Toward an algebraic multigrid method for the indefinite Helmholtz equation

Robert D Falgout, Matthieu Lecouvez, Pierre Ramet, Clément Richefort

TOWARD AN ALGEBRAIC MULTIGRID METHOD FOR THE INDEFINITE HELMHOLTZ EQUATION *

ROBERT D. FALGOUT†, MATTHIEU LECOUVEZ‡, PIERRE RAMET§, AND CLÉMENT RICHEFORT‡

Abstract. It is well known that multigrid methods are very competitive in solving a wide range of SPD problems. However achieving such performance for non-SPD matrices remains an open problem. In particular, three main issues may arise when solving a Helmholtz problem : some eigenvalues may be negative or even complex, requiring the choice of an adapted smoother for capturing them, and because the near-kernel space is oscillatory, the geometric smoothness assumption cannot be used to build efficient interpolation rules. Moreover, the coarse correction is not equivalent to a projection method since the indefinite matrix does not define a norm. We present some investigations about designing a method that converges in a constant number of iterations with respect to the wavenumber. The method builds on an ideal reduction-based framework and related theory for SPD matrices to improve an initial least squares minimization coarse selection operator formed from a set of smoothed random vectors. A new coarse correction is proposed to minimize the residual in an appropriate norm for indefinite problems. We also present numerical results at the end of the paper.

Key words. Algebraic Multigrid, Helmholtz Equation, Linear Algebra, Polynomial Smoother, Indefinite matrix

1. Introduction.

The numerical simulation of various physical phenomena leads to potentially very large linear systems of equations written $A\boldsymbol{x} = \boldsymbol{b}$ in matrix form. These systems can be solved directly by a convenient factorization of A, or iteratively by computing and refining an approximation of the solution $\boldsymbol{x}$ starting from an initial guess $\boldsymbol{x}_0$. Multigrid methods [6, 25] work iteratively and are known to be scalable and quasi-optimal for solving sparse linear systems of equations for many classes of problems. Each multigrid iteration combines a projection method on a coarser space to capture the eigenvectors associated with the small eigenvalues, and a few iterations of a smoothing method to capture the remaining eigenvectors generally associated with the large eigenvalues.

To simplify the discussion in what follows, we use the term "small/large eigenvector" to designate an eigenvector with small/large eigenvalue. We similarly say "positive/negative eigenvector" when referring to the eigenvalue sign. Additionally, capital italic Roman letters (A, E, P) denote matrices and bold lowercase letters denote vectors $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r}, \boldsymbol{\alpha})$. Other lowercase letters denote scalars (σ, λ), while capital calligraphic letters denote sets and spaces $(\mathcal{C}, \mathcal{F}, \mathcal{K})$.

1.1. Multigrid methods.

While errors composed of small eigenvectors are known to be more difficult to eliminate for most iterative methods, multigrid methods accelerate the convergence to the solution by projecting them onto a coarser space. The coarse projection of those difficult eigenvectors is repeated recursively until reaching a small enough coarse matrix for which the factorization by a direct solver is fast.

*This work was funded by CEA. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-xxxxxx).

†Lawrence Livermore National Laboratory, Livermore, CA (rfalgout@llnl.gov)

‡Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) (matthieu.lecouvez@cea.fr, richefort.clement@protonmail.com)

§Univ. Bordeaux, CNRS, Bordeaux INP, INRIA, LaBRI, UMR 5800, F-33400 Talence, France (pierre.ramet@inria.fr)

Assuming the matrix is symmetric positive definite (SPD), the best approximation of the solution within the coarse projection space is computed by minimizing the approximation error in A-norm. The core idea in multigrid methods is to make this projection practical by recursively defining smaller subspaces by way of sparse operators P_l, called interpolation operators. The computation of $\boldsymbol{x}$ is accelerated by way of a hierarchy of coarse problems $A_l \boldsymbol{x}_l = \boldsymbol{r}_l$, where $\boldsymbol{r}_l$ is the residual of the level l in the grid hierarchy. P_l determines the coarse projection subspace of the level l, and transfers the information from level $l+1$ to l. In most symmetric applications, coarse matrices are constructed following the Galerkin formula $A_{l+1} = P_l^T A_l P_l$. The two-level coarse correction operator denoted by

$$\Pi_A(P) := P(P^T A P)^{-1} P^T A \tag{1.1}$$

is an A-orthogonal projector onto range(P) and coincides with a minimization problem in the SPD case such that

$$\underset{\tilde{\boldsymbol{x}} \in \text{span}\{P\}}{\arg\min} ||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_A = \Pi_A(P)\,\boldsymbol{r}. \tag{1.2}$$

Two-level methods actually need both types of solvers. The coarse correction (1.1) requires a direct method for factorizing the coarsest matrix whereas the remaining error is eliminated on the fine level through a few iterations of an iterative method called a smoother. From Equation (1.1), the error propagation matrix for the coarse correction of a two-level method is

$$E = I - \Pi_A(P). \tag{1.3}$$

Likewise, the error propagation matrix for the smoother is

$$E_M = I - M^{-1} A \tag{1.4}$$

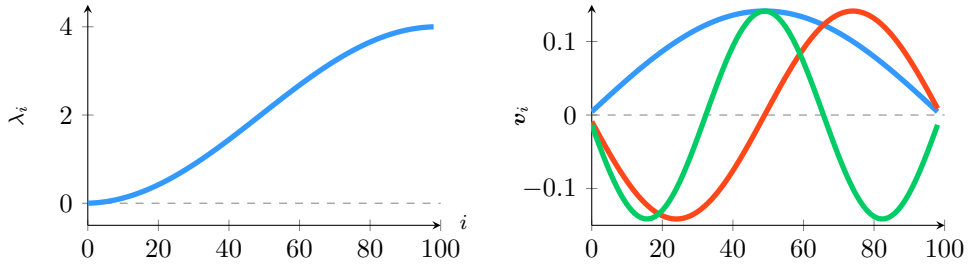
where M^{-1} is an approximation of A^{-1}. The smoother is applied before each restriction and after each interpolation, as illustrated in Algorithm 1.1.

Algorithm 1.1 Two-level cycle

Inputs : $\boldsymbol{b}$ right-hand side, $\tilde{\boldsymbol{x}}$ approximation of $\boldsymbol{x}$ or initial guess, $\boldsymbol{r} = b - A\tilde{\boldsymbol{x}}$ residual
  A initial matrix, M smoother, P interpolation operator
for $j = 1, \nu$ **do**
  $\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + M^{-1}\boldsymbol{r}$
  $\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
end for
$\boldsymbol{r}_{\mathcal{C}} \leftarrow P^T \boldsymbol{r}$
$\tilde{\boldsymbol{e}}_{\boldsymbol{\mathcal{C}}} \leftarrow \text{Solve}(P^T A P, \boldsymbol{r}_{\mathcal{C}})$
$\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + P\tilde{\boldsymbol{e}}_{\mathcal{C}}$
$\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
for $j = 1, \nu$ **do**
  $\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + M^{-1}\boldsymbol{r}$
  $\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
end for
Output : $\tilde{\boldsymbol{x}}$ approximation of $\boldsymbol{x}$ at the end of the cycle

Finding a smoother and a coarse correction that are complementary is a major concern in the design of the method. Moreover, the context in which a multigrid method is applied determines what kind of operators should be used in the method. In

particular, the near-kernel space of smallest eigenvectors is especially important in the design of interpolation. In elliptic problems such as the Laplace equation whose spectrum is illustrated in Figure 1.1, the convergence of multigrid methods is well known. The matrix A is SPD, so smoothers like w-Jacobi or Gauss-Seidel are known to be good smoothers since they damp the large eigenvectors without modifying the small ones. In this elliptical context, these small and large eigenvectors are characterized by low and high frequency oscillations respectively. Hence, while the smoother damps the oscillatory modes, the interpolation must target the slowly varying modes associated with small eigenvalues (see Figure 1.1b). For this reason, the geometric smoothness of the near-kernel space is generally a key assumption, and makes the construction of good interpolation rules more convenient in the initialization of the method.



(a) Eigenvalues (b) Eigenvectors

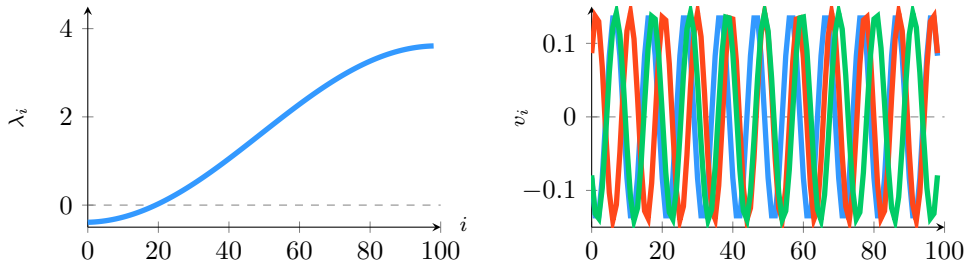
Fig. 1.1: Laplace eigenvalues and three smallest eigenvectors

Likewise in classic algebraic multigrid [26, 23, 11], the interpolation operators are designed to target what is called algebraically smooth components. The smoothed aggregation method [9] is particularly efficient for solving problems with an *a priori* known near-kernel space, for instance in diffusion [28] or elasticity [27] where the target small eigenvectors are the constant vector and rigid body modes respectively. Those vectors are split between disjoint aggregates over the entire domain to initiate a tentative block interpolation operator. A few smoothing iterations are applied to the tentative interpolation operator to extend its pattern, but especially in order to clean the tentative interpolation range from high frequencies. Usually, a few iterations of the Jacobi relaxation method are enough, but this step of energy minimization has been generalized to Krylov methods such as the conjugate gradient [21] by enforcing sparsity constraints in the minimization space to keep a practical interpolation operator. If near-kernel space information is lacking, test vectors can still be computed algebraically, as in adaptive smoothed aggregation [5]. Furthermore, because the choice of the interpolation strategy is essential in the convergence of the method, an ideal framework maximizing the complementarity between the smoother and the coarse correction [12] has been established to guide the algorithm development. While this idealistic scenario of convergence is mostly used as a theoretical tool, some reduction-based methods enable a good approximation of the ideal interpolator given some initial coarse and fine variable splitting [19, 29].

1.2. Why Helmholtz problems are difficult for multigrid. The Helmholtz equation (1.5) involves indefinite matrices with potentially wide and oscillatory near-kernel spaces [10]. This equation is our target in this paper.

$$\text{(Continuous Helmholtz problem)} \Leftrightarrow \begin{cases} -\Delta \boldsymbol{u} - k^2 \boldsymbol{u} = \boldsymbol{f} & \text{on } \Omega \\ \quad + \text{b. c.} & \text{on } \partial\Omega \end{cases} \tag{1.5}$$

In fact, the Helmholtz equation can be seen as a shifted Poisson equation, where geometrically smooth eigenvectors (i.e., low Fourier modes, see Figure 1.1b) can be negative eigenvectors because of the shift. In the same way, the smallest eigenvectors of the shifted Laplacian are higher in frequency (see Figure 1.2b).



(a) Eigenvalues (b) Eigenvectors

Fig. 1.2: Helmholtz eigenvalues and three smallest eigenvectors

This complication breaks the near-kernel space geometric smoothness assumption, a keystone of many multigrid methods. To design a coarse correction and smoothers that are complementary in this context, interpolation rules must reproduce the near-kernel oscillation, and contrary to usual relaxation methods, smoothers have to deal with both positive and negative eigenvalues. More importantly, the coarse correction is not equivalent to a minimization problem anymore since the indefinite matrix does not define a norm (i.e., the equality (1.2) is not valid for Helmholtz). Whereas the coarse correction is guaranteed to not amplify the error for SPD matrices, the approximation error can be amplified in the indefinite case because the spectrum of the matrix has both signs.

