university of groningen

faculty of science and engineering

Solving the Helmholtz equation numerically

Master Project Mathematics

January 2018

Student: H.T. Stoppels

First supervisor: Dr. ir. F.W. Wubs

Second supervisor: Prof. dr. Arjan van der Schaft

Abstract

Linear systems $Ax = b$ involving large, sparse, indefinite and nearly singular matrices A naturally arise in interior eigenvalue problems. Classical iterative methods such as Krylov subspace methods are known to have difficulty with these problems. In this thesis we explore the possibility of obtaining cheap low-dimensional approximations to problematic eigenspaces in an attempt to deflate them. We show that approximate Schur complement techniques can be exploited to not only obtain these approximations, but to construct a preconditioner as well. The Helmholtz equation will be a guiding example throughout this work.

Contents

1 Introduction

In this thesis we will look at large linear systems

$$Ax = b \tag{1}$$

where the matrix $A \in \mathbb{C}^{n\times n}$ is sparse, indefinite and potentially nearly singular, and the solution x has components in the direction of the eigenvectors of A associated to the eigenvalues closest to the origin. This type of problem is tough for classical iterative methods, yet it arises naturally in interior eigenvalue problems.

In this introduction we will assess why these problems are considered so hard for Krylov subspace methods. Subsequently, we will see how these problems arise in the context of eigenvalue problems. Finally, we introduce the Helmholtz equation as an instance of this, as its discretization leads to the same kind of problems.

We will explore the literature surrounding the Helmholtz equation in Chapter 2, in the hope to find fruitful ideas that could carry over to solving interior eigenvalue problems. Then in Chapter 3 we will revisit problem (1) once more, and present an original analysis and some results.

1.1 Large, indefinite problems and iterative methods

Classical iterative methods such as Krylov subspace methods rely on the fact that matrix-vector products with sparse matrices A are a cheap $O(n)$ operation. By repeated multiplication with A, they build an ℓ-dimensional Krylov subspace

$$K_\ell(A, b) = span\{b, Ab, \ldots, A^{\ell-1}b\} \subset \mathbb{C}^n$$

and solve the problem $Ax = b$ approximately in the (Petrov-)Galerkin sense by imposing

$$Ax_\ell - b \perp \mathcal{W} \text{ for } \mathcal{V}$$

for the *search subspace* $\mathcal{V} = K_\ell(A, b)$ and a *test subspace* $\mathcal{W} \subset \mathbb{C}^n$. If A were symmetric positive-definite and $\mathcal{W} = \mathcal{V}$, then orthogonality in the inner product induced by A can lead to x_ℓ's with minimal error. For indefinite problems however, the only optimality property we can aim for is selecting $x_\ell \in \mathcal{V}$ such that the residual $r_\ell = b - Ax_\ell$ is minimized in the Euclidean norm. This is achieved by setting $\mathcal{W} = A\mathcal{V}$ as in least-squares problems and results in methods like GMRES and MINRES [18].

To see what makes the residual small, we write $x_\ell = p_\ell(A)b$ where p_ℓ is a polynomial of degree $(\ell - 1)$. The residual at iteration ℓ takes the form

$$r_\ell = b - Ax_\ell = [I - Ap(A)]b = q_\ell(A)b$$

where q_ℓ is a polynomial of order ℓ with the property $q_\ell(0) = 1$. For now we assume A is normal and write its orthonormal eigendecomposition as $AY = Y\Lambda$ where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$ the ith eigenvalue. The residual norm can hence be written as

$$\|r_\ell\|_2^2 = \sum_{i=1}^{n} q_\ell(\lambda_i)^2 (y_i, b)^2$$

and will be small in size whenever our polynomial q has its zeros near eigenvalues λ_j of A whenever b has large components in the direction of the corresponding eigenvector y_j.

Clustering of eigenvalues. Keeping the number of iterations small is equivalent to finding a low-order polynomial q that produces a small residual norm. Therefore it must be so that the eigenvalues are all clustered in $\mathbb{C}$, so that a few zeros of the polynomial q within this cluster wil suffice. However, when A is the direct discretization of a differential operator such as the Helmholtz operator, A's eigenvalues cannot be clustered as they approximate the eigenvalues of and unbounded operator. The usual trick is to precondition problem (1) with a mapping $M \in \mathbb{C}^{n\times n}$ as

$$M^{-1}Ax = M^{-1}b$$

such that $\|I - M^{-1}A\|$ is small in a sense.

Small residuals, yet large errors. This is however not the full story. Partly because our analysis of the residual only holds whenever A is normal, but more importantly because minimization of the residual is not equivalent to minimization of the error. Let us refer to eigenpairs of A corresponding with eigenvalues close to the origin *problematic* eigenpairs. Suppose we expand our solution in the eigenbasis of A as $x = YY^*x$ then the projection

$$(y_i, b) = (y_i, Ax) = \lambda_i (y_i, x)$$

shows that the contribution to the residual norm

$$\|r_\ell\|_2^2 = \sum_{i=1}^{n} q_\ell(\lambda_i)^2 \lambda_i^2 (y_i, x)^2$$

of components of x in the direction of problematic eigenvectors y_j is small by virtue of λ_j^2 being small. This is very undesirable, since the error in these directions can still be large.

Dealing with small eigenvalues The problem sketched above is not necessarily a problem of the extraction criterion (the choice of $\mathcal{W}$), but more generally a problem with the Krylov subspace as a search space. The combined facts that the Krylov subspace is shift-invariant in the sense that $K_\ell(A, b) = K_\ell(A-\tau I, b)$ for any $\tau \in \mathbb{C}$ and that it is constructed by iterates of the power method, make that good approximations to the "exterior" eigenvectors occur first in it. This observation has made many authors incorporate so-called *deflation techniques*, which come in a variety of forms, but are all centered around the idea that problematic eigenspaces must be removed from the operator A or explicitly appended to the search space $\mathcal{V}$ [6]. Deflation requires us to obtain a low-dimensional subspace $\mathcal{P} \subset \mathbb{C}^n$ that approximates the problematic eigenspaces well.

Assuming we *can* find such an approximation, suppose the columns of $P \in \mathbb{C}^{n\times m}$ with $m \ll n$ form a basis for $\mathcal{P}$ and the columns of $Q \in \mathbb{C}^{n\times(n-m)}$ form a basis for $\mathcal{P}^\perp$. Then we can recast the problem in these new bases as

$$\begin{bmatrix} Q^*AQ & Q^*AP \\ P^*AQ & P^*AP \end{bmatrix} \begin{bmatrix} x_q \\ x_p \end{bmatrix} = \begin{bmatrix} b_q \\ b_p \end{bmatrix}$$

for $x = Qx_q + Px_p$ and $b = Qb_q + Pb_p$. Note that P^*AP is $m \times m$, which is assumed to be small enough for direct methods to be applicable. Elimination of x_q however requires us to solve (among other things) the system

$$Q^*AQx_q = b_q, \tag{2}$$

which is not yet very attractive, because it is large and not suited for iterative methods as Q is not available and otherwise large and dense. We can circumvent this problem by lifting (2) back into $\mathbb{C}^n$, meaning that we have to solve

$$(I - PP^*)Ax_Q = (I - PP^*)b \text{ for } x_Q \perp P \tag{3}$$

Equations (2) and (3) are equivalent in the sense that x_q solves (2) iff $x_Q = Qx_q$ solves (3), but the advantage of (3) is that a matrix-vector product only requires m additional inner products and axpy's, leaving the total costs for a matrix-vector product at $O(n)$ complexity, where the hidden constant depends on m. Indeed, note that the Krylov subspace

$$K_\ell((I - PP^*)A, (I - PP^*)b)$$

is perpendicular to $\operatorname{Ran} P$ for all ℓ, making Krylov subspace methods very suitable to solve (3).

However, the catch is that deflation will only work well whenever we are able to construct a good enough approximate basis for the problematic eigenspaces at virtually no additional costs. This seems to run into circular reasoning, because as we will soon see, eigenproblems themselves can give rise to equations of the form (3) where P is initially of low quality. However, in Chapter 3 we will see that cheap approximations can sometimes be obtained.

1.2 Eigenvalue problems

The Arnoldi and Jacobi-Davidson methods are popular iterative algorithms to find a few solution to the eigenvalue problem

$$Ax = \lambda x \tag{4}$$

of a large and sparse matrix A for eigenvalues λ near a specified target $\tau \in \mathbb{C}$. The *interior eigenvalue problem* concerns the situation where τ is chosen well within the convex hull of eigenvalues of A. It is the interior eigenvalue problem that leads in both methods to systems (1) that are indefinite and nearly singular.

Arnoldi. The Arnoldi method solves (4) in the (Petrov)-Galerkin sense

$$Ax - \lambda x \perp \mathcal{W} \text{ for } x \in \mathcal{V}$$

where the search subspace $V = K_\ell(A, x_0)$. As mentioned in Section 1.1, the "exterior" eigenvectors enter the search subspace first, and therefore the eigenvalue problem (4) is recasted to shifted and inverted problem

$$(A - \tau I)^{-1}x = (\lambda - \tau)^{-1}x = \theta x,$$

so that in this formulation θ is large whenever λ is close to the target τ. The construction of the search subspace

$$\mathcal{V} = K_\ell((A - \tau I)^{-1}, x_0)$$

requires us to solve indefinite linear systems

$$(A - \tau I)y^{n+1} = y^n.$$

It is necessary to solve these systems accurately, since internally the method relies upon this relation in the Arnoldi decomposition [2].

Jacobi-Davidson. The Jacobi-Davidson method leads to similar indefinite systems, although they are not necessarily aimed to be solved up to high accuracy, as the method is not based on the Arnoldi decomposition. The solver can, in fact, be derived as a Newton method applied to the non-linear equation

$$f(x, \lambda) = \begin{bmatrix} Ax - \lambda x \\ \frac{1}{2}(\|x\|^2 - 1) \end{bmatrix} = 0.$$

Given an initial guess $\begin{bmatrix} \hat{x} & \hat{\lambda} \end{bmatrix}^T$, the Newton method prescribes a correction

$$\begin{bmatrix} \hat{x} \\ \hat{\lambda} \end{bmatrix} \leftarrow \begin{bmatrix} \hat{x} \\ \hat{\lambda} \end{bmatrix} - Df(\hat{x}, \hat{\lambda})^{-1} f(\hat{x}, \hat{\lambda})$$

where Df is the Jacobian. If we write the correction itself as the vector $\begin{bmatrix} y & \theta \end{bmatrix}^T$ such that the update reads $\hat{x} \leftarrow \hat{x} + y$ and $\hat{\lambda} \leftarrow \hat{\lambda} + \theta$, we obtain the system of equations

$$\begin{bmatrix} A - \hat{\lambda} I & -x \\ x^* & 0 \end{bmatrix} \begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} \hat{\lambda}\hat{x} - A\hat{x} \\ \frac{1}{2}(\|\hat{x}\|^2 - 1) \end{bmatrix}$$

The Jacobi-Davidson method does in fact not perform the update to $\hat{x}$, but rather enriches a search subspace with the correction y. Therefore we can discard θ altogether. As we have the freedom to pick $\|\hat{x}\| = 1$, and $\hat{\lambda} = \hat{x}^* A \hat{x}$ as the Rayleigh quotient, the problem for y is equivalent to solving *the correction equation*

$$(I - \hat{x}\hat{x}^*)(A - \hat{\lambda} I)y = -r \text{ for } y \perp \hat{x}, \tag{5}$$

where $r = A\hat{x} - \hat{\lambda}\hat{x}$. Here we use that $r \perp \hat{x}$ precisely because $\hat{\lambda}$ is the Rayleigh quotient. Note that (5) is identical to (3) in that it shares the deflation idea.

Initially, however, when $\hat{x}$ is not yet a good approximation to an eigenvector of an eigenvalue near τ, the method will have trouble to converge. Therefore one typically replaces $\hat{\lambda}$ with τ in the correction equation (5), which is more or less the same as doing a couple iterations of the Arnoldi method.

It is precisely in the early stage of Jacobi-Davidson where the correction equation (5) is nothing but a linear system involving the indefinite matrix $A - \tau I$ deflated with a virtually random vector $\hat{x}$.

1.3 The Helmholtz equation

As eigenvalue problems (4) are a broad subject, we narrow our research down to (discretizations of) a particular PDE: the Helmholtz equation. In what follows we will first introduce the Helmholtz equation and assess some of its properties before we discuss standard discretizations. This allows us to get some insights into the size of the linear systems and the behaviour of the solutions. Our notation and tools of analysis (Sobolev spaces in particular) are based on [9]. The only non-standard notation we use is the following.

Definition 1. We write $a \lesssim b$ and $b \gtrsim a$ whenever there exists a constant $C > 0$ such that $a \leq Cb$. If both $a \lesssim b$ and $b \lesssim a$, then we write $a \sim b$.

Definition 2. The $\|\cdot\|_{z,U}$ norm for $H^1(U)$ is defined as

$$\|v\|_{z,U}^2 := \|\nabla v\|_{L^2(U)}^2 + z^2 \|v\|_{L^2(U)}^2$$

for any non-zero constant $z \in \mathbb{R}$. It is obviously equivalent to the standard $\|\cdot\|_{H^1(U)}$ norm.

Let's consider the scalar wave equation for an unknown $v = v(t, x)$, which reads

$$v_{tt} = \nabla \cdot A(x)\nabla v + g \text{ in } (-\infty, \infty) \times U,$$

where $U \subset \mathbb{R}^d$ is the spatial domain, A is of size $d \times d$, real and positive definite matrix uniformly in x in the sense that there is a $\gamma > 0$ such that

$$y^* A(x) y \geq \gamma \|y\|_2^2 \text{ a.e. for } x \in U \text{ and all } y \in \mathbb{R}^d.$$

Finally $g = g(t, x)$ is a forcing term. This equation forms without doubt the simplest wave propagation model. We will consider time-harmonic solutions and forcings of the form

$$v(t, x) = e^{-ikt} u(x) \text{ and } g(t, x) = e^{-ikt} f(x)$$

where the *wave number* $k \neq 0$. Substitution gives rise to an elliptic PDE called the *Helmholtz equation*:

$$Lu = f \text{ in } U \tag{6}$$

for the *Helmholtz operator*

$$Lu := -\nabla \cdot A(x)\nabla u - k^2 u. \tag{7}$$

This equation is studied in a variety of domains including acoustics, seismology, electromagnetics and quantum mechanics. In literature the term Helmholtz equation is sometimes reserved for the special case $A = I$ and $f = 0$, which makes (6) actually the eigenproblem for the Laplacian:

$$-\Delta u - k^2 u = 0. \tag{8}$$

Note that any plane wave $e^{ik\hat{a}\cdot x}$ satisfies (8) when $\|\hat{a}\| = 1$. To ensure uniqueness on unbounded domains, one typically imposes the Sommerfeld radiation condition

$$\lim_{r \to \infty} r^{(d-1)/2}(u_r - iku) = 0, \tag{9}$$

which has the interpretation that waves should be "out-going." This interpretation is most pronounced in the representation formula of Appendix A.

Lemma 1. If u is a classical solution of (8) on $\mathbb{R}^d$ satisfying (9) and $\operatorname{Im} k \geq 0$, then $u = 0$.

The proof of this is in Appendix A. Lemma 1 shows that $-\Delta$ can only have eigenvalues with negative imaginary part when the radiation condition is imposed. Note that k^2 can therefore not always be interpreted as the "target" τ, and hence the situation is slightly different from eigenvalue problems when a radiation condition is imposed.

Scattering problems Of interest are so-called scattering problems, where we are given an incoming field $u^i(x)$ satisfying (6) on $U := \mathbb{R}^d \setminus \overline{D}$, with $D \subset \mathbb{R}^d$ an open, bounded and connected domain called the *scatterer* or obstacle. It has a boundary $\Gamma_D := \partial D$. Our goal is to find the scattered field $u^s(x)$ satisfying (6) such that the *total field*

$$u := u^i + u^s$$

satisfies the zero Dirichlet or *sound-soft scattering* problem

$$\begin{aligned} Lu &= f \text{ on } U, \\ u &= 0 \text{ on } \Gamma_D, \\ \lim_{r\to\infty} r^{(d-1)/2}(u^s_r - iku^s) &= 0 \end{aligned} \tag{10}$$

The trivial solution $u^s = -u^i$ is ruled out by the radiation condition; we do not assume u^i satisfies the radiation condition itself.

