preconditioned system $\widehat{A}$ using κ = 0.0625*

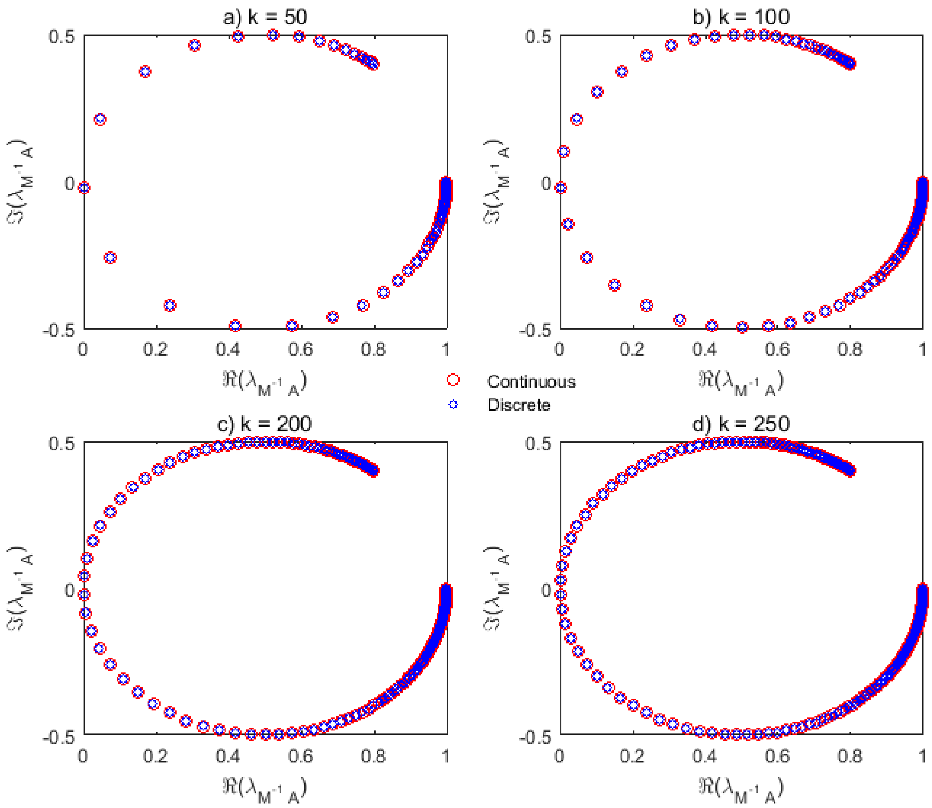


Figure 3.2 reveals that utilizing a more refined grid decreases the error between the analytical and discrete eigenvalues. Although the discrete eigenvalues approach the analytical eigenvalues as the number of grid points per wave length is increased, the occurrence of the clustering eigenvalues near the origin remains undeterred. These results are in coherence with Erlangga et al.. Erlangga et al. reiterate that the differences in convergence behavior are primarily determined by the magnitude of the smallest eigenvalue. The authors come to the conclusion that the smallest eigenvalues are independent of the step size h.

3.2.2 Near-null Eigenvalues

A vital part of this study focuses on the behavior of the near-nullspace eigenvalues of $\widehat{A}$. The conclusions from this section are of paramount importance for the ADEF-preconditioner, as the small eigenvalues reappear for large k, surpassing the effect of the intended deflation, see Chapter 4. This section focuses on revealing where the problem with respect to the small clustering eigenvalues originates. Figure 3.1 from the previous section shed light on two important features. Firstly, we indeed observed the small eigenvalues crowding up around the origin as k increases. Secondly, on a coarse resolution grid, the discrete eigenvalues appear to be deviating from their analytical counterparts. In section 1.5.1, we discovered that, though the analytical eigenvalues approach zero when the index j of λ_j is approximately equal to $\frac{k}{\pi}$, the corresponding statement does not apply unequivocally to the discrete eigenvalues. In fact, as k increases, the smallest discrete eigenvalues λ_l emerge at some index $l > j = \lfloor \frac{k}{\pi} \rfloor$.

We will proceed with a similar analysis for the CSLP preconditioned matrix $\widehat{A}$. Sheikh et al. (2016) have confirmed that both the real and imaginary part of the discrete eigenvalues of $\widehat{A}$ are small in magnitude whenever the same holds for the eigenvalues of the original coefficient matrix A. More precisely, the preconditioner maps the near-nullspace eigenvalues of A to near-nullspace eigenvalues of $\widehat{A}$. As a result, the inscalability of the CSLP-preconditioned solver is due to the eigenvalues of A approaching the nullspace as k increases, causing the the eigenvalues of $\widehat{A}$ to become even smaller. This effect translates into an increasing number of iterations needed by the Krylov solver to converge.

Apart from some eigenvalues clustering around the origin, the number of small eigenvalues is reported to increase along with k. In order to predetermine the scope of potential scalability of any solver based on the CSLP-preconditioner, it would practical to get an estimate to what extent the number of small eigenvalues are increasing related to k. Moreover, it would be insightful if we could find an one-to-one translation between the magnitude of the eigenvalues A and to what extend the magnitude of the eigenvalues of $\widehat{A}$ drop. We will adapt the approach followed in (Sheikh et al., 2016), by finding an explicit upperbound for the smallest eigenvalue of A. We take β_1 to be 1 and and start by spliting the expressions for the analytical eigenvalues into their respective real and imaginary part

$$\begin{aligned}
\lambda_j &= \frac{j^2\pi^2 - k^2}{j^2\pi^2 - k^2 - i\beta_2 k^2},\ j = 1, 2, 3, \ldots \\
&= \frac{\gamma}{\gamma - i\beta_2 k^2},\ \gamma = j^2\pi^2 - k^2 \\
&= \frac{\gamma}{\gamma - i\beta_2 k^2}\frac{\gamma + i\beta_2 k^2}{\gamma + i\beta_2 k^2} \\
&= \frac{\gamma}{\gamma^2 + \beta_2^2 k^4} + \mathbf{i}\frac{\beta_2 k^2}{\gamma^2 + i\beta_2^2 k^4} \\
&= \Re(\lambda_j) + \Im(\lambda_j),\ j = 1, 2, 3, \ldots
\end{aligned}$$

The factor k^4 already gives a glimpse as to why the eigenvalues of $\widehat{A}$ approach zero as k increases. Note that if $\beta_2 = 0$, the factor k^4 drops out. On another note, if $j = \frac{k}{\pi}$, then γ is zero, leading to $\Re(\lambda_j)$ being zero. However, due to rounding error, we generally have $|\gamma| > 0$ due to the fact that j needs to be integer.

3.2.3 Convergence Behavior

In the previous section we noted that roughly 1 percent of the total eigenvalues is shifting towards the origin. It is widely acknowledged that a clustered set of eigenvalues around $(1, 0)$ is, without loss of generality, favorable for the convergence of Krylov subspace methods. Various studies have treated the departing shift of a set of eigenvalues and its effect on the performance of Krylov subspace methods, as they start to cluster around the origin.

For the CSLP-preconditioned Helmholtz problem in particular, the effect of such eigenvalues on the performance of the GMRES-method has been investigated as well. Erlangga et al., Erlangga (2005) and (van Gijzen et al., 2007a) mention that for the current model problem, the convergence behavior is independent of the stepsize. Another feature mentioned is the fact that the number of iterations seems to grow linearly with k and is bounded above in case damping is introduced (van Gijzen et al., 2007a). A similar observation has been made in Sheikh (2014) and Sheikh et al. (2016).

In this section, we will briefly look at the convergence behavior of the GMRES-method in solving the system $\widehat{A}u = \widehat{f}$, for various k. The results are presented in Table 3.1.

Table 3.1: *Number of iterations and relative residual. Est. gives the estimated number of small eigenvalues between* $[magn(\frac{1}{k}), 0.01]$*, while true provides the actual account of the eigenvalues. Rel. Res. represents the relative residual.*

k	Iterations	Est.	True	Rel. Res.
100	28	2	1	3.230682e-08
500	89	8	8	6.074832e-08
1000	159	16	16	4.987300e-08
2000	294	32	33	6.477246e-08
3000	428	50	49	5.734978e-08
4000	561	66	68	6.739762e-08
5000	688	82	85	7.305645e-08
5500	752	90	93	8.011798e-08
5600	769	94	94	6.897173e-08
5700	1493	96	96	3.428289e-03

We can confirm that the number of iterations grows with k. The clustering of the eigenvalues around zero appears to be responsible for these results. For wave numbers $k \geq 1000$ the number of small eigenvalues seems to be roughly between 10 and 12 percent of the number of iterations. A convenient ballpoint for the number of iterations can consequently be deduced.
An interesting observation is that the method diverges for $k = 5700$, as the dependency on k becomes non-linear. Unfortunately, employing a finer grid resolution would not render any benefits due to the convergence being independent of the stepsize.