For these reasons, finding a recurring process to build a scalable multilevel method is still an open question. Multiple correction [18], wave-ray [4, 17], and Complex-Shifted Laplacian [8] approaches have already been investigated to address this issue. In this paper, we present a fully algebraic approach built on ideal reduction-based ideas, and demonstrate its potential for solving the Helmholtz problem with constant iteration count independent of the wavenumber k. Certain discretization matrices resulting from the continuous problem (1.5) can be non-symmetric due to the boundary conditions. To center the discussion on the indefinite nature of Helmholtz, the next approaches address the symmetric indefinite shifted laplacian matrix arising from the following 5-pts stencil

$$\hat{A} = \begin{bmatrix} & -1 & \\ -1 & 4-(kh)^2 & -1 \\ & -1 & \end{bmatrix}. \tag{1.6}$$

In Section 2, we start by presenting a normal equation polynomial smoother specifically designed to damp the desired proportion of largest eigenvalues independently of their signs, while interpolation rules for propagating oscillatory near-kernel information are established in Section 3. The Section 4 gives more details on why the indefiniteness can corrupt the coarse correction by introducing a concept of pollution and Section 5 exposes an alternative coarse correction to the classical one which avoids the divergence scenarios. Finally, Section 6 presents benchmarks of this new multigrid

method for different Helmholtz problems, with varying shift kh and wavenumber k.

Along the different approaches presented in this paper, $\boldsymbol{v}_i$ denotes the i^{th} eigenvector of A associated with the eigenvalue λ_i. Moreover, we always assume the eigenvalues to be ordered in magnitude (i.e., $\forall i < n\ ,\ |\lambda_i| \leq |\lambda_{i+1}|$) such that $V_c := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_c}]$ and $V_f := \left[\boldsymbol{v}_{n_{c+1}}, \ldots, \boldsymbol{v}_n\right]$ contain the small and large eigenvector sets of size n_c and n_f respectively. Naturally, the full set of eigenvectors are given by $V = [V_c, V_f]$, and $n = n_c + n_f$.

2. Polynomial Smoothers for Indefinite Problem.

Working with a smoothing method whose behavior on the spectrum is *a priori* known is interesting to guarantee the effectiveness of the cycle. Here, the smoother must damp large positive and negative eigenvalues, which is problematic for most standard methods. Generally, a polynomial method with degree greater than one can work. Krylov iterations are good polynomial smoothers in the indefinite case but they minimize the global residual norm regardless of the eigenvalues and are non-linear because of their right-hand side dependence. A linear polynomial is more convenient for generating the set of smoothed candidates vectors needed to construct the interpolation operator described in Section 3.

2.1. General considerations on polynomial smoothers.

One way to ensure that both positive and negative eigenvectors are damped is to consider a normal equation polynomial smoother. In general, the degree m of the polynomial must be greater than one to damp positive and negative eigenvectors, as the polynomial illustrated in Figure 2.1 does. Resorting to normal equations enables the polynomial to treat eigenvalues with respect to their magnitude rather than their sign, which is equivalent to work with even powers of A if the matrix is hermitian, which is what we assume in this section. In the future, it might be interesting to investigate more general polynomials to avoid normal equations and consider odd exponents. In this first approach, we use the convenient symmetry property enabled by normal equations in the Chebyshev framework.

Let $p_m(A^2)$ be a polynomial of degree m that approximates A^{-2}. From Equation (1.4), let $q_{m+1}(A^2)$ be the associated error propagation matrix of the polynomial smoother such that

$$q_{m+1}(A^2) := I - p_m(A^2)A^2. \tag{2.1}$$

Additionally, let $\boldsymbol{v}$ be an eigenvector of A associated with the eigenvalue λ. Hence,

$$q_{m+1}(\lambda^2)\boldsymbol{v} = (1 - p_m(\lambda^2)\lambda^2)\boldsymbol{v}. \tag{2.2}$$

In multigrid methods, a good smoother eliminates the large eigenvalues that the coarse correction does not capture and vice-versa. Let a and b be real scalars such that $0 < a < b$. Assume these large squared eigenvalues are contained in the interval $[a, b]$. The construction of a relevant interval will be discussed in the next. Since the polynomial smoother $p_m(A^2)$ is an inverse approximate of A^{-2}, the polynomial function $p_m(x)$ can be constructed to approximate the function x^{-1} [15] from $m + 1$ interpolation points x_i selected within the interval of large eigenvalues $[a, b]$. In particular, selecting the scaled first kind Chebyshev polynomial roots as interpolation points

$$x_i := \frac{b+b}{2} + \frac{b-a}{2}\cos\left(\frac{(2i+1)\pi}{2(m+1)}\right)\ ,\ i = 1, \ldots, m+1. \tag{2.3}$$

gives the minimal error propagation function $q_{m+1}(x)$ on the interval $[a,b]$. The polynomial is constructed to satisfy the $m+1$ following constraints

$$x_i \in [a,b]\ ,\ p_m(x_i) = \frac{1}{x_i} \Leftrightarrow q_m(x_i) = 0\ ,\ i = 1,\dots,m+1. \tag{2.4}$$

Because the selected nodes x_i are the roots of q_{m+1} and $q_{m+1}(0) = 1$, then the Lagrange formula yields

$$p_m(x) := \sum_{i=1}^{m+1} \frac{1}{x_i} \prod_{j=1\,,\,j\neq i}^{m+1} \frac{x - x_j}{x_i - x_j}\ ,\ q_{m+1}(x) = \prod_{i=1}^{m+1} \frac{x - x_i}{-x_i}. \tag{2.5}$$

First kind Chebyshev polynomials are defined by the three-terms recurrence relation

$$\forall t \in [-1,1]\ ,\ C_0(t) = 1\ ,\ C_1(t) = t\ ,\ C_{m+1}(t) = 2tC_m(t) - C_{m-1}(t). \tag{2.6}$$

The roots of q_{m+1} are the roots of C_{m+1} but scaled on $[a,b]$, the error propagation function q_{m+1} can be derived as the following re-scaled Chebyshev polynomial

$$q_{m+1}(x) = \frac{C_{m+1}\left(\frac{b+a-2x}{b-a}\right)}{C_{m+1}\left(\frac{b+a}{b-a}\right)}. \tag{2.7}$$

As explained in [2], the upper bound of $C_{m+1}(t)$ on $[-1,1]$ equals one for $t = 1$ and is strictly monotonically increasing for $t > 1$. Accordingly, the supremum of the numerator on $[a,b]$ equals one for $x = a$, and the denominator is strictly greater than one because $\frac{b+a}{b-a} > 1$. Last, $q_{m+1}(0) = 1$ and q_{m+1} is strictly monotonically decreasing for $x \in [0,a]$. As a consequence, $|q_{m+1}(x)| < 1$ on the interval $(0,b]$. Assuming $b \geq \lambda_{\max}^2$, then the spectral radius $\rho\left(q_{m+1}(A^2)\right) < 1$. In other words, the smoother is a convergent iterative method and does not amplify any region of the spectrum.

2.2. Constructing an appropriate target interval. One way to determine an interval $[a,b]$ without preliminary information [2, 1] is to compute a few power iterations to determine b by an overestimation of the largest eigenvalue, and choose the lower bound a according to b, for example $a = \frac{1}{2}b$. However, to respect the complementarity principle, the percentage of damped eigenvalues by the smoother must approximate the proportion of non-coarse variables (i.e. the n_f largest eigenvalues in our case). For instance, if a coarse level is one quarter the size of the finer level, then three-quarters of the largest amplitude eigenvectors should be damped by the smoother, while the coarse correction deals with the remaining small eigenvectors. Consequently, since eigenvalues are not necessarily uniformly separated, a should be determined so that a proportion of eigenvalues belongs to the interval $[a,b]$. Moreover, the spectral distribution of coarse matrices are unknown in a multi-level setting. Therefore, a good interval should satisfy

$$\lambda_i^2 \in [a,b] \ \Leftrightarrow\ \lambda_i \in \left[-\sqrt{b}, -\sqrt{a}\right] \cup \left[\sqrt{a}, \sqrt{b}\right]\ ,\ i = n_c,\dots,n_f. \tag{2.8}$$

While this interval can be fixed using geometric information, we first compute a rough approximation of the matrix *spectral density* as detailed in [16]. This spectral density permits to determine which portion of the spectrum should be damped by

the smoother, and is defined by the distribution function $\phi(t)$ that represents the probability of finding an eigenvalue at each point $t \in [-1, 1]$. We set the lower bound a of the Chebyshev node interval in a second step so that the probability within the interval equals the target proportion, for instance half of the total area in a scenario of exact balance between coarse and non-coarse variables. As defined in (2.6), the distribution function ϕ is approximated by a linear combination of orthogonal Chebyshev polynomial functions, such that

$$\phi(t) = \sum_{k=1}^{\infty} \mu_k C_k(t) \approx \sum_{k=1}^{n_\mu} \mu_k C_k(t). \tag{2.9}$$

Because Chebyshev functions are naturally defined over $[-1, 1]$, the spectral density function must evaluate the spectral density of the scaled matrix $B = \frac{2}{b}A^2 - I$. Since b is assumed to bound the eigenvalues of A^2, the spectrum of B belongs to $[-1, 1]$. The coefficients μ_k are then determined by a moments matching procedure, which gives

$$\mu_k = \frac{2 - \delta_{k0}}{n\pi} \times \mathrm{Trace}(C_k(B)). \tag{2.10}$$

Here, n corresponds to the matrix size and δ_{k0} the Kronecker symbol. The trace can be approximated by a stochastic trace estimation from a set of n_{vec} random and orthogonal vectors $\boldsymbol{z}_l$, where each element of these vectors is chosen following a normal distribution with zero mean and a unit standard deviation. Therefore, the trace approximations are given by

$$\mathrm{Trace}(C_k(B)) \approx \frac{1}{n_{\mathrm{vec}}} \sum_{l=1}^{n_{\mathrm{vec}}} \boldsymbol{z_l}^T C_k(B) \boldsymbol{z_l} \ , \ k = 1, \dots \ n_\mu. \tag{2.11}$$

According to (2.11), each trace can be estimated by a sample mean of n_{vec} products $\boldsymbol{z_l}^T C_k(B)\boldsymbol{z_l}$, and the n_μ vectors $C_k(B)\boldsymbol{z_l}$ can be computed from the three-term recurrence defined in (2.6). Once the distribution function ϕ is approximated following Equation (2.9), a rough area approximation by trapezoid rule yields a correct lower bound that satisfies a proportion around $\frac{n_f}{n}$. This lower bound only needs to be remapped on the initial interval to return the correct value for a. The interval $[a, b]$ constitutes a purely algebraic interval in which the polynomial smoother is the most efficient. The bounds a and b are represented in Figure 2.1, where $x_{50\%}$ illustrates a theoretical lower bound target for the shifted laplacian matrix resulting from (1.6). Last, the total number of matrix vector products required by the spectral density approximation step for the construction of a relevant interval is $n_{\mathrm{vec}} \times n_\mu$.

3. Constructing good interpolation rules. Interpolators are used both to construct the coarse level matrices and to transfer information across levels. SPD and geometric smoothness assumptions cannot be used to determine appropriate interpolation operators in our case. Some methods such as smoothed aggregation [9, 20] and bootstrap-AMG [3] use candidate vectors that are close to the near-kernel space to design the interpolation rules. These test vectors are either deduced from geometric information [4, 22] or algebraically as in adaptive multigrid methods [5]. Here, we prefer to stick to a fully algebraic and recurring process to create our interpolation operators. Candidate vectors will be generated from random vectors smoothed by the polynomial presented in Section 2, and used by the least squares minimization framework to determine good fine variable interpolation rules. This initial least squares interpolation operator is used as a coarse variable operator in the ideal reduction-based framework [12].

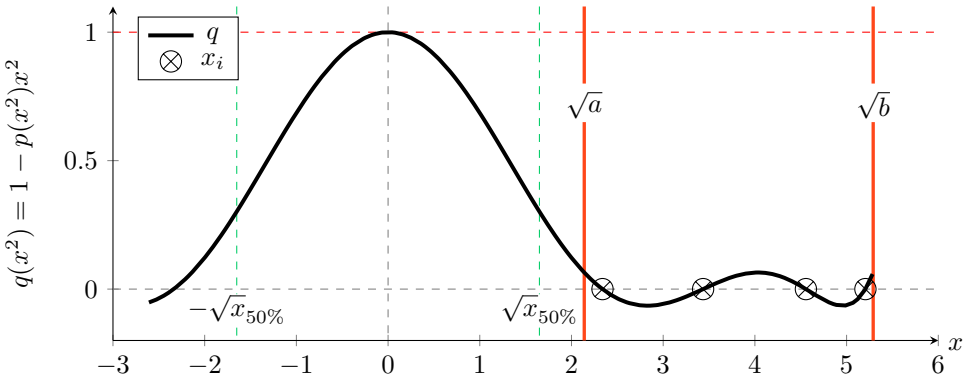


Fig. 2.1: Spectrum of the polynomial smoother error propagation matrix for $kh = 1.65$

3.1. Ideal framework. Even though the ideal framework requires an SPD assumption and has not been generalized to indefinite problems, the idea of removing irrelevant information from the interpolation range is of particular interest for capturing the near-kernel space of oscillatory problems, and will be our guiding principle in this section.