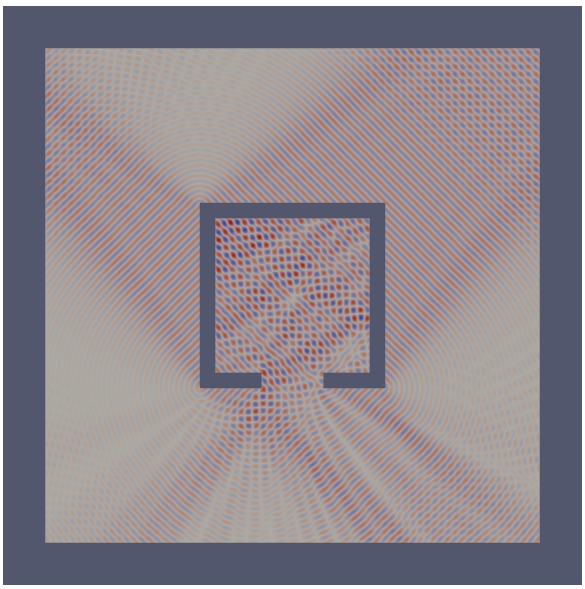
Truncated scattering problems The unboundedness of U is unattractive for direct discretizations, and therefore U is often truncated to finite size. Let $\Gamma_E = \partial U \setminus \Gamma_D$ denote the new (exterior) boundary that is introduced. We hope to impose a boundary condition on Γ_E that is satisfied by any u solving problem (10). However, in numerical methods we also want a condition that is *local* to ensure sparsity, and therefore it is popular to simply take a first-order approximation to the radiation condition (9). This leads to the *truncated sound-soft scattering problem*

$$\begin{aligned} Lu &= f \text{ on } U \\ u &= 0 \text{ on } \Gamma_D \\ \partial_n u - i\eta u &= \partial_n u^i - i\eta u^i \text{ on } \Gamma_E \end{aligned} \tag{11}$$

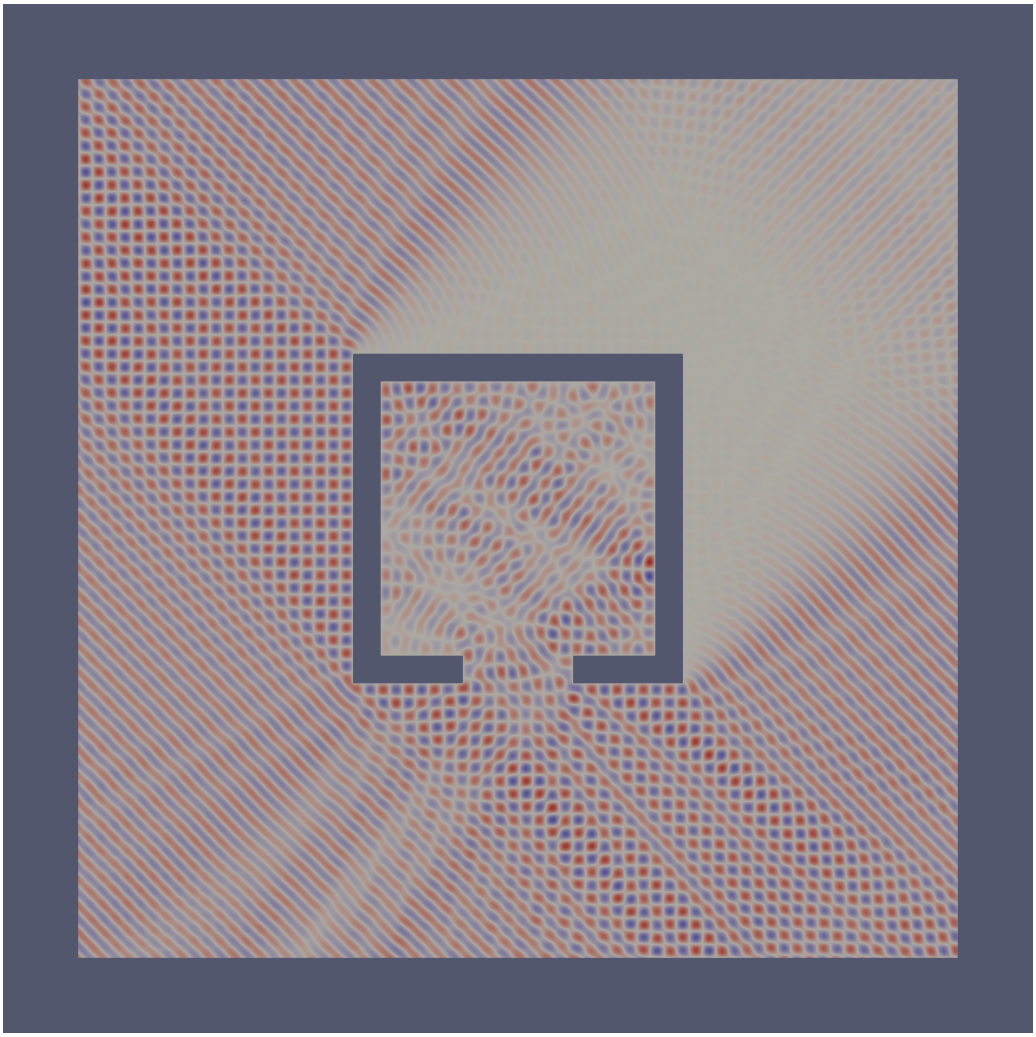
where $\eta \sim k$. If $\eta = k$, we see that outgoing waves with wave number k travelling in a direction perpendicular to the boundary are diminished. The solution u^s might however suffer from reflections when the outgoing waves do not make a sharp angle with the boundary. This might happen when $dist(\Gamma_D, \Gamma_E)$ is too small, or when the coefficients of L vary near Γ_E. Figure 1 shows an example of a scattering problem.

We will incidentally see a proof of uniqueness of the truncated scattering problem for star-shaped domains when $\operatorname{Im} k \geq 0$ is small enough (Lemma 6), and general domains with wave numbers $\operatorname{Im} k > 0$ (Lemma 5). However, for in-depth results on existence, uniqueness and regularity of elliptic PDEs on bounded and unbounded domains, we refer to the excellent reference [14]. We only remark that there exist so-called trapping domains, for which eigenvalues of L have nearly zero imaginary part.

(a) Incoming plane wave $u^i(x) = e^{ik\hat{\alpha}\cdot x}$



(b) Scattered field $u^s(x)$.



(c) Total field $u(x)$.

Figure 1: Example truncated scattering problem (11) with $L = -\Delta - k^2$, $f = 0$ and wave number $k = 300$ on $[0,1]^2$. Disretized with finite-volumes on a 2048×2048 grid. Solved with a direct method. Notice the unphysical reflections in the shadow region of the total field (due to the approximate Sommerfeld radiation condition).

1.3.1 (Lack of) quasi-optimality in FEM

We consider the weak formulation of the truncated scattering problem (11) in the Sobolev space

$$\hat{H} = \{v \in H^1(U) : v = 0 \text{ on } \Gamma_D\} \text{ with the norm } \|\cdot\|_{H^1(U)}.$$

For ease we take $\eta = k$ and write $g := \partial_n u^i - iku^i$. Via partial integration and substitution of the boundary conditions we obtain the sesquilinear form

$$B[u,v] = \int_U A\nabla u \cdot \nabla \overline{v}\, dx - k^2 \int_U u\overline{v}\, dx - ik \int_{\Gamma_E} u\overline{v}\, dS \tag{12}$$

and the linear functional $F \in \hat{H}'$:

$$F(v) = \int_U f\overline{v}\, dx + \int_{\Gamma_E} g\overline{v}\, dS.$$

Definition 3 (Weak formulation)**.** The weak formulation of the truncated scattering problem is to find $u \in \hat{H}$ such that

$$(Lu - f, v) = 0 \text{ for all } v \in \hat{H}.$$

This is equivalent to

$$B[u,v] = F(v) \text{ for all } v \in \hat{H}.$$

The finite-element method (FEM) weakens the problem of Definition 3 to the following.

Definition 4. Let $\mathcal{P} \subset \hat{H}$ be a finite-dimensional linear subspace. The FEM solution to (3) is the solution to the Galerkin problem:

$$\text{Find } u \in \mathcal{P} \text{ such that } (Lu - f, v) = 0 \text{ for all } v \in \mathcal{P}.$$

We briefly state the usual (complex variants of) tools of analysis for elliptic problems

Definition 5. For a Hilbert space H with norm $\|\cdot\|$, a sesquilinear form $B : H \times H \to \mathbb{C}$ is **continuous** when

$$|B[u,v]| \leq \alpha \|u\| \|v\|,$$

and **coercive** when

$$|B[u,u]| \geq \beta \|u\|^2,$$

for constants $\alpha > 0$ and $\beta > 0$.

Theorem 1 (Lax-Milgram)**.** If the sesquilinear form B on H is continuous and coercive, then for any bounded linear functional $F \in H'$ there exists a unique solution $u \in H$ to the problem

$$B[u, v] = F(v) \text{ for all } v \in H. \tag{13}$$

Lax-Milgram guarantees existence and uniqueness of finite-element problems to find $u \in \mathcal{P}$ such that

$$B[u, v] = F(v) \text{ for all } v \in \mathcal{P} \tag{14}$$

as well, since any finite-dimensional subspace $\mathcal{P} \subset H$ is a Hilbert space on its own equiped with the same inner product.

Definition 6. If, for any sesquilinear form B, problem (13) has a unique solution $u \in H$, and (14) has a unique FEM solution $\hat{u} \in \mathcal{P}$, then $\hat{u}$ is said to be **quasi-optimal** when

$$\|u - \hat{u}\| \leq C\|u - v\| \text{ for all } v \in \mathcal{P}$$

for a constant $C > 0$.

Quasi-optimality for a FEM solution is to say that it is only a constant away from the best approximation in the finite-element space. For truncated scattering problems we hope to find a constant C that is independent of the wave number k, so that FEM is robust.

Corollary 1 (Cea's lemma)**.** If the sesquilinear form B is coercive and continuous, then the unique FEM solution $\hat{u}$ to (14) is quasi-optimal with constant $C = \alpha/\beta$.

This is enough machinery to assess our truncated scattering problem. Indeed, it is not hard to see that we cannot guarantee coercivity of the bilinear form.

Lemma 2. The bilinear form (12) is not coercive uniformly in the wave number k, when k is large enough.

Proof. For the principle eigenvalue λ_1 of $-\nabla \cdot A\nabla u$ on U with $u = 0$ on ∂U it holds [9]

$$\lambda_1 = \min\{\int_U Au \cdot u \, dx : u \in H_0^1(U), \|u\|_{L^2(U)} = 1\}.$$

Since $H_0^1(U) \subset \hat{H}$, the minimizer u_1 is in $\hat{H}$ as well, and it satisfies

$$B[u_1, u_1] = \lambda_1 - k^2,$$

showing that B cannot be coercive on $\hat{H}$ when $k^2 = \lambda_1$. □

As a result of Lemma 2, Cea's lemma does not apply, and we cannot guarantee quasi-optimality of FEM. This is of course not to say we cannot prove quasi-optimality, but it typically requires us to use properties of the finite-element space or the domain itself. In what follows we will only consider h-FEM.

Intuitively one would expect that exact solution to the problem of Definition 3 can be represented with an error bounded independently from k when a constant number of grid points per wavelength is chosen, or equivalently, kh is small enough. The number of unknowns N in the discretization would then grow as $N \sim k^d$. This already seems demanding for large k, yet quasi-optimality of h-FEM has not been proven under this condition. The current best result on quasi-optimality with constant independent of k for h-FEM in dimension $d = 2, 3$ requires that hk^2 is small enough [15]. In that case we even have $N \sim k^{2d}$, but the estimate could be too pessimistic. Numerical results of [3] indicate that the L^2 error of the solution in 2D problems can be bounded independently of k if h^2k^3 is small enough. This would lead to $N \sim k^{2d/3}$ unknowns.

The stringent conditions for (or lack of) quasi-optimality in FEM for the Helmholtz equation is a phenonemon often referred to as the *pollution effect*. In what follows we will try to characterize it.

1.3.2 Direct discretization and the pollution effect in 1D

In one dimension scattering problems are trivial since any wave e^{ikx} satisfies the Sommerfeld radiation condition and hence the total field is identically zero. The one-dimensional case is however instructive and for ease of exposition we will therefore consider the problem:

$$-u_{xx} - k^2 u = f \text{ on } U = (0,1) \text{ with } u(0) = 0 \text{ and } u_x(1) = iku(1).$$

The bilinear form (12) and the functional now become

$$B[u,v] := \int_0^1 u_x \overline{v}_x - k^2 u\overline{v}\, dx - iku(1)\overline{v}(1) \text{ for } u, v \in \hat{H} \text{ and } F(v) := \int_0^1 f\overline{v}\, dx$$

We construct a uniform grid $x_j := jh$ where h is a constant mesh-width with hat-likebasis functions

$$\phi_j := \begin{cases} \frac{x - x_{j-1}}{h} & x_{j-1} \leq x \leq x_j \\ \frac{x_{j+1} - x}{h} & x_j < x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

The finite-element space $\mathcal{P} \subset \hat{H}$ is defined as $\mathcal{P} := span\{\phi_i\}_{i=1}^{n-1}$ where $h = 1/(n+1)$. The FEM problem is to find

$$\hat{u} := \sum_{j=1}^{n-1} \alpha_j \phi_j \in \mathcal{P}$$

such that $B[u_h, \phi_k] = F(\phi_k)$ for all $\phi_k \in \mathcal{P}$. This can be recasted into a system of equations

$$A\alpha = b \tag{15}$$

where $\alpha := (\alpha_1, \ldots, \alpha_{n-1})$, $A_{ij} := hB[\phi_j, \phi_i]$ and $b_i := hF(\phi_i)$. The elements of A can be found by simply working out the integrals. If we set $q := kh$, then we can write our matrix A as

$$A = \text{diag}\begin{bmatrix} r(q) & 2s(q) & r(q) \end{bmatrix} \text{ with the exception } A_{nn} = s(q) - iq$$

where $q := kh$ and

$$r(q) := -1 - \tfrac{1}{6}q^2 \text{ and } s(q) := 1 - \tfrac{1}{3}q^2.$$

We ask ourselves the question: which *discrete* fundamental solutions exist to problem (15)? To work this out, we plug in a Fourier mode $e^{ik'x}$ with a *discrete* wave number k', that must satisfy the homogeneous free-space Helmholtz problem. Note that $u_h(x_j) = \alpha_j$, so we set $\alpha_j = e^{ik'jh}$ to obtain the equation

$$r(q)e^{ik'(j-1)h} + 2s(q)e^{ik'jh} + r(q)e^{ik'(j+1)h} = 0,$$

which is equivalent to

$$e^{ik'h} = -\frac{s(q)}{r(q)} \pm \sqrt{\frac{s^2(q)}{r^2(q)} - 1} \tag{16}$$

Now if $\left|\frac{s(q)}{r(q)}\right| < 1$, which happens when $q \in (0, \sqrt{12})$, then the solutions of (16) form a complex conjugate. Considering only the real part in that case gives us

$$\cos k'h = -\frac{s(q)}{r(q)} \text{ or } k'h = \arccos\left(-\frac{s(q)}{r(q)}\right).$$

Using the Taylor expansion of the arccos we find eventually

$$k' = k - \frac{k^3h^2}{24} + O(k^5h^4).$$

What we see is that *under the assumption* that kh is small enough, the fundamental solutions to the Helmholtz problem are in fact waves that travel slightly *slower*

compared to the continuous fundamental solution. The discrete waves develop a phase lag. So what we are looking at is the pre-asymptotic behaviour of the FEM solution, which can be expected to lag in phase with respect to the exact solution.

It allows us to conclude that for large k we can expect the phase lag to be worse for fixed mesh-widths, as the term k^3h^2 might dominate. Indeed in [13] it was shown that the relative error of $\hat{u}$ in the semi-norm $\|\partial_x \cdot\|_{L^2(U)}$ can be estimated by $C_1kh + C_2k^3h^2$ for constants $C_1, C_2 > 0$ independent from k and h. The former term is due to the local approximation error, while the latter term is a global pollution error.