3.3 Concluding Remarks and Summary

In this chapter we looked into the literature and the results related to the CSLP-preconditioned Helmholtz problem, which we can summarize into the following main points:

- The CSLP-preconditioned Helmholtz problem shifts the eigenvalues of the original coefficient matrix onto the complex plane
- The eigenvalues are circularly distributed, having their real part bounded between 0 and 1
- As the wave number grows, the eigenvalues start to cluster around the origin, impeding the rate of convergence
- The convergence and shifting of eigenvalues towards the origin is independent of the step size, yet the discrete eigenvalues approach the analytical ones as the step size is reduced
- Both the discrete and analytical near-null eigenvalues arise synergistically at the index where the eigenvalues of the original coefficient matrix A intersects the origin
- Approximately 1 percent of the total number of eigenvalues starts shifting to zero as k increases

Chapter 4

Deflation

4.1 Deflation

The previous chapter revealed that Krylov subspace methods are adversely affected by close to zero eigenvalues. While the application of the CSLP preconditioner was successful in confining the eigenvalues between 0 and 1, the Krylov solver remains defenseless against the hampering convergence behavior caused by these small eigenvalues for large k. Deflation is a technique, intrinsically designed, to "deflate" these unwanted eigenvalue onto zero. By means of projection, it is possible to alleviate the adverse effects on the Krylov solver by either explicitly modifying the operator of the linear system (Nicolaides (1987b)) or by augmenting the eigenvectors corresponding to the troublesome eigenvalues (Morgan (1995), Morgan (2002)). For large systems, the latter option defeats its purpose as the computation of eigenvectors is computationally expensive. As a consequence, most applications in the literature are based on approximations of invariant subspaces obtained from Jordan decompositions. Deflation for large scale problems relies on multiplying the linear system by a projection matrix P and applying the Krylov subspace method to the projected system PA, rendering the projection matrix P to act as a preconditioner at the same time [1]

$$\begin{aligned} &PA\widehat{u} = Pf \\ &A \in \mathbb{C}^{n \times n},\ P \in \mathbb{C}^{n \times n},\ \widehat{u} \in \mathbb{C}^{n} \\ &m = \dim(P) < n \end{aligned}$$

Consider $A \in \mathbb{C}^{n \times n}$. Then its Jordan decomposition is given by

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} J_1 & \emptyset \\ \emptyset & J_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1}$$

where $J_1 \in \mathbb{C}^{m \times m}$ and $J_2 \in \mathbb{C}^{(n-m)\times(n-m)}$ with $m \leq n$ represent the square Jordan blocks. Letting $P_{\{U_1,U_2\}}$ denote the projection onto $U_1 \subseteq \mathbb{R}^{m \times m}$ along $U_2 \subseteq \mathbb{R}^{(n-m)\times(n-m)}$, the projected system can be decomposed as

$$P_{\{U_1,U_2\}}A = A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \emptyset & \emptyset \\ \emptyset & J_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1}$$

The resulting system PA will have a zero eigenvalue with algebraic multiplicity m. The spectrum contained in the Jordan block J_1 appears invisible to the Krylov solver, ameliorating the conditions for convergence. Analytically, the invariant subspaces are based on (generalized) eigenvectors, creating the necessity for approximations to these subspaces in order to meet practical purposes. As a result, the remaining part of the spectrum will typically differ from $\sigma(J_2)$. However, Gaul (2014) argues that small perturbations to an invariant subspace lead to small perturbations in the remaining spectrum, as long as the subspaces are well-conditioned.

[1] Please note that the subsequent interpretation is for theoretical purposes only. In practice, the preconditioner is not implemented directly. Also note that the projected system PA can be singular in case P is not equal to the identity matrix. Additionally the application of a deflation preconditioner is accompanied by a correction scheme as the solution $\widehat{u}$ is not necessarily a solution to the original system, see (Gaul, 2014).

4.1.1 Deflation Based Preconditioning for GMRES

Consider a general complex valued linear system. The projection matrix $\widehat{P}$ and its complementary projection P can be defined as

$$\begin{aligned}&\widehat{P} = AQ \text{ where } Q = ZE^{-1}Z^T \text{ and } E = Z^T AZ\\&A \in \mathbb{C}^{n\times n},\ Z \in \mathbb{R}^{m\times n}\\&P = I - AQ\end{aligned}$$

where Z functions as the deflation matrix whose $m < n$ columns are considered the deflation vectors and I is the $n \times n$ identity matrix. Matrix P is also known as the projection preconditioner. Note that the explicit requirement of Z being an A-invariant subspace or an approximation of an A-invariant subspace has been discarded.

In the literature, this type of projection preconditioning is rarely used in combination with the GMRES method, mainly because it requires the original coefficient matrix A to be self-adjoint and positive-semi-definite ((Nicolaides, 1987a), (Dostál, 1988)). Gaul (2014) poses that in case the original coefficient matrix A is self-adjoint and positive-semi-defininte, there exists a well-defined projection $\widehat{P}$. Violations of these assumptions could potentially lead to ill-defined projections. At first consideration, such a choice would seem counterintuitive for the Helmholtz problem. However, Gaul (2014) provides two conditions in order for the projection matrix P and the deflated GMRES-method to be well-defined.

Theorem 4.1.1 (Deflated GMRES) *Let $Au = f$ be a linear system with A non-singular, $A \in \mathbb{C}^{n\times n}, b \in \mathbb{C}^n$ and Z a subspace of $\mathbb{C}^{n\times n}$ with dimension $\dim(Z) = m < n$. Furthermore, let $\theta_{(Z,AZ)} < \frac{\pi}{2}$, where $\theta_{(Z,AZ)}$ denotes the principal angle between the subspaces Z and AZ. Then the projection matrix $P := Pu = (I - AQ)u = u - AZ\langle Z, AZ\rangle^{-1}\langle Z, u\rangle$, $u \in \mathbb{C}^n$ is well defined. Moreover, for all initial guesses u_0, the GMRES method applied to the deflated system is well-defined.*

Proof A proof has been given by Gaul (2014) in section 3.3, theorem 3.9

As a consequence, as long as the original coefficient matrix A is non-singular and the principal angle between the subspaces is smaller than $\frac{\pi}{2}$, the deflated GMRES can be applied to the projection matrix $\widehat{P}$. This result is of paramount importance, as the ADEF-preconditioner is defined on the coarse space spanned by the grid-interpolation vectors and is implemented using the GMRES method.

One of the preliminary applications of the abovementioned projection preconditioner P in conjunction with the GMRES-method has been studied by Erlangga and Nabben (2008) and Yeung et al. (2010) whom used an exact A-invariant subspace Z. Their main findings stipulate that deflated-GMRES converges faster than non-deflated-GMRES in the absence of breakdowns caused by ill-conditioned deflation subspaces. On the contrary, utilizing general deflation vectors does not seem to lead to satisfactory results.

4.1.2 Deflation Based Preconditioning for the Helmholtz problem

For linear systems arising from the CSLP-preconditioned Helmholtz equation in particular, it is possible to take geometrically constructed multigrid vectors as deflation vectors ((Sheikh, 2014), (A. Sheikh, 2011)). The rationale behind this approach exploits the fact that multigrid inter-grid operators project small frequencies onto coarser levels. The first works to explore the application of domain decomposition methods to deflation techniques are (Nabben and Vuik, 2008a) and (Nabben and Vuik, 2004). Specifically for Krylov subspace methods, Nabben and Vuik pointed out similarities between the deflated CG method and domain decomposition methods. An extension on this perspective was the implementation of a deflated Krylov subspace method through a standard multigrid method in (Nabben and Vuik, 2008b). Such a perspective provides a practical way of implementing deflation as a multigrid method.
Generally, standard multigrid methods are not suitable for the Helmholtz problem, unless the wave number is small enough relative to the step size. Within the hierarchical context of multigrid methods, this translates into the requirement that the waves according to wave number k must be resolved on the coarsest level (Ernst and Gander, 2012). When a multilevel deflation operator P is applied to a preconditioned system $M^{-1}A$, the deflation operator P acts as a second level preconditioner, allowing

for the application of multigrid methods to the indefinite Helmholtz equation. Despite enabling the application of multigrid methods to the indefinite Helmholtz equation, there still remain some difficulties. First of all, their exists a duality between optimizing the complex shift β_2 in order to limit the clustering of eigenvalues around zero and keeping β_2 large enough in order for the multigrid method to perform robustly (Ernst and Gander, 2012). Furthermore, using multilevel Krylov methods requires the inversion of the matrices M and E on several coarse levels and is done approximately. Sheikh (2014), Sheikh et al. (2016) and Erlangga and Nabben (2008) mention that inexact inversion of these matrices disperses the deflated eigenvalues in the vicinity of the origin. These additionally occurring eigenvalues, which appear near the origin, should not be confused with the near-zero eigenvalues which are intrinsic to the linear system itself. As a result, deflating to the largest eigenvalue instead of zero would be able to mitigate the effect of additional near zero eigenvalues appearing due to the approximate inversion. The latter implementation is only suitable in combination with a flexible-Krylov method, such as the flexible-GMRES, GCR or flexible-IDR method. Another alternative would be to use IDR-methods. In this interim thesis the main focus lies with the near-zero eigenvalues that are not related to the approximate inversion. Consequently, unless stated otherwise, inversion is done using a direct method and all linear systems are solved using the GMRES-method.

