SIAM J. SCI. COMPUT.
Vol. 42, No. 2, pp. A901–A928

SCALABLE CONVERGENCE USING TWO-LEVEL DEFLATION PRECONDITIONING FOR THE HELMHOLTZ EQUATION*

VANDANA DWARKA† AND CORNELIS VUIK†

Abstract. Recent research efforts aimed at iteratively solving the Helmholtz equation have focused on incorporating deflation techniques for accelerating the convergence of Krylov subpsace methods. The requisite for these efforts lies in the fact that the widely used and well-acknowledged complex shifted Laplacian preconditioner (CSLP) shifts the eigenvalues of the preconditioned system towards the origin as the wave number increases. The two-level-deflation preconditioner combined with CSLP showed encouraging results in moderating the rate at which the eigenvalues approach the origin. However, for large wave numbers the initial problem resurfaces and the near-zero eigenvalues reappear. Our findings reveal that the reappearance of these near-zero eigenvalues occurs if the near-singular eigenmodes of the fine-grid operator and the coarse-grid operator are not properly aligned. This misalignment is caused by accumulating approximation errors during the inter-grid transfer operations. We propose the use of higher-order approximation schemes to construct the deflation vectors. The results from rigorous Fourier analysis and numerical experiments confirm that our newly proposed scheme outperforms any other deflation-based preconditioner for the Helmholtz problem. In particular, the spectrum of the adjusted preconditioned operator stays fixed near one. These results can be generalized to general shifted indefinite systems with random right-hand sides. For the first time, the convergence properties for very large wave numbers ($k = 10^6$ in one dimension and $k = 10^3$ in two dimensions) have been studied, and the convergence is close to wave number independence. Wave number independence for three dimensions has been obtained for wave numbers up to $k = 75$. The new scheme additionally shows very promising results for the more challenging Marmousi problem. Irrespective of the strongly varying wave number, we obtain a constant number of iterations and a reduction in computational time as the results remain robust without the use of the CSLP preconditioner.

Key words. Helmholtz equation, convergence, scalable, deflation, CSLP

AMS subject classifications. 65F08, 65F10, 65F15

DOI. 10.1137/18M1192093

1. Introduction.

From investigating the earth's layers in seismology to assessing the effect of electromagnetic scattering in the presence of human tissue through MRI, the Helmholtz equation finds its application in many engineering practices. Many efforts have been rendered in order to obtain accurate and computationally feasible solutions. Two major problem arise in trying to solve the Helmholtz equation numerically. First of all, for large wave numbers the numerical solution suffers from the so called "pollution error," which intrinsically is a phase difference between the exact and numerical solution. The second issue relates to the convergence behavior of the underlying solver. For medium to large wave numbers, the linear system becomes indefinite due to the negative eigenvalues. In order to balance the accuracy for such large wave numbers the linear system becomes very large and thus preconditioned iterative solvers are preferred, especially when considering higher-dimensional problems [7]. As the wave number increases the eigenvalues of the preconditioned matrix start to shift towards the origin. These near-zero eigenvalues have a detrimental ef-

*Submitted to the journal's Methods and Algorithms for Scientific Computing section June 5, 2018; accepted for publication (in revised form) January 13, 2020; published electronically April 1, 2020.

https://doi.org/10.1137/18M1192093

Funding: This work was supported by DIAM, Delft University of Technology.

†Department of Applied Mathematics, Delft University of Technology, Delft, the Netherlands (v.n.s.r.dwarka@tudelft.nl, c.vuik@tudelft.nl, http://ta.twi.tudelft.nl/users/vuik/).

fect on the convergence speed of Krylov-based iterative solvers. In order to mitigate these effects, many preconditioners for the Helmholtz problem have been proposed throughout the years.