1.3.3 Ritz values of self-adjoint elliptic operators

So far we have seen a quantitative analysis of the pollution effect, which shows for a specific finite element space the phase lag in terms of n and k. Here we will revisit the pollution effect qualitatively in terms of the approximate eigenvalues and eigenvectors of a self-adjoint elliptic operator acting on a finite element space.

We consider the situation where our Helmholtz operator L is self-adjoint, which is for instance the case with Dirichlet zero boundary conditions. Hence we assume L acts on $H_0^1(U)$ for a bounded domain $U \subset \mathbb{R}^d$. This section is based on the proof of Theorem 6.5.2 in [9], but we generalize to non-coercive operators L.

Definition 7. Let (λ_i, w_i) denote the eigenpairs of L satisfying

$$\begin{aligned} Lw_i &= \lambda_i w_i \text{ in } U, \\ w_i &= 0 \text{ on } \partial U. \end{aligned} \tag{17}$$

in the variational sense in $H_0^1(U)$.

Definition 8. Let $\mathcal{P} \subset H_0^1(U)$ be an n-dimensional, linear subspace. Then the pair (θ, v) is called a *Ritz pair of L with respect to $\mathcal{P}$* if $v \in \mathcal{P}$ such that

$$(Lv - \theta v, w) = 0 \text{ for all } w \in \mathcal{P}.$$

We will always assume $\|v\|_{L^2(U)} = 1$ and denote the set of all Ritz pairs of as (θ_i, v_i) for $i = 1, \ldots, n$.

Theorem 2. If (θ, v) is a Ritz pair of L, then

1. the Ritz value is the Rayleigh quotient $\theta = (Lv, v)/(v, v)$;
2. the Ritz value is a convex combination of eigenvalues of L. In particular

$$\theta = \sum_{i=1}^{\infty} (u, w_i)^2 \lambda_i.$$

The first statement follows immediately from Definition 8, but the second statement requires more care. To prove it we first study the shifted operator

$$L_k := L + k^2 I$$

making use of the fact that its bilinear form defines an inner product. The standard bilinear forms of L and L_k are respectively

$$B[u,v] := \int_U A\nabla u \cdot \nabla v - k^2 uv \, dx \text{ and } B_k[u,v] := \int_U A\nabla u \cdot \nabla v \, dx$$

where $u, v \in H_0^1(U)$. In particular B_k is coercive, since:

$$B_k[u,u] = (A\nabla u, \nabla u) \geq \gamma \|\nabla u\|_{L^2}^2 \gtrsim \|u\|_{H_0^1}^2.$$

The last inequality follows from the Poincaré inequality. This means B_k is an inner product for $H_0^1(U)$. Consider the eigenvalue problems for the shifted operator

$$\begin{aligned} L_k w_i &= \vartheta_i^2 w_i \text{ in } U \\ w_i &= 0 \text{ on } \partial U \end{aligned} \tag{18}$$

Obviously (ϑ_i^2, w_i) solves (18) if and only if $(\vartheta_i^2 - k^2, w_i)$ solves (17). Without loss assume $\|w_i\|_{L^2(U)} = 1$. We use that the spectrum of L_k is discrete,

$$0 < \vartheta_1 < \vartheta_2 \leq \vartheta_3 \leq \dots,$$

and $\{w_i\}_1^\infty$ forms an orthonormal basis for $L^2(U)$ [9]. Hence for any $u \in H_0^1(U)$ with $\|u\|_{L^2(U)} = 1$ we can write

$$u = \sum_{i=1}^{\infty} d_i w_i \text{ in } L^2(U) \tag{19}$$

where $d_i := (u, w_i)$. Furthermore

$$\sum_{i=1}^{\infty} d_i^2 = \|u\|_{L^2(U)}^2 = 1.$$

Lemma 3. The series (19) converges as well in $H_0^1(U)$ equiped with the inner product $B_k[u,v]$.

Proof. We claim $\{\frac{w_i}{\vartheta_i}\}_1^\infty$ is an orthonormal basis for $H_0^1(U)$ with this new inner product. Indeed

$$B_k[\tfrac{w_i}{\vartheta_i}, \tfrac{w_i}{\vartheta_i}] = \tfrac{1}{\vartheta_i^2}(L_k w_i, w_i) = (w_i, w_i) = 1$$

and

$$B_k[w_i, w_j] = (L_k w_i, w_j) = \vartheta_i^2 (w_i, w_j) = 0$$

show that the elements of $\{\frac{w_i}{\vartheta_i}\}_1^\infty$ are orthonormal. To show they form a basis it's enough to verify that if $u \in H_0^1(U)$ and

$$B_k[w_i, u] = 0 \text{ for all } i = 1, 2, \ldots$$

then $u = 0$. But clearly

$$0 = B_k[w_i, u] = (L_k w_i, u) = \vartheta_i^2 (w_i, u)$$

implies $u = 0$, since $\{w_i\}_1^\infty$ is a basis for $L^2(U)$. Hence we have

$$u = \sum_{i=1}^{\infty} B_k[u, \tfrac{w_i}{\vartheta_i}] \tfrac{w_i}{\vartheta_i} \text{ in } H_0^1(U)$$

Finally, computing (u, w_j) for any w_j then gives that $B_k[u, \frac{w_j}{\vartheta_j}] = \vartheta_j d_j$. So the series (19) converges in $H_0^1(U)$ as well. □

Proof of Theorem 2. Take a Ritz pair (θ, u) and assume $(u, u) = 1$. Then

$$\theta + k^2 = (Lu, u) + k^2 (u, u) = B_k[u, u].$$

Now we apply Lemma 3 and write

$$u = \sum_{i=1}^{\infty} d_i w_i \text{ in } H_0^1(U)$$

with $d_i = (u, w_i)$. Hence

$$\theta + k^2 = B_k[u, u] = \sum_{i=1}^{\infty} d_i^2 \vartheta_i^2 = \sum_{i=1}^{\infty} d_i^2 (\lambda_i + k^2) = \sum_{i=1}^{\infty} d_i^2 \lambda_i + k^2$$

This shows that

$$\theta = \sum_{i=1}^{\infty} (u, w_i)^2 \lambda_i,$$

proving the second statement of Theorem 2. □

1.3.4 Ritz values and the pollution effect

We will now apply the theory of Theorem 2 to a FEM problem. Suppose a non-trivial $f \in L^2(U)$ is given and we tackle the problem

$$\begin{aligned} Lu &= f \text{ in } U; \\ u &= 0 \text{ on } \partial U, \end{aligned}$$

using the finite element space is $\mathcal{P}$. This means we have to find $\hat{u} \in \mathcal{P}$ such that

$$(L\hat{u}, v) = (f, v) \text{ for all } v \in \mathcal{P}.$$

We can explicitly write the solution (if it exists) in terms of the Ritz pairs of L with respect to $\mathcal{P}$, since the Ritz functions $\{v_i\}_1^n$ form an orthonormal basis for $\mathcal{P}$ in the $L^2(U)$ norm. The solution reads

$$\hat{u} = \sum_{i=1}^{n} \tfrac{1}{\theta_i}(f, v_i)v_i$$

The main insight is that Ritz functions corresponding to Ritz values close to the origin contribute strongly to the finite element solution. Whenever a Ritz function v_i approximates an eigenfunction w_j relatively well, while its Ritz value θ_i does *not* approximate the eigenvalue λ_j well, then the FEM solution can be polluted. In particular so when θ_i is close to 0 while λ_j is not.

The main question then is: when does a Ritz value θ_i being close to an eigenvalue λ_j imply that the Ritz function v_i is a good approximation of the eigenfunction w_j? Theorem 2 tells us that for the principle eigenvalue λ_1 it must hold by convexity of the Ritz values that whenever $\theta_1 \approx \lambda_1$, then $(v_1, w_1) \approx 1$. So in that case a good approximation of the eigenvalue amounts to the Ritz function being a good approximate eigenfunction. This argument can be repeated: if $\theta_2 \approx \lambda_2$, then it must be so that $(v_2, w_2) \approx 1$, since v_1 and v_2 are orthogonal.

In practice however, we do not know whether the first Ritz values are close to the first eigenvalues. More importantly, the preceding argument piles approximation upon approximation and therefore loses validity exactly for eigenvalues λ_n with n large — the interior eigenvalues. Hence, for high wave numbers k, we might expect the Ritz values near the origin not to approximate corresponding eigenvalues well, causing pollution.

1.4 Summary

Linear systems involving indefinite and nearly singular matrices occur naturally in interior eigenvalue problems. They are troublesome for Krylov subspace methods: in the normal case we show that problematic eigenspaces enter the Krylov subspace

only after many iterations, and components of the error in these directions produce small residuals.

Convergence can be improved by preconditioning and deflation, or a combination of both. Deflation requires availability of approximations to problematic eigenspaces, yet the sole goal of eigenproblem solvers is to obtain these as well. Hence, deflation can only be successful when cheap, heuristic approximations of problematic eigenspaces can be formed — an idea pursued in Chapter 3.

At the continuous level we study the Helmholtz operator, as standard discretizations of it lead exactly to these indefinite matrices. We see that the infinite-dimensional analog of indefiniteness is lack of coercivity of the sesquilinear form. Standard analysis tools do not apply in this case, and we cannot immediately conclude that the approximate h-FEM solution is near the best approximation from the search space. The intuitive idea that hk should be small enough to obtain good h-FEM solutions proves false, a phenomenon known as the pollution effect. We characterize this behaviour in the self-adjoint case in terms of approximate eigenvalues (Ritz values). If the Ritz values do not approximate problematic eigenvalues well, then the FEM solution cannot be accurate.

2 Numerical methods from literature

In what follows we will look at a subset of the vast amount of literature surrounding numerical methods for the Helmholtz equation. What we will see however, is that many methods rely on assumptions we do not encounter in general eigenvalue problems.

2.1 Multigrid

One of the main conclusions of Section 1.1 is that Krylov subspace methods applied to (1) take many iterations to reduce the components of the error in the direction of problematic eigenvectors. The short explanation is that these components produce small contributions to the residual.

This difficulty is not unique to the indefinite Helmholtz equation, as the same problem shows up in direct discretizations of self-adjoint diffusion operators (the case $k = 0$). In this case the Conjugate Gradients method applies, which is a Krylov subspace method that minimizes the error itself in the norm induced by the matrix. However, this norm is skew with eigenvalues serving as weights for the components of the error in the direction eigenvectors. The same problem occurs, as the method will not immediately reduce the error in the direction of problematic eigenvectors in the Euclidean norm.

A popular solution to this problem in the *definite* or coercive case $k = 0$ is to apply geometric multigrid. In this case, problematic eigenfunctions manifest as "low-frequency" or slowly oscillating components on the geometric grid. Hence, the error components that are not reduced quickly enough by the iterative solver are in fact well represented on a coarse grid. Since a coarse grid reduces the dimensionality of the problem, the computational work is reduced as well. The V-cycle of restricting the error equation to the coarse grid, solving it there, and interpolating back to the fine grid lies at the core of the geometric multgrid method. For convergence proofs of the case $k = 0$ we refer to [5]. Here we describe the V-cyle simply as follows:

Pre-smoothing. A (few iterations of a) Krylov subspace method gives us an approximate solution $\hat{u} \in \mathcal{P}_1 \subset H^1(U)$ to the Galerkin problem

$$\text{Find } u \in \mathcal{V}_i \text{ such that } (Lu - f, v) = 0 \text{ for all } v \in \mathcal{P}_1.$$

The error $e := u - \hat{u} \in \mathcal{P}_1$ and the residual $r := f - L\hat{u}$ satisfy the Galerkin problem

$$(Le - r, v) = 0 \text{ for all } v \in \mathcal{P}_1. \tag{20}$$

Coarsening. Problem (20) is solved *approximately* for e, by restricting it on a coarser grid $\mathcal{P}_2 \subset \mathcal{P}_1$:

$$\text{Find } \hat{e} \in \mathcal{P}_2 \text{ such that } (L\hat{e} - r, v) = 0 \text{ for all } v \in \mathcal{P}_2. \tag{21}$$

In practice, basis functions for the finite element subspace $\mathcal{P}_2$ are formed sparsely from basis functions of $\mathcal{P}_1$, as is shown in Figure 2 for one-dimensional piece-wise linear basis functions.

Interpolation and post-smoothing. The approximate error $\hat{e}$ is lifted back to $\mathcal{P}_1$, and the solution is updated as $\hat{u} \leftarrow \hat{u} + \hat{e}$. The Krylov subspace method is run again with the updated solution as initial guess.

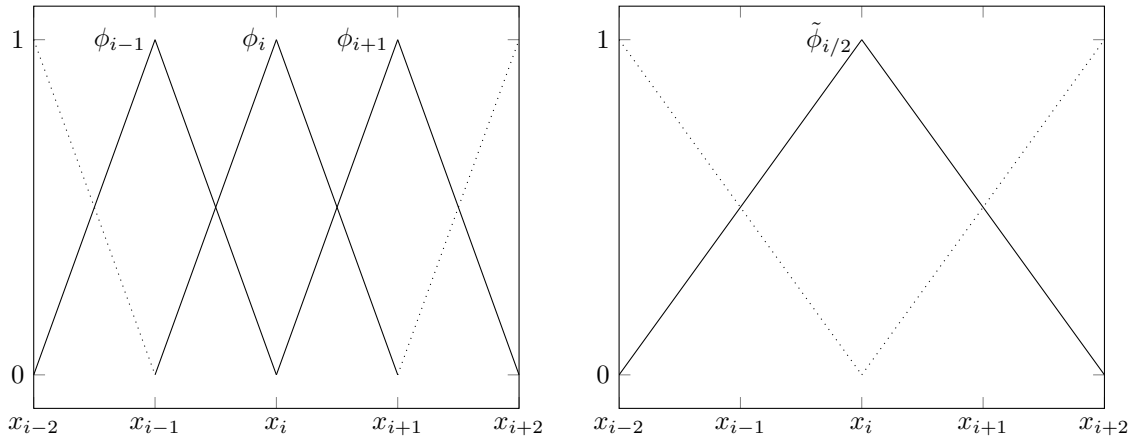


Figure 2: Three fine grid basis functions (left) are combined to a single coarse grid basis function (right) of twice the mesh-width: $\hat{\phi}_{i/2} = \frac{1}{2}(\phi_{i-1} + 2\phi_i + \phi_{i+1})$.

However, the geometric multigrid method as is *will not work* for the Helmholtz problem with high wave numbers for two reasons.

Lack of quasi-optimality The coarse grid problem (21) is susceptible to the pollution effect. The exact solution to (21) will contain large errors precisely in the directions of functions it was meant to capture: the problematic eigenfunctions.

Approximation error If the grid is too coarse in the sense that kh is too large, the error $\|e - \hat{e}\|_{H^1}$ is large for any $\hat{e} \in \mathcal{P}_2$, as the oscillations cannot be represented.

It is however worth pointing out that the first problem precedes the second: multigrid can fail even when there is an $\hat{e} \in \mathcal{P}_2$ that has a small approximation

error $\|e - \hat{e}\|_{H^1}$. To put it differently: the problem is initially only the Galerkin formulation of (21), not the quality of the search space $\mathcal{P}_2$.

In light of this, some authors have suggested to modify the differential operator L in problem (21): we retain the Galerkin formulation, yet replace the wave number k by a suitable discrete wave number $\hat{k}$ as in Section 1.3.2, so that the Ritz values match the eigenvalues better. This is indeed an attempt to avoid the pollution effect altogether. However, in [1] it was proven that the pollution effect can be minimized this way, but not diminished in two dimensions and higher.

Petrov-Galerkin condition. Another idea that comes to mind is to replace the Galerkin condition of (21) with a Petrov-Galerkin condition, which is often done in eigenvalue problems. To keep computational costs and memory usage fixed, the Arnoldi and Jacobi-Davidson method incorporate restarts, in which they shrink the dimension of their search space and retain only the current best approximate eigenfunctions [2]. "Current best" can be defined as those Ritz functions v_i that have Ritz values θ_i close to the target τ. However, this criterion is flawed by the pollution effect.