Variants for the Helmholtz problem

In this section we describe the available choices of deflation vectors tailored for the Helmholtz equation. Most of this section contains a summary of §4.2 of Sheikh (2014). We define the following general deflated system

$$\mathcal{P}_h = I - \mathcal{A}_h Q_h \text{ where } Q_h = Z_h E_{2h}^{-1} Z_h^T \text{ and } E_{2h} = Z_h^T \mathcal{A}_h Z_h \tag{4.1}$$

where $\mathcal{A}_h$ and E_{2h} denote the representative notation for the system matrix to which deflation will be applied and the coarse grid approximation respectively. Various schemes distinguish themselves by allowing different choices for $\mathcal{A}_h$ and E_{2h}, which we will list below

1. *Deflation using the CSLP-preconditioned operator (ideal).* As the near nullspace eigenvalues arise after preconditioning by CSLP, deflation is aimed at the preconditioned system. Already in constructing E_{2h}, it requires the inversion of the matrix $M_{h,(\beta 1,\beta 2)}$ making it computationally expensive.

$$\mathcal{A}_h = M_{h,(\beta 1,\beta 2)}^{-1} A_h \text{ and } E_{2h} = Z_h^T \mathcal{A} Z_h$$

2. *Deflation using the CSLP-preconditioned operator (practical).* This is in fact the first variant, where the exact inversion of $M_{h,(\beta 1,\beta 2)}$ has been replaced by an approximation $M_{2h,(\beta_1,\beta_2)}$. $\Theta = Z_h^T Z_h$ is an approximation term as well.

$$\begin{aligned}
E_{2h} &= X_h^T \mathcal{A}_h X_h,\ X_h \in \mathbb{R}^{n\times n} \\
&= X_h^T (M_{h,(\beta 1,\beta 2)}^{-1} A_h) X_h \\
&= X_h^T X_h (X_h^{-1} M_{h,(\beta 1,\beta 2)}^{-1} (X_h^T)^{-1} X_h^T A_h) X_h \\
&= X_h^T X_h (X_h^{-1} M_{h,(\beta 1,\beta 2)}^{-1} (X_h^T)^{-1})(X_h^T A_h X_h) \\
&= \Theta M_{2h,(\beta_1,\beta_2)}^{-1} A_{2h} \\
E_{2h} &= Z_h^T \mathcal{A}_h Z_h \approx \Theta M_{2h,(\beta_1,\beta_2)}^{-1} A_{2h}
\end{aligned}$$

3. *Deflation using the Helmholtz operator.* Here the deflation preconditioner is directly applied to the Helmholtz operator. A first level preconditioning by the CSLP-preconditioner may precede the deflation preconditioner.

$$\mathcal{A}_h = A_h \text{ and } E_{2h} = Z_h^T \mathcal{A}_h Z_h.$$

4. *Deflation using the CSLP-preconditioned Helmholtz operator.* Here deflation is applied solely to the CSLP operator.

$$\mathcal{A}_h := \Delta_h - (\hat{\beta}_1 - i\hat{\beta}_2)k^2 I \text{ and } E_{2h} = Z_h^T \mathcal{A}_h Z_h$$

5. *Rediscretization* Here the coefficient matrix is rediscretisized on the coarse grid with step size $2h$.

$$E_{2h} = \text{ re-discretization}(A_h).$$

4.2 ADEF-Preconditioner

The two-level ADEF-preconditioner can be categorized into the third variant. Officially, the ADEF-preconditioner is defined by including a shift term γ to counteract the effect of the approximate inversion of E_{2h}^{-1}, i.e.

$$\mathcal{P}_h = I - \mathcal{A}_h Q_h + \gamma Q_h. \tag{4.2}$$

For the CSLP-preconditioned system $\widehat{A}$, the largest eigenvalue is 1. Deflating onto the largest eigenvalue of $\widehat{A}$ instead of zero (i.e. $\max_{j=1,2,..n-1} \lambda_j(\widehat{A}) = \mu = 1$) leads to the two-level ADEF1-preconditioner. We will follow the same approach as Sheikh (2014) and disregard the shift by taking $\mu = 0$, i.e. we are deflating all unwanted eigenvalues onto the origin instead of one. Note that in this case, the operator $\mathcal{P}_h$ is officially known as the *deflation preconditioner*. However, for the sake of convenience and in order to keep the notation in line with Sheikh (2014), we will refer to P_h as the *ADEF-preconditioner*, while keeping the shift term zero.
Based on theory above, the ADEF-preconditioner is defined by taking the coarse correction operator I_h^{2h} from a multigrid setting as the deflation subspace Z in equation 4.1. I_h^{2h} can be interpreted as interpolating from grid Ω_{2h} to grid Ω_h. As a result, the ADEF-preconditioner is commonly referred to as a two-level method and we obtain

$$\widehat{P_h} = A_h Q_h \text{ where } Q_h = Z_h E_h^{-1} Z_h^T \text{ and } E_h = Z_h^T A_h Z_h \tag{4.3}$$

$$P_h = I_h - A_h Q_h \text{ where } Q_h = I_h^{2h} A_{2h}^{-1} I_{2h}^h \text{ and } A_{2h} = I_h^{2h} A_h I_{2h}^h \tag{4.4}$$

For spectral improvement, the ADEF-preconditioner is applied to the CSLP preconditioned system, which leads to solving the following linear systems (Sheikh, 2014)

$$\begin{aligned} M_{(\beta_1,\beta_2)}^{-1} A_h u_h &= M_{(\beta_1,\beta_2)}^{-1} b_h \\ M_{(\beta_1,\beta_2)}^{-1} P_h A_h u_h &= M_{(\beta_1,\beta_2)}^{-1} P_h b_h \\ P_h^T M_{(\beta_1,\beta_2)}^{-1} A_h u_h &= P_h^T M_{(\beta_1,\beta_2)}^{-1} b_h \end{aligned} \tag{4.5}$$

As a result of equation 4.5, the spectrum of both systems are equivalent

$$\sigma(M_{(\beta_1,\beta_2)}^{-1} P_h A_h) = \sigma(P_h^T M_{(\beta_1,\beta_2)}^{-1} A_h)$$

Due to the equivalent spectra, the order of implementation should not lead to mutually differentiating results.

4.2.1 Spectral Analysis

Sheikh (2014) and Yeung et al. (2010) have provided analytical expressions for the eigenvalues of the ADEF-preconditioner using rigorous Fourier analysis. The following section contains an excerpt of Sheikh's paragraph §4.2 and §5.2 from (Sheikh, 2014) covering the one-dimensional two-grid spectral analysis of the ADEF-preconditioner. We start by using a second order accurate stencil, writing the coefficient matrix A_h as follows

$$\left[A_h\right] = \frac{1}{h^2}\left[-1 \quad 2-\kappa^2 \quad -1\right] \text{ where } \kappa = kh,$$

Recall that the basic eigenfunctions are sines, and thus we can define the following eigenvectors

$$\phi_h^l = \sin(l\pi \underline{x}) \text{ for } 1 \leq l \leq n-1 \tag{4.6}$$

which correspond to the eigenvalues

$$\lambda^l(A_h) = \frac{1}{h^2}(2 - 2c_l - \kappa^2),\ c_l = \cos(l\pi h)$$

The eigenvectors form an orthonormal set and provide a basis for block diagonalization. Defining the following blocks

$$B^l_{h,2h,(\beta_1,\beta_2)} = (P^T_{h,2h})^l (S_h)^l$$

$$S_h^l = (M_h{}^{-1}_{(\beta_1,\beta_2)} A_h)^l$$

we can represent the spectrum in terms of these block matrices. Each index l defines a block up to $n/2$ and we obtain a representation of the spectrum in terms of these blocks

$$\sigma(B_{2,2h,(\beta_1,\beta_2)}) = \sigma\Big(\big[B^l_{2,2h,(\beta_1,\beta_2)}\big]_{1\leq n/2}\Big)$$

We now reorder the eigenvectors into a basis V_h, which will provide the basis for bringing the coefficient matrix and the preconditioner matrices into block diagonal form

$$V_h = \big[\theta_h^1, \theta_h^{n-1}, \theta_h^2, \theta_h^{n-2}, ..., \theta_h^{n/2-1}, \theta_h^{n/2}\big].$$

$$A_h = \Big[A_h^l\Big]_{1\leq n/2},$$

$$A_h^l = \begin{pmatrix} \frac{1}{h^2}(2-2c_l-\kappa^2) & 0 \\ 0 & \frac{1}{h^2}(2-2c_l-\kappa^2) \end{pmatrix}$$

$$A_h^{n/2} = \frac{2}{h^2} - k^2$$

We have shown in Chapter 3 section 3.2 that the eigenvectors from equation 4.6 coincide with the eigenvectors of the CSLP-preconditioned coefficient matrix given that $M_{h,(\beta_1,\beta_2)}$ and A_h commute. Thus, we can use V_h as a basis for diagonalization of $M_{h,(\beta_1,\beta_2)}$ and $S_{h,(\beta_1,\beta_2)}$. Given the eigenvalues of the respective matrices

$$\lambda^l(M_{h,(\beta_1,\beta_2)}) = \frac{1}{h^2}\Big(2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)\Big),$$

$$\lambda^l(S_{h,(\beta_1,\beta_2)}) = \frac{2 - 2c_l - \kappa^2}{2 - 2c_l - \kappa^2(\beta_1 - i\beta_2)}$$

we can block diagonalize them using the following blocks

$$M^l_{h,(\beta_1,\beta_2)} = \frac{1}{h^2}\begin{pmatrix} \frac{1}{h^2}(2-2c_l-\kappa^2(\beta_1-i\beta_2)) & 0 \\ 0 & \frac{1}{h^2}(2-2c_l-\kappa^2(\beta_1-i\beta_2)) \end{pmatrix},$$

$$S^l_{h,(\beta_1,\beta_2)} = \frac{1}{h^2}\begin{pmatrix} \frac{2-2c_l-\kappa^2}{2-2c_l-\kappa^2(\beta_1-i\beta_2)} & 0 \\ 0 & \frac{2+2c_l-\kappa^2}{2+2c_l-\kappa^2(\beta_1-i\beta_2)} \end{pmatrix}.$$

$$M_h^{n/2} = \frac{2}{h^2} - (\beta_1 - i\beta_2)k^2$$

$$S_h^{n/2} = \frac{2 - k^2}{2 - (\beta_1 - i\beta_2)k^2}$$

The basis V_h further diagonalizes the linear interpolation operator I_h^{2h} into blocks

$$(I_h^{2h})^l = \Big[\frac{1}{2}(1 + c_l) \quad -\frac{1}{2}(1 - c_l)\Big],$$

As a result, the block diagonal form of the diagonal Galerkin coarse grid operator A_{2h} and M_{2h} are equal to

$$A_{2h}^l = (I_h^{2h})^l A_h^l (I_{2h}^h)^l = \frac{2(1 - c_l^2) - \kappa^2(1 + c_l^2)}{2h^2}$$

$$M_{2h}^l = (I_h^{2h})^l M_h^l (I_{2h}^h)^l = \frac{2(1 - c_l^2) - (\beta_1 - i\beta_2)\kappa^2(1 + c_l^2)}{2h^2}$$