A specific class of preconditioners focuses on the operator in question and shows performance gains for medium sized wave numbers. In [1] the preconditioner matrix is equal to the discretized Laplacian operator, and variations of this include a real and/or complex shift. A widely known preconditioner is the *complex shifted Laplacian preconditioner* (CSLP). The method was originally proposed by [10]. By including a complex shift, the original Helmholtz operator was modified to work as a preconditioner. Unless the shift is kept $\mathcal{O}(k)$ and the preconditioner is inverted exactly, the small eigenvalues of the preconditioned system still rush to zero as the wave number increases [14]. In order to properly manage the computational costs, one multigrid iteration was proposed to obtain an approximation of the inverse. Using rigorous Fourier analysis, the authors showed that the use of the multigrid method to obtain a cost effective preconditioner came at the price of having to keep the complex shift rather large, i.e., of $\mathcal{O}(k^2)$. A more recent analysis provided a generalization for this claim without having to restrict to Dirichlet boundary conditions [3]. A different approach can be found by using preconditioning techniques based on domain decomposition methods applied to the corresponding (shifted) problem, which is largely based on the work in [15]. These methods split the computational domain into subdomains and solve a local subproblem of smaller dimension using a direct method [22, 20, 21]. The performance of these preconditioners depends on the accuracy of the transmission conditions, which currently is robust for constant wave number model problems [18, 17]. While this resulted in a reduced number of iterations, the number of iterations still mildly grows with the constant wave number k. Balancing between wave number independent convergence and practical constraints created the opportunity to consider a deflation strategy, which was first proposed in [8]. Deflation, in essence, aims to move the unwanted eigenvalues to zero or one and has been studied widely; see [35, 36, 37]. While being able to improve the convergence and performance significantly, the near-zero eigenvalues keep reappearing for large wave numbers. In this work we present an adapted deflation scheme based on higher-order approximations in order to obtain an efficient and fast solver for very large wave numbers. While convergence is boosted when we combine the deflation preconditioner with the multigrid approximate inverse of the CSLP preconditioner, our results also show wave number independent convergence when we exclude the CSLP preconditioning step. The structure of this paper is as follows. A problem description and some background on the underlying Krylov-based solver are introduced in sections 2 and 3, respectively. Here we also briefly introduce the problem of the near-zero eigenvalues. In section 4 we embark on examining the root cause of these near-zero eigenvalues in full detail. We proceed by introducing the adapted deflation scheme in section 5, together with a spectral analysis of both methods in subsection 5.2. Finally we present numerical results for benchmark problems in section 6.

2. Problem description. We start by focusing on a one-dimensional mathematical model using a constant wave number $k > 0$:

$$-\frac{d^2u}{dx^2} - k^2\, u = \delta(x - x'), x \in \Omega = [0, L] \subset \mathbb{R}, \tag{2.1}$$
$$u(0) = 0, u(L) = 0.$$

We will refer to this model problem as MP 1-A. As mentioned previously, using Dirichlet boundary conditions, the resulting coefficient matrix is normal and hence GMRES convergence after preconditioning is completely determined by the spectrum. While this allows for extensive analysis of the convergence behavior, no true wave number independent convergence has been reported for this model problem unless the shift in the CSLP preconditioner is kept very small and exact inversion is utilized [14]. This motivates us to start with the study of this simple model problem in order to create a foundation for obtaining wave number independent convergence. Moving on, we introduce MP 1-B as the model problem where Sommerfeld radiation conditions have been implemented. In this case, the boundary conditions become

$$\left(\frac{\partial}{\partial n} - ik\right) u(x) = 0, x \in \partial[0, L].$$

For both model problems, discretization using second-order finite differences with step size $h = \frac{1}{n}$ leads to

$$\frac{-u_{l-1} + 2u_l - u_{l+1}}{h^2} - k^2 u_l = f_l, l = 1, 2, \ldots, n.$$

Using a lexicographic ordering, the linear system can be formulated exclusively on the internal grid points due to the homogeneous Dirichlet boundary conditions. We obtain the following linear system and eigenvalues with indices $l = 1, 2, \ldots n$:

$$\begin{aligned} Au &= \frac{1}{h^2}\text{tridiag}[-1 \;\; 2 - k^2h^2 \;\; -1]u = f, \\ \hat{\lambda}^l &= \frac{1}{h^2}\left(2 - 2\cos(l\pi h)\right) - k^2. \end{aligned} \tag{2.2}$$

In order to investigate the scalability of the convergence in higher dimensions (section 6), we define MP 2 and MP 3 to be the two-dimensional and three-dimensional versions of the one-dimensional model problem MP 1-A defined above (2.1). The discretization using second-order finite differences goes accordingly for higher dimensions, with the resulting matrices being penta- and hepta-diagonal for two dimensions (2D) and three dimensions (3D), respectively.