Rather, the standard Galerkin projection is rejected in favour of the least-squares formulation, leading to *harmonic Ritz values and functions*. To be brief, we present the theory in finite dimensions for normal matrices A.

Definition 9. The pair (θ, u) is a *harmonic Ritz pair of* $A \in \mathbb{C}^{n\times n}$ *with respect to* $\mathcal{P} \subset \mathbb{C}^n$ if $Au - \theta u \perp A\mathcal{P}$.

Lemma 4. The pair (θ, u) is a harmonic Ritz pair of A with respect to $\mathcal{P}$ if and only if (θ^{-1}, v) is a Ritz pair of A^{-1} with respect to $A\mathcal{P}$ where $v = Au$.

Proof.

$$A^{-1}v - \theta^{-1}v \perp A\mathcal{P} \iff Au - \theta u \perp A\mathcal{P}.$$

□

Corollary 2. Harmonic Ritz values θ of A with respect to $\mathcal{P}$ are a weighted harmonic mean of the eigenvalues of A.

Proof. By Theorem 2, any Ritz value θ^{-1} of A^{-1} is a convex combination eigenvalues of A^{-1}. The eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A. By Lemma 4 it follows that θ is a harmonic Ritz value, and hence a harmonic mean of eigenvalues of A. □

However, replacing the Galerkin projection (21) with the least-squares formulation does not help us. Consider the extreme case where a Ritz value is shifted such that it is identically zero. Its Ritz function does not contribute to the residual, and the least-squares formulation does therefore not contain this direction. The corresponding harmonic Ritz value is "at infinity".

2.2 Shifted Laplacian preconditioner in the interior domain

For moderate wave numbers k there has been some success with the so-called Shifted Laplacian Preconditioner in truncated scattering problems [8]. The idea is to precondition a Helmholtz problem with a Helmholtz operator with a different wave number. There is some indirection here: the shifted operator is chosen at the continuous equations and only then discretized. The idea is that with an appropriate shift, the preconditioner can be (approximately) applied with efficient methods that are not feasible for the Helmholtz operator itself. We will see an instance of such a method in Section 2.3.

In fact we have already seen a shifted operator in Section 1.3.3, namely L_k. This operator serves well to explain the concept. Let us denote

$$u = L_k^{-1} g \text{ whenever } B_k[u, v] = (g, v) \text{ for all } v \in H_0^1(U).$$

Now if (λ_i, w_i) with $w_i \in H_0^1(U)$ is a weak solution to the eigenvalue problem $Lu = \lambda u$, then so is $(\lambda_i + k^2, w_i)$ a weak solution to $L_k u = \lambda u$. Therefore

$$L_k^{-1} L w_i = \frac{\lambda_i}{\lambda_i + k^2} w_i.$$

Since $\lambda_i \to \infty$ as $i \to \infty$, we see that the eigenvalues of our preconditioned operator $L_k^{-1} L$ can only accumulate at 1. This means that for fixed k, we might expect grid-independent convergence of iterative methods. The drawback of this approach is clear as well: for large k, problematic eigenvalues get mapped even closer to the origin, and the quality of the preconditioner is questionable.

However, the k-dependence of the preconditioner might be fixed if we "shift" the wave number to be complex-valued with positive imaginary part. This is equivalent to adding damping to the problem as was noted in Appendix A. We will analyze this idea following the lines of [10].

Define

$$k_\delta := k + i\delta \text{ and } L_\delta := -\Delta - k_\delta^2$$

with $k > 0$ and $\delta \geq 0$ and consider the problem

$$\begin{aligned} L_\delta u &= f \text{ in } U \\ \partial_n u - iku &= g \text{ on } \partial U \end{aligned} \tag{22}$$

With $\delta = 0$ we get the original Helmholtz equation with approximate Sommerfeld boundary conditions. The associated sesquilinear form to (22)

$$B_\delta[u, v] := \int_U \nabla u \cdot \nabla \overline{v} \, dx - k_\delta^2 \int_U u \overline{v} \, dx - ik \int_{\partial U} u \overline{v} \, dS$$

together with the linear functional

$$F(v) := \int_U f\overline{v}\,dx + \int_{\partial U} g\overline{v}\,dS$$

defines the variational problem to find $u \in H^1(U)$ such that

$$B_\delta[u,v] = F(v) \text{ for all } v \in H^1(U).$$

Here we assume that $f \in L^2(U)$ and $g \in L^2(\partial U)$ so that F is bounded. Let $\mathcal{P}$ denote an n-dimensional linear subspace of H^1 with basis elements $\{\phi_i\}_1^n$. The finite-element problem then comes down to finding $u = \sum_{i=1}^n x_i\phi_i \in \mathcal{P}$ such that

$$A_\delta x = b \text{ where } A_\delta := S - k_\delta^2 M - ikN \in \mathbb{C}^{n\times n}$$

with elements

$$S_{ij} := \int_U \nabla\phi_i \cdot \nabla\overline{\phi_j}\,dx, \quad M_{ij} := \int_U \phi_i\overline{\phi_j}\,dx,$$
$$N_{ij} := \int_{\partial U} \phi_i\overline{\phi_j}\,dS, \qquad b_i := F(\phi_i).$$

In what follows we discuss how the problem $A_0x = b$, can be left-preconditioned as $A_\delta^{-1}A_0x = A_\delta^{-1}b$ for some choice of δ such that $\|I - A_\delta^{-1}A_0\|_2$ is small and independent of k. We assume the cost of applying A_δ^{-1} is smaller when δ is large enough. Note that

$$I - A_\delta^{-1}A_0 = A_\delta^{-1}(A_\delta - A_0) = (\delta^2 - 2k\delta i)A_\delta^{-1}M, \tag{23}$$

so we just have to estimate $\|A_\delta^{-1}M\|_2$. The way to do so is to come up with a variational problem that discretizes to $A_\delta x = My$. Then we use quasi-optimality of the shifted operator to relate the FEM solution to the continuous one. Finally we need estimates on the continuous solution operator. For the latter, good estimates exploit properties of the domain.

Definition 10. Let $U \subset \mathbb{R}^d$ be a bounded and connected domain. Then U is said to be **star-shaped** with respect to the origin when for a given $c > 0$,

$$x \cdot n \geq c \tag{24}$$

for almost all $x \in \partial U$.

2.2.1 Boundedness of the solution operator

Lemma 5 (General domain)**.** If $u \in C^2(\overline{U})$ satisfies (22), ∂U is C^1 and $\delta > 0$ then

$$\|u\|_{k,U}^2 \lesssim \left(\frac{1}{\delta^2} + \frac{1}{k\delta} + \frac{1}{k^2}\right)\|f\|_{L^2(U)}^2 + \left(\frac{1}{\delta} + \frac{1}{k}\right)\|g\|_{L^2(\partial U)}^2$$

where the hidden constants do not depend on k.

Lemma 6 (Star-shaped)**.** If $u \in C^2(\overline{U})$ satisfies (22), ∂U is C^1, U is star-shaped with respect to the origin such that (24) holds, then for small enough $\delta > 0$

$$\|u\|_{k,U}^2 \lesssim \left(1 + \frac{1}{k^2} + \frac{1}{k^4}\right) \|f\|_{L^2(U)}^2 + \left(1 + \frac{1}{k^2}\right) \|g\|_{L^2(U)}^2$$

where the hidden constants do not depend on k. In particular, this estimate holds for $\delta = 0$ as well.

The proofs of these Lemma's are delegated to Appendix B.

2.2.2 Continuity and coercivity of B_δ when $\delta > 0$

Our shifted operator does satisfy the Lax-Milgram conditions as we will see shortly. We are interested what the continuity and coercivity constants are in terms of k and δ, so get a quasi-optimality estimate from Cea's lemma.

Continuity Using the Cauchy inequality we get

$$\begin{aligned} |B_\delta[u,v]| &\leq \|\nabla u\|_{L^2(U)} \|\nabla v\|_{L^2(U)} + (k^2 + \delta^2) \|u\|_{L^2(U)} \|v\|_{L^2(U)} + J \\ &\leq \frac{k^2 + \delta^2}{k^2} \left[\|\nabla u\|_{L^2(U)} \|\nabla v\|_{L^2(U)} + k^2 \|u\|_{L^2(U)} \|v\|_{L^2(U)} \right] + J \end{aligned}$$

where $J := k\|u\|_{L^2(\partial U)} \|v\|_{L^2(\partial U)}$. For positive a, b, c, d we use the inequality

$$(ac + bd)^2 \leq (a^2 + b^2)(c^2 + d^2)$$

with $a = k\|u\|$, $b = \|\nabla u\|$, $c = k\|v\|$, $d = \|\nabla v\|$ so that

$$|B_\delta[u,v]| \leq \frac{k^2 + \delta^2}{k^2} \|u\|_{k,U} \|v\|_{k,U} + J.$$

To estimate the J term, we employ the trace theorem [12]

$$\|u\|_{L^2(\partial U)}^2 \leq C \|u\|_{L^2(U)} \|u\|_{H^1(U)}$$

for a constant C depending only on U. Therefore

$$J \lesssim k \left(\|u\|_{L^2(U)} \|u\|_{H^1(U)} \|v\|_{L^2(U)} \|v\|_{H^1(U)} \right)^{1/2}.$$

Using the Cauchy inequality with $\varepsilon > 0$ we get

$$J \lesssim k \left(\frac{\varepsilon}{2} \|u\|_{L^2(U)} \|v\|_{L^2(U)} + \frac{1}{2\varepsilon} \|u\|_{H^1(U)} \|v\|_{H^1(U)} \right)$$

Take $\varepsilon = k$ to obtain the estimate

$$J \lesssim \left(\|\nabla u\|^2_{L^2(U)} + (1+k^2)\|u\|^2_{L^2(U)}\right)^{1/2} \left(\|\nabla v\|^2_{L^2(U)} + (1+k^2)\|v\|^2_{L^2(U)}\right)^{1/2}$$

Hence

$$J \lesssim \frac{1+k^2}{k^2}\|u\|_{k,U}\|v\|_{k,U}.$$

Therefore

$$|B_\delta[u,v]| \lesssim \alpha\|u\|_{k,U}\|v\|_{k,U} \text{ where } \alpha = \left(\frac{k^2+\delta^2}{k^2} + \frac{1+k^2}{k^2}\right).$$

Coercivity Showing coercivity of B_δ requires the same trick with the complex part of the wave number as employed in Lemma 12 of Appendix A:

$$B[v, k_\delta v] = \overline{k_\delta}\|\nabla v\|^2_{L^2(U)} - k_\delta|k_\delta|^2\|v\|^2_{L^2(U)} - ik\overline{k_\delta}\|u\|^2_{\partial U}.$$

The imaginary part of this expression is now sign-definite:

$$-\operatorname{Im}(B[v, k_\delta v]) = \delta\left(\|\nabla v\|^2_{L^2(U)} + |k_\delta|^2\|v\|^2_{L^2(U)}\right) + k^2\|u\|^2_{\partial U}.$$

Hence

$$\begin{aligned} |B_\delta[v,v]| = \tfrac{1}{|k_\delta|}|B[v,k_\delta v]| &\geq \tfrac{1}{|k_\delta|}(-\operatorname{Im}(B[v,k_\delta v])) \\ &\geq \frac{\delta}{|k_\delta|}\left(\|\nabla v\|^2_{L^2(U)} + |k_\delta|^2\|v\|^2_{L^2(U)}\right) \\ &\geq \beta\|v\|^2_{k,U} \end{aligned}$$

where

$$\beta = \frac{\delta}{\sqrt{k^2+\delta^2}}.$$

This shows coerciveness of B_δ when $\delta \neq 0$.

Quasi-optimality Since B_δ satisfies the conditions of Lax-Milgram (Theorem 1), Cea's lemma (Corollary 1) applies, and we obtain a quasi-optimality constant $C = \alpha/\beta$.

2.2.3 Boundedness of $\|I - A_\delta^{-1}A_0\|_2$

For any given $\tilde{y} \in \mathbb{C}^n$, we must construct a variational problem involving B_δ that results in a discretization $A_\delta\tilde{x} = M\tilde{y}$ for $\tilde{x} \in \mathbb{C}^n$. That way we can use our previous estimates to obtain a bound on $\|A_\delta^{-1}M\|_2$. Let

$$\tilde{f} := \sum_{i=1}^{n} \tilde{y}_i\phi_i$$

and define the variational problem to find $\tilde{u} \in H^1(U)$ such that

$$B_\delta[\tilde{u}, \tilde{v}] = \int_U \tilde{f}\overline{\tilde{v}}\, dx \text{ for all } \tilde{v} \in H^1(U).$$

Since $\tilde{f} \in H^1(U)$ by construction, the right-hand side defines a bounded, linear functional in $\tilde{v}$. Let

$$\tilde{u}_n := \sum_{i=1}^{n} \tilde{x}_i \phi_i \in \mathcal{P} \text{ for } \tilde{x} \in \mathbb{C}^n$$

be the FEM approximation to the variational problem:

$$B_\delta[\tilde{u}_n, \tilde{v}] = \int_U \tilde{f}\overline{\tilde{v}}\, dx \text{ for all } \tilde{v} \in \mathcal{P}.$$

Expanding the definitions of $\tilde{u}_n$ and $\tilde{f}$ shows this is indeed equivalent to

$$A_\delta \tilde{x} = M\tilde{y}.$$

We assume the mesh is such that $\|\tilde{u}_n\|^2_{L^2(U)} \sim h^d \|\tilde{x}\|_2^2$ where h is maximum mesh width and d the dimension. First note

$$k^2 h^d \|\tilde{x}\|_2^2 \lesssim k^2 \|\tilde{u}_n\|^2_{L^2(U)} \leq \|\tilde{u}_n\|^2_{k,U}$$

so that $kh^{d/2}\|\tilde{x}\|_2 \lesssim \|\tilde{u}_n\|_{k,U}$. Next, by quasi-optimality, we get

$$\|\tilde{u}_n\|_{k,U} \leq \|\tilde{u}_n - \tilde{u}\|_{k,U} + \|\tilde{u}\|_{k,U} \leq (1 + \alpha/\beta)\, \|\tilde{u}\|_{k,U}.$$

Depending on the domain we consider, we get a constant C_{sol} either from Lemma 5 or from Lemma 6 such that

$$\|\tilde{u}\|_{k,U} \leq C_{sol} \|\tilde{f}\|_{L^2(U)}.$$

Lastly, since $\|\tilde{f}\|_{L^2(U)} \sim h^{d/2}\|\tilde{y}\|_2$ as well, we get

$$\|A_\delta^{-1} M \tilde{y}\|_2 = \|\tilde{x}\|_2 \lesssim k^{-1}\, (1 + \alpha/\beta)\, C_{sol} \|\tilde{y}\|_2. \tag{25}$$

Lemma 7. It holds that

$$\|I - A_\delta^{-1} A_0\|_2 \lesssim k^{-1}(\delta^2 + k\delta)\, (1 + \alpha/\beta)\, C_{sol}$$

Proof. Follows from (23) combined with (25). □

From Lemma 7 it follows[1] that whenever $\delta \sim k$, then on general domains $\|I - A_\delta^{-1} A_0\|_2 \lesssim 1 + k^{-2}$.

[1]This seems to be erroneous, although we cannot point the finger at the mistake.

2.3 Domain decomposition techniques

Domain decomposition can be seen as a divide and conquer technique to reduce a large problem into smaller ones that can be solved independently, reducing complexity and allowing parallellism.

However, standard domain decomposition with Dirichlet or Neumann conditions on the interfaces will fail for the Helmholtz equation, since k^2 can correspond with a Neumann or Dirichlet eigenvalue of $-\nabla \cdot A\nabla$ on a subdomain. In literature we see attempts to fix this by means of an approximate Sommerfeld radiation condition on the interfaces, such that each subproblem is a well-posed truncated scattering problem on its own. This idea was first pursued in [4], and we follow its core idea. As we will see, convergence relies on the presence of an "energy sink", in the form of damping ($\operatorname{Im} k > 0$) or an approximate Sommerfeld boundary condition on Γ_E. If the Helmholtz problem does not satisfy these conditions, we can still employ domain decomposition as an implementation of the Shifted Laplacian Preconditioner of Section 2.2.