Accordingly, we assume A is SPD in this section dedicated to the ideal framework. Following [12], let $\mathcal{C}$ and $\mathcal{F}$ be complementary coarse and fine variable subsets of Ω respectively of size n_c and n_f. Let $R^T : \mathbb{R}^{n_c} \to \mathbb{R}^n$ and $S : \mathbb{R}^{n_f} \to \mathbb{R}^n$ be coarse and fine variable operators respectively, such that $RS = 0$. The space defined by the coarse variable operator R^T must be handled by the coarse correction, whereas the fine variable operator S defines a space where smoothing must operate in order to respect the complementarity principle. The *Ideal Interpolation* operator is a theoretical operator that is the best that satisfies $RP = I_c$, in the sense that it minimizes the difference between variables and interpolated coarse variables, within a space that is the most complementary to the range of the smoother M. The ideal interpolation operator is given by

$$P_* = \arg\min_P \left(\max_{\boldsymbol{e} \neq 0} \frac{||(I - PR)\boldsymbol{e}||_M}{||\boldsymbol{e}||_A} \right) = (I - S(S^T AS)^{-1} S^T A) R^T. \tag{3.1}$$

Let $P_{:,i}$ and $R^T_{:,i}$ be the i^{th} columns of P and R^T respectively. Each column of the ideal interpolation operator is therefore defined by

$$P_{:,i} = R^T_{:,i} - \boldsymbol{s}_i, \text{ with } \boldsymbol{s}_i = \arg\min_{\tilde{\boldsymbol{s}} \in \text{Range}(S)} ||R^T_{:,i} - \tilde{\boldsymbol{s}}||_A = S(S^T AS)^{-1} S^T A R^T_{:,i} \tag{3.2}$$

In fact, the matrix that multiplies each column of R^T in (3.2) and (3.1) is a projection operator onto the A-orthogonal complement of the range of S. The ideal interpolation operator is constructed by extracting from R^T the information that can already be solved in the subspace S. Such information is irrelevant at a coarse level and should be handled by the smoother. Under the assumption that the smoother captures the space spanned by S, the best coarse matrix is therefore a matrix where S-related information is subtracted. Even if applying $(S^T AS)^{-1}$ is too expensive, it gives insight for building a more practical method.

3.2. Least Squares Minimization Interpolation Operator. As mentioned at the beginning of Section 3.1, demonstrating that the interpolation operator (3.1) is ideal in the theoretical framework of [12] requires A to be symmetric positive-definite. However, the reduction viewpoint which consists in cleaning the range of interpolation by extracting irrelevant information at a coarse level perspective is of interest. In addition, numerical experiments reveal that the classical coarse variable operator $R^T = [0\ I_c]^T$ does not have good approximation property for the oscillatory near-kernel space that characterizes Helmholtz. Therefore, a new coarse variable operator has to be designed algebraically. Using the smallest eigenvectors V_c from Section 3.1 to enforce the representation of the near-kernel space within the interpolation range is not practical. Instead, we construct a set of vectors approximating an oscillatory and potentially large near-kernel space by using the normal equations polynomial smoother developed in Section 2.

In this section, we present a coarse variable operator $\hat{R}^T$ of size $n \times n_c$ constructed by a least squares minimization strategy [3]. Let the columns of T be a set of κ smoothed random vectors $\boldsymbol{z}_l$ that approximates the near-kernel space such that

$$T_{:,l} = q_{m+1}(A^2)\boldsymbol{z}_l\ ,\ l = 1, \dots, \kappa. \tag{3.3}$$

where $T_{:,l}$ designate the l^{th} column of the set T. We assume a $\mathcal{C}/\mathcal{F}$ splitting with n_c and n_f their respective size. $\mathcal{C}$-points are interpolated to the finer level with a simple injection rule, while interpolation rules of $\mathcal{F}$-points are determined by the least squares minimization method presented in this section. Due to this splitting, the coarse interpolation block in $\hat{R}^T$ corresponds to a $n_c \times n_c$ identity matrix denoted by I_c, while R_f^T designate the block of interpolation for the $\mathcal{F}$-points.

Let i be an $\mathcal{F}$-point and $\hat{\boldsymbol{r}}_i$ the vector containing the non-zero elements of the i^{th} row of $\hat{R}^T$. The idea consists of constructing each $\mathcal{F}$-point interpolation rule by minimizing the squared difference between $\mathcal{F}$-values of the near-kernel candidate vectors and the interpolation from their connected $\mathcal{C}$-points in $\mathcal{C}_i$. Denote by $T_{i,:}$ a row vector containing the i^{th} values of each test vector, and $T_{\mathcal{C}_i,l}$ a vector containing the values in $T_{:,l}$ of the $\mathcal{C}$-points that are connected to variable i. Then

$$\forall i \in \mathcal{F}\ ,\ \hat{\boldsymbol{r}}_i = \underset{\hat{\boldsymbol{r}} \in \mathbb{C}^{\text{card}(\mathcal{C}_i)}}{\arg\min} \sum_{l=1}^{\kappa} w_l \left(T_{i,l} - \hat{\boldsymbol{r}} \cdot T_{\mathcal{C}_i,l}\right)^2 =: \underset{\hat{\boldsymbol{r}} \in \mathbb{C}^{\text{card}(\mathcal{C}_i)}}{\arg\min}\ \mathcal{L}_i(\hat{\boldsymbol{r}}) \tag{3.4}$$

where w_l are scaling weights (for instance $w_l = 1/|\lambda_l|$ if T contains near-kernel eigenvectors). Finding the minimum of the convex loss function $\mathcal{L}_i$ is equivalent to solving

$$\nabla \mathcal{L}_i(\hat{\boldsymbol{r}}_i) = 0. \tag{3.5}$$

Equation (3.5) can be rewritten element-wise

$$\frac{\partial \mathcal{L}_i(\hat{\boldsymbol{r}}_i)}{\partial \hat{\boldsymbol{r}}_{ij}} = \sum_{l=1}^{\kappa} 2w_l(T_{i,l} - \hat{\boldsymbol{r}}_i \cdot T_{\mathcal{C}_i,l})T_{\mathcal{C}_{ij},l} = 0\ ,\ \forall j = 1, \dots, \text{card}(\mathcal{C}_i). \tag{3.6}$$

Finally, (3.6) leads to a system of linear equations to solve for each fine variable i

$$\hat{\boldsymbol{r}}_i T_{\mathcal{C}_i} W T_{\mathcal{C}_i}^T = T_i W T_{\mathcal{C}_i}^T \tag{3.7}$$

The matrix is full rank and the solution of Equation (3.7) is unique if we have at least $\kappa = \max_i \{\text{Card}(\mathcal{C}_i)\}$ locally linearly independent test vectors. Even if it is statistically always the case when starting from random candidate vectors, the matrix singularity can be detected during the factorization. In that special case, a pseudo-inverse can be computed to find an optimal solution in the least squares sense.

3.3. Ideal approximation from least squares coarse operator. In Section 3.2, we presented a coarse variable operator for Helmholtz designed by a least squares minimization strategy. Using the framework presented in 3.1, define

$$\hat{R}^T = \begin{bmatrix} R_f^T \\ I_c \end{bmatrix} \text{ and } \hat{S} = \begin{bmatrix} I_f \\ -R_f \end{bmatrix} \tag{3.8}$$

where $\hat{R}^T$ is the least squares coarse variable operator presented in Section 3.2 and R_f^T is its $\mathcal{F}$-points interpolation block. Note that $\hat{R}\hat{S} = 0$ as required. Hence, since the least squares operator is designed to propagate the candidate vectors that are composed of small eigenvectors due to the Chebyshev polynomial smoother of Section 2, the space spanned by $\hat{S}$ is, by orthogonality, mostly composed of large eigenvectors. Accordingly, the aim of using the ideal framework in this oscillatory context is to improve the coarse variable operator by extracting the irrelevant information related to these large eigenvectors that can be solved in $\hat{S}$.

However, two major issues arise in the use of the ideal interpolation operator (3.1). The first is a general concern related to the fine block $\hat{S}^T A \hat{S}$, which is usually not practical to invert, and would lead to a dense interpolation operator $\hat{P}$. To circumvent this problem, an approximation based on sparsity constraints must be applied. The second issue is related to the indefiniteness of the initial matrix. Indeed, as shown by the equation (3.2), applying the left operator of the ideal formula removes the information contained in the range of $\hat{S}$ by minimizing an approximation error in A-norm. However, such a norm does not exist in the indefinite case. Ignoring this problem may still give interesting results in practice, but we consider instead the $A^T A$-norm to ensure the effectiveness of the interpolation operator. Since $\hat{S}$ is sparse, we control the sparsity of $\hat{P}$ by restricting the search space to a few columns of $\hat{S}$ only. Define X_i to be the injection operator of ones and zeros of size $n_f \times n_i$ with $n_i \leq n_f$ that selects n_i columns of $\hat{S}$, $\hat{S}X_i$. From (3.2), let $\boldsymbol{s}_i$ be the solution of the ideal minimization problem such that

$$\boldsymbol{s}_i := \underset{\tilde{\boldsymbol{s}} \in \text{Range}(\hat{S}X_i)}{\arg\min} ||\hat{R}_{:,i}^T - \tilde{\boldsymbol{s}}||_{A^T A} = \hat{S}X_i \left(X_i^T \hat{S}^T A^T A \hat{S} X_i \right)^{-1} X_i^T \hat{S}^T A^T A \hat{R}_{:,i}^T. \tag{3.9}$$

Accordingly, columns of the reduction-based interpolation operator are computed by

$$\hat{P}_{:,i} = \hat{R}_{:,i}^T - \boldsymbol{s}_i = \hat{R}_{:,i}^T - \hat{S}X_i \rho_{n_i}, \tag{3.10}$$

where ρ_{n_i} is the solution of the $n_i \times n_i$ linear system

$$X_i^T \hat{S}^T A^T A \hat{S} X_i \rho_{n_i} = X_i^T \hat{S}^T A^T A \hat{R}_{:,i}^T. \tag{3.11}$$

The choice of the non-zero pattern of $\hat{P}$ must satisfy a good trade-off between approximation properties of the near-kernel space and complexity. While improving the sparsity of this interpolation operator is a topic of future research, one strategy is to choose the columns of $\hat{S}$ based on the entries of $\hat{S}^T A^T A \hat{R}_{:,i}^T$. In fact, each entry

corresponds to the scalar product between a column of $\hat{S}$ and $\hat{R}^T_{:,i}$ in A^TA-norm. A large entry designates a column of $\hat{S}$ that contributes a lot in the solution of the minimization problem (3.9). The column selection phase iterates until the entries associated with the selected columns represent a percentage τ of the entire set of non-zero entries. At each iteration, the column associated with the largest entry of $\hat{S}^TA^TA\hat{R}^T_{:,i}$ is selected, which is equivalent to extending X_i with the euclidean basis vector with one at the index of the chosen column and zeros elsewhere. Because the columns with the largest entries in $\hat{S}^TA^TA\hat{R}^T_{:,i}$ are selected first, the set of selected columns is the smallest set that satisfies

$$||X_i\hat{S}^TA^TA\hat{R}^T_{:,i}||^2 \geq \tau \times ||\hat{S}^TA^TA\hat{R}^T_{:,i}||^2, \text{ with } \tau \in [0,1]\,. \tag{3.12}$$

We note that even though setting $\tau = 1$ selects all the column associated with non-zero entries in the right-hand side, the remaining columns associated with zero entries are omitted, and therefore the matrix $X_i^T\hat{S}^TA^TA\hat{S}X_i$ still correspond to a principle sub-matrix of $\hat{S}^TA^TA\hat{S}$. The Figure 3.3 represents the error of interpolation of every eigenvector for two different shifted problems resulting from (1.6) with respect to τ. The red dots correspond to the error when no ideal approximation is used at all (i.e. $\tau = 0$ and therefore $\hat{P} = \hat{R}^T$), whereas blue and green dots represent the error of interpolation for $\tau = 0.5$ and $\tau = 1$ respectively. The legend for each color associates the percentage τ with the average number of non-zero entries $\frac{\text{nnz}}{n}$ in the resulting interpolation operator. Because the subspace $\hat{S}X_i$ grows with τ, larger values of τ leads to denser interpolation operators. For both shifts, the portion of the spectrum for which the least-squares minimization interpolation operator is the most accurate corresponds to the smallest eigenvalues in magnitude. This feature is an expected and desired effect of generating the set of test vectors from the polynomial smoother introduced in Section 2. However, the interpolation error increases with the shift. Therefore, the ideal approximation correction becomes necessary as the problem gets more indefinite. In particular, Figure 3.3 shows that the interpolation error decreases as more columns of $\hat{S}$ are added to approximate the ideal interpolation operator. One drawback of this gain in accuracy is the fill-in of the matrix.