Subsequently the 1×1 block for $M_{2h}^{-1}A_{2h}$ will be

$$(M_{2h}^{-1}A_{2h})^l = \frac{2(1 - c_l^2) - \kappa^2(1 + c_l^2)}{2(1 - c_l^2) - (\beta_1 - i\beta_2)\kappa^2(1 + c_l^2)}$$

for $1 \leq l \leq n/2 - 1$ and $(M_{2h}^{-1}A_{2h})^{n/2} = \frac{2-\kappa^2}{2-\kappa^2(\beta_1 - i\beta_2)}$. Also the 1×1 block for the approximation term $B_h = I_h^{2h} I_{2h}^h$ can be simplified as

$$B_h^l = (I_h^{2h} I_{2h}^h)^l = (1 + c_l^2),\ 1 \leq l \leq n/2$$

Finally, we consider the deflation preconditioner $P_{h,2h}^T$ which can be written in a block diagonal form, using the blocks

$$P_{h,2h}^l = I - (I_{2h}^h)^l (A_{2h}^l)^{-1} (I_h^{2h})^l A_h^l,\ 1 \leq l \leq n/2$$

A standard computation gives the 2×1 blocks of bilinear interpolation where for $1 \leq l \leq n/2 - 1$

$$(I_{2h}^h)^l = \frac{1}{2}\begin{pmatrix} (1 + c_l) \\ -(1 - c_l) \end{pmatrix},$$

and where

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

The Galerkin coarsening then results in the 1×1 blocks where for $1 \leq l \leq n/2 - 1$

$$A_{2h}^l = (I_h^{2h})^l A_h^l (I_{2h}^h)^l = \frac{1}{2h^2}\big[2(1 - c_l^2) - \kappa^2(1 + c_l^2)\big]$$

A straightforward computation subsequently allows to obtain for $1 \leq l \leq n/2 - 1$ the following 2×2 blocks of $(P_{h,2h}^l)^T$

$$(P_{h,2h}^l)^T = \frac{1}{C}\begin{pmatrix} -(c_l + 1)(c_l^2 - 1) + \frac{1}{2}\kappa^2(c_{l_2}^2 - 1) & \frac{1}{2}(c_l^2 - 1)(2 + 2c_l - \kappa^2) \\ \frac{1}{2}(c_l^2 - 1)(-2 + 2c_l + \kappa^2) & (c_l^2 - 1)(3 + c_l) + \frac{1}{2}\kappa^2(c_l^2 + 3) \end{pmatrix}$$

where $c = \frac{1}{2(1-c_l^2)+\kappa^2(c_l^2+1)}$ and 1×1 block

$$(P_{h,2h}^{n/2})^T = 1$$

The basis V_h can therefore be used to block diagonalize the deflated preconditioner operator, i.e., we can write

$$B_{h,2h,(\beta_1,\beta_2)} = \big[B_{h,2h,(\beta_1,\beta_2)}^l\big]_{1 \leq l \leq n/2}$$

where for $1 \leq l \leq n/2 - 1$, $B^l_{h,2h,(\beta_1,\beta_2)}$ is the 2×2 matrix

$$B^l_{h,2h,(\beta_1,\beta_2)} = (P^l_{h,2h})^T \text{diag}\begin{pmatrix} \lambda^l(S_{h,(\beta_1,\beta_2)}) \\ \lambda^{n-l}(S_{h,(\beta_1,\beta_2)}) \end{pmatrix}$$

and where

$$B^{n/2}_{h,2h,(\beta_1,\beta_2)} = \lambda^{n/2}(S_{h,(\beta_1,\beta_2)}) = \frac{2 - \kappa^2}{2 - \kappa^2(\beta_1 - i\beta_2)} \tag{4.7}$$

This block diagonal form renders an analytical computation of the eigenvalues of $B_{h,2h,(\beta_1,\beta_2)}$ feasible and results in the conclusion that $B_{h,2h,(\beta_1,\beta_2)}$ has a zero eigenvalue of multiplicity $n/2-1$, the eigenvalue 4.7 and $n/2 - 1$ eigenvalues of the form

$$\lambda^l(B_{h,2h,(\beta_1,\beta_2)}) = \frac{a_l + ib_l}{c_l + id_l} \text{ for } 1 \leq l \leq n/2 - 1$$

where a_l, b_l, c_l and d_l are third order polynomials in κ^2 and given by

$$\begin{aligned} a_l =& (-1 - c_l^2)\beta_1\kappa^6 + (4\beta_1 + 2 - 2c_l^2 + 4c_l^2\beta_1)\kappa^4 \\ &+ (8c_l^2 - 4\beta_1 - 8 + 4c_l^4\beta_1)\kappa^2 + (8 - 16c_l^2 + 8_l^4) \qquad (4.8) \\ b_l =& (1 + c_l^2)\beta_2\kappa^6 + (-4c_l^2\beta_2 - 4\beta_2)\kappa^4 + (4\beta_2 - 4c_l^4\beta_2)\kappa^2 \\ c_l =& (\beta_2^2 - \beta_1^2 + c_l^2\beta_2^2 - c_l^2\beta_1^2)\kappa^6 \qquad (4.9) \\ &+ (4\beta_1 - 2c_l^2\beta_1^2 + 2c_l^2\beta_2^2 + 2\beta_1^2 - 2\beta_2^2 + 4c_l^2\beta_1)\kappa^4 \\ &+ (8\beta_1 c_l^2 - 8\beta_1 - 4 + 4c_l^4)\kappa^2 + 8c_l^4 + 8 - 16c_l^2 \\ d_l =& (2\beta_1\beta_2 + 2c_l^2\beta_1\beta_2)\kappa^6 + (4c_l^2\beta 1\beta 2 - 4\beta_1\beta_2 - 4\beta_2 - 4c_l^2\beta_2)\kappa^4 \\ &+ (8\beta_2 - 8c_l^2\beta_2)\kappa^2 \qquad (4.10) \end{aligned}$$

Additionally, we obtain the following expression for the eigenvalues of $P^l_{h,2h})^T A_h$

$$\lambda^l(P^T_{h,2h}A_h) = -\frac{(c_l^2 + 1)\kappa^4 + (-4c_l^2 - 4)\kappa^2 - 4(c_l^4 - 1)}{((c_l^2 + 1)\kappa^2 + 2(c_l^2 - 1))h^2} \tag{4.11}$$

4.2.2 Near-null Eigenvalues

Using the expressions above, we can investigate the behavior of the eigenvalues. So far we are familiar with the occurrence of small eigenvalues for the CSLP preconditioned system. However, for large wave numbers k these small eigenvalues reappear and seem to ponder as to what causes the ADEF-preconditioned solver to remain non-scalable for large wavenumbers. In order to gain insight into this phenomena, we will compare the real part of the spectrum of the ADEF-preconditioner applied to the coefficient matrix A_h and CSLP-preconditioned matrix $\widehat{A} = M_{h,(\beta_1,\beta_2)}A_h$. For ease of notation, we now denote the matrices by A, M and P respectively. We will use the expressions obtained in equation 4.10 and 4.11.

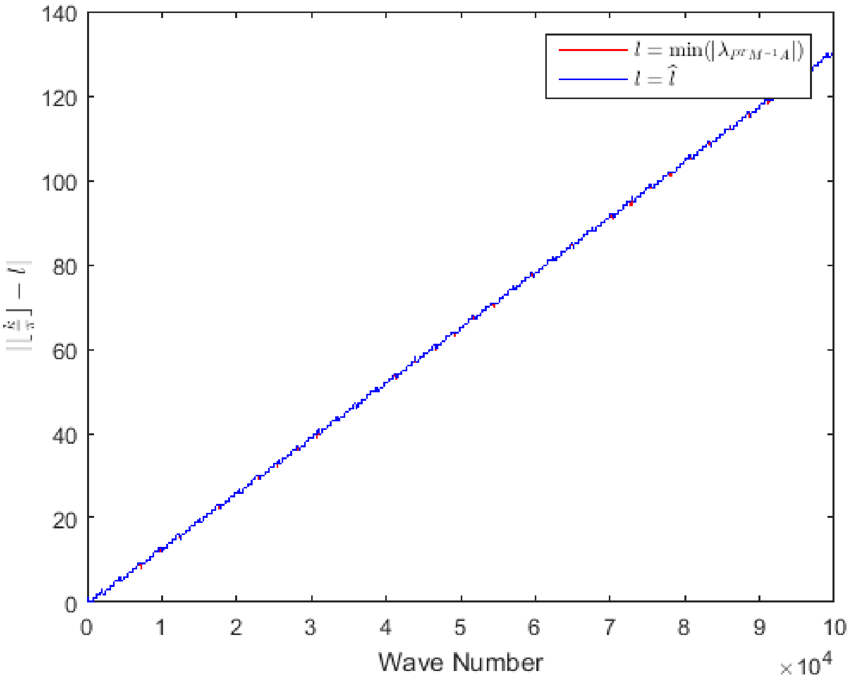
Previously we noted that the index at which the smallest eigenvalues appear in the discrete case are different compared to the index for the analytical case, which always centers around $\frac{k}{\pi}$. In the previous case, we discovered that the difference between the analytical index $j = \lfloor \frac{k}{\pi} \rfloor$ and the discrete index l at which these eigenvalues occur increases linearly. In order to locate where the problem originates, we therefore proceed with a similar analysis, starting by investigating where the smallest eigenvalues appear for the deflated, ADEF-preconditioned and CSLP-preconditioned systems respectively.