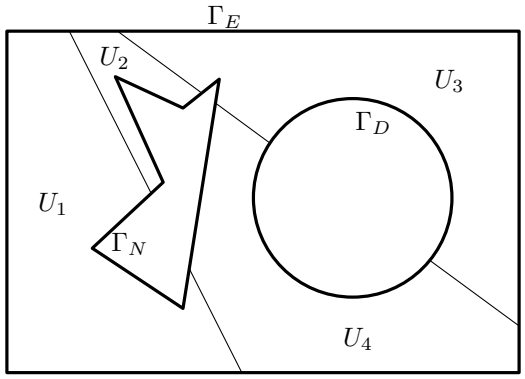


Figure 3: Problem (26) *with four subdomains.*

Consider the problem

$$\begin{aligned} -\Delta u - k^2 u &= f \text{ in } U, \\ u &= -u^{inc} \text{ on } \Gamma_D, \\ \partial_n u &= -\partial_n u^{inc} \text{ on } \Gamma_N, \\ \partial_n u - iku &= 0 \text{ on } \Gamma_E. \end{aligned} \tag{26}$$

Partition $U = \sum_{j=1}^{m} U_j$ in m non-overlapping subdomains U_j with interfaces $\Gamma_{ij} := \overline{U_i} \cap \overline{U_j}$ as depicted in Figure 3. Then we define the iterative process for each

$n = 0, 1, \ldots$ in each subdomain

$$\begin{aligned} -\Delta u_i^{n+1} - k^2 u_i^{n+1} &= f \text{ in } U_i, \\ \partial_n u_i^{n+1} - iku_i^{n+1} &= -\partial_n u_j^n - iku_j^n \text{ on all } \Gamma_{ij} \\ u_i^{n+1} &= -u^{inc} \text{ on } \Gamma_D, \\ \partial_n u_i^{n+1} &= -\partial_n u^{inc} \text{ on } \Gamma_N, \\ \partial_n u_i^{n+1} - iku_i^{n+1} &= 0 \text{ on } \Gamma_E. \end{aligned} \tag{27}$$

We interpret $\partial_n u_i^{n+1}$ as being the normal derivative pointing outward from U_i, while $\partial_n u_j^n$ is the normal derivative pointing outward from U_j. We define the error in each subdomain $e_i^n := u_i^n - u$ with u restricted to U_i. By linearity the error e_i^n satisfies (27) with $f = u^{inc} = 0$, so we restrict ourself to the homogeneous problem.

Lemma 8. The error e_i^n in (27) satisfies

$$\sum_{j\neq i} \int_{\Gamma_{ij}} |\partial_n e_i^n|^2 + |k|^2 |e_i^n|^2 \, dS = \sum_{j\neq i} \left(\int_{\Gamma_{ij}} |\partial_n e_i^n \pm ike_i^n|^2 \, dS \right) \pm 2|k|^2 \int_{\Gamma_E} |e_i^n|^2 \, dS \pm 2 \operatorname{Im} k \|e_i^n\|_{|k|,U_i}^2$$

Proof. This follows by Lemma 12 from Appendix A and substituting the boundary conditions.[2] □

Definition 11. Define the positive energy-like quantity

$$E^n := \sum_{i=1}^m \sum_{j\neq i} \int_{\Gamma_{ij}} |\partial_n u_i^n|^2 + |k|^2 |u_i^n|^2 \, dS$$

Lemma 9. The sequence $\{E^n\}_1^\infty$ decreases monotonically when either $\operatorname{Im} k > 0$ or both $\operatorname{Im} k \geq 0$ and $\Gamma_E \neq \varnothing$. To be precise $E^n = E^{n-1} - \delta^n$ where

$$\delta_n := \sum_{i=1}^m 2|k|^2 \int_{\Gamma_E} |e_i^n|^2 + |e_i^{n-1}|^2 \, dS + 2 \operatorname{Im} k (\|e_i^n\|_{|k|,U_i}^2 + \|e_i^{n-1}\|_{|k|,U_i}^2).$$

Proof. For ease of exposition we will only prove this in the case of two sub-domains. The generalization to m subdomains is clear. Take e_1^n in Lemma 8 and find

$$\int_{\Gamma_{12}} |\partial_n e_1^n|^2 + |k|^2 |e_1^n|^2 \, dS = \int_{\Gamma_{12}} |\partial_n e_1^n - ike_1^n|^2 \, dS - 2|k|^2 \int_{\Gamma_E} |e_1^n|^2 \, dS - 2 \operatorname{Im} k \|e_1^n\|_{|k|,U_1}^2$$

[2] Lemma 12 was obtained under strong regularity conditions that we cannot guarantee, but the result can hold under much weaker conditions [14].

Then substitute the boundary condition on the interface Γ_{12} and apply Lemma 8 once more, now with the sign changed, to arrive at

$$\int_{\Gamma_{12}} |\partial_n e_1^n|^2 + |k|^2 |e_1^n|^2 \, dS = \int_{\Gamma_{12}} |\partial_n e_1^{n-1}|^2 + |k|^2 |e_1^{n-1}|^2 \, dS$$
$$- 2|k|^2 \left(\int_{\Gamma_E} |e_1^n|^2 \, dS + \int_{\Gamma_E} |e_2^{n-1}|^2 \, dS \right) - 2 \operatorname{Im} k \left(\|e_1^n\|_{|k|,U_1}^2 + \|e_2^n\|_{|k|,U_2}^2 \right). \quad (28)$$

Lastly, equation (28) holds with e_1^n and e_2^n interchanged. Add (28) to itself with the roles of e_1^n and e_2^n reversed to get the identity of the lemma. □

Theorem 3. The iteration (27) is convergent in the sense that $e_i^n \to 0$ in the $\| \cdot \|_{|k|,U_i}$ norm for all i as $n \to \infty$ when either $\operatorname{Im} k > 0$ or $\operatorname{Im} k = 0$ and $\Gamma_E \neq \varnothing$.

Proof. Since the sequence $\{E^n\}_1^\infty$ is bounded and monotone under the conditions of the theorem, it is convergent. Since $E^n = E^0 + \sum_{i=1}^n \delta_n$ it follows that

$$\lim_{n \to \infty} \delta^n = 0.$$

Hence, for the case $\operatorname{Im} k > 0$ we have

$$\|e_i^n\|_{|k|,U_i}^2 \to 0 \text{ for all } i = 1, \ldots, m.$$

When $\operatorname{Im} k = 0$, then we only have $\int_{\Gamma_E} |e_i^n|^2 \, dS \to 0$ on each subdomain. Our tools are not enough to prove convergence here. The gist is to concatenate the e_i^n's on each subdomain to a total error e^n defined on U. The limit $e^n \to e$ satisfies $e = 0$ on Γ_E, and then by uniqueness of the truncated scattering problem it should follow that $e = 0$ on U altogether. □

Note that we heavily rely upon having either an "exterior" boundary Γ_E which has an approximate Sommerfeld boundary condition or a positive imaginary part to the wave number. Indeed having this approximate Sommerfeld boundary condition is necessary for convergence. Consider the one-dimensional problem $-u'' - k^2 u = f$ on $(0, 1)$ such that k^2 is real and not an eigenvalue for the boundary conditions $u(0) = u(1) = 0$. Take two subdomains $U_1 := (0, \frac{1}{2})$ and $U_2 := (\frac{1}{2}, 1)$. The errors have the form $e_i^n(x) = c_i^n(e^{ikx} - e^{-ikx})$ for some constants c_i^n. The boundary condition on the interface is then equivalent to $c_1^n = c_2^{n-1}$ and $c_2^n = c_1^{n-1}$, meaning there is no reduction of the error.

2.3.1 Recent improvements

A currently popular approach that has taken the domain decomposition with Sommerfeld-like interface conditions a step further is the so-called *Sweeping Preconditioner* [7]. It has two components: it improves the approximation quality

of the Sommerfeld boundary condition on the interfaces by replacing it with a Perfectly Matched Layer (an analytic continutation technique). Secondly it makes many very thin subdomains so that direct methods apply; this is to obtain (nearly) linear-time complexity when applied as preconditioner.

Other ideas (so-called *Optimized Schwarz Methods*) optimize the constants in the Robin-condition on the interfaces to improve convergence [11].

Unfortunately all domain decomposition methods only apply in the presence of either exponential damping ($\operatorname{Im} k > 0$) or a radiation condition on Γ_E as in (26).

2.4 Boundary integral formulations

A very successful idea that tackles the issue of indefiniteness of discretizations is to reformulate the Helmholtz equation in terms of boundary integral equations (BIEs). We will only touch on it very briefly, because its scope is very limited: a fundamental solution must be available, and to be computationally feasible, the problem must be homogeneous.

The motivation for integral equations is that the differential operator L is unbounded on $H^1(U)$, yet its inverse or solution operator is typically an integral operator that is Fredholm (a "compact perturbation" of the identity) [9]. And in some cases we have explicit representations of the solution operator, such as those in Appendix A: whenever $-\Delta u - k^2 u = 0$ in U, we have

$$u(x) = \int_{\partial U} \Phi \partial_n u - u \partial_n \Phi \, dS \text{ for } x \in U \tag{29}$$

where $U \subset \mathbb{R}^3$ is open and bounded domain with ∂U of class C^1. As (29) does not hold for $x \in \partial U$, one takes the limit as $x \to x_0$ for $x_0 \in U$ and $x_0 \in \partial U$ uniformly in x; for details we refer to [14]. The resulting equality lives only on the boundary ∂U. The main idea is then to view $\partial_n u$ as an unknown whenever u is given on ∂U and vice versa.

2.5 Asymptotic approximations

An interesting attempt to reduce the dimensionality of the discretization is to identify parts of the solution that do not oscillate at the scale of the wave length. For instance, geometrical optics tells us that light rays in a homogeneous medium propagate in straight lines. Slightly more general, Fermat's principle states that a ray of light between two points follows the path of least time. If the propagation speed of the medium varies slowly, these rays can be expected to vary smoothly as well, even though the solution along them is highly oscillatory.

Geometrical optics is supported from asymptotics of the Helmholtz equation in the limit $k \to \infty$ by posing the ansatz

$$u(x) = a(x)e^{ikp(x)} \tag{30}$$

where $p(x)$ is the real-valued *phase* and $a(x)$ is the real-valued *amplitude*. The oscillations are captured in $e^{ikp(x)}$, and we expect a and p to be smoothly varying. Formally substituting (30) into

$$Lu = 0 \text{ in } \mathbb{R}^d$$

where L is as in (7) and dividing by $k^2 e^{ikp}$ gives us the identity

$$a\left[\nabla p^T A \nabla p - 1\right] - \frac{i}{k}\left[2\nabla p^T A \nabla a + a\nabla \cdot A \nabla p\right] - \frac{1}{k^2}\nabla \cdot A \nabla a = 0. \tag{31}$$

If we take the real part of (31) and *discard* the order k^{-2} term, we get the *Eikonal equation*

$$\nabla p^T A \nabla p = 1$$

just for the phase. If we consider the imaginary part of (31), we obtain a transport equation

$$\nabla p^T A \nabla a + a\nabla \cdot A \nabla p = 0$$

for the amplitude. Of course the amplitude and phase depend on k as well, so we have to be careful in calling the term $\frac{1}{k^2}\nabla \cdot A \nabla a$ an order k^{-2} term. A justification via the Stationary Phase Method is found in [9].

We remark that efficient, exact solvers exist for the discretized Eikonal equation; in particular the Fast Marching Method, which is effectively a shortest path algorithm based on Dijkstra's method [16] and therefore runs in $O(N \log N)$ complexity where the number of unknowns N is independent from k.

It is worth noting that characteristics of p may intersect, for instance due to caustics (lense effects as a result of varying propagation speeds). Rather than the expected interference pattern, we obtain nothing but a discontinuity in ∇p. In unbounded domains with slowly varying coefficients of L, this might not be problematic; it is however unclear how these asymptotic approximations are helpful in trapping domains, where waves can reflect indefinitely between the boundaries of the domain.

2.6 Summary

Although considered optimal for the case $k = 0$, standard multigrid does not apply to indefinite Helmholtz problems for two reasons: coarse grid approximations suffer

from the pollution effect, and even if they did not, oscillatory behaviour cannot be represented accurately whenever kh is small enough. The standard Galerkin formulation lies at the heart of the former problem. Attempts in literature to circumvent it include modifying the differential operator to use a "discrete wave number". This however relies on apriori knowledge, which we only have in special cases. Another attempt to replace the Galerkin condition with a least-squares (Petrov-Galerkin) condition does not work either. In Chapter 3 we will address the two problems of multigrid and construct both coarse grid operators that suffer less from pollution *and* FEM basis functions that incorporate oscillations in an attempt to coarsen even when kh is small.

The Shifted Laplacian Preconditioner applies preconditioning via a Helmholtz operator with shifted wave number, assuming that operator is cheap to invert. The problem however is that for large shifts problematic eigenvalues are mapped even closer to the origin, leading to slow convergence of Krylov subspace methods. The non-standard domain decomposition technique of Section 2.3 provides a potential implementation of the Shifted Laplacian Preconditioner, where we see that convergence is indeed improved when exponential damping is added ($\operatorname{Im} k > 0$).

We briefly remark that BIEs, although very effective, only apply in the case of homogeneous problems whenever a fundamental solution is available; this is generally not the case. Finally, asymptotic approximations as $k \to \infty$ are an attempt to make the complexity of the problem independent from k, yet it remains unclear how to apply these techniques in trapping domains.

3 Transform and drop for indefinite matrices

What we will try in this chapter is to revisit the multigrid procedure of Section 2.1. First, we will address the lack of quasi-optimality of the Galerkin formulation by constructing a coarse grid operator that keeps good approximations to problematic eigenpairs. Secondly, we will deal with the approximation error of problematic eigenfunctions on coarse grids by relaxing the geometric aspect of multigrid; rather we try to automatically construct a FEM basis of oscillatory functions.

Our basic assumption will be that the initial finite-element space is large enough for the FEM solution to be accurate enough. For clarity we will work directly with the corresponding systems of linear equations

$$Ax = b \text{ where } A \in \mathbb{C}^{n\times n} \text{ and } x, b \in \mathbb{C}^n. \tag{32}$$

We will assume $P \in \mathbb{C}^{n\times m}$ to be an orthonormal[3] interpolation operator for a coarse grid on which problematic eigenvectors of A are well-approximated. The operator P is complemented by an orthonormal $Q \in \mathbb{C}^{n\times(n-m)}$ such that

$$\mathbb{C}^n = \operatorname{Ran} P \oplus \operatorname{Ran} Q.$$

For any $z \in \mathbb{C}^n$ we write $z = Pz_p + Qz_q$ for $z_p = P^*z$ and $z_q = Q^*z$. We denote A in the new basis for $\mathbb{C}^n$ as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} := \begin{bmatrix} Q^*AQ & Q^*AP \\ P^*AQ & P^*AP \end{bmatrix}. \tag{33}$$

The ideas are based on the block decomposition

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ & \tilde{A}_{22} \end{bmatrix}$$

where

$$\tilde{A}_{22} := A_{22} - A_{21}A_{11}^{-1}A_{12}$$

is the *Schur complement*, assuming A_{11} is invertible.

3.1 Dealing with lack of quasi-optimality

The Schur complement shows up when we solve (32) approximately for the best approximation $x_P \in \operatorname{Ran} P$ (in the Euclidean norm):

$$\tilde{A}_{22}x_p = b_p - A_{21}A_{11}^{-1}b_q \text{ where } x_P = Px_p.$$

The Schur complement is hence our way to avoid quasi-optimality on a coarse grid.

[3] In multigrid P is usually sparse; orthonormality would destroy that. We only assume orthonormality for notational convenience.

3.1.1 Approximating the Schur complement

There is however an obvious issue with working with the Schur complement: A_{11}^{-1} is not sparse, and even if it is sparsely approximated, the product with A_{21} and A_{12} creates additional fill-in. We must therefore approximate $\tilde{A}_{22}$ to avoid increased computational complexity. Note that the Galerkin projection itself can be seen as the cheapest approximation where the A_{21} term is discarded. In our work we will simply approximate A_{11} by its diagonal $\operatorname{diag} A_{11}$, which is effective if A_{11} is diagonally dominant. Secondly we simply drop fill-in of $\tilde{A}_{22}$, so that its sparsity pattern matches that of A_{22}.