The final test problem is a representation of an industrial problem and is widely referred to as the two-dimensional Marmousi problem, which we denote by MP 4. We consider an adapted version of the original Marmousi problem developed in [35]. The original domain has been truncated to $\Omega = [0, 8192] \times [0, 2048]$ in order to allow for efficient geometric coarsening of the discrete velocity profiles given that the domain remains a power of 2. The original velocity $c(x, y)$ is also adapted by letting $2587.5 \leq c \leq 3325$. We will use the adjusted domain in order to benchmark against the results from [35]. On the adjusted domain Ω, we define

$$\begin{aligned} -\Delta u(x, y) - k(x, y)^2 u(x, y) &= \delta(x - 4000, y), (x, y) \in \Omega \setminus \partial\Omega \subset \mathbb{R}^2, \\ \left(\frac{\partial}{\partial \mathbf{n}} - ik\right) u(x, y) &= 0, (x, y) \in \partial\Omega, \end{aligned} \tag{2.3}$$

where n denotes the outward normal unit vector in the x- and y-directions, respectively. Note that we now have a nonconstant wave number $k(x, y) = \frac{2\pi freq}{c(x,y)}$, where the frequency is given in Hertz.

2.1. Effect of nonnormality. By assuming Dirichlet boundary conditions for our first model problem, we are able to simplify the analysis and perform rigorous Fourier analysis, which shows that the new scheme is able to align the near-zero eigenvalues of the fine- and coarse-grid coefficient matrices. Having a higher-order approximation scheme for the deflation vectors enables us to reach wave number independent convergence in one dimension (1D) and close to wave number independent convergence in 2D and 3D for very large wave numbers. The difficulty in using Sommerfeld radiation conditions is that the resulting coefficient matrix becomes complex and non-self-adjoint. Therefore, there are no closed-form expressions for the eigenvalues. Additionally, it has been noted that in the case of nonnormal matrices, spectral analysis becomes less meaningful in order to assess convergence properties [23, 39]. If A is normal, then the condition number of the eigenvector matrix is one. In the case of a nonnormal matrix, the condition number of the eigenvector matrix is larger than one. As a result it has been shown that arbitrary matrices can be created with arbitrary eigenvalues and right-hand sides which give the same GMRES residual norms [23]. While this often has been interpreted as "spectral analysis for a nonnormal matrix is insufficient," the original authors also mentioned that even for a matrix which is far from normal, GMRES can converge very well and the eigenvalues can still primarily govern its convergence in some specific cases. For example it may be the case that the eigenvector matrix is well conditioned, A is close to Hermitian despite having ill-conditioned eigenvectors, or zero is outside the field of values (FOV) of A. While the latter approach has received great attention in the past years to explain convergence behavior of the Helmholtz equation, its use is very sensitive to having zero inside the FOV, which often seems to be the case for indefinite systems [6]. A more recent and detailed analysis showed that the dependence on the condition number of the eigenvectors is often a large overestimation of the actual error [27]. In fact, it has been shown that for diagonalizable matrices, eigenvalues close to the origin indeed hamper GMRES convergence and GMRES convergence does not explicitly depend on the condition number of the eigenvector matrix [28]. While the latter may be large, convergence is still predominantly governed by the eigenvalues if the eigenvector matrix is not too far from unitary. Similarly for nondiagonalizable matrices such as a highly non-normal single, plain Jordan block, GMRES convergence can still be strongly governed by an eigenvalue with a large modulus [28, 6, 27, 25, 26]. An important implication of this for a diagonalizable matrix is that convergence for a nonnormal A can behave as convergence for a normal A. While the literature does not quantify terms as a "small" condition number or "not too far from normality/unitary" for this particular application, there exists vast numerical evidence showing that altering the spectrum leads to better GMRES convergence. This corroborates the acceleration of GMRES convergence using deflation preconditioning techniques [19, 5, 29, 31]. In fact, in [31] the authors state that "deflated GMRES can be effective even when the eigenvectors are poorly defined ... and for highly nonnormal matrices," where convergence is boosted after removing small (pseudo)eigenvalues. Therefore, in order to fully understand the efficiency of our proposed deflation preconditioner, we start conducting spectral and convergence behavior analysis of the proposed preconditioner for the self-adjoint case. We then provide numerical evidence to investigate the performance of the preconditioner for non-self-adjoint problems. Our numerical results support the notion that similar conclusions as regards wave number independent convergence holds for the non-self-adjoint Helmholtz equation and the heterogeneous Marmousi problem.