We mark the index l at either the least negative or smallest positive real part of the eigenvalues of $P^T M^{-1} A$. Here we explicitly compute the real part of spectrum and take the minimum over the total set. We also mark the approximated index $\widehat{l}$ obtained in section 3.2.2 by

$$\widehat{l} = \text{round}(\frac{\arccos(1 - \frac{h^2(k^2)}{2})}{\pi h}), \widehat{l} \in \mathbb{N} \setminus \{0\}. \tag{4.12}$$

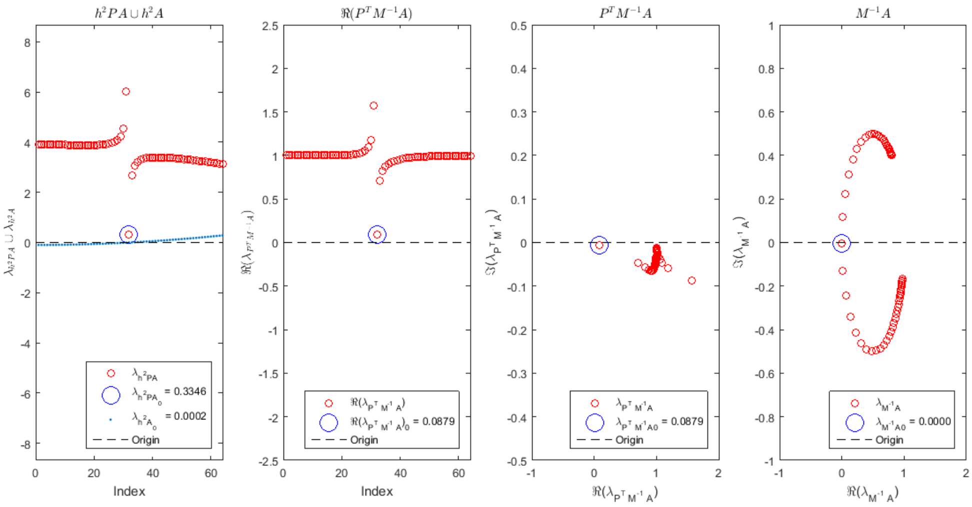
We will use equation 4.12 in order to investigate to what extent the location of the troubling eigenvalues of $\widehat{A}$ and $P^TM^{-1}A$ coincide. Equation 4.12 provides an alternative to locating these eigenvalues without computing them, compared to using the arccos function in conjunction with equation 4.10. The latter option would be cumbersome due to the complexity of the expressions.
Figure 4.1 contains a plot of the difference between $|\lfloor \frac{k}{\pi} \rfloor - l|$ and resembles the departure of index $\widehat{l}$ from $\lfloor \frac{k}{\pi} \rfloor$ as k increases.

Figure 4.1: *Difference between l and $\lfloor \frac{k}{\pi} \rfloor$ as a function of k using κ = 0.3125..* $\widehat{l}$ has been calculated using equation 4.12.



The results are comparable to the ones obtained for the CSLP-preconditioned system $\widehat{A}$ in section 3.2.2. For increasing k, the difference between the index at which the smallest eigenvalue occurs and $\lfloor \frac{k}{\pi} \rfloor$ increases linearly. Furthermore, equation 4.12 appears to locate the index at which the smallest eigenvalue of $P^TM^{-1}A$ occurs quite adequately, supporting the preliminary conclusion that the troubling eigenvalues of $\widehat{A}$ and $P^TM^{-1}A$ seem to originate from the same source.
Based on the results from Figure 4.1, we now let $\widehat{l}$ be in accordance with equation 4.12 and use this index to plot the eigenvalues of PA, A, $P^TM^{-1}A$ and $M^{-1}A$. We investigate the behavior for $k = 100, 1000$ and 10000 complementary to Sheikh (2014), keeping κ fixed at 0.3125, unless stated otherwise. Note that all the eigenvalues of the deflated coefficient matrix PA are real. From Chapter 3, section 3.2.2 we know that the troubling eigenvalues for the CSLP-preconditioned system $\widehat{A}$ are located at the same index where the eigenvalues of A are closest to zero. In order to assess whether a similar conclusion can be drawn for the system PA, we include the eigenvalues of A in the plot as a benchmark. In order to simplify the analysis, we have scaled the eigenvalues of PA and A by h^2. The blue marker represents the eigenvalue corresponding to the index $\widehat{l}$, which has been calculated using equation 4.12

Figure 4.2: *Eigenvalues of h^2PA, $\Re(P^TM^{-1}A)$ and $M^{-1}A$ resp. for $k = 100$. The marker indicates the eigenvalue corresponding the index $\widehat{l}$, which has been calculated using equation 4.12.*



For $k = 100$, we observe that the real part of the spectrum of $P^TM^{-1}A$ exhibits a similar pattern relative to the spectrum of h^2PA. The index corresponding to the eigenvalue of h^2A approaching zero coincides with the index of the eigenvalue of $\Re(P^TM^{-1}A)$ approaching zero, which supports the notion from Figure 4.1. Consequently, as the eigenvalues of h^2PA approach the eigenvalues of h^2A near zero, so do the eigenvalues of $\Re(P^TM^{-1}A)$. As a result, there seems to be a one-to-one correspondence between the zero-approaching eigenvalues. Compared to the smallest eigenvalue of $M^{-1}A$, the magnitude of the smallest eigenvalue of h^2PA and h^2A is considerably larger. Also, more small eigenvalues appear to be clustering around zero for the CSLP-preconditioned system compared to the ADEF-preconditioned system. We now proceed by repeating the previous experiments for $k = 1000$ and $k = 10000$.

Figure 4.3: *Eigenvalues of h^2PA, $\Re(P^TM^{-1}A)$ and $M^{-1}A$ resp. for $k = 1000$. The marker indicates the eigenvalue corresponding the index $\widehat{l}$, which has been calculated using equation 4.12.*

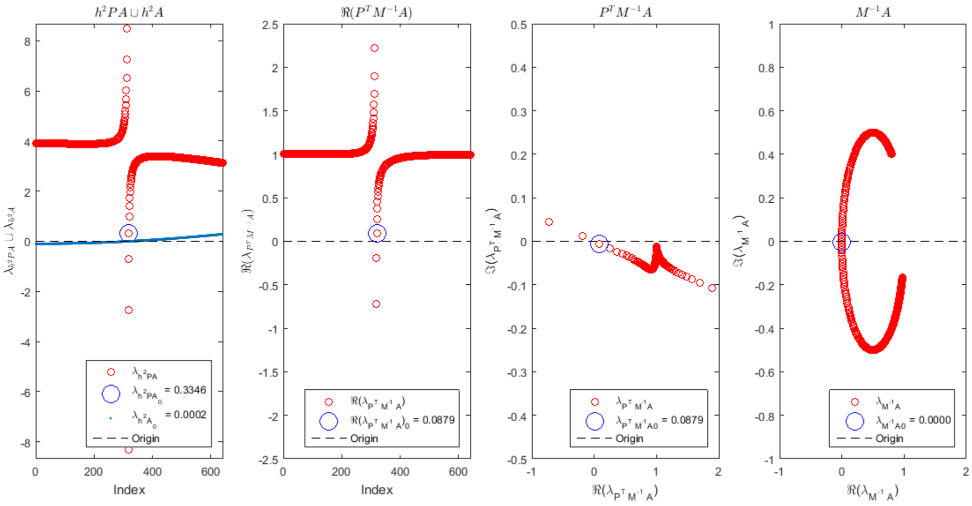


Figure 4.4: *Eigenvalues of h^2PA, $\Re(P^TM^{-1}A)$ and $M^{-1}A$ resp. for $k = 10000$. The marker indicates the eigenvalue corresponding the index $\widehat{l}$, which has been calculated using equation 4.12.*