3.1.2 Spectral analysis of the Schur complement

As we have learned from Section 1.3.3, the pollution effect can partly be explained by the Ritz values of A with respect to P not approximating problematic eigenvalues of A well. The natural question is how the (exact) Schur complement deals with these problematic eigenvalues. Consider the eigenvalue problem

$$Ax = \lambda x$$

By change of basis under P and Q we get

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_q \\ x_p \end{bmatrix} = \lambda \begin{bmatrix} x_q \\ x_p \end{bmatrix}.$$

We eliminate x_q and obtain

$$A_{22}x_p = \lambda x_p + A_{21}(A_{11} - \lambda I)^{-1} A_{12} x_p. \tag{34}$$

Since (34) is nonlinear with respect to λ, we expand $(A_{11} - \lambda I)^{-1}$ around $\lambda = 0$ via a truncated Neumann series, and obtain that way other approximate eigenvalue problems that involve the Schur complement. One elegant way to do so is via the resolvent operator and its identities.

Definition 12. Denote the *resolvent operator* for A_{11} by

$$R_\lambda := (A_{11} - \lambda I)^{-1}$$

for values of λ such that the inverse is well-defined.

Property 1. For $\lambda, \mu \in \mathbb{C}$ we have

$$R_\lambda = R_\mu + (\lambda - \mu) R_\lambda R_\mu \tag{35}$$

and

$$R_\lambda R_\mu = R_\mu R_\lambda.$$

Proof. The first identity is true since

$$R_\lambda - R_\mu = R_\lambda(A_{11} - \mu I)R_\mu - R_\lambda(A_{11} - \lambda I)R_\lambda = (\lambda - \mu)R_\lambda R_\mu.$$

The second follows by subtracting Equation (35) from itself with the roles of μ and λ interchanged. □

Using the resolvent operator we write (34) as

$$A_{22}x_p = \lambda x_p + A_{21}R_\lambda A_{12}x_p. \tag{36}$$

We apply Property 1 with $\mu = 0$ to obtain

$$\tilde{A}_{22}x_p = \lambda x_p + \lambda A_{21}R_0 R_\lambda A_{12}x_p, \tag{37}$$

where

$$\tilde{A}_{22} := A_{22} - A_{21}A_{11}^{-1}A_{12}$$

is the *Schur complement* of (33). We generate another identity by applying the resolvent property once more to (37):

$$\tilde{A}_{22}x_p = \lambda\left[I + A_{21}A_{11}^{-2}A_{12}\right]x_p + \lambda^2 A_{21}R_0^2 R_\lambda A_{12}x_p. \tag{38}$$

Dropping the last term in equations (36), (37) and (38) gives us the approximate eigenvalue problems:

Definition 13 (Approximate eigenproblems)**.** We define approximations (θ, u) to the pair (λ, x_p) of problem (34) as solutions of the following three problems.

1. The Galerkin projection:
$$A_{22}u = \theta u. \tag{39}$$

2. The eigenproblem for the Schur complement:
$$\tilde{A}_{22}u = \theta u. \tag{40}$$

3. The generalized eigenproblem for the Schur complement:
$$\tilde{A}_{22}u = \theta\left[I + A_{21}A_{11}^{-2}A_{12}\right]u. \tag{41}$$

We address the quality of approximations by answering the question: given an eigenpair (λ, x) of A, do the three approximations of Definition 13 have an approximate eigenvalue θ near λ? The Bauer–Fike theorem is the natural tool for this.

Theorem 4 (Bauer–Fike)**.** Suppose $B \in \mathbb{C}^{n\times n}$ is diagonalizable such that $BV = VD$ for $V, D \in \mathbb{C}^{n\times n}$ with D diagonal. Let $\kappa(V) := \|V\|\|V^{-1}\|$ denote the condition number of the matrix V. If (λ^a, x^a) is an approximate eigenpair of B with a residual $r := Ax^a - \lambda^a x^a$, then there exists an eigenvalue λ of B such that

$$|\lambda - \lambda^a| \leq \kappa(V)\frac{\|r\|}{\|x^a\|}.$$

Lemma 10 (Approximation quality)**.** If A_{11} is diagonalizable, then there exists an eigenvalue θ_1 of problem (39) such that

$$|\lambda - \theta_1| \lesssim \|A_{21}\|\frac{\|x_q\|}{\|x_p\|}.$$

If $\tilde{A}_{22}$ is diagonalizable, there exists an eigenvalue θ_2 of problem (40) such that

$$|\lambda - \theta_2| \lesssim |\lambda|\|A_{21}\|\|A_{11}^{-1}\|\frac{\|x_q\|}{\|x_p\|}.$$

If $\left[I + A_{21}A_{11}^{-1}A_{12}\right]^{-1}\tilde{A}_{22}$ exists and is diagonalizable, then there exists an eigenvalue θ_3 of problem (41) such that

$$|\lambda - \theta_3| \lesssim |\lambda|^2\|A_{21}\|\|A_{11}^{-2}\|\frac{\|x_q\|}{\|x_p\|}.$$

The hidden constants are the condition numbers of the respective eigenvector matrices.

Proof. We view the pair (λ, x_p) as approximate eigenpair to equations (39), (40) and (41). Bauer–Fike then applies with residual terms as in equations (36), (37) and (38). Finally, substitute $x_q = -R_\lambda A_{12}x_p$. □

Lemma 10 has the interpretation that the Schur complement $\tilde{A}_{22}$ approximates problematic eigenvalues (where $|\lambda|$ is small) better than the Galerkin projection A_{22} of A. Note that the fraction $\|x_q\|/\|x_p\|$ is a measure of how well the eigenvector x can be approximated within $\operatorname{Ran} P$. Finally, since A_{11} is the Galerkin projection of A on $\operatorname{Ran} Q$, the quantity $\|A_{11}^{-1}\|$ should be small as $\operatorname{Ran} P$ is assumed to capture the problematic eigenvectors.

3.2 Constructing coarse grids

We will now discuss how to obtain a good interpolation operator P, which we have so far assumed to be given. Optimally P is sparse, approximates problematic

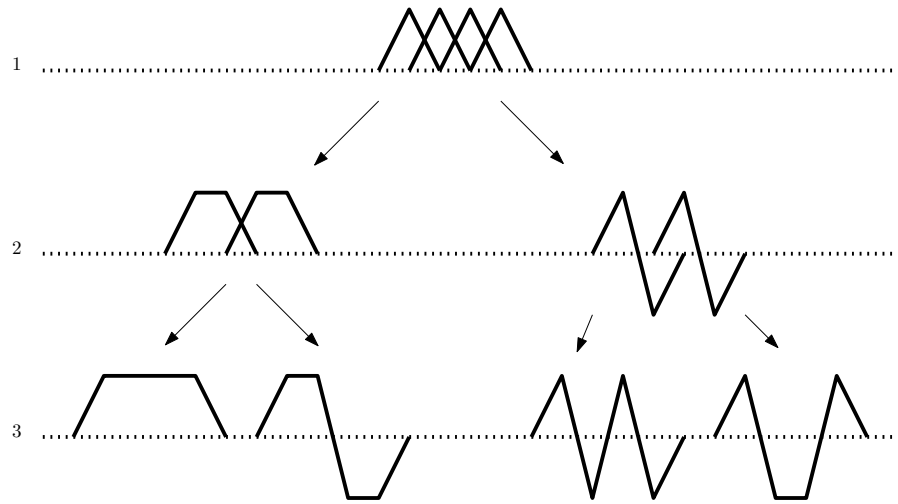


Figure 4: "Algebraic" multigrid can form oscillatory basis functions. "Geometric" multigrid can be interpreted as following the left-most branch. Each left and right child correspond with the interpolation operator P_+ and P_- respectively.

eigenvectors well and reduces the dimensionality sufficiently (so that $m \approx n/2$). In *one dimension*, assuming lexicographical ordering of the unknowns, we can define

$$P_+ := I \otimes \begin{bmatrix} 1 & 1 \end{bmatrix}^T / \sqrt{2} \text{ and } P_- := I \otimes \begin{bmatrix} 1 & -1 \end{bmatrix}^T / \sqrt{2}$$

of the appropriate size. We take $P \leftarrow P_+$ whenever $P_-^T A P_-$ is diagonally dominant and vice versa in an attempt to minimize $\|A_{11}^{-1}\|_2$ in Lemma 10. Slightly more general, we can transform A under the so-called Haar basis $\begin{bmatrix} P_+ & P_- \end{bmatrix}$ and subsequently reorder the unknowns so that A_{11} is diagonally dominant; this implicitly defines P and Q of (33).

We remark that the approximation quality of P_+ and P_- is of lower order, but the upside is simplicity.

3.2.1 Multilevel coarsening: transform & drop

The coarsening process can be repeated recursively to reduce the dimensionality even further. We hope to obtain a sequence of nested linear subspaces

$$\mathbb{C}^n = \mathcal{P}_0 \supset \mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_\ell$$

of exponentially shrinking dimension such that $\mathcal{P}_\ell$ at level ℓ has good approximations to problematic eigenpairs. In one-dimensional h-FEM this could correspond to following a single branch from root to leaf in the tree of Figure 4. Algorithm 1 shows such a procedure where

$$\mathcal{P}_i = \operatorname{Ran} P_1 \cdots P_i \text{ for } i = 1, \ldots, \ell.$$

We refer to this algorithm as *transform & drop*, as we make a change of basis and only then approximate the Schur complement sparsely.

Algorithm 1 Multilevel coarsening: transform & drop

function COARSEN(A, ℓ)
  Let $A^{(1)} := A$
  for $i = 1, \ldots, \ell$ **do**
    Find a sparse, orthonormal basis $V_i = \begin{bmatrix} Q_i & P_i \end{bmatrix}$ that partitions

$$V_i^T A^{(i)} V_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix} \text{ such that } A_{11}^{(i)} \text{ is diagonally dominant.}$$

    Set $A^{(i+1)} \leftarrow A_{22}^{(i)} - A_{21}^{(i)}(A_{11}^{(i)})^{-1} A_{12}^{(i)}$ (or a suitable approximation).
  end for
end function

Multiple ideas are possible now: the orthonormal matrix $P_1 \cdots P_\ell$ can potentially be used for deflation techniques. If the dimension of $\mathcal{P}_\ell$ is still too large for deflation, one can use dense linear algebra to form the Schur decomposition $A^{(\ell+1)}U = US$ where S is upper triangular with diagonal values sorted by absolute magnitude. Subsequently one can construct the orthonormal matrix $P_1 \cdots P_\ell[u_1, \ldots, u_{\tilde{\ell}}]$ with $\tilde{\ell} < \ell$ as an even lower-dimensional basis for deflation.

Finally, we can also use Algorithm 1 to obtain a preconditioner, as it constructs an incomplete block-LU factorization of A. To see this, define

$$L^{(i)}U^{(i)} := \begin{bmatrix} I & \\ A_{21}^{(i)}(A_{11}^{(i)})^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ & A^{(i+1)} \end{bmatrix} = V_i^T A^{(i)} V_i. \tag{42}$$

where no approximations are made to the Schur complement $A^{(i+1)}$, and define

$$W_i := \begin{bmatrix} I & \\ & V_i \end{bmatrix} \in \mathbb{R}^{n \times n} \text{ for all } i = 1, \ldots, \ell$$

where the identity matrix is of the appropriate size.

Lemma 11. Algorithm 1 forms the block-LU decomposition

$$W_\ell^T \cdots W_1^T A W_1 \cdots W_\ell = LU$$

for some matrices L and U where L is lower block-triangular and U upper block-triangular with diagonally dominant blocks on the diagonals of L, apart from that last one.

Proof. This can be shown by induction, as

$$W_{i-1}^T \cdots W_1^T A W_1 \cdots W_{i-1} = \begin{bmatrix} * & \\ * & I \end{bmatrix} \begin{bmatrix} * & * \\ & A^{(i-1)} \end{bmatrix}$$

implies

$$W_i^T \cdots W_1^T A W_1 \cdots W_i = \begin{bmatrix} * & \\ * & L^{(i)} \end{bmatrix} \begin{bmatrix} * & * \\ & U^{(i)} \end{bmatrix}$$

where $L^{(i)}$ and $U^{(i)}$ are as in (42). □

Hence, if we *approximate* the Schur complements and replace $A_{11}^{(i)}$ with its diagonal for all $i = 1, \ldots, \ell$, we obtain an approximate block-LU decomposition

$$A \approx M := W_1^T \cdots W_\ell^T L U W_\ell^T \cdots W_1,$$

where only the lower-right block $A^{(\ell+1)}$ of U is potentially ill-conditioned. Finally we refer to [17] for a recipe to combine the idea of deflation with application of M^{-1}, to avoid problems when the block $A^{(\ell+1)}$ is very ill-conditioned or even (numerically) singular. In short the idea is to apply $M^{-1}A$ only to the complement of (a subspace of) $\mathcal{P}_\ell$.

3.3 Numerical illustration

To illustrate the theory we consider the eigenvalue problem

$$\begin{aligned} -u'' - k^2 u &= 0 \text{ on } (0, \pi), \\ u(0) = u(\pi) &= 0. \end{aligned} \tag{43}$$

We denote again $q = kh$. Discretization with second-order finite-differences yields after multiplication by h^2 a matrix

$$A = \operatorname{diag}\begin{bmatrix} -1 & (2 - q^2) & -1 \end{bmatrix}$$

of size $n \times n$ with $h := \pi/(n+1)$. The first level ($i = 1$ in Algorithm 1) with $A^{(1)} = A$, $P_1 = P_-$ and $Q_1 = P_+$ yields the blocks

$$\begin{aligned} A_{11}^{(1)} &= \operatorname{diag}\begin{bmatrix} -\frac{1}{2} & (3 - q^2) & -\frac{1}{2} \end{bmatrix}, & A_{12}^{(1)} &= \operatorname{diag}\begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \\ A_{21}^{(1)} &= \operatorname{diag}\begin{bmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}, & A_{22}^{(1)} &= \operatorname{diag}\begin{bmatrix} -\frac{1}{2} & (1 - q^2) & -\frac{1}{2} \end{bmatrix}. \end{aligned}$$

We see $A_{11}^{(1)}$ is diagonally dominant for sensible values of q, and

$$\|A_{12}^{(1)}\| < 1,$$

while $A_{22}^{(1)}$ is indefinite. We consider the operators

$$\begin{aligned} S_1^{(i+1)} &:= A_{22}^{(i)}, & S_1^{(1)} &= A, \\ S_2^{(i+1)} &:= A_{22}^{(i)} - \text{tridiag}\left[A_{21}^{(i)}(\text{diag}\, A_{11}^{(i)})^{-1} A_{12}^{(i)}\right], & S_2^{(1)} &= A, \\ S_3^{(i+1)} &:= A_{22}^{(i)} - A_{21}^{(i)} A_{11}^{(i)})^{-1} A_{12}^{(i)}, & S_3^{(1)} &= A, \end{aligned}$$

as approximateions to the coarse grid operators $A^{(i+1)}$ in Algorithm 1 for $i = 1, \ldots, \ell$. They correspond respectively to the Galerkin approximation, the approximate Schur complement of Section 3.1.1 and the exact Schur complement.

We remark that since A_{12} and A_{21} have zeros on the diagonal, the first approximate Schur complement $S_2^{(1)}$ has the form

$$\text{diag}\begin{bmatrix} -\frac{1}{2} & \left(1 - q^2 - \frac{1}{2(3-q^2)}\right) & -\frac{1}{2} \end{bmatrix},$$

with the exception of the first and last entries on the diagonal. We remark the similarity to the idea of replacing the wave number on coarse grids with a discrete wave number as proposed in Section 2.1; the difference is however that we obtain a new "discrete" wave number without apriori knowledge.