3. Deflated Krylov methods. Starting with the simple MP 1-A, (2.2) reveals that the spectrum contains both positive and negative eigenvalues for

$$k > \frac{2\sin(\pi\frac{h}{2})}{h} \approx \pi.$$

This indefiniteness narrows the choice of potential Krylov-based solvers due to the conjugate gradient type methods being ineffective. While the application of the CSLP preconditioner was successful in confining the eigenvalues between 0 and 1, the Krylov solver remains defenseless against the hampering convergence behavior caused by the small eigenvalues for large k, which is why deflation was introduced to boost the convergence behavior of the Krylov solver. Deflation is a technique which aims to move near-zero eigenvalues to zero by using an orthogonal projection. It can also be used to move these unwanted eigenvalues to 1 or the largest eigenvalue. In both cases, the eigenvalues are mapped to the desired value when the exact eigenvectors are utilized. Due to practical considerations within the context of Krylov solvers, it is possible to alleviate the adverse effects of near-zero eigenvalues using deflation by either explicitly modifying the operator of the linear system [32] or by finding approximations to the eigenvectors corresponding to the troublesome eigenvalues. For example, such approximations are used in [29] and [30], where harmonic Ritz vectors serve as eigenvector approximations to augment the Krylov subspace in order to gain faster convergence. Deflation for large scale problems relies on multiplying the linear system by a projection matrix P and applying the Krylov subspace method to the projected system PA, rendering the projection matrix P to act as a preconditioner at the same time.

3.1. Deflation-based preconditioning for GMRES. Consider a general real valued linear system. The projection matrix $\widehat{P}$ and its complementary projection P can be defined as

$$\begin{aligned} &\widehat{P} = AQ, \text{ where } Q = ZE^{-1}Z^T \text{ and } E = Z^TAZ, \\ &A \in \mathbb{R}^{n\times n},\ Z \in \mathbb{R}^{m\times n}, \\ &P = I - AQ, \end{aligned} \tag{3.1}$$

where Z functions as the deflation matrix whose $m < n$ columns are considered the deflation vectors and I is the $n \times n$ identity matrix. Additionally, the coarse-grid coefficient matrix E is assumed to be invertible. Matrix P is also known as the projection preconditioner. In Algorithm 3.1 we present the preconditioned deflated GMRES algorithm, which will be used for numerical testing in section 6 and includes a preconditioner matrix M. In our case, we use the CSLP preconditioner, which is defined by

$$M = -\Delta - (\beta_1 + \beta_2 i)k^2,$$

where $i = \sqrt{-1}$ and $(\beta_1, \beta_2) \in [0, 1]$. The CSLP preconditioner is included in order to obtain a more favorable spectrum. Unless stated otherwise, we use one $V(1,1)$-multigrid cycle to obtain an approximate inverse of the CSLP preconditioner.

We now proceed by discussing the computational cost of the preconditioner. Before discussing the computational costs, we need to establish that a state-of-the-art scalable solver should comprise two components: wave number independent convergence and $\mathcal{O}(n)$ time complexity with respect to the number of grid points. Many

Algorithm 3.1 Preconditioned deflated GMRES for system $Au = b$.