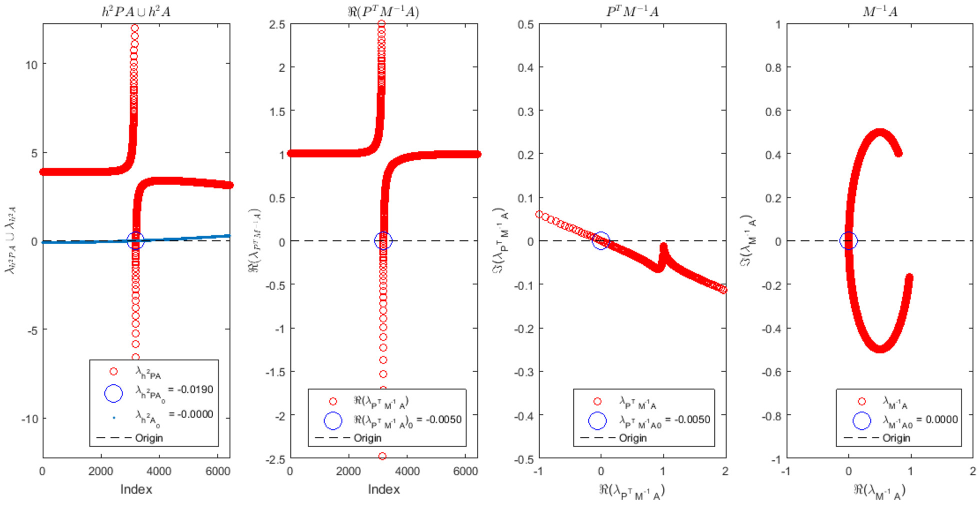


Figure 4.3 and Figure 4.4 are illustrative of the fact that as k is increased, the eigenvalues of the ADEF-preconditioned system are starting to cluster around zero. Interestingly, the magnitude of the eigenvalues does not necessarily drop. This effect, on the contrary, was reported for the CSLP-preconditioned system. For example, for both $k = 100$ and $k = 1000$, the real part of the smallest eigenvalue of $P^TM^{-1}A$ is 0.0879. For $k = 10000$ we get a value of 0.005. Despite the smallest eigenvalue becoming even smaller for $k = 10000$ relatively, this does not necessarily imply that the order of magnitude drops solely as k increases. As we can see from Figure 4.5, already for $k = 1233$ we obtain an eigenvalue of similar magnitude compared to the case where $k = 10000$, which indicates that the magnitude of the smallest eigenvalue in relation to increasing wave number k is not necessarily monotone.

Figure 4.5: *Eigenvalues of $\Re(P^TM^{-1}A)$ for $k = 1200, 1220, 1233, 1250$. The marker indicates the eigenvalue corresponding the index $\widehat{l}$, which has been calculated using equation 4.12.*

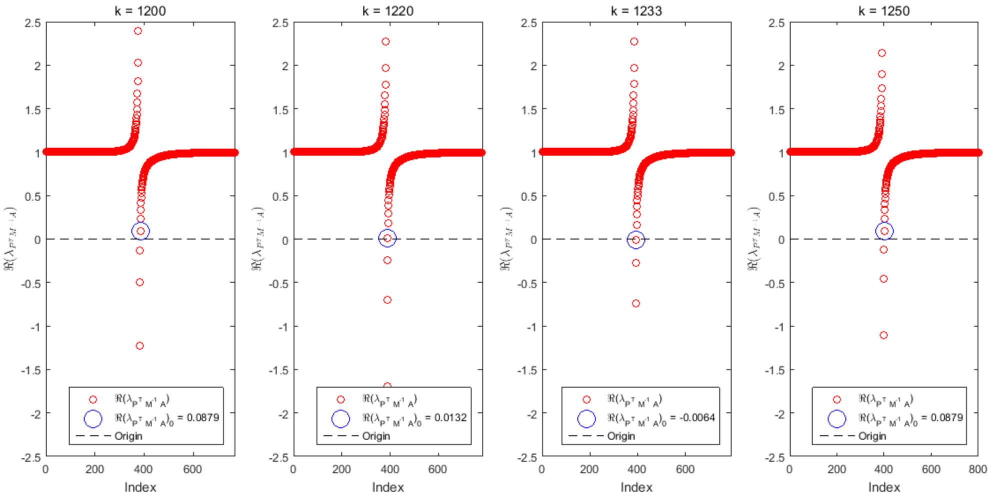
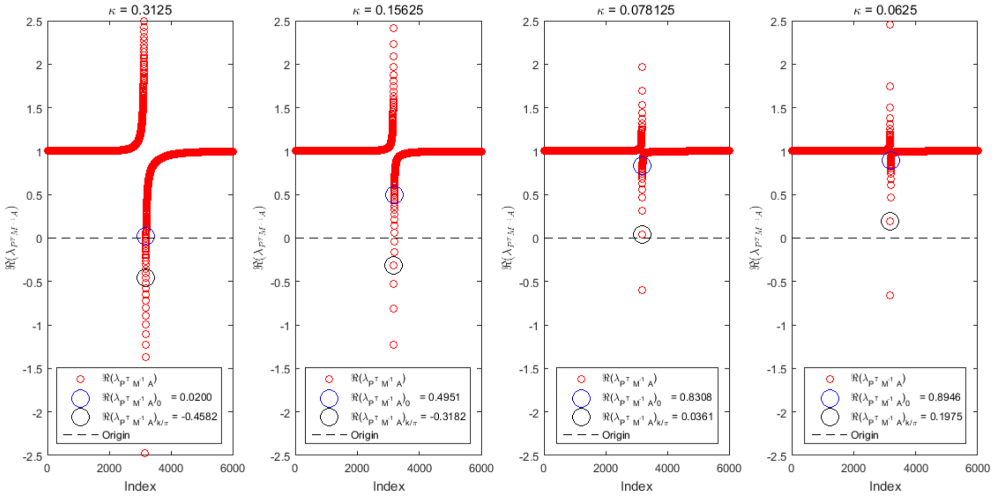


Figure 4.6 plots the eigenvalues of the $\Re(P^TM^{-1}A)$ for $k = 10000$, while increasing the number of grid points per wavelength. The eigenvalues of $M^{-1}A$ were reported to be independent of the grid and

step size, whereas the eigenvalues of $\Re(P^T M^{-1} A)$ appear to be dependent on these spatial parameters. Increasing the grid resolution by letting h go to zero dissolving the cluster of small eiegenvalues near zero. However, this comes at a price as increasing the number of grid points per wavelength also increases the problem size. As a result, more and more eigenvalues get deflated as the dimension of the deflation subspace increases along the problem size. Next to dissolving the clustering small eigenvalues, increasing the number of grid points per wavelength additionally enforces the smallest eigenvalue to be located at the index $\lfloor \frac{k}{\pi} \rfloor$ (black marker). The discrete eigenvalues become better approximations of their continuous counterpart. A similar effect was also reported for the eigenvalues of $M^{-1}A$, see section 3.2.1.

Figure 4.6: *Eigenvalues of* $\Re(P^T M^{-1} A)$ *for* $k = 10000$ *and number of grid points per wavelength* $20, 40, 80, 100$. *The blue marker indicates the index calculated using equation 4.12. The black marker indicates the index* $\lfloor \frac{k}{\pi} \rfloor$.



4.2.3 Convergence Behavior

In this section we will explore the convergence behavior of the ADEF-preconditioned system versus the CSLP-preconditioned system. One of the reported findings in both the literature and our results thus far indicate that the ADEF-preconditioner suffers from clustering near nullspace eigenvalues for high wave numbers. In order to see the effect of these eigenvalues on the convergence behavior, Table 4.1 contains the results after implementing the two-level deflation preconditioner ADEF on our model problem. We have inverted all matrices exactly using a direct solver.

Recall that we only implemented Dirichlet boundary conditions, resulting in a system which does not introduce any damping and should be considered a spectrally worst case scenario (Sheikh, 2014). We also compare the performance to the CSLP preconditioner.

Table 4.1: *Number of full GMRES iterations using the ADEF and CSLP preconditioner. Iter. gives the number of GMRES iterations and Rel. Res. represents the relative residual.*

	$\kappa = 0.625$				$\kappa = 0.3125$			
k	ADEF		CSLP		ADEF		CSLP	
	Iter.	Rel. Res.	Iter.	Rel. Res.	Iter.	Rel. Res.	Iter.	Rel. Res.
100	16	1.6650e-08	44	7.6988e-08	10	2.2809e-08	44	7.4631e-08
500	41	4.1160e-08	160	9.6237e-08	19	6.2831e-08	155	6.1181e-08
1000	67	5.1416e-08	296	4.7705e-08	27	7.0026e-08	283	8.5375e-08
5000	251	8.6474e-08	1200	3.3587e-02	81	5.9620e-08	1200	6.9000e-03

These preliminary findings support the notion obtained from observing the spectrum of $P^T M^{-1} A$ and $\widehat{A}$; the convergence behavior of the ADEF preconditioner seems to be highly dependent on the grid resolution κ and as a result the step size h. For $\kappa = 0.625$, the convergence behavior appears to be sublinear. Lowering κ by increasing the number of grid points per wave length reduces the iterations needed by the ADEF-solver to reach convergence. These findings do not apply analogously to the CSLP-solver, which was already reported to be independent of the step size, see section 3.2.2. Moreover, the CSLP-solver diverges for $k = 5000$, see section 3.2.3.
The effect of the deflated eigenvalues is perceptible at both levels of κ. Removing most of the bad eigenvalues and dissolving the cluster near zero immediately results in a substantial reduction in the number of iterations. Note that in case Sommerfeld radiation conditions are implemented as boundary conditions, the original coefficient matrix A will contain a damping term and we expect the number of iterations to drop even further (Sheikh, 2014).

Finally, we investigate to what extent the convergence behavior of the ADEF-solver is impacted by allowing the complex shift β_2 to vary. In section 3.2.1 we noted that the convergence behavior of the CSLP-solver is positively impacted by lowering β_2 as this reduces the amount of near nullspace eigenvalues. However, implementing a smaller complex shift leads to the preconditioner resembling the original coefficient matrix, making the inversion more difficult.
Table 4.2 presents the result of implementing the ADEF-preconditioner using different complex shifts and shows that the ADEF-preconditioner is less sensitive to varying β_2, especially for $\kappa = 0.3125$. A similar conclusion has been rendered in Sheikh (2014) for the two-dimensional constant wave number problem.