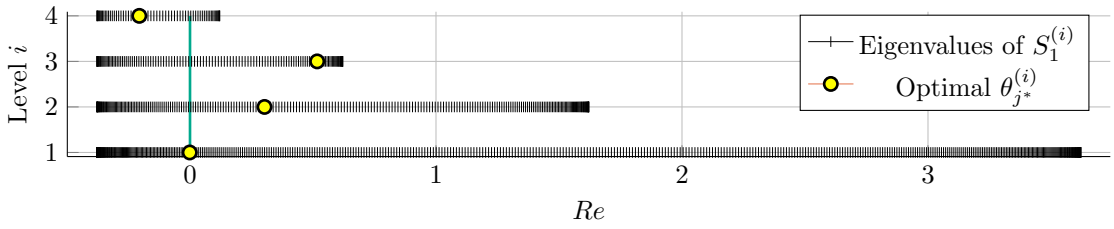
Example 1. We take $n = 512$ and $k = 100.5$, so that problem (43) is discretized with approximately 10.2 unknowns per wavelength. When applying Algorithm 1 we pick P_i either P_+ or P_-, depending on diagonal dominance of $A_{11}^{(i)}$. We compute the exact, normalized problematic eigenvector x of A with eigenvalue closest to 0. Then, on each level i, we compute the normalized eigenpairs $(\theta_j^{(i)}, u_j^{(i)})$ of the coarse grid operator $A^{(i+1)}$ and interpolate $\hat{u}_j^{(i)} := P_1 \cdots P_i u_j^{(i)}$ to $\mathbb{C}^n$. Finally, we define the *optimal approximate eigenpair* $(\theta_{j^*}^{(i)}, \hat{u}_{j^*}^{(i)})$ *for level* i, which has an interpolated eigenvector with minimal angle with x :

$$j^* = \arg\min_j \angle(\hat{u}_j^{(i)}, x).$$

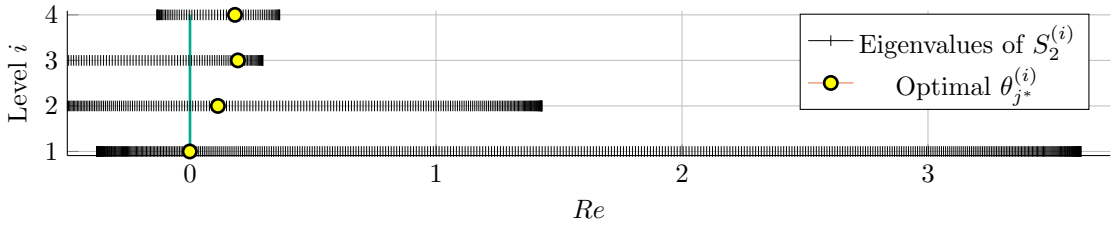
In Figure 5 we plot all eigenvalues $\theta_j^{(i)}$, and highlight the optimal eigenvalue. The caption also shows the corresponding angles $\angle(\hat{u}_{j^*}^{(i)}, x)$. We immediately see that that the Galerkin projection in Figure 5a has a reasonable optimal approximate eigenvector on levels 2 and 3, yet its corresponding eigenvalue drifts away (this can be understood via Theorem 2 and Lemma 10). The moving eigenvalue results in Algorithm 1 choosing a suboptimal operator P_3, as diagonal dominance becomes an incorrect criterium. This problem is alleviated by the approximate Schur complement in Figure 5b, which selects a better coarse grid operator P_3. Finally, the exact Schur complement of Figure 5c keeps the optimal approximate

eigenvalues close to the origin. We see that the approximate Schur complement results in virtually the same quality approximate eigenvector as the exact Schur complement.

(a) Galerkin projection: $\angle(\hat{u}_j^{(i)}, x) \approx (1.00, 0.93, 0.67, 0.20)$ *resp.*



(b) Approximate Schur complement: $\angle(\hat{u}_j^{(i)}, x) \approx (1.00, 0.95, 0.76, 0.72)$ *resp.*



(c) Exact Schur complement: $\angle(\hat{u}_j^{(i)}, x) \approx (1.00, 0.95, 0.77, 0.74)$ *resp.*

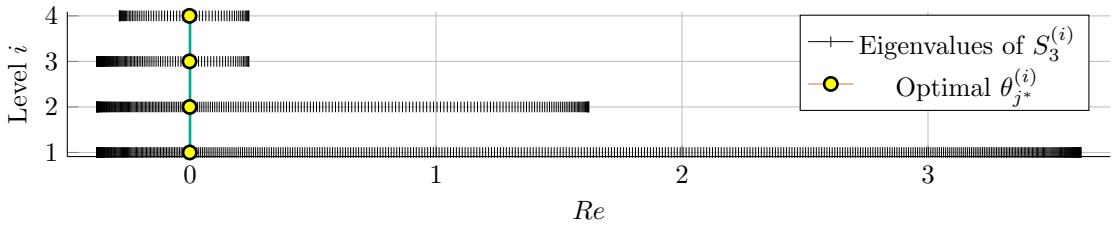


Figure 5: The spectrum of the coarse grid operator $A^{(i)}$ *of Algorithm 1 for different approximations of the Schur complement; see Example 1.*

3.4 Discussion

The one-dimensional example looks promising, but have to address the following points to make the idea feasbile in general:

Variable coefficients. Consider the operator $Lu = -u'' - \frac{k^2}{c(x)^2}u$ where the propagation speed $c(x) \geq 1$ varies over the domain. In this case we still use the

Haar-transform, but we have to decide for each pair of basis functions whether to retain the sum or the difference of the two.

Indecisiveness. Suppose our coarse grid operator takes form

$$A = \operatorname{diag}\begin{bmatrix} -1 & 0 & -1 \end{bmatrix},$$

then the Galerkin projection of the Haar basis produces

$$P_{\pm}^T A P_{\pm} = \mp \operatorname{diag}\begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} \end{bmatrix};$$

neither of the two is diagonally dominant. It seems best to retain both types of basis functions on this level and only decide on the next level which to discard. Effectively, we build

$$P = I \otimes \left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right),$$

compute $P^T A P$, and reorder this matrix such that the upper left block is definite and the lower right block indefinite block.

Multiple dimensions. The one-dimensional example is trivial, because waves travel in a single direction. In two and three dimensions, plane waves $u = e^{ik\hat{\alpha}\cdot x}$ with any direction $\hat{\alpha}$ such that $\|\hat{\alpha}\| = 1$ satisfy $-\Delta u - k^2 u = 0$ (disregarding boundary conditions). As an example, take the eigenvalue problem $-\Delta u = k^2 u$ on $[0, \pi]^2$ with $u = 0$ on the boundaries and $k^2 = 5525$. The solutions are separable and of the form $u(x, y) = \sin(\alpha_1 x) \sin(\alpha_2 y)$ where

$$(\alpha_1, \alpha_2) \in \{(55, 50), (62, 41), (70, 25), (71, 22), (73, 14), (74, 7)\}.$$

The conclusion is two-fold: on local information alone we cannot decide which directions to keep, and the example shows many directions are possible.

3.5 Summary

In this chapter we have seen the *transform & drop* procedure that overcomes two of the issues of the standard multigrid method of Section 2.1. Firstly, the coarse grid operator is chosen as an approximate Schur complement, which is as sparse as the operator obtained in a Galerkin projection, yet less susceptible to pollution. Secondly, oscillatory basis functions are automatically built so that coarsening can go beyond the requirement that kh should be small enough.

The process constructs an incomplete block-LU decomposition of the coefficient matrix A, which can be used for preconditioning in Krylov subspace methods. Furthermore it constructs a low-dimensional subspace that contains good approximations to problematic eigenpairs, which is useful for deflation techniques.

4 Discussion and conclusion

In this thesis we have studied linear systems involving large, sparse, indefinite and potentially nearly singular matrices A that typically arise in interior eigenvalue problems. In the case where A is normal, we relate slow convergence of Krylov subspace methods to the presence of problematic eigenvalues near the origin. Preconditioning and deflation techniques are proposed to improve convergence, yet deflation relies on the availability of good approximations to problematic eigenspaces, which seems to run into circular reasoning.

We study the Helmholtz operator as a motivating example, as standard discretizations lead to aforementioned matrices. At the continuous level indefiniteness corresponds to lack of coercivity of the sesquilinear form. As a result of this, the Galerkin condition of h-FEM is not guaranteed to yield solutions near the best approximation in the search space. In fact, the h-FEM solution is *polluted* for high wave numbers k, meaning that keeping hk small enough is not sufficient to retain the same accuracy of the solution as k grows larger. We relate this inn the self-adjoint case to Ritz values not approximating interior eigenvalues well.

Literature surrounding the Helmholtz equation is studied to glean insights into numerical methods for it, yet many of the ideas proposed (the Shifted Laplacian Preconditioner, adaptations of multigrid & domain decomposition, and more) do not directly apply to linear systems found in eigenvalue problems. However, the multigrid procedure with a modified coarse grid operator to overcome the pollution effect seems promising.

Finally, in Chapter 3, we propose a method that improves the multigrid procedure in two ways: the coarse grid operator is chosen as an approximate Schur complement and oscillatory basis functions are automatically built so that coarsening can continue even when kh is small. We show that the exact Schur complement is less susceptible to pollution.

The *transform & drop* procedure of Chapter 3 effectively builds a stable, incomplete block-LU decomposition of the indefinite matrix which is useful for preconditioning. Even more, it resolves the paradox of deflation, as it cheaply yields a low-dimensional linear subspace that approximately spans the problematic eigenfunctions. The method is illustrated with a one-dimensional Helmholtz problem.

Future directions of research include generalization to two and three dimensions, which seems difficult due to the directionality of waves.

A Properties of the Helmholtz equation

We will list some properties of the Helmholtz equation. Throughout this section, let L denote the operator

$$Lu := -\Delta u - k^2 u$$

and assume $U \subset \mathbb{R}^d$ is an open, bounded and connected domain with a boundary of class C^1.

Fundamental solution. We will derive the fundamental solution of (8) on $\mathbb{R}^3$, exploiting rotational invariance. Although the same method works in $d = 2$, we stick to $d = 3$, because its fundamental solution has a more attractive form. For $d = 2$ see [14]. Suppose $u = u(r)$ is a function of $r = |x|$. Note that $r_{x_i} = x_i/r$. Hence

$$u_{x_i} = u_r \tfrac{x_i}{r} \text{ and } u_{x_i x_i} = u_{rr}\frac{x_i^2}{r^2} + u_r\left(\frac{1}{r} - \frac{x_i^2}{r^3}\right).$$

Therefore

$$Lu = -u_{rr} - \frac{d-1}{r}u_r - k^2 u. \tag{44}$$

Substitute $u(r) = e^{\lambda r}/r$ in (44) and equate to zero:

$$e^{\lambda r}\left[\lambda^2 + k^2 + \lambda(d-3)(r^{-1} + r^{-2})\right] = 0.$$

This expression simplifies only in three dimensions where we get a solution of the form:

$$\Phi_\pm(x) = Ce^{\pm ik|x|}/|x|.$$

We will use Φ_+ because it satisfies the radiation condition (9).

Representation formula. Next, we will construct a representation formula for the solution to

$$Lu = f \text{ on } U.$$

Fix $x \in U$ and define $U_\varepsilon := U \setminus B(x, \varepsilon)$. Assuming u is $C^2(\overline{U})$ we can apply Green's (real-valued) identity with $u(y)$ and $\Phi(y) = \Phi_+(y - x)$:

$$\int_{U_\varepsilon} Lu\,\Phi - u\,L\Phi\,dy = \int_{\partial U_\varepsilon} u\partial_n\Phi - \Phi\partial_n u\,dS.$$

Since $L\Phi(y) = 0$ whenever $y \neq x$ and $x \notin U_\varepsilon$, we have

$$\int_{U_\varepsilon} \Phi f\,dy = \int_{\partial U_\varepsilon} u\partial_n\Phi - \Phi\partial_n u\,dS.$$

With an inward pointing normal $n = (x-y)/\varepsilon$ over the ball $B(x,\varepsilon)$ we can explicitly compute $\partial_n \Phi$:

$$-\int\limits_{\partial B(x,\varepsilon)} u\partial_n \Phi \, dS = e^{ik\varepsilon}(ik\varepsilon - 1)\frac{C}{\varepsilon^2} \int\limits_{|x-y|=\varepsilon} u(y)\, dy.$$

But the integral is nearly a mean value, since

$$\frac{C}{\varepsilon^2}\int_{|x-y|=\varepsilon} u(y) \to 4\pi C u(x) \text{ as } \varepsilon \to 0.$$

Therefore $C = 1/4\pi$ is a natural choice for the constant. This way we get

$$-\int_{\partial B(x,\varepsilon)} u\partial_n \Phi \to u(x) \text{ as } \varepsilon \to 0.$$

As $\Phi(y) \sim \varepsilon^{-1}$ on $B(x,\varepsilon)$ the other term vanishes:

$$\int_{\partial B(x,\varepsilon)} \Phi \partial_n u \to 0 \text{ as } \varepsilon \to 0.$$

Hence, in sending $\varepsilon \to 0$, we get for any $x \in U$ that

$$u(x) = \int_U \Phi_+(y-x) f(y)\, dy + \int_{\partial U} \Phi \partial_n u - u\partial_n \Phi \, dS.$$

Note that we have not assumed anywhere that k is real. In particular, if the wave number k is replaced with $k + i\delta$ with $k, \delta > 0$, we see that $\Phi_+(x)$ decays as $O(|x|^{-\infty})$ rather than $O(|x|^{-1})$ as $|x| \to \infty$.

The eigenvalue problem. To prove Lemma 1 we need some machinery loosely based on [19].

Lemma 12. For u satisfying $Lu = 0$ in U we have

$$\int_{\partial U} |\partial_n u \pm iku|^2 = \int_{\partial U} |\partial_n u|^2 + |k|^2|u|^2 \mp 2\,\mathrm{Im}\, k \|u\|^2_{|k|,U}.$$

Proof. Note that

$$\int_{\partial U} |\partial_n u \pm iku|^2 = \int_{\partial U} |\partial_n u|^2 + |k|^2|u|^2 \pm 2\,\mathrm{Im}(\partial_n u \overline{k}\overline{u}). \tag{45}$$

Next, partial integration of $-\Delta u - k^2 u$ against $\overline{ku}$ gives

$$\int_{\partial U} \partial_n u \overline{k}\overline{u} = \overline{k}\|\nabla u\|^2_{L^2(U)} - k|k|^2\|u\|^2_{L^2(U)} \tag{46}$$

Finally, take the imaginary part of (46) to see that

$$\operatorname{Im} \int_{\partial U} \partial_n u \overline{k} \overline{u} = -\operatorname{Im} k \left[\|\nabla u\|^2 + |k|^2 \|u\|^2 \right] = -\operatorname{Im} k \|u\|^2_{|k|,U}.$$

Substituting this in (45) gives the desired result. □

Lemma 13. Let $B(0,R) \subset \mathbb{R}^d$ denote a ball around the origin of radius R. If u satisfies the radiation condition (9), then

$$\lim_{R\to\infty} \int_{\partial B(0,R)} |u_r - iku|^2 \, dS = 0$$

Proof. This follows by noting that

$$\int_{\partial B(0,R)} |u_r - iku|^2 \, dS \lesssim \max_{x \in \partial B(0,R)} R^{d-1} |(u_r - iku)|^2 \to 0$$

as $R \to \infty$. □

Finally we can prove Lemma 1.

Proof of Lemma 1. Using Lemma 12 with $U = B(0,R)$ and Lemma 13 we get

$$\int_{\partial B(0,R)} |u_r - iku|^2 \, dS = 2 \operatorname{Im} k \|u\|^2_{|k|,B(0,R)} + \int_{\partial B(0,R)} |u_r|^2 + |k|^2 |u|^2 \, dS \to 0$$

as $R \to \infty$. If $\operatorname{Im} k > 0$ all terms are positive, and hence immediately $u \equiv 0$ on $\mathbb{R}^d$. If the wave number is purely real, we can only conclude

$$\lim_{R\to\infty} \int_{\partial B(0,R)} |u_r|^2 \, dS = \lim_{R\to\infty} \int_{\partial B(0,R)} |u|^2 \, dS = 0. \tag{47}$$

Take any $x_0 \in \mathbb{R}^d$ and select $R > |x_0|$. Then by the representation formula

$$u(x_0) = \int_{\partial B(0,R)} \Phi u_r - u \Phi_r \, dS$$

where $\Phi(x) = \Phi_+(x - x_0)$. Using Hölder's inequality we can bound $|u(x_0)|$ by

$$\left(\int_{\partial B(0,R)} |\Phi|^2 \, dS \int_{\partial B(0,R)} |u_r|^2 \, dS \right)^{1/2} + \left(\int_{\partial B(0,R)} |u|^2 \, dS \int_{\partial B(0,R)} |\Phi_r|^2 \, dS \right)^{1/2}. \tag{48}$$

Because both u and Φ satisfy the radiation condition, we let $R \to \infty$ and employ (47) in (48) to conclude that $u(x_0) = 0$ for any $x_0 \in \mathbb{R}^d$. □

B Rellich & Morawetz-Ludwig identities

The standard procedure of obtaining a weak formulation of a PDE is to multiply by a test function, integrate over the domain and apply partial integration. For the Helmholtz equation with real wave numbers k, this produces a sign-indefinite formulation.