Choose u_0 and compute $r_0 = b_0 - Au_0$ and $v_1 = r_0/\|r_0\|$
for for $j = 1, 2, \ldots, k$ or until convergence **do**
  $\tilde{v}_j := Pv_j$
  $w = M^{-1}A\tilde{v}_j$
  for $i := 1, 2, \ldots, j$ **do**
    $h_{i,j} := w^T v_i$
    $w := w - h_{i,j}v_i$
  end for
  $h_{j+1,j} := \|w\|$
  $v_{j+1} := w/h_{j+1,j}$
end for
Store $V_k = \left[\tilde{v}_1, \ldots, \tilde{v}_k\right]$; $H_k = \{h_{i,j}\}$, $1 \le i \le j+1$, $1 \le j \le m$
Compute $y_k = argmin_y \|b_0 - H_k y\|$ and $u_k = u_0 + V_k y_k$
The entries of upper $k+1, k$ Hessenberg matrix H_k are the scalars $h_{i,j}$
Update approximated solution $\mathbf{u_k = Qb + P^T u_k}$

TABLE 1
Number of extra flops above standard preconditioned GMRES. The cost of the sparse factorization and backward solve have been discussed in [2]. $A \in \mathbb{R}^{n\times n}$ *and* $E \in \mathbb{R}^{m\times m}$. $Z \in \mathbb{R}^{m\times n}$. *Bold indicates costs of initialization only.*

Operation	Linear			Quadratic		
	1D	2D	3D	1D	2D	3D
$\mathbf{E = LU}$	$\mathbf{m}$	$\mathbf{m^{1\frac{1}{2}}}$	$\mathbf{m^2}$	$\mathbf{4m}$	$\mathbf{4m^{1\frac{1}{2}}}$	$\mathbf{4m^2}$
$w = (LU)^{-1}y$	$2m$	$2m\log(m)$	$2m^{1\frac{1}{3}}$	$4m$	$4m\log(m)$	$4m^{1\frac{1}{3}}$
$v = Zw$	$3n$	$3n$	$3n$	$5n$	$5n$	$5n$
$y = Z^T w$	$3n$	$3n$	$3n$	$5n$	$5n$	$5n$
$z = Av$	$6n$	$10n$	$14n$	$6n$	$10n$	$14n$
$Px = x - z$	n	n	n	n	n	n

works have either focused on scalable convergence, obtaining linear time scaling results, or finding a balance between the trade-off. While our primary focus lies on establishing the theory and numerical evidence for obtaining true wave number independent convergence, our current solver is the first which allows for a reduction of the computational costs while maintaining true wave number independent convergence for very high wave numbers. In Table 1, we provide the extra cost of the initialization phase and the iterative phase, using the fact that A, E, Z are all sparse. Apart from the deflation operator, the inversion of the CSLP preconditioner by one multigrid iteration costs $\mathcal{O}(n)$ flops [9]. The extra computational cost of the deflation method depends on the design of the deflation matrix Z. The largest cost in the initialization phase is attributed to the one-time factorization of E leading up to the backward solve on the coarse grid. Similarly to the approach taken in the literature with respect to two-level methods, we only count the backward solve in order to determine the computational complexity of the iterative phase [2, 17, 38, 13, 20, 22]. The work in [14] appears to be one of the few to report absolute wave number independent convergence using the CSLP preconditioner as a basis. However, the convergence is subjected to keeping the shift small ($\mathcal{O}(\frac{1}{k})$) and inverting the precondtioner exactly, which is why the authors predominantly emphasize its theoretical contribution. For

a three-dimensional application, exact inversion of the CSLP preconditioner using a small shift costs $\mathcal{O}(n^2)$ flops for the factorization stage and $\mathcal{O}(n^{1\frac{1}{3}})$ flops for the solve stage. By using the two-level deflation preconditioner, we obtain wave number independent convergence for very high wave numbers by reducing the sequential factorization and solve stage by a factor of 16 and 8, respectively.[1] In fact, the complexity of our two-level method is comparable to the sequential timings reported in [22]. There, a two-dimensional domain decomposition preconditioner is implemented having an $\mathcal{O}(n^{1\frac{1}{2}})$ time complexity with convergence being moderately dependent on k. A natural extension from a two-level method to a multilevel method has additionally been researched in the aforementioned and other works, and provides promising directions for future research.