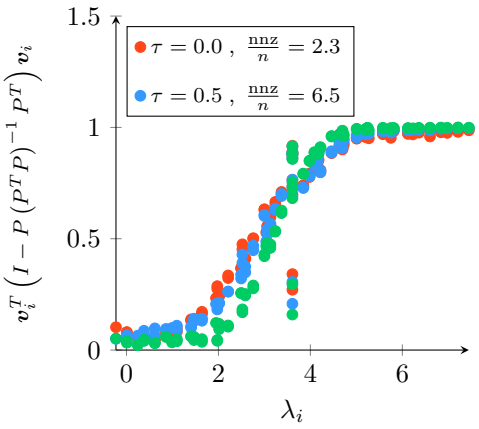


Fig. 3.1: $kh = 0.625$

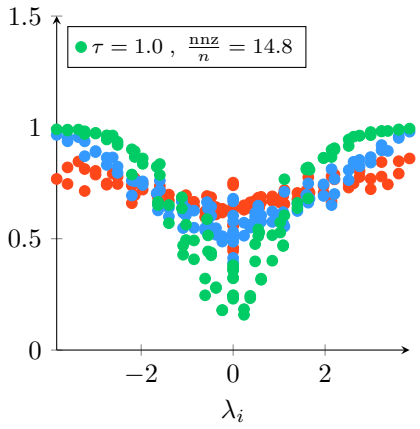


Fig. 3.2: $kh = 2.0$

Fig. 3.3: Error of interpolation with respect to the shift and sparsity

4. Alteration of the coarse correction in the indefinite case. While both smoothers and interpolation operators are now designed to face two inconvenient properties of the Helmholtz equation, signed eigenvalues and oscillatory near-kernel space, the effectiveness of the classical coarse correction is not guaranteed in an indefinite context. Worse still, the classical coarse correction can amplify the error associated with small eigenvectors although $\hat{P}$ has good approximation properties. Before discussing an alternative coarse correction, let us highlight how the matrix indefiniteness can corrupt the classical coarse correction with a simple illustration.

The Figure 4.3 plots the smallest eigenvector of a two-dimensional shifted Laplacian matrix in blue for two different shifts. The shift of 4.2 is greater than the shift of 4.1. As expected, the higher the shift, the more oscillatory the problem. In red are plotted the results of the coarse correction when applied to the blue eigenvectors. In this example, the coarse correction is implemented with the reduction-based interpolation operator introduced in Section 3. Additionally, the green curves represent the best representation of both eigenvectors in the interpolation range. First, note that the blue and green curves align almost perfectly in both sub-figures, which means that the interpolation range introduced in Section 3 offers a good approximation to the potentially oscillatory smallest eigenvector. In both cases, $\hat{P}$ has good approximation properties. In Figure 4.1, where the problem is discretized with 10 points per wavelength, the red coarse correction vector is relatively close to the blue eigenvector. The slight difference between both is only a matter of amplitude. In contrast, while the oscillations of the coarse correction vector illustrated in Figure 4.2 are synchronized with the oscillations of the smallest eigenvector, its direction is reversed. In that case, while the interpolation range is almost perfect, the error of the smallest eigenvector is not reduced by the coarse correction, but amplified.
At this stage, let us define a concept of pollution to better understand how the matrix indefiniteness can corrupt the coarse correction.

THEOREM 4.1. *Let A be an $n \times n$ matrix, and V its orthonormal set of eigenvectors, each associated with the corresponding element of the diagonal eigenvalue matrix Λ. Also, let P be an $n \times n_c$ interpolation operator. Assuming $V_c^T P$ is non-singular, we write the linear decomposition of the post-scaled interpolation operator as $P\left(V_c^T P\right)^{-1} = VK$, where K is the following $n \times n_c$ matrix of coefficients*

$$K := V^T P (V_c^T P)^{-1} = \begin{bmatrix} I_c \\ K_f \end{bmatrix}. \tag{4.1}$$

The block I_c corresponds to the identity matrix of size $n_c \times n_c$, and the block K_f is a $n_f \times n_c$ matrix such that $K_f := V_f^T P (V_c^T P)^{-1}$. The interpolation error of the eigenvector $\boldsymbol{v}_i$ of V_c is given by

$$\boldsymbol{v}_i^T \left(I - \Pi(P)\right) \boldsymbol{v}_i = 1 - \left[\left(I_c + K_f^T K_f\right)^{-1}\right]_{i,i}, \tag{4.2}$$

where $[\cdot]_{j,k}$ denotes the entry (j,k) of the bracketed matrix.

Proof. First, note that post-multiplying P by any non-singular matrix M_c of size $n_c \times n_c$ does not change the l_2-projection

$$\begin{aligned} (PM_c)((PM_c)^T(PM_c))^{-1}(PM_c)^T &= PM_cM_c^{-1}(P^TP)^{-1}M_c^{-T}M_c^TP^T \\ &= P(P^TP)^{-1}P^T = \Pi\left(P\right). \end{aligned} \tag{4.3}$$

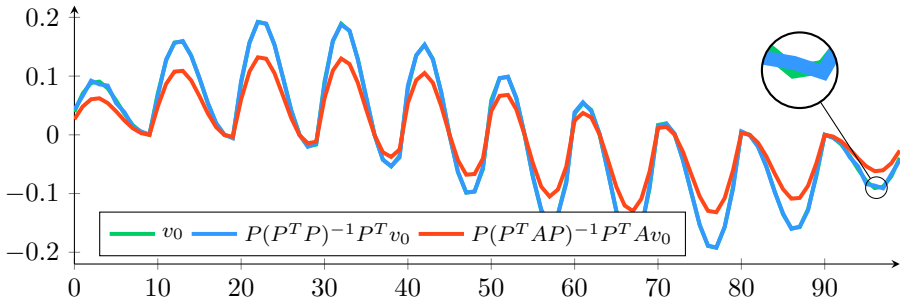


Fig. 4.1: $kh = 0.625$

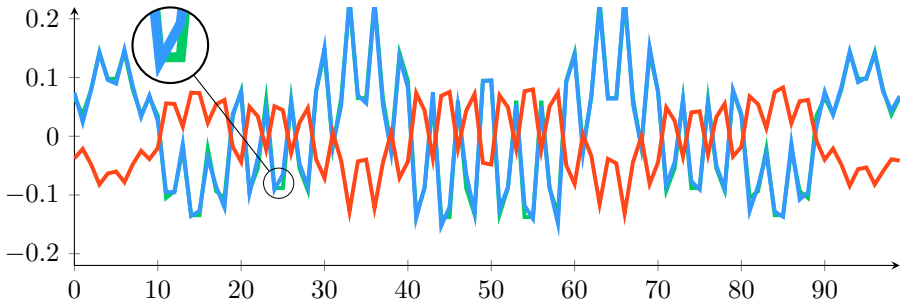


Fig. 4.2: $kh = 1.71$

Fig. 4.3: Layering of : $\boldsymbol{v}_1$ (blue) vs. $P(P^TP)^{-1}P^T\boldsymbol{v}_1$ (green) vs. $P(P^TAP)^{-1}P^TA\boldsymbol{v}_1$ (red), for two different shifts

In particular for $M_c = (V_c^TP)^{-1}$,

$$I - \Pi(P) = I - P(P^TP)^{-1}P^T$$

$$= I - P(V_c^TP)^{-1}\left(P(V_c^TP)^{-1}\right)^T P(V_c^TP)^{-1}\Big)^{-1}\left(P(V_c^TP)^{-1}\right)^T. \tag{4.4}$$

Since $P(V_c^TP)^{-1} = VK$, it follows that

$$I - \Pi(P) = I - (VK)((VK)^T(VK))^{-1}(VK)^T$$

$$= I - VK(K^TK)^{-1}K^TV^T. \tag{4.5}$$

For any eigenvector $\boldsymbol{v}_i$ of A, let $\boldsymbol{e}_i := V^T\boldsymbol{v}_i$ be the canonical unit vector with a one at the i^{th} position and zero elsewhere. Assuming $\boldsymbol{v}_i \in V_c$ $(i \le n_c)$, the vector $\boldsymbol{c}_i := K^T\boldsymbol{e}_i$ of size n_c is also a unit vector with a one at the i^{th} position. Consequently, the damping factor of $\boldsymbol{v}_i \in V_c$ is

$$\boldsymbol{v}_i^T(I - \Pi(P))\boldsymbol{v}_i = \boldsymbol{v}_i^TV(I - K(K^TK)^{-1}K^T)V^T\boldsymbol{v}_i$$

$$= \boldsymbol{e}_i^T(I - K(K^TK)^{-1}K^T)\boldsymbol{e}_i$$

$$= 1 - \boldsymbol{c}_i^T(K^TK)^{-1}\boldsymbol{c}_i = 1 - \left[(I_c + K_f^TK_f)^{-1}\right]_{i,i}. \tag{4.6}$$ ∎

Since the l_2-projection is unchanged by post-multiplication of P, we assume for what follows that K has the form (4.1). The block K_f designates what we call "pollution". This block of pollution causes the slight difference between an eigenvector

$\boldsymbol{v}_i$ of V_c and its best representation in the range of P. When a column of K_f is null, the interpolation error of the associated eigenvector equals zero, such that blue and green curves align perfectly. In practice however, this property is unlikely to be satisfied for Helmholtz, because P should be sparse for cost considerations and the smallest eigenvectors are usually unknown. Moreover, the near-kernel space of the Helmholtz equation is oscillatory. This makes the construction of good interpolation rules more difficult, and tends to pollute the interpolation range. In fact, this pollution is probably unavoidable and the columns of K_f are unlikely to be zero. While the pollution decreases the convergence speed of multigrid methods for SPD problems, we demonstrate that it can corrupt the coarse correction and make the method diverge in the indefinite case, as illustrated by the reversed red vector of Figure 4.3(b).

In that direction, let us discuss the effectiveness of the coarse correction by looking at the contraction of the n_c small eigenvectors V_c only, assuming the n_f large eigenvectors V_f are damped by the smoother.

THEOREM 4.2. *Define A and P as in the setting of Theorem 4.1. Also, let the matrix K be defined as in (4.1). The contraction of an eigenvector $\boldsymbol{v}_i$ of V_c after the coarse correction is given by*

$$\boldsymbol{v}_i^T E \boldsymbol{v}_i = 1 - \lambda_i \left[\left(\Lambda_c + K_f^T \Lambda_f K_f \right)^{-1} \right]_{i,i}. \tag{4.7}$$

Proof. By the same reasoning of the proof for Theorem 4.1, we note that post-multiplying P by any non-singular matrix M_c of size $n_c \times n_c$ does not change the coarse correction

$$\begin{aligned} (PM_c)((PM_c)^T A (PM_c))^{-1} (PM_c)^T &= PM_c M_c^{-1} (P^T A P)^{-1} M_c^{-T} M_c^T P^T \\ &= P (P^T A P)^{-1} P^T \end{aligned} \tag{4.8}$$

In particular for $M_c = (V_c^T P)^{-1}$,

$$\begin{aligned} E &= I - P(P^T A P)^{-1} P^T A \\ &= I - P(V_c^T P)^{-1} \left(P(V_c^T P)^{-1})^T A P (V_c^T P)^{-1} \right)^{-1} (P(V_c^T P)^{-1})^T A. \end{aligned} \tag{4.9}$$

Similar to (4.5), the equality $P(V_c^T P)^{-1} = VK$ leads to

$$E = I - (VK)((VK)^T A (VK))^{-1} (VK)^T A = V(I - K(K^T \Lambda K)^{-1} K^T \Lambda) V^T. \tag{4.10}$$

Define the euclidean basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{c}_i$ as in the proof of Theorem 4.1. Subsequently, the contraction of $\boldsymbol{v}_i \in V_c$ is

$$\begin{aligned} \boldsymbol{v}_i^T E \boldsymbol{v}_i &= \boldsymbol{v}_i^T V (I - K(K^T \Lambda K)^{-1} K^T \Lambda) V^T \boldsymbol{v}_i \\ &= \boldsymbol{e}_i^T (I - K(K^T \Lambda K)^{-1} K^T \Lambda) \boldsymbol{e}_i \\ &= 1 - \lambda_i \boldsymbol{c}_i^T (K^T \Lambda K)^{-1} \boldsymbol{c}_i = 1 - \lambda_i \left[\left(\Lambda_c + K_f^T \Lambda_f K_f \right)^{-1} \right]_{i,i}. \end{aligned} \tag{4.11}$$

∎

Theorem 4.2 shows that the damping factors rely on a combination of the small eigenvalues Λ_c plus the large eigenvalues Λ_f, such that the mix is given by the entries of the pollution K_f.