However, there are other ways to obtain a weak form that might be sign-definite, using test functions of a special form. To start out, let's consider the Laplace operator and multiplier $(x \cdot \nabla v)$, which gives the identity

$$(x \cdot \overline{\nabla v})\Delta u = \nabla \cdot \left[(x \cdot \overline{\nabla v})\nabla u\right] - \nabla u \cdot \overline{\nabla v} - \nabla u \cdot ((x \cdot \nabla)\overline{\nabla v}). \tag{49}$$

It can be verified by working out the divergence term. Note that for $v = u$ the $\nabla u \cdot \overline{\nabla v}$ is quadratic, yet the other term $\nabla u \cdot ((x \cdot \nabla)\overline{\nabla v})$ is not. To fix this, we add (49) to itself with the roles of v and u interchanged. By employing the identity

$$\nabla u \cdot ((x \cdot \nabla)\overline{\nabla v}) + \overline{\nabla v} \cdot ((x \cdot \nabla)\nabla u) = \nabla \cdot \left[x \nabla u \cdot \overline{\nabla v}\right] - d \nabla u \cdot \overline{\nabla v}$$

we obtain

$$\begin{aligned}(x \cdot \overline{\nabla v})\Delta u + (x \cdot \nabla u)\overline{\Delta v} =& \nabla \cdot \left[(x \cdot \overline{\nabla v})\nabla u + (x \cdot \nabla u)\overline{\nabla v} - x \nabla u \cdot \overline{\nabla v}\right] \\ &+ (d-2)\nabla u \cdot \overline{\nabla v}.\end{aligned} \tag{50}$$

To relate this back to the Helmholtz operator $L := \Delta + k^2$ we use the fact

$$(x \cdot \overline{\nabla v})u + (x \cdot \nabla u)\overline{v} = \nabla \cdot [x u \overline{v}] - d u \overline{v}$$

and add it k^2 times to (50) to obtain

$$\begin{aligned}(x \cdot \overline{\nabla v})Lu + (x \cdot \nabla u)\overline{Lv} =& \nabla \cdot [(x \cdot \overline{\nabla v})\nabla u + (x \cdot \nabla u)\overline{\nabla v} + x(k^2 u \overline{v} - \nabla u \cdot \overline{\nabla v})] \\ &+ (d-2)\nabla u \cdot \overline{\nabla v} - d k^2 u \overline{v}.\end{aligned} \tag{51}$$

Definition 14. With $v = u$ this is the **Rellich identity**:

$$\begin{aligned}2\,\mathrm{Re}(x \cdot \overline{\nabla u})Lu =& \nabla \cdot \left[2\,\mathrm{Re}(x \cdot \overline{\nabla u})\nabla u + x(k^2|u|^2 - |\nabla u|^2)\right] \\ &+ (d-2)|\nabla u|^2 - d k^2 |u|^2.\end{aligned}$$

However, there is again a sign-indefiniteness in the three-dimensional case, and in two dimensions we drop the term $|\nabla u|^2$. To remedy these problems we take a *linear combination* of the $x \cdot \nabla v$ and v multipliers. Define the multiplier

$$\mathcal{M}v := x \cdot \nabla v + \alpha v.$$

By Green's identity it holds for any $\alpha \in \mathbb{C}$ that

$$\overline{\alpha v} Lu + \alpha u \overline{Lv} = \nabla \cdot (\overline{\alpha v} + \alpha u \overline{\nabla v}) - 2\,\mathrm{Re}(\alpha) \nabla u \cdot \overline{\nabla v} - 2\,\mathrm{Re}(\alpha) k^2 u \overline{v}. \tag{52}$$

Combining equation (51) with (52) suggests taking $\alpha = \frac{d-1}{2}$ so that the sign-indefiniteness in the non-divergence term is gone:

$$\begin{aligned} \overline{\mathcal{M} v} Lu + \mathcal{M} u \overline{Lv} =& \nabla \cdot \left[\overline{\mathcal{M} v} \nabla u + \mathcal{M} u \overline{\nabla v} + x(k^2 u \overline{v} - \nabla u \cdot \overline{\nabla v}) \right] \\ & - \nabla u \cdot \overline{\nabla v} - k^2 u \overline{v}. \end{aligned}$$

Definition 15. With $v = u$ we arrive at **Morawetz identity**:

$$2\,\mathrm{Re}(\overline{Mu} Lu) = \nabla \cdot \left[2\,\mathrm{Re}(\overline{\mathcal{M} u} \nabla u) + x(k^2 |u|^2 - |\nabla u|^2) \right] - |\nabla u|^2 - k^2 |u|^2. \tag{53}$$

Proof of Lemma 5. Once again we use Green's identity with u and $\overline{u}$ over U to obtain

$$\int_U \overline{u}\Delta u + |\nabla u|^2\,dx = \int_{\partial U} \overline{u}\partial_n u\,dS.$$

We then substitute $\Delta u = f - (k+\delta i)^2 u$ in U together with the boundary condition on ∂U and arrive at

$$(k+i\delta)^2\|u\|^2_{L^2(U)} - \|\nabla u\|^2_{L^2(U)} + ik\|u\|^2_{L^2(\partial U)} = -\int_U \overline{u}f\,dx - \int_{\partial U}\overline{u}g\,dS. \tag{54}$$

Now we look at the imaginary part of (54). Estimate the integrals on the right-hand side with the Hölder inequality followed by the weighted Cauchy inequality $2ab \le \varepsilon a^2 + b^2/\varepsilon$ for $a, b, \varepsilon > 0$:

$$(2k\delta - \tfrac{\varepsilon_1}{2})\|u\|^2_{L^2(U)} + (k - \tfrac{\varepsilon_2}{2})\|u\|^2_{L^2(\partial U)} \le \tfrac{1}{2\varepsilon_1}\|f\|^2_{L^2(U)} + \tfrac{1}{2\varepsilon_2}\|g\|^2_{L^2(\partial U)} \tag{55}$$

Naturally we take $\varepsilon_1 = 2k\delta$ and $\varepsilon_2 = k$ and obtain the estimate

$$2k\delta\|u\|^2_{L^2(U)} + k\|u\|^2_{L^2(\partial U)} \le \tfrac{1}{2k\delta}\|f\|^2_{L^2(U)} + \tfrac{1}{k}\|g\|^2_{L^2(\partial U)}. \tag{56}$$

Next, take the real part of (54), which is

$$(\delta^2 - k^2)\|u\|^2_{L^2(U)} + \|\nabla u\|^2_{L^2(U)} = \operatorname{Re}\left(\int_U \overline{u}f + \int_{\partial U}\overline{u}g\right) \tag{57}$$

The only negative term on the left-hand side of (57) is $-k^2\|u\|^2_{L^2(U)}$, which we get rid of by adding it twice to both sides of the equation. Again we estimate the right-hand side using the Hölder and Cauchy inequalities:

$$\|u\|^2_{k,U} \le (2k^2 + \tfrac{\varepsilon_1}{2})\|u\|^2_{L^2(U)} + \tfrac{\varepsilon_2}{2}\|u\|^2_{L^2(\partial U)} + \tfrac{1}{2\varepsilon_1}\|f\|^2_{L^2(U)} + \tfrac{1}{2\delta_2}\|g\|^2_{L^2(\partial U)} \tag{58}$$

Take $\varepsilon_1 = k^2$ and $\varepsilon_2 = k$. Finally estimate $\|u\|^2_{L^2(U)}$ and $\|u\|^2_{L^2(\partial U)}$ on the right-hand of (58) side using (56), so that

$$\|u\|^2_{k,U} \le \left(\frac{5}{8\delta^2} + \frac{1}{4k\delta} + \frac{1}{2k^2}\right)\|f\|^2_{L^2(U)} + \left(\frac{3}{4\delta} + \frac{1}{k}\right)\|g\|^2_{L^2(\partial U)}$$

□

Proof of Lemma 6. Note: we assume $|x| \le 1$ for $x \in U$. Integrating Morawetz identity (53) yields

$$2\operatorname{Re}\int_U \overline{Mu}L_\delta u\,dx + \|u\|^2_{k,U} = \int_{\partial U} 2\operatorname{Re}(\overline{Mu}\partial_n u) + (k^2|u|^2 - |\nabla u|^2)(x\cdot n)\,dS$$

The main trick to estimate the boundary integral is to split the gradient into a tangential and perpendicular part

$$\nabla u = (\nabla u - n\partial_n u) + n\partial_n u =: \nabla_{\partial U} u + n\partial_n u \text{ on } \partial U.$$

Then using the fact that the domain is star-shaped and bounded, we can write

$$\begin{aligned}\|u\|_{k,U}^2 + c\|\nabla_{\partial U} u\|_{L^2(\partial U)}^2 \leq -2\,\mathrm{Re}\int_U \overline{Mu} L_\delta u\,dx + 2\,\mathrm{Re}\int_{\partial U} (\overline{x\cdot\nabla_{\partial U}u + \alpha u}\partial_n u)\,dS \\ + \left(k^2\|u\|_{L^2(\partial U)}^2 + \|\partial_n u\|_{L^2(\partial U)}^2\right).\end{aligned} \tag{59}$$

Next substitute $L_\delta u = f + (\delta^2 - 2k\delta i)u$ in the volume integral and note that

$$\mathrm{Re}\int_U \overline{Mu} L_\delta u\,dx = \mathrm{Re}\int_U \overline{Mu} f + (\delta^2 - 2k\delta i)(x\cdot\nabla\overline{u})u\,dx + \alpha\delta^2\|u\|_{L^2(U)}^2 \tag{60}$$

where the last term has the correct sign. For brevity define $\gamma := |\delta^2 - 2k\delta i|$. Substitute (60) in (59) and apply Hölder and Cauchy with ε on all terms of the integrands to obtain

$$\begin{aligned}&(\tfrac{1}{2} + \tfrac{\alpha\delta^2}{k^2} - \tfrac{\varepsilon\gamma}{k^2})k^2\|u\|_{L^2(U)}^2 + (\tfrac{1}{2} - \tfrac{\gamma}{\varepsilon})\|\nabla u\|_{L^2(U)}^2 + \tfrac{1}{2}c\|\nabla_{\partial U}u\|_{L^2(\partial U)}^2 \\ &\leq (2 + \tfrac{2\alpha^2}{k^2})\|f\|_{L^2(U)}^2 + (2 + \tfrac{1}{c} + \alpha)\|\partial_n u\|_{L^2(\partial U)}^2 + (k^2 + \alpha)\|u\|_{L^2(\partial U)}^2\end{aligned} \tag{61}$$

for any $\varepsilon > 0$. To get positive factors in the left-hand side, we impose both

$$\tfrac{\gamma}{\varepsilon} \leq \tfrac{1}{4} \text{ and } \tfrac{\varepsilon\gamma}{k^2} - \tfrac{\alpha\delta^2}{k^2} \leq \tfrac{1}{4}.$$

Choosing $\varepsilon = 4\gamma$ satisfies the first condition, but requires for the second inequality that

$$16\delta^4 + 64k^2\delta^2 \leq 4\alpha\delta^2 + k^2.$$

This condition is satisfied for δ *small enough*. Asymptotically, by keeping δ fixed, dividing by k^2 and sending $k \to \infty$, we find that

$$\delta < \frac{1}{2\sqrt{2}}.$$

For smaller k we can always find a smaller $\delta > 0$ such that the inequality holds. Next, we drop the positive $\frac{c}{2}\|\nabla_{\partial U}u\|_{L^2(\partial U)}$ term from (61), gather the constants that are independent from k and find

$$\|u\|_{k,U}^2 \lesssim (1 + \tfrac{1}{k^2})\|f\|_{L^2(U)}^2 + \|\partial_n u\|_{L^2(\partial U)}^2 + (k^2 + 1)\|u\|_{L^2(\partial U)}^2. \tag{62}$$

Lastly, we employ the boundary condition. Via the Cauchy inequality we get

$$\|\partial_n u\|^2_{L^2(\partial U)} \lesssim \|g\|^2_{L^2(\partial U)} + k^2\|u\|^2_{L^2(\partial U)}.$$

Hence (62) can be estimated by

$$\|u\|^2_{k,U} \lesssim (1 + \tfrac{1}{k^2})\|f\|^2_{L^2(U)} + \|g\|^2_{L^2(\partial U)} + (k^2+1)\|u\|^2_{L^2(\partial U)}. \tag{63}$$

Finally, to get rid of the $\|u\|^2_{L^2(\partial U)}$ term, we use our estimate (55) with $\varepsilon_2 = k$ to get

$$\|u\|^2_{L^2(\partial U)} \leq \tfrac{\varepsilon_1}{k}\|u\|^2_{L^2(U)} + \tfrac{1}{k\varepsilon_1}\|f\|^2_{L^2(U)} + \tfrac{1}{k^2}\|g\|^2_{L^2(U)} \tag{64}$$

for any $\varepsilon_1 > 0$. Combining (63) and (64) shows that choosing $\varepsilon_1 = \frac{1}{2}\frac{k^3}{k^2+1}$ leads to

$$\|u\|^2_{k,U} \lesssim (1 + \tfrac{1}{k^2} + \tfrac{1}{k^4})\|f\|^2_{L^2(U)} + (1 + \tfrac{1}{k^2})\|g\|^2_{L^2(U)},$$

which is the desired inequality.

□

References

[1] Ivo Babuška, Frank Ihlenburg, Ellen T Paik, and Stefan A Sauter. A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution. *Computer methods in applied mechanics and engineering*, 128(3-4):325–359, 1995.

[2] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst. *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM, 2000.

[3] Alvin Bayliss, Charles I Goldstein, and Eli Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59(3):396–404, 1985.

[4] Jean-David Benamou and Bruno Desprès. A domain decomposition method for the Helmholtz equation and related optimal control problems. *Journal of Computational Physics*, 136(1):68–82, 1997.

[5] Susanne Brenner and Ridgway Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[6] Andrew Chapman and Yousef Saad. Deflated and augmented Krylov subspace techniques. *Numerical linear algebra with applications*, 4(1):43–66, 1997.

[7] Björn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling & Simulation*, 9(2):686–710, 2011.

[8] Yogi A Erlangga, Cornelis Vuik, and Cornelis Willebrordus Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3-4):409–425, 2004.

[9] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[10] Martin J Gander, Ivan G Graham, and Euan A Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, 131(3):567–614, 2015.

[11] Martin J Gander, Frédéric Magoules, and Frédéric Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM Journal on Scientific Computing*, 24(1):38–60, 2002.

[12] Pierre Grisvard. *Elliptic problems in nonsmooth domains*. SIAM, 2011.

[13] Frank Ihlenburg and Ivo Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.

[14] William Charles Hector McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge university press, 2000.

[15] Jens Markus Melenk. *On generalized finite element methods*. PhD thesis, Research directed by Dept. of Mathematics. University of Maryland at College Park, 1995.

[16] Olof Runborg. Mathematical models and numerical methods for high frequency waves. *Commun. Comput. Phys*, 2(5):827–880, 2007.

[17] Gerard LG Sleijpen and Fred W Wubs. Exploiting multilevel preconditioning techniques in eigenvalue computations. *SIAM Journal on Scientific Computing*, 25(4):1249–1272, 2004.

[18] Henk A Van der Vorst. *Iterative Krylov methods for large linear systems*, volume 13. Cambridge University Press, 2003.

[19] Calvin H Wilcox. A generalization of theorems of Rellich and Atkinson. *Proceedings of the American Mathematical Society*, 7(2):271–276, 1956.