3.2. The deflation (DEF) preconditioner. Based on the above, the DEF preconditioner has been defined by taking the coarse correction operator I_{2h}^h from a multigrid setting as the DEF subspace Z in (3.1). I_{2h}^h can be interpreted as interpolating from grid Ω_{2h} to grid Ω_h. As a result, the DEF preconditioner is commonly referred to as a two-level method and we obtain

$$\begin{aligned} &\widehat{P} = A_h Q, \text{ where } Q = Z {A_{2h}}^{-1} Z^T \text{ and } A_{2h} = Z^T A_h Z, \\ &P = I_h - A_h Q, \text{ where } Z = I_{2h}^h. \end{aligned} \tag{3.2}$$

In the literature a distinction is made with respect to the two-level DEF operator. On the one hand we have the DEF preconditioner as defined above. On the other hand we have the ADEF preconditioner, which is defined by taking $P_{ADEF} = P + \gamma Q$. The inclusion of the shift γ ensures that the coarse-grid solve with respect to A_{2h} can be approximated, for example, by considering a multilevel implementation [38, 35]. When considering approximate inversion, γ is generally either set to 1 or the largest eigenvalue of the original coefficient matrix. In this work we solely focus on the DEF preconditioner in a two-level setting, and thus we can take $\gamma = 0$.

3.2.1. Inscalability and spectral analysis. We now shift our focus to the study of the eigenvalues of the DEF operator without inclusion of CSLP. To study the eigenvalues, we use the analytical derivations and expressions for the spectrum of the DEF operator applied to the coefficient matrix A from [19]. The authors have provided concise analytical expressions for the eigenvalues of the standard two-level DEF operator. We use these expressions to perform a preliminary analysis of the spectrum. For $l = 1, 2, \ldots, \frac{n}{2}$, the eigenvalues of the system PA are given by

$$\lambda^l(PA) = \lambda^l(A)\left(1 - \frac{\lambda^l(A)\cos(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right) + \lambda^{n+1-l}(A)\left(1 - \frac{\lambda^{n+1-l}(A)\sin(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right). \tag{3.3}$$

Inspection of (3.3) leads to the observation that the eigenvalues of the DEF operator P are given by

$$\lambda^l(P) = \left(1 - \frac{\lambda^l(A)\cos(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right) + \left(1 - \frac{\lambda^{n+1-l}(A)\sin(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right). \tag{3.4}$$

[1] The size of E in a three-dimensional application is $\frac{n}{8}$.

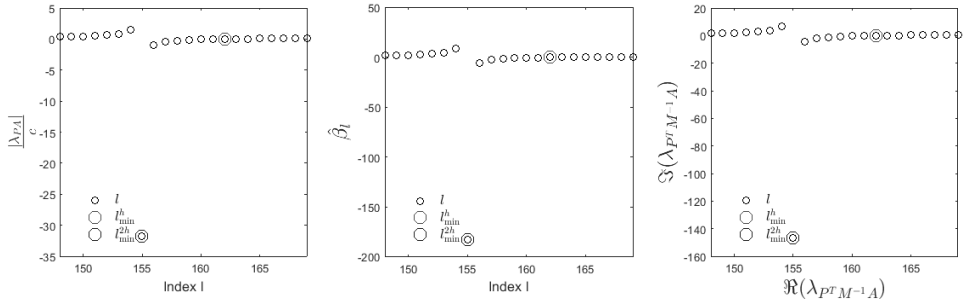


FIG. 1. *$kh = 0.625, k = 500$. Left: eigenvalues of PA scaled by magnitude of the largest eigenvalue (c). Center: Ratio between eigenvalues of the fine-grid and coarse-grid operator ($\hat{\beta}$ from equation (3.6)). Right: real part of eigenvalues $P^T M^{-1} A$.*

By introducing the following coefficients, we can rewrite (3.3) as

$$\alpha^l = \left(1 - \frac{\lambda^l(A)\cos(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right) = \frac{\lambda^{n+1-l}(A)\sin(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})},$$

$$\beta^l = \left(1 - \frac{\lambda^{n+1-l}(A)\sin(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})}\right) = \frac{\lambda^l(A)\cos(l\pi\frac{h}{2})^4}{\lambda^l(A_{2h})},$$

$$\lambda^l(PA) = \lambda^l(A)\alpha^l + \lambda^{n+1-l}(A)\beta^l, l = 1, 2, \ldots, \frac{n}{2}. \tag{3.5}$$