The effectiveness of the coarse correction is well-known in the SPD case. If all eigenvalues are positives, one can remark that

$$\forall i \leq n_c \ , \ 0 \leq \left[\left(\Lambda_c + K_f^T \Lambda_f K_f\right)^{-1}\right]_{i,i} \leq \left[\Lambda_c^{-1}\right]_{i,i} = \lambda_i \Rightarrow 0 \leq \boldsymbol{v}_i^T E \boldsymbol{v}_i \leq 1. \tag{4.12}$$

Hence, the coarse correction always operates a contraction on $\boldsymbol{v}_i$ regardless the block of pollution K_f. In the indefinite case however, the property (4.12) does not hold. In fact, a necessary condition for the coarse correction to be a contraction is

$$\forall i \leq n_c \ , \ \left|\boldsymbol{v}_i^T E \boldsymbol{v}_i\right| \leq 1 \Rightarrow 0 \leq \lambda_i \left[\left(\Lambda_c + K_f^T \Lambda_f K_f\right)^{-1}\right]_{i,i} \leq 2. \tag{4.13}$$

From Equation (4.13), it follows that each diagonal entry must have the same sign as its associated eigenvalue, and be smaller than twice the inverse of the eigenvalue in magnitude. Nothing guarantee such conditions to be satisfied in the case where small and large and either negative or positive eigenvalues are mixed. Especially for very small eigenvalues, the mix can easily lead to a diagonal entry of the opposite sign even though K_f is small, because its entries are weighted by the large eigenvalues Λ_f. Therefore, a good interpolation operator can still cause the coarse correction to amplify the error. For very near-zero eigenvalues, even a round-off error can eventually lead to divergence in the indefinite case. The following example better depicts how the pollution can cause divergence in the indefinite setting for a 2×2 matrix.

EXAMPLE 4.3. *Let A be a 2×2 matrix, and $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ its eigenvectors respectively associated with eigenvalues $|\lambda_1| < |\lambda_2|$. Let P be an interpolation operator of size 2×1 targeting the smallest eigenvector $\boldsymbol{v}_1$, such that*

$$P = \boldsymbol{v}_1 + \epsilon \boldsymbol{v}_2. \tag{4.14}$$

From definition (4.1), the K matrix can be derived by

$$K = V^T P \left(\boldsymbol{v}_1^T P\right)^{-1} = \left[\boldsymbol{v}_1, \boldsymbol{v}_2\right]^T \cdot \left[\boldsymbol{v}_1 + \epsilon \times \boldsymbol{v}_2\right] = \begin{bmatrix} 1 \\ \epsilon \end{bmatrix}. \tag{4.15}$$

From Theorem 4.2, the action of the coarse correction on $\boldsymbol{v}_1$ is given by

$$\boldsymbol{v}_1^T E \boldsymbol{v}_1 = 1 - \lambda_1 \left[\left(\Lambda_c + K_f^T \Lambda_f K_f\right)^{-1}\right]_{1,1} = 1 - \frac{\lambda_1}{\lambda_1 + \epsilon^2 \lambda_2} \tag{4.16}$$

The figure 4.4 depicts the action of the coarse correction on $\boldsymbol{v}_1$ with respect to the pollution block $K_f^T \Lambda_f K_f = \epsilon^2 \lambda_2$. A first observation is that the coarse correction does not amplify the smallest eigenvector if eigenvalues have the same sign. If the eigenvalues are oppositely signed, then the coarse correction amplifies $\boldsymbol{v}_1$ for $\epsilon^2 \lambda_2 < -\lambda_1/2$. Therefore, the condition on the pollution $K_f = \epsilon$ that drives the error of interpolation is particularly difficult respectively for small and large values of λ_1 and λ_2.

The next theorem derives a more general condition for the coarse correction to be a contraction of the smallest eigenvalues in the indefinite case based on the concept of pollution.

THEOREM 4.4. *If A is indefinite, then*

$$\left|\lambda_{n_c}\left(K_f^T \Lambda_f K_f\right)\right| \leq \frac{1}{2}|\lambda_1| \ \Rightarrow \ \forall \boldsymbol{v}_i \in V_c \ , \ \left|\boldsymbol{v}_i^T E \boldsymbol{v}_i\right| \leq 1 \tag{4.17}$$

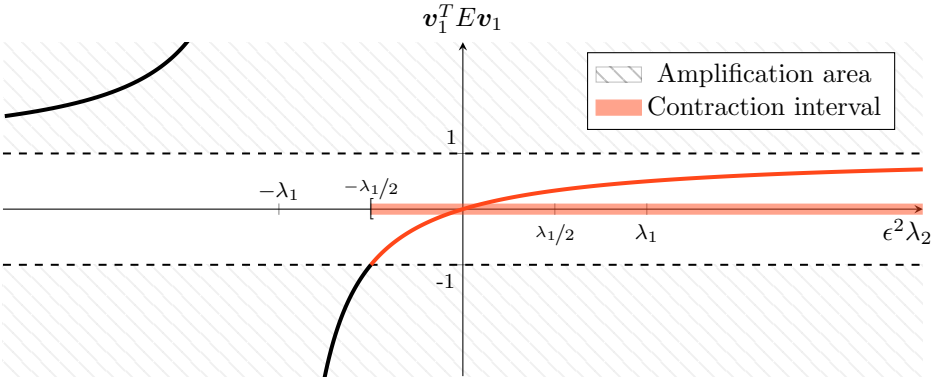


Fig. 4.4: Contraction of the coarse correction with respect to the pollution

Proof. Define $M_K = I_c + \Lambda_c^{-1} K_f^T \Lambda_f K_f$. From the shape of the matrix K defined in Equation (4.10), we have

$$\begin{aligned} V_c^T E V_c &= V_c^T V (I - K(K^T \Lambda K)^{-1} K^T \Lambda) V^T V_c \\ &= I_c - (K^T \Lambda K)^{-1} \Lambda_c \\ &= I_c - (I_c + \Lambda_c^{-1} K_f^T \Lambda_f K_f)^{-1} \Lambda_c^{-1} \Lambda_c \\ &= I_c - M_K^{-1}. \end{aligned} \tag{4.18}$$

Hence, it follows that

$$\forall \boldsymbol{v}_i \in V_c \ , \ \boldsymbol{v}_i^T E \boldsymbol{v}_i = \boldsymbol{e}_i^T V_c^T E V_c \boldsymbol{e}_i = 1 - \boldsymbol{e}_i^T M_K^{-1} \boldsymbol{e}_i. \tag{4.19}$$

where $\boldsymbol{e}_i$ is the i^{th} vector of the euclidean basis in $\mathbb{R}^{n_c}$. Therefore, $|\boldsymbol{v}_i^T E \boldsymbol{v}_i| \leq 1$ if

$$\forall \boldsymbol{v}_i \in V_c \ , \ -1 \leq \boldsymbol{v}_i^T E \boldsymbol{v}_i \leq 1 \ \Leftrightarrow \ 0 \leq \boldsymbol{e}_i^T M_K^{-1} \boldsymbol{e}_i \leq 2. \tag{4.20}$$

We begin by deriving a condition for the right bound of (4.20), and will show, in a second time, that it also satisfies the left one. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ two vectors of $\mathbb{R}^n$ linked by the relation $\boldsymbol{x} = M_k \boldsymbol{y}$. The right bound is satisfied if

$$\max_{x \neq 0} \frac{||M_K^{-1} \boldsymbol{x}||}{||\boldsymbol{x}||} = \max_{y \neq 0} \frac{||\boldsymbol{y}||}{||M_K \boldsymbol{y}||} = \left(\min_{y \neq 0} \frac{||M_K \boldsymbol{y}||}{||\boldsymbol{y}||} \right)^{-1} \leq 2. \tag{4.21}$$

Therefore, the condition (4.21) is equivalent to

$$\min_{y \neq 0} \frac{||M_K \boldsymbol{y}||}{||\boldsymbol{y}||} \geq \frac{1}{2}. \tag{4.22}$$

Let $\sigma_i(M)$ be the i^{th} largest singular value of a given matrix M (we omit the matrix between parenthesis when referring to the singular values of the initial matrix A). In addition, let us recall the following triangle inequality $||\boldsymbol{y} + \boldsymbol{z}|| \geq ||\boldsymbol{y}|| - ||\boldsymbol{z}||$, $\forall \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^{n_c}$. Thus, we have that

$$\begin{aligned} \min_{y \neq 0} \frac{||M_K \boldsymbol{y}||}{||\boldsymbol{y}||} = \min_{y \neq 0} \frac{||\boldsymbol{y} + \Lambda_c^{-1} K_f^T \Lambda_f K_f \boldsymbol{y}||}{||\boldsymbol{y}||} &\geq \min_{y \neq 0} \left(1 - \frac{||\Lambda_c^{-1} K_f^T \Lambda_f K_f \boldsymbol{y}||}{||\boldsymbol{y}||} \right) \\ &= 1 - \max_{y \neq 0} \frac{||\Lambda_c^{-1} K_f^T \Lambda_f K_f \boldsymbol{y}||}{||\boldsymbol{y}||} \\ &= 1 - \sigma_{n_c} \left(\Lambda_c^{-1} K_f^T \Lambda_f K_f \right). \end{aligned} \tag{4.23}$$

It follows that the condition (4.22) is satisfied if $\sigma_{n_c}(\Lambda_c^{-1}K_f^T\Lambda_f K_f) \leq \frac{1}{2}$. Finally, since $\sigma_{n_c}(\Lambda_c^{-1}K_f^T\Lambda_f K_f) \leq \sigma_1^{-1}\sigma_{n_c}(K_f^T\Lambda_f K_f)$ and the singular values coincide with eigenvalues in magnitude because both Λ_c and $K_f^T\Lambda_f K_f$ are hermitian, the right bound of (4.20) is satisfied if

$$|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)| \leq \frac{1}{2}|\lambda_1|. \tag{4.24}$$

We now address the left bound of (4.20) assuming the condition (4.24) holds. Our goal is to prove that all diagonal entries of M_K^{-1} are positives. In that direction, let $F(M)$ be the field of values of a given matrix M of size n_c such that

$$F\left(M\right) := \left\{\boldsymbol{x}^* M\boldsymbol{x} \mid \forall \boldsymbol{x} \in \mathbb{C}^{n_c}\ ,\ \boldsymbol{x}^*\boldsymbol{x} = 1\right\}. \tag{4.25}$$

If M is hermitian, one can show that (e.g. [14, chapter 4])

$$\min_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^* M\boldsymbol{x} = \lambda_{\min}\left(M\right) \text{ and } \max_{\boldsymbol{x}^*\boldsymbol{x}=1} \boldsymbol{x}^* M\boldsymbol{x} = \lambda_{\max}\left(M\right). \tag{4.26}$$