Table 4.2: *Number of GMRES iterations for* $\kappa = 0.625$ *and* $\kappa = 0.3125$ *using the ADEF-preconditioner with shifts* $(\beta_1, \beta_2) = (1, 0.25)$ *and* $(1, 1)$.

	$\kappa = 0.625$		$\kappa = 0.3125$	
k	ADEF(1,0.25)	ADEF(1,1)	ADEF(1,0.25)	ADEF(1,1)
	Iter.	Iter.	Iter.	Iter.
100	14	17	10	10
200	22	25	11	12
300	27	29	15	16
400	32	35	17	17
500	38	42	19	19
600	43	46	20	21
700	48	51	21	21
800	53	57	24	24

4.3 Concluding Remarks and Summary

This chapter evolved around the literature and main results as regards the application of geometric deflation techniques to the Helmholtz equation. Before we summarize our principal findings, we would like to remark that so far we have been comparing the clustering eigenvalues of the CSLP-preconditioned systems to the ADEF-preconditioned system. For the latter, we also recorded clustering eigenvalues for large k. Given that the range of both spectra is not equivalent, the conclusion as regards the clustering eigenvalues should be perhaps be adjusted to reflect this notion. For example, suppose for the CSLP-preconditioned system a cluster of 10 eigenvalues reside between 0 and 0.01, which accounts for 1 percent of the range. Despite, the ADEF-preconditioned system having less eigenvalues between 0 and 0.01, the clustering of eigenvalues with larger magnitude around zero seem to be responsible for the increase in the number of iterations. Thus, for the ADEF-preconditioned system, 10 eigenvalues between 0.01 and 0.5 may cause more harm, see Figure 4.3 and Table 4.1.

We conclude this chapter by summarizing our main findings into the following points

- Deflation based preconditioning applied to the CSLP-preconditioned Helmholtz equation leads to the ADEF-preconditioner, where multigrid interpolation vectors are chosen as basis vectors
- The ADEF+CSLP-preconditioned system has $n/2$ eigenvalues deflated towards zero
- The ADEF+CSLP-preconditioner outperforms the CSLP-preconditioner
- As k increases clustering eigenvalues near zero reappear. Their magnitude, however, is not necessarily monotone as k increases.
- The center of the problem for both the CSLP-preconditioned and ADEF-preconditioned systems originates from the same point; the index of the smallest absolute eigenvalue of the original coefficient matrix A.
- The eigenvalues of the ADEF-preconditioned and ADEF+CSLP-preconditioned system are grid dependent, unlike the eigenvalues of the CSLP-preconditioned system. Increasing the number of grid points per wave length reduces the clustering eigenvalues around zero
- The ADEF+CSLP-preconditioned system is insensitive to changing β_2

Chapter 5

Research Proposal

5.1 Preliminary Findings and Conclusions

5.2 Test Problems

For the sake of theoretical thoroughness, most of the analysis and cited literature in this interim thesis was based on our simple one-dimensional model problem from Chapter 1, section 1.3. In this subsection, we will define some more complicated models next to our original model problem in order to serve our future research purposes and to test the robustness and validity of the results that will be obtained. These model problems coincide with the test problems used by Sheikh (2014).

5.2.1 One-Dimensional Test Problems

Next to our original model problem as defined in Chapter 1 section 1.3, we will additionally include another one-dimensional test problem where we implement Sommerfeld radiation conditions along with Dirichlet boundary conditions. Similar to our original model problem, we place the source at the center of the numerical domain.

Constant Wave Number

On the standard unit domain $\Omega = [0, 1]$ with constant wave number k we consider

$$\begin{aligned}
&-\Delta u(x) - k^2 u(x) = \delta(x - \frac{1}{2}), x \in \Omega \setminus \partial\Omega,\\
&u(x) = 0, x \in \partial\Omega,\\
&\left(\frac{\partial}{\partial n} - ik\right) u(x) = 0, x \in \partial\Omega,
\end{aligned}$$

where n denotes the outward normal unit vector in the x-direction.

5.2.2 Two-Dimensional Test Problems

Here we describe the possible two-dimensional test problems, which can be used for future research purposes.

Constant Wave Number

On the standard two-dimensional square unit domain $\Omega = [0,1] \times [0,1]$ with constant wave number k we consider

$$
\begin{aligned}
&-\Delta u(x,y) - k^2 u(x,y) = \delta(x - \frac{1}{2}, y - \frac{1}{2}),\ (x,y) \in \Omega \setminus \partial\Omega,\\
&u(x,y) = 0,\ (x,y) \in \partial\Omega\\
&\left(\frac{\partial}{\partial \mathbf{n}} - ik\right) u(x,y) = 0, (x,y) \in \partial\Omega,
\end{aligned}
$$

where n denotes the outward normal unit vector in the x- and y-direction respectively. Sheikh (2014) uses this test problem as the main model to conduct numerical experiments for large k.

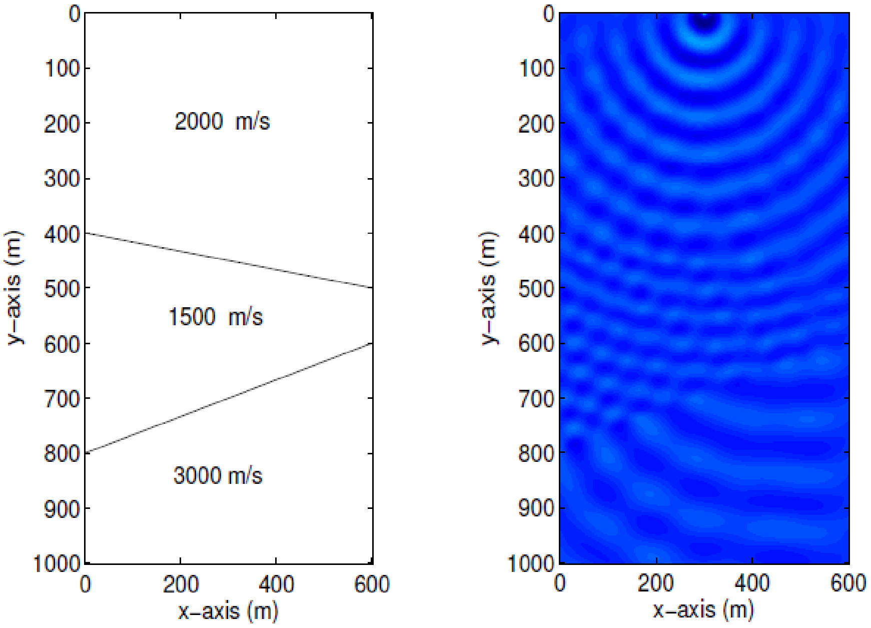
Wedge Problem

This model problem accounts for various sources of heterogeneity, which is common in physical problems arising in geophysical seismic imaging. We consider an inhomogeneous medium, where the domain is subdivided into three layers with different velocities and consequently different wave numbers. We define the domain $\Omega = [0,600] \times [0,1000]$. For a visual representation of the domain for the wedge problem, see Figure 5.1 a. To model wave diffraction, a point source is placed at $(x,y) = (300,0)$, which leads to the pattern as depicted in Figure 5.1 b. On Ω, we define

$$
\begin{aligned}
&-\Delta u(x,y) - k(x,y)^2 u(x,y) = \delta(x - 300, y),\ (x,y) \in \Omega \setminus \partial\Omega,\\
&\left(\frac{\partial}{\partial \mathbf{n}} - ik\right) u(x,y) = 0, (x,y) \in \partial\Omega,
\end{aligned}
$$

where n denotes the outward normal unit vector in the x- and y-direction respectively. Moreover, $k(x,y) = \frac{2\pi freq}{c(x,y)}$ is given in terms of the velocity profile as shown in Figure 5.1 a. Sheikh (2014) uses a set of four different frequencies 10, 20, 40 and 80 Hz for numerical testing.