Accordingly, let $F(\Lambda_c)$ and $F(K_f^T\Lambda_f K_f)$ be the field of values of Λ_c and $K_f^T\Lambda_f K_f$ respectively. Since A is non-singular, then $0 \notin F\left(\Lambda_c\right)$. Therefore, the spectrum of $\Lambda_c^{-1}K_f^T\Lambda_f K_f$ is included as follows (e.g. [13, chapter 1])

$$\forall j \leq n_c\ ,\ \lambda_j\left(\Lambda_c^{-1}K_f^T\Lambda_f K_f\right) \in F\left(K_f^T\Lambda_f K_f\right)/F\left(\Lambda_c\right). \tag{4.27}$$

The set ratio in (4.27) has the usual algebraic interpretation such that

$$\forall \alpha \in \frac{F\left(K_f^T\Lambda_f K_f\right)}{F\left(\Lambda_c\right)}\ ,\ -\frac{\max\limits_{\boldsymbol{x}^*\boldsymbol{x}=1}\left|\boldsymbol{x}^* K_f^T\Lambda_f K_f\boldsymbol{x}\right|}{\min\limits_{\boldsymbol{x}^*\boldsymbol{x}=1}\left|\boldsymbol{x}^*\Lambda_c\boldsymbol{x}\right|} \leq \alpha \leq \frac{\max\limits_{\boldsymbol{x}^*\boldsymbol{x}=1}\left|\boldsymbol{x}^* K_f^T\Lambda_f K_f\boldsymbol{x}\right|}{\min\limits_{\boldsymbol{x}^*\boldsymbol{x}=1}\left|\boldsymbol{x}^*\Lambda_c\boldsymbol{x}\right|}. \tag{4.28}$$

Furthermore, matrices Λ_c and $K_f^T\Lambda_f K_f$ are hermitian so the property (4.26) holds for both of them. Because the spectrum belongs to the set ratio as in (4.27), we have

$$-|\lambda_1|^{-1}\cdot\left|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)\right| \leq \lambda_j\left(\Lambda_c^{-1}K_f^T\Lambda_f K_f\right) \leq \left|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)\right|\cdot|\lambda_1|^{-1}. \tag{4.29}$$

Therefore, assuming the condition (4.24) is satisfied, it follows

$$\lambda_j\left(\Lambda_c^{-1}K_f^T\Lambda_f K_f\right) \geq -\left|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)\right| \times |\lambda_1|^{-1} \geq -\frac{1}{2}. \tag{4.30}$$

Adding one to each member of the inequality (4.30) finally gives

$$\lambda_j\left(M_K\right) = \lambda_j\left(I + \Lambda_c^{-1}K_f^T\Lambda_f K_f\right) \geq \frac{1}{2} \tag{4.31}$$

The condition (4.24) implies that all eigenvalues of M_K are positives. Subsequently, $\det\left(M_K\right) > 0$. The adjugate formula for the inverse of M_K shows that diagonal entries are positives if the determinant of principal sub-matrices are also positives. In that direction, denote by $\left[\ \cdot\ \right]_{\Omega_{-i}}$ the principal sub-matrix obtained by deleting the i^{th} row and column of a matrix. Since Λ_c is diagonal, one can show that

$$\left[\Lambda_c^{-1}K_f^T\Lambda_f K_f\right]_{\Omega_{-i}} = \left[\Lambda_c\right]_{\Omega_{-i}}^{-1}\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}. \tag{4.32}$$

As in Equation (4.27), the spectrum is included such that

$$\forall j \leq n_c - 1\ ,\ \lambda_j\left([\Lambda_c]_{\Omega_{-i}}^{-1}\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right) \in F\left(\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right)/F\left([\Lambda_c]_{\Omega_{-i}}\right),$$

and therefore the following bound holds

$$\lambda_j\left([\Lambda_c]_{\Omega_{-i}}^{-1}\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right) \geq -\left|\lambda_{n_c-1}\left(\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right)\right| \times |\lambda_1|^{-1}. \tag{4.33}$$

The matrix $K_f^T\Lambda_f K_f$ being hermitian, Cauchy's interlace theorem states that

$$\lambda_j\left(K_f^T\Lambda_f K_f\right) \leq \lambda_j\left(\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right) \leq \lambda_{j+1}\left(K_f^T\Lambda_f K_f\right),\ j = 1, \cdots, n_c - 1. \tag{4.34}$$

As a consequence, and from the inequality (4.30), we have

$$\lambda_j\left([\Lambda_c]_{\Omega_{-i}}^{-1}\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right) \geq -\left|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)\right| \times |\lambda_1|^{-1} \geq -\frac{1}{2}. \tag{4.35}$$

Hence, eigenvalues of principal sub-matrices also satisfy

$$\lambda_j\left([M_K]_{\Omega_{-i}}\right) = \lambda_j\left(I_{n_c-1} + [\Lambda_c]_{\Omega_{-i}}^{-1}\left[K_f^T\Lambda_f K_f\right]_{\Omega_{-i}}\right) \geq \frac{1}{2}. \tag{4.36}$$

Because eigenvalues of the principal sub-matrices are positives, so are the determinants. From the adjugate formula of M_K^{-1}, it follows that

$$\boldsymbol{e}_i^T M_K^{-1}\boldsymbol{e}_i = \left[M_K^{-1}\right]_{i,i} = \frac{\det\left([M_K]_{\Omega_{-i}}\right)}{\det\left(M_K\right)} \geq 0,\ i = 1, \cdots, n_c \tag{4.37}$$

As a consequence, both left and right bounds of (4.20) are satisfied. Finally,

$$\left|\lambda_{n_c}\left(K_f^T\Lambda_f K_f\right)\right| \leq \frac{1}{2}|\lambda_1| \ \Rightarrow\ \forall \boldsymbol{v}_i \in V_c\ ,\ \left|\boldsymbol{v}_i^T E \boldsymbol{v}_i\right| \leq 1 \tag{4.38}$$

∎

The condition provided by Theorem 4.4 is that the amplitude of the block $K_f^T\Lambda_f K_f$ never exceeds half of the smallest eigenvalue in magnitude. No assumption can be made on the sign of eigenvalues in the indefinite case, so that the condition prevents the coarse correction from amplifying the error in the case where eigenvalues are oppositely signed. Applied to the previous example 4.4, Theorem 4.4 states that $|\epsilon^2\lambda_2| < |\lambda_1|/2$. That said, the condition is extremely strict and probably impossible to satisfy in practice for very small eigenvalues. In a practical method, the block K_f will never be sufficiently small for solving all type of indefinite problems because of a potentially very near-zero eigenvalue. As illustrated by Figure 4.3, a good interpolation operator with small K_f can still cause divergence although it satisfies good approximation properties. The classical coarse correction is hopeless for indefinite problems.

5. Alternative coarse correction for indefinite problems. As discussed in the previous section, the classical coarse correction is not equivalent to a minimization problem in the indefinite case, and improving P will never be enough to remedy this loss of equivalence. Moreover, because the interpolation operator developed in Section 3 targets the smallest eigenvectors of each level, every coarser matrix is more indefinite than its fine parent. Then, as the number of coarse levels increases, the

balance between negative and positive eigenvalues reaches an equilibrium, and makes the effectiveness of the classical coarse correction difficult to predict. Nevertheless, Figure 4.3 shows that the interpolation operator has good approximation properties for the oscillatory near-kernel space. In particular, the Figure 4.3(b) suggests that only the direction of the coarse correction vector has to be changed; the shape is correct. Hence, a coarse correction that amplifies or flips the smallest eigenvectors can still provide pertinent information for solving the system. In this section, we propose to minimize the approximation error in a proper norm for indefinite problems and within a space composed of vectors returned by the classical coarse correction. Moreover, to decrease the eigenvector pollution, each coarse correction vector is smoothed by the polynomial smoother of Section 2.

5.1. Notations and general considerations on GMRES. The *Generalized Minimal RESidual* (GMRES) method [24] approximates the solution in a Krylov subspace by minimizing the residual in the Euclidean norm. The method can solve any class of matrix system since the norm is valid independent of the context, which is of particular interest for the indefinite case. Let us first define some notation before introducing the alternative coarse correction. Let W_p be the $n \times p$ rectangular matrix containing the p orthonormalized Krylov vectors such that

$$\text{range}\,(W_p) = \text{span}\left\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}, \ldots, A^{p-1}\boldsymbol{b}\right\}. \tag{5.1}$$

Each column of W_p is orthonormalized following a Gram-Schmidt process. The coefficients of the orthonormalization are stored in the rectangular Hessenberg matrix $\bar{H}_p$ of size $p+1 \times p$. The square matrix H_p is of size $p \times p$ and obtained from $\bar{H}_p$ by deleting its last row. Both matrices W_p and H_p are linked by

$$AW_p = W_{p+1}\bar{H}_p \text{ and } W_p^T A W_p = H_p, \tag{5.2}$$

which leads to the following equality

$$\begin{aligned}\underset{\tilde{\boldsymbol{x}}\in\text{range}(W_p)}{\arg\min} \|\boldsymbol{b} - A\tilde{\boldsymbol{x}}\|_2 &= \underset{\boldsymbol{\rho}_m\in\mathbb{C}^p}{\arg\min}\|\boldsymbol{b} - AW_p\boldsymbol{\rho}_p\|_2 = \underset{\boldsymbol{\rho}_p\in\mathbb{C}^p}{\arg\min}\|W_p^T\boldsymbol{b} - H_p\boldsymbol{\rho}_p\|_2 \\ &= W_p H_p^{-1} W_p^T \boldsymbol{b}. \end{aligned} \tag{5.3}$$

In practice, GMRES takes advantage of the convenient Hessenberg shape of $\bar{H}_p$ to construct an upper triangular matrix by applying Given's rotations. The minimization of the residual then relies on a backward substitution. The relation (5.2) can be generalized [7] to any arbitrary subspace $W_p = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p]$ such that

$$\underset{\tilde{\boldsymbol{x}}\in\text{range}(W_p)}{\arg\min} ||\boldsymbol{b} - A\tilde{\boldsymbol{x}}||_2 = W_p H_p^{-1} Z_p^T \boldsymbol{b} \text{ with } AW_p = Z_p H_p \tag{5.4}$$

and where Z_p denotes the orthonormalized basis of AW_p. Note that the Arnoldi relation (5.4) does not define any particular recurrence relation since W_p is arbitrary and not necessarily designed by successive matrix vector products. In addition, the only matrix that needs to be orthonormal in the generalized setting is Z_p.

5.2. Minimization within a space of coarse correction vectors. As mentioned in the introduction of this section, the interpolation operator has good approximation properties for the oscillatory near-kernel space. Even though the small eigenvectors that constitute each coarse correction vector are likely to be oriented in

the wrong direction or amplified because of the pollution effect introduced in Section 4, they still provide useful information about the near-kernel space. For ease of discussion, we present this idea on a two-level method. The multi-level case will be depicted in the next section dedicated to numerical experiments.

In that direction, let W_i be the set of coarse correction vectors of the i^{th} iteration linked by the Arnoldi relation (5.4) with its orthonormal counterpart Z_i. Accordingly, let $\boldsymbol{w}_j \in W_i$ and $\boldsymbol{z}_j \in Z_i$ denote the j^{th} vectors of the set W_i and Z_i respectively. At each iteration i, the classical coarse correction returns a new coarse correction vector that is smoothed by the Chebyshev polynomial smoother presented in Section 2. This new smoothed coarse correction vector is therefore added to the previous set such that

$$W_i = [W_{i-1}\ ,\ \boldsymbol{w}_i] \text{ with } \boldsymbol{w}_i = q_{m+1}^{\nu}(A^2)\Pi_A(P)\,\boldsymbol{r}^{(i)}, \tag{5.5}$$

where $\boldsymbol{r}^{(i)}$ designates the residual at the i^{th} iteration. From the Arnoldi relation (5.4), we have

$$H_i = Z_i^T A W_i = H_i^{-T} W_i^T A^T A W_i\ ,\ Z_i = A W_i H_i^{-1}. \tag{5.6}$$

Hence, solving the minimization problem (5.4) is equivalent to solve the normal equations within the subspace spanned by W_i

$$\begin{aligned} W_i H_i^{-1} Z_i^T A &= W_i \left(W_i^T A^T A W_i\right)^{-1} H_i^T Z_i^T A \\ &= W_i \left(W_i^T A^T A W_i\right)^{-1} W_i^T A^T A = \Pi_{A^TA}(W_i)\,. \end{aligned} \tag{5.7}$$

The concept of pollution also drives convergence in the alternative setting. Section 4 demonstrated that the block K_f pollutes the range of P and therefore impacts the classical coarse correction. Because the minimization W_i resorts to the classical coarse correction by way of Equation (5.5), the block of pollution still impact the capture of the small eigenvectors. Resorting to euclidean norm in (5.4) prevents from divergence, but it also squares the eigenvalues of the initial problem because of the equivalence with an A^TA-orthogonal projection. This naturally increases the gap between small and large eigenvalues, and therefore decreases the contraction of the smallest over the largest.

Smoothing the classical coarse correction vectors by way of the polynomial $q_{m+1}^{\nu}(A^2)$ compensates this effect by decreasing the distribution of large eigenvectors in the minimization space. This idea of damping the large eigenvalues to reveal the smaller ones is also used to generate a relevant set of test vectors for the construction of the least-squares minimization operator introduced in Section 3. Once the coarse correction vector is smoothed and included in W_i, the set Z_i is extended as follows

$$Z_i = [Z_{i-1}\ ,\ \boldsymbol{z}_i] \text{ with } \boldsymbol{z}_i = \frac{1}{h_{i,i}}\left(A\boldsymbol{w}_i - \sum_{j=1}^{i-1} h_{j,i}\cdot \boldsymbol{z}_j\right), \tag{5.8}$$

where coefficients $h_{j,i}$ result from the orthogonalization process of the new vector $A\boldsymbol{w}_i$. Those coefficients are stored in the squared upper triangular matrix

$$H_i = \left[\begin{array}{ccc|c} & H_{i-1} & & \begin{matrix} h_{1,p} \\ \vdots \\ h_{i-1,i} \end{matrix} \\ \hline 0 & \cdots & 0 & h_{i,i} \end{array}\right] \text{ with } h_{j,i} = \begin{cases} \langle \boldsymbol{z}_j, \boldsymbol{z}_i\rangle & \text{if } j < i \\ ||\boldsymbol{z}_i||_2 & \text{if } j = i \end{cases}. \tag{5.9}$$

The algorithm 5.1 presents the alternative two-level cycle, and can be compared with the classic one in Algorithm 1.1.

Algorithm 5.1 Two-level cycle with the alternative coarse correction

Inputs : $\boldsymbol{b}$ right-hand side, $\tilde{\boldsymbol{x}}$ approximation of $\boldsymbol{x}$, $\boldsymbol{r} = b - A\tilde{\boldsymbol{u}}$ residual
A initial matrix, M smoother, P interpolation operator

for $j = 1, \nu$ **do**
$\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + p(A^2)\boldsymbol{r}$
$\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
end for
$\boldsymbol{r}_{\mathcal{C}} \leftarrow P^T\boldsymbol{r}$
$\boldsymbol{e}_{\mathcal{C}} \leftarrow \text{Solve}(P^TAP, \boldsymbol{r}_{\mathcal{C}})$
$\boldsymbol{w} \leftarrow q^{\nu}_{m+1}(A^2)P\boldsymbol{e}_{\mathcal{C}}$
$\hat{\boldsymbol{w}}, H_i \leftarrow \text{Orthonormalize}(\boldsymbol{w}, Z_{i-1})$
$W_i, Z_i \leftarrow [W_{i-1}\ ,\ \boldsymbol{w}], [Z_{i-1}\ ,\ \hat{\boldsymbol{w}}]$
for $j = 1, \nu$ **do**
$\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + p(A^2)\boldsymbol{r}$
$\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
end for
$\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} + W_iH_i^{-1}Z_i^T\boldsymbol{r}$
$\boldsymbol{r} \leftarrow \boldsymbol{b} - A\tilde{\boldsymbol{x}}$
Output : $\tilde{\boldsymbol{x}}$ approximation of $\boldsymbol{x}$ at the end of the cycle

EXAMPLE 5.1. *Let us pursue Example 4.3, where A is a 2×2 matrix, with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ its eigenvectors respectively associated with eigenvalues $|\lambda_1| < |\lambda_2|$. The interpolation operator P targets $\boldsymbol{v}_1$ as defined by (4.14). Let W_1 be the minimization space of dimension 1 constructed following (5.5) such that*

$$W_1 = q_{m+1}(A^2)\Pi_A(P)\boldsymbol{v}_1 = \frac{\lambda_1^2}{\lambda_1 + \epsilon^2\lambda_2}\left(q_{m+1}(\lambda_1^2)\boldsymbol{v}_1 + q_{m+1}(\lambda_2^2)\epsilon\boldsymbol{v}_2\right). \tag{5.10}$$

Furthermore, define E_{W_1} to be the error propagation matrix of the alternative coarse correction. One can show that

$$\boldsymbol{v}_1^T E_{W_1}\boldsymbol{v}_1 = \boldsymbol{v}_1^T\left(I - \Pi_{A^TA}(W_1)\right)\boldsymbol{v}_1 = 1 - \frac{q_{m+1}^2(\lambda_1^2)\lambda_1^2}{q_{m+1}^2(\lambda_1^2)\lambda_1^2 + q_{m+1}^2(\lambda_2^2)\epsilon^2\lambda_2^2}. \tag{5.11}$$

To simplify the discussion, let us assume that the smallest eigenvector is preserved by the smoother, such that $q_{m+1}(\lambda_1^2) = 1$. The following figure illustrates the contraction of $\boldsymbol{v}_1$ after applying the alternative coarse correction with respect to the pollution and the polynomial. As expected, the smoother increases the contraction and counter balance the squared large eigenvalue λ_2 that weights the pollution $K_f = \epsilon$ when minimizing in euclidean norm.

6. Numerical Experiments. In the following numerical experiments, the interval of the Chebyshev polynomial smoother is determined following the spectral density approximation method presented in Section 2.2. The number n_ν of coefficients μ_k in the moment matching procedure is fixed to 15, and n_{vec} fixed to 5. The degree m of the polynomial is 3. Regarding the construction of the interpolation operator, the number of smoothed test vectors is fixed to 15. Last, the number of interpolation points in the least square minimization strategy used to construct the coarse grid selection operator $\hat{R}^T$ never exceeds 4 (i.e., $\max_{i\in\mathcal{F}}\{\text{Card}(C_i)\} = 4$).

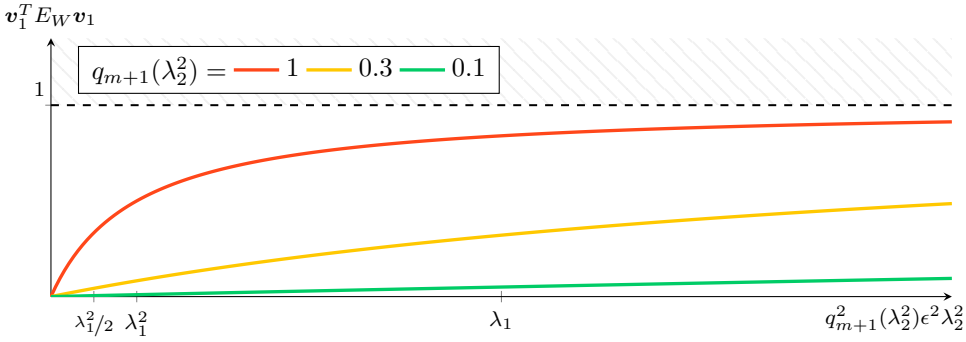


Fig. 5.1: Illustration of the contraction of a small eigenvector with respect to the pollution and the polynomial smoother

6.1. Two-level experiment on the Two Dimensional Shifted Laplacian. Let us first apply this new multigrid setting to the two-dimensional shifted laplacian problem associated with the stencil matrix (1.6). The size of the shifted laplacian matrix is fixed to $n = 100$. The following figures depict the number of iterations with respect to the shift kh and using either the classical or the alternative coarse correction. Recall that the matrix is the most indefinite (exact balance between negative and positive eigenvalues) when $kh = 2$, and that the near-kernel space becomes more oscillatory as kh increases. Those number of iterations are also presented with respect to the percentage τ that governs the number of selected columns of $\hat{S}$ in the approximation of ideal interpolation. The resulting operator complexity defined by $\phi := \frac{\sum_l \text{nnz}(A_l)}{\text{nnz}(A_0)}$ for different values of τ is provided by Table 6.1.

τ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ϕ	1.81	3.00	3.47	3.69	3.93	4.16	4.35	4.55	4.73	4.87	5.15

Table 6.1: Operators complexity of the two-level method with respect to τ

Last, the tolerance of the relative residual norm is set to 10^{-6}, and the maximal number of iterations is fixed to 100. Peak values of the standard multigrid setting on the left column denote divergence, whereas they stand for slow convergence in the alternative setting plotted on the right column. Both left figures 6.3 and 6.1 correspond to a two-level method built on the classical coarse correction respectively for $\nu = 2$ and $\nu = 4$. Whereas increasing the number of selected columns in $\hat{S}$ for approximating the ideal interpolation operator by way of the parameter τ generally helps the convergence, the method remains likely to diverge for the reasons explained in Section 4. Still, the best setting for the classical coarse correction is naturally $\tau = 1$ and $\nu = 4$. Certain divergence scenarios that happens for $\nu = 2$ (for instance around $kh = 0.8$) are fixed by doubling the number of smoothing iterations. Doing so improves the set of test vectors in approximating the near-kernel space, and therefore leads to a better least-squares minimization coarse variable operator that decreases the pollution K_f. It remains however impossible to derive a general setting that ensures the convergence of the standard method in all cases. Both right figures 6.2 and 6.4 represent the same experiment with the alternative coarse correction. The peaks around $kh = 2$ depict slow convergence situation where the relative residual norm is stuck around 10^{-5} because of very near-zero eigenvalues. Beside those extremely

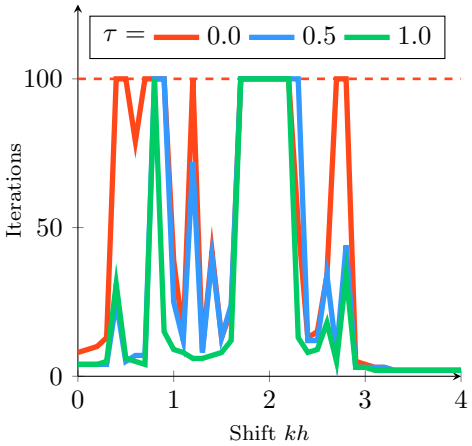


Fig. 6.1: Classical CC, $\nu = 2$

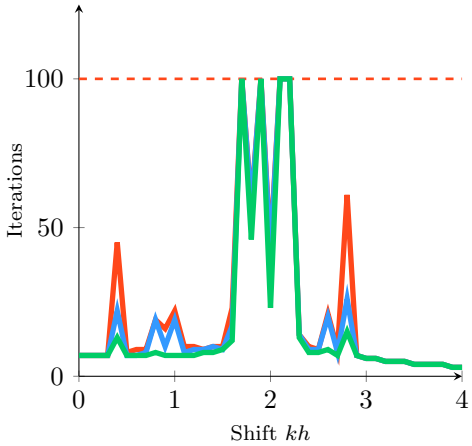


Fig. 6.2: Alternative CC, $\nu = 2$

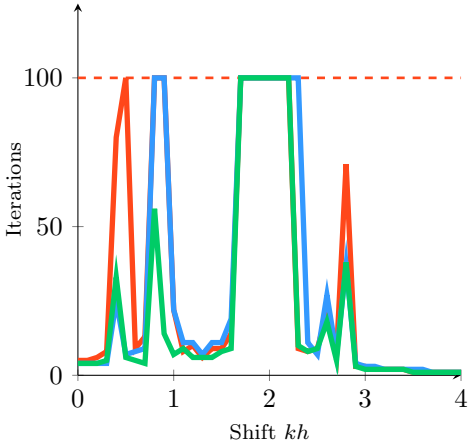


Fig. 6.3: Classical CC, $\nu = 4$

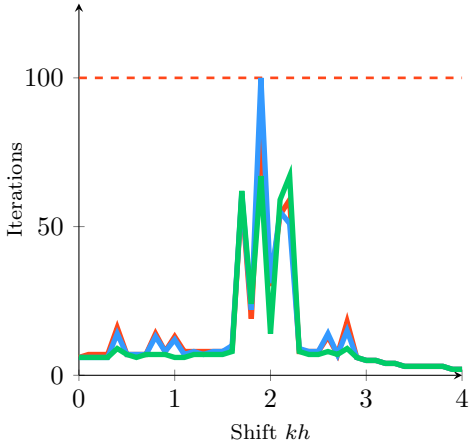


Fig. 6.4: Alternative CC, $\nu = 4$

Fig. 6.5: Number of iterations of two-level methods with respect to kh and τ

indefinite cases, the method converges in all cases. We also remark that the divergence of the standard method correlate with more iterations in the alternative setting. At the cost of complexity, increasing τ or ν provides a better convergence factor.