Figure 5.1: *(a) Domain of the wedge problem with velocity distribution over 3 layers. (b) Pattern of wave diffraction through layers of different velocity.*



Marmousi Problem

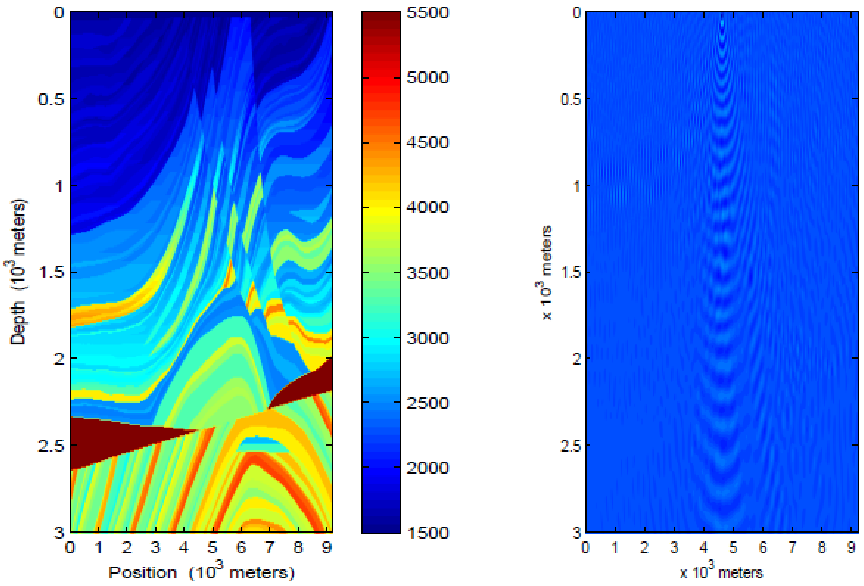
The final test problem is a representation of an industrial problem and is widely referred to as the Marmousi Problem. The original Marmoussi problem is defined on a rectangular domain $\Omega = [0, 9200] \times 0, 3000]$. Unlike the wedge-problem which consist of three layers, there are now 158 layers with velocities ranging from 1500 m/s to 5500 m/s. Sheikh (2014) considers a slightly adapted version of the original Marmousi problem. The original domain has been truncated to $\Omega = [0, 8192] \times [0, 2048]$ in order to allow for efficient geometric coarsening of the discrete velocity profiles given that the domain remains in powers of 2. The original velocity $c(x, y)$ is also adapted by considering $2587.5 \leq c \leq 3325$.
On the adjusted domain Ω, we define

$$-\Delta u(x,y) - k(x,y)^2 u(x,y) = \delta(x - 4000, y), (x,y) \in \Omega \setminus \partial\Omega,$$

$$\left(\frac{\partial}{\partial \mathbf{n}} - ik\right) u(x,y) = 0, (x,y) \in \partial\Omega,$$

where n denotes the outward normal unit vector in the x- and y-direction respectively. Similar to the wedge problem, we have a non-constant wave number $k(x,y) = \frac{2\pi freq}{c(x,y)}$, where in this particular case $c(x,y)$ ranges between 2587.5 and 3325. For this adjusted version of the Marmousi problem, numerical experiments have been conducted by Sheikh (2014) using the frequencies 1, 10, 20 and 40 Hz, where the grid has been resolved in such a way that the maximum wave number k at $freq = 1$ has a grid resolution of $kh \leq 0.039$. For the remaining frequencies, a grid resolution of $kh \leq 0.39$ is utilized.
A visual representation of the domain of the original Marmousi problem and the results obtained using a frequency of 20 Hz. are illustrated in Figure 5.2 a and b.

Figure 5.2: *(a) Domain of the Marmousi problem with velocity distribution over 158 layers. (b) Pattern of wave diffraction through layers of different velocity for freq = 20.*



5.3 Research Questions

Using the findings from both the analysis and literature survey in this interim thesis, we will conduct future research regarding the scalability of an iterative Helmholtz solver. In order to determine whether we can obtain a scalable iterative Helmholtz solver, we will answer the following main questions:

1. To what extent is the ADEF-solver scalable?

(a) What causes the current solver to remain inscalable?

(b) What parameters can be identified to influence the solver scalability?

- Which parameters can be defined on the operator level?
- Which parameters can be defined on the geometric level?

2. Is there a relation between the pollution error and the cause for inscalability? (i.e. the close to zero eigenvalues). If affirmative, will reducing the pollution error lead to better scalability of the solver?

   - How does the spectrum of $P^T M^{-1} A$ alter after introducing a perturbed wave number k to model numerical pollution?
   - What is the effect of the perturbation on the coarse resp. fine grid?

3. What alternatives are available to obtain a scalable iterative Helmholtz solver?

If time permits, we will additionally briefly look into the effects of implementing different discretization schemes on the propagation of the numerical pollution error.

Bibliography

D. A. Sheikh, C.Vuik. A scalable helmholtz solver combining the shifted laplace preconditioner with multigrid deflation. Technical report, DIAM, jan 2011.

W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics*, 9(1):17–29, 1951.

A. Bayliss, C. I. Goldstein, and E. Turkel. An iterative method for the helmholtz equation. *Journal of Computational Physics*, 49(3):443–457, 1983.

A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the fem solution for the helmholtz equation in one, two and three dimensions. *International journal for numerical methods in engineering*, 46(4):471–499, 1999.

Z. Dostál. Conjugate gradient method with preconditioning by projector. *International Journal of Computer Mathematics*, 23(3-4):315–323, 1988.

M. Eiermann and O. G. Ernst. Geometric aspects of the theory of krylov subspace methods. *Acta Numerica 2001*, 10:251–312, 2001.

Y. Erlangga, C. Vuik, and C. Oosterlee. On a class of preconditioners for solving the discrete Helmholtz equation.

Y. A. Erlangga. *A robust and efficient iterative method for the numerical solution of the Helmholtz equation*. PhD thesis, 2005.

Y. A. Erlangga and R. Nabben. Deflation and balancing preconditioners for krylov subspace methods applied to nonsymmetric matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(2):684–699, 2008.

Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous helmholtz problems. *SIAM Journal on Scientific Computing*, 27(4):1471–1492, 2006a.

Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. Comparison of multigrid and incomplete lu shifted-laplace preconditioners for the inhomogeneous helmholtz equation. *Applied numerical mathematics*, 56(5):648–666, 2006b.

O. G. Ernst and M. J. Gander. Multigrid methods for helmholtz problems: A convergent scheme in 1d using standard components.

O. G. Ernst and M. J. Gander. Why it is difficult to solve helmholtz problems with classical iterative methods. In *Numerical analysis of multiscale problems*, pages 325–363. Springer, 2012.

M. J. Gander and F. Nataf. Ailu: a preconditioner based on the analytic factorization of the elliptic operator. *Numerical linear algebra with applications*, 7(7-8):543–567, 2000.

M. J. Gander and F. Nataf. Ailu for helmholtz problems: a new preconditioner based on the analytic parabolic factorization. *Journal of Computational Acoustics*, 9(04):1499–1506, 2001.

A. Gaul. Recycling krylov subspace methods for sequences of linear systems: analysis and applications. 2014.

K. Gerdes and F. Ihlenburg. On the pollution effect in fe solutions of the 3d-helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 170(1):155–172, 1999.

A. Hannukainen. Convergence analysis of gmres for the helmholtz equation via pseudospectrum. *arXiv preprint arXiv:1505.08072*, 2015.

F. Ihlenburg and I. Babuska. Finite element solution of the helmholtz equation with high wave number part ii: the hp version of the fem. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997.

A. L. Laird and M. Giles. Preconditioned iterative solution of the 2d helmholtz equation. Technical report, 2002.

C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.

J. Liesen and P. Tichỳ. Convergence analysis of krylov subspace methods. *GAMM-Mitteilungen*, 27(2): 153–173, 2004.

F. N. M. Gander. An incomplete lu preconditioner for problems in acoustics. *Journal of Computational Acoustics*, 13(3):1–22, 2005.

G. Meurant and J. D. Tebbens. The role eigenvalues play in forming gmres residual norms with non-normal matrices. *Numerical Algorithms*, 68(1):143–165, 2015.

R. B. Morgan. A restarted gmres method augmented with eigenvectors. *SIAM Journal on Matrix Analysis and Applications*, 16(4):1154–1171, 1995.

R. B. Morgan. Gmres with deflated restarting. *SIAM Journal on Scientific Computing*, 24(1):20–37, 2002.

R. Nabben and C. Vuik. A comparison of deflation and coarse grid correction applied to porous media flow. *SIAM Journal on Numerical Analysis*, 42(4):1631–1647, 2004.

R. Nabben and C. Vuik. A comparison of abstract versions of deflation, balancing and additive coarse grid correction preconditioners. *Numerical Linear Algebra with Applications*, 15(4):355–372, 2008a.

R. Nabben and C. Vuik. A comparison of abstract versions of deflation, balancing and additive coarse grid correction preconditioners. *Numerical Linear Algebra with Applications*, 15(4):355–372, 2008b.

R. A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM Journal on Numerical Analysis*, 24(2):355–365, 1987a.

R. A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM Journal on Numerical Analysis*, 24(2):355–365, 1987b.

O. Runborg. Helmholtz equation and high frequency approximations. *Lecture notes for Numerical Solutions of Differential Equations*, 2012.

Y. Saad. *Numerical Methods for Large Eigenvalue Problems: Revised Edition*, volume 66. Siam, 2011.

A. Sheikh. *Development Of The Helmholtz Solver Based On A Shifted Laplace Preconditioner And A Multigrid Deflation Technique.* TU Delft, Delft University of Technology, 2014.

A. Sheikh, C. Vuik, and D. Lahaye. Fast iterative solution methods for the helmholtz equation. Technical report, Delft University of Technology, Faculty of Electrical and Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, 2009.

A. Sheikh, D. Lahaye, L. G. Ramos, R. Nabben, and C. Vuik. Accelerating the shifted laplace preconditioner for the helmholtz equation by multilevel deflation. *Journal of Computational Physics*, 322: 473–490, 2016.

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*, 29:1942–1958, 2007a.

M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete helmholtz operator preconditioned with a shifted laplacian. *SIAM Journal on Scientific Computing*, 29(5):1942–1958, 2007b.

C. Vuik and D. Lahaye. Scientific computing (wi4201). *Lecture notes for wi4201*, 2012.

K. Wang and Y. S. Wong. Pollution-free finite difference schemes for non-homogeneous helmholtz equation. *Int. J. Numer. Anal. Model*, 11(4):787–815, 2014.

M. Yeung, J. Tang, and C. Vuik. On the convergence of gmres with invariant-subspace deflation, report 10-14, delft institute of applied mathematics. *Delft University of Technology*, 2010.