6.2. Multi-level experiment on the Two Dimensional Helmholtz problem with absorbing boundary conditions.
The following numerical experiments depict the convergence for a two dimensional Helmholtz problem using absorbing boundary conditions and with a discretization coefficient set to $kh = 0.625$ (i.e. 10 points per wavelength, where k corresponds to the wavenumber). Therefore, the discretization matrix is indefinite, complex and non-hermitian, and grows with k. As a consequence, the restriction operation is made through the transpose conjugate $\hat{P}^*$. Moreover, the squared matrix in the polynomial setting is replaced by A^*A. Also note that those numerical experiments result from the alternative coarse correction only, and that Z^T is replaced by Z^* in (5.4). The first benchmark illustrated by Figure 6.6 exposes the convergence of the method by fixing the number of selected column of $\hat{S}$ to the maximum (i.e., $\tau = 1$). Each curve corresponds to a method following its number of levels. The y-axis corresponds to

the number of iterations, while the wavenumber varies along the x-axis. The number of iterations is constant until the fourth level. The number of iterations of both the five-level and six-level methods increase with the wavenumber.

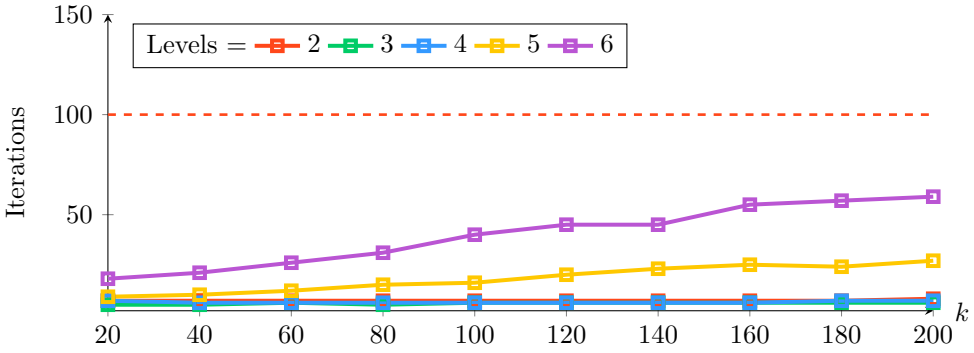


Fig. 6.6: Number of iterations following the wavenumber k, $\nu = 2$, $\tau = 1$

While setting $\tau = 1$ enables the method to converge almost constantly up to five levels, the operator complexity is too high for practical implementation. Therefore, the second benchmark exposes the number of iterations of a two-level method with respect to the parameter τ. Figure 6.7 shows that the plain least-squares minimization

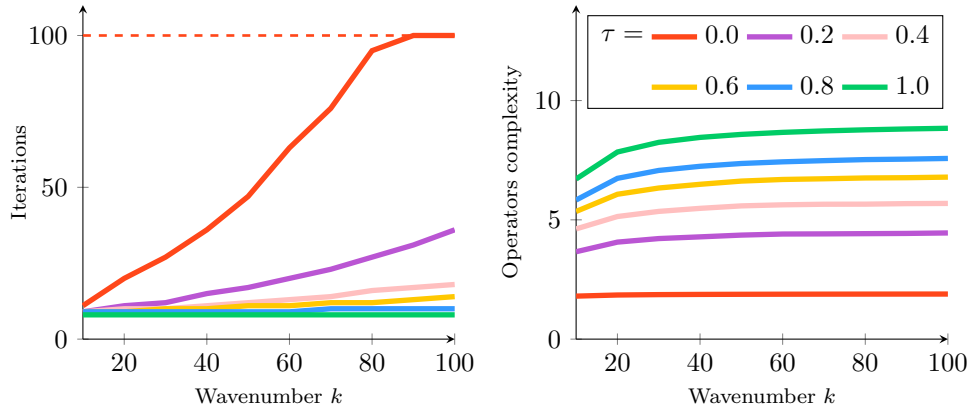


Fig. 6.7: Two-level method with alternative coarse correction - number of iterations and operators complexity with respect to k and τ, $\nu = 2$

operator (i.e. $\tau = 0$) is not a suitable choice as k is growing. Even though larger subspaces $\hat{S}X_i$ in the approximation of the ideal interpolation operator yields denser matrices, the number of iterations tends to size independence as τ grows. A trade-off between convergence and complexity may be possible depending on the problem-size. More generally, Figure 6.7 reveals how important is the role that plays the ideal approximation step in the convergence. A better sparsification strategy is a topic of further research.

7. Conclusions. Indefinite and oscillatory problems are difficult for multigrid methods. The negative eigenvalues require an adapted smoother, and the interpolator

should capture the oscillatory near-kernel space. More importantly, the coarse correction should be adapted to the indefiniteness of the initial matrix, which does not define a norm. The normal equation polynomial smoother is designed to target a desired proportion of eigenvalues according to their amplitude, and the range of our interpolator offers a good approximation of the near-kernel space despite its oscillations. The alternative coarse correction space proposed in the paper permits to minimize the global residual in a proper norm for indefinite problems, in a space approximating the set of smallest eigenvectors known to be difficult for most iterative methods. Finding a better trade-off between sparsity and accuracy of interpolation, and constructing a polynomial without resorting to normal equations will be important points in our future investigations.

REFERENCES

[1] M. F. Adams, M. Brezina, J. J. Hu, and R. S. Tuminaro, *Parallel multigrid smoothing: polynomial versus gauss–seidel*, Journal of Computational Physics, 188 (2003), pp. 593–610.

[2] A. H. Baker, R. D. Falgout, T. V. Kolev, and U. M. Yang, *Multigrid smoothers for ultra-parallel computing*, SIAM Journal on Scientific Computing, 33 (2011), pp. 2864–2887, https://doi.org/10.1137/100798806, https://doi.org/10.1137/100798806, https://arxiv.org/abs/https://doi.org/10.1137/100798806.

[3] A. Brandt, J. Brannick, K. Kahl, and I. Livshits, *Bootstrap amg*, SIAM Journal of Scientific Computing, 33 (2011), pp. 612–632, https://doi.org/10.1137/090752973.

[4] L. I. Brandt A., *Wave-ray multigrid method for standing wave equations.*, ETNA. Electronic Transactions on Numerical Analysis [electronic only], 6 (1997), pp. 162–181, http://eudml.org/doc/119506.

[5] M. Brezina, R. Falgout, S. MacLachlan, T. Manteuffel, S. McCormick, and J. Ruge, *Adaptive smoothed aggregation (αsa)*, SIAM Journal on Scientific Computing, 25 (2004), pp. 1896–1920, https://doi.org/10.1137/S1064827502418598, https://doi.org/10.1137/S1064827502418598, https://arxiv.org/abs/https://doi.org/10.1137/S1064827502418598.

[6] W. Briggs, V. Henson, and S. McCormick, *A Multigrid Tutorial, 2nd Edition*, 01 2000.

[7] O. Coulaud, L. Giraud, P. Ramet, and X. Vasseur, *Deflation and augmentation techniques in krylov subspace methods for the solution of linear systems*, 2013, https://arxiv.org/abs/1303.5692.

[8] V. Dwarka and C. Vuik, *Stand-alone multigrid for helmholtz revisited: Towards convergence using standard components*, 2023, https://arxiv.org/abs/2308.13476.

[9] P. Ek, M. Brezina, and J. Mandel, *Convergence of algebraic multigrid based on smoothed aggregation*, Computing, 56 (1998), https://doi.org/10.1007/s002110000226.

[10] O. G. Ernst and M. J. Gander, *Why it is difficult to solve helmholtz problems with classical iterative methods*, (2010).

[11] R. D. Falgout, *An introduction to algebraic multigrid*, Computing in Science and Engineering, vol. 8, no. 6, November 1, 2006, pp. 24-33, (2006), https://www.osti.gov/biblio/897960.

[12] R. D. Falgout and P. S. Vassilevski, *On generalizing the amg framework*, SIAM J. NUMER. ANAL, 42 (2003), pp. 1669–1693.

[13] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1 ed., Apr. 1991, https://doi.org/10.1017/CBO9780511840371, https://www.cambridge.org/core/product/identifier/9780511840371/type/book (accessed 2024-05-24).

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2 ed., Oct. 2012, https://doi.org/10.1017/CBO9781139020411, https://www.cambridge.org/highereducation/product/9781139020411/book (accessed 2024-05-24).

[15] J. K. Kraus, P. S. Vassilevski, and L. T. Zikatanov, *Polynomial of best uniform approximation to x^{-1} and smoothing in two-level methods*, 2012, https://arxiv.org/abs/1002.1859.

[16] L. Lin, Y. Saad, and C. Yang, *Approximating spectral densities of large matrices*, SIAM Review, 58 (2016), pp. 34–65, https://doi.org/10.1137/130934283, https://doi.org/10.1137/130934283, https://arxiv.org/abs/https://doi.org/10.1137/130934283.

[17] I. Livshits, *A scalable multigrid method for solving indefinite helmholtz equations with constant wave numbers*, Numerical Linear Algebra with Applications, 21 (2014), https://doi.org/10.1002/nla.1926.

[18] I. Livshits, *Multiple galerkin adaptive algebraic multigrid algorithm for the helmholtz equa-*

tions, SIAM Journal on Scientific Computing, 37 (2015), pp. S195–S215, https://doi.org/10.1137/140975310, https://doi.org/10.1137/140975310, https://arxiv.org/abs/https://doi.org/10.1137/140975310.

[19] S. MacLachlan and Y. Saad, *A greedy strategy for coarse-grid selection*, SIAM Journal on Scientific Computing, 29 (2007), pp. 1825–1853, https://doi.org/10.1137/060654062, https://doi.org/10.1137/060654062, https://arxiv.org/abs/https://doi.org/10.1137/060654062.

[20] L. Olson and J. Schroder, *Smoothed aggregation for helmholtz problems*, Numerical Linear Algebra with Applications, 17 (2010), pp. 361 – 386, https://doi.org/10.1002/nla.686.

[21] L. N. Olson, J. B. Schroder, and R. S. Tuminaro, *A general interpolation strategy for algebraic multigrid using energy minimization*, SIAM Journal on Scientific Computing, 33 (2011), pp. 966–991, https://doi.org/10.1137/100803031, https://doi.org/10.1137/100803031, https://arxiv.org/abs/https://doi.org/10.1137/100803031.

[22] E. Parolin, D. Huybrechs, and A. Moiola, *Stable approximation of helmholtz solutions by evanescent plane waves*, 2022, https://doi.org/10.48550/ARXIV.2202.05658, https://arxiv.org/abs/2202.05658.

[23] J. W. Ruge and K. Stüben, *4. Algebraic Multigrid*, https://doi.org/10.1137/1.9781611971057.ch4, https://epubs.siam.org/doi/abs/10.1137/1.9781611971057.ch4, https://arxiv.org/abs/https://epubs.siam.org/doi/pdf/10.1137/1.9781611971057.ch4.

[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, https://www-users.cs.umn.edu/~saad/IterMethBook_2ndEd.pdf.

[25] G. Strang, *Multigrid methods*, tech. report, MIT, 2006, https://math.mit.edu/classes/18.086/2006/am63.pdf.

[26] K. Stüben, *Algebraic multigrid (amg). an introduction with applications*, (1999).

[27] P. Vanek, J. Mandel, and M. Brezina, *Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems*, tech. report, USA, 1995.

[28] B. M. Vanvek Petr and M. Jan, *Convergence of algebraic multigrid based on smoothed aggregation*, (2001), https://doi.org/10.1007/s211-001-8015-y.

[29] T. U. Zaman, S. P. MacLachlan, L. N. Olson, and M. West, *Coarse-grid selection using simulated annealing*, Journal of Computational and Applied Mathematics, 431 (2023), p. 115263, https://doi.org/https://doi.org/10.1016/j.cam.2023.115263, https://www.sciencedirect.com/science/article/pii/S0377042723002078.