Since the sine and cosine terms are always strictly less than 1, the eigenvalues of the system PA are essentially the product of eigenvalues of A multiplied by the scaled ratio of the eigenvalues of A and A_{2h}. In order to simplify the analysis, we therefore proceed by analyzing

$$\hat{\beta}^l = \left|\frac{\lambda^l(A)}{\lambda^l(A_{2h})}\right|, l = 1, 2, \ldots, \frac{n}{2}, \tag{3.6}$$

which provides an upperbound to the previously defined coefficients. It is easy to see that the eigenvalues of PA will approach the origin if the factor $\hat{\beta}^l$ becomes small for some l. If we define the constant c to be the magnitude of the largest eigenvalue of A, then we can scale the eigenvalues of PA by c and compare them to the eigenvalues $P^T M^{-1} A$ and $\hat{\beta}$. In Figure 1 we have plotted a selected range of eigenvalues of PA scaled by c and compared these to the eigenvalues of $P^T M^{-1} A$ (right) and $\hat{\beta}^l$ (center). On the x-axis we have the relevant indices l corresponding to the respective close to zero eigenvalues. The figure provides affirmative support for our remark that the behavior of the eigenvalues of both PA and $P^T M^{-1} A$ are, apart from a scaling factor, determined by the behavior of $\hat{\beta}^l$ as all three figures exhibit the same shape and pattern. $\hat{\beta}^l$ approaches the origin whenever $|\lambda^l(A)|$ becomes small, which is at $l = l^h_{\min}$ (marker). If $l^h_{\min} \neq l^{2h}_{\min}$ and $l^{2h}_{\min} < l^h_{\min}$, then we are dividing a relatively small number $|\lambda^{l^h_{\min}}(A)|$ by a larger number $|\lambda^{l^h_{\min}}(A_{2h})|$, which brings the resulting fraction closer to zero. The further apart $l^h_{\min}$ and $l^{2h}_{\min}$ are, the closer to zero the resulting term will be. The outlier appointed by the marker is the result of exactly the opposite effect. At $l = l^{2h}_{\min}$, $|\lambda^l(A_{2h})|$ will be at its smallest, while the magnitude of $|\lambda^l(A)|$ will still be large. In like manner, we get a large term, which explains the typical outliers we often encounter when the spectra of the operators PA and $P^T M^{-1} A$ are plotted.

4. Eigenvector perturbations. The next question which needs to be answered is what is causing the near-zero eigenvalues of the coarse grid operator to shift. It has been reported that interpolating coarse-grid functions always introduces high-frequency modes, which can be interpreted as an aliasing phenomenon [11, 12]. These high-frequency modes are the main cause for interpolation errors [11]. The effect becomes more severe as index l increases. If the high-frequency eigenmodes are activated by interpolating from a coarse to a fine grid, then the coarse-grid eigenvectors will not be approximated accurately. This affects the eigenvalues of A_{2h} as A_{2h} is obtained by first restricting the fine-grid elements onto the coarse-grid and then transferring the result back onto the fine grid. To measure the extent of this effect, we make use of Lemma 4.1 and Corollary 4.2.

LEMMA 4.1 (intergrid transfer I). *Let B be the $\frac{n}{2} \times \frac{n}{2}$ matrix given by $B = Z^T Z$, where $Z = I_{2h}^h$ is the prolongation matrix and let $l_{\min}$ be the index of the smallest eigenvalue of A in terms of magnitude. Then there exists a constant C_h, depending on h, such that*

$$B\phi_{l_{\min},2h} = C_h \phi_{l_{\min},2h} \text{ and } \lim_{h\to 0} C_h = \lambda_{l_{\min}}(B) = 2,$$

where $\phi_{l,h}$ is the lth eigenvector on the fine-grid of A and $\lambda_l(B)$ is the lth eigenvalue of B.

Proof. We use the method from [12]. For $i = 1, 2, \ldots n$ we have

$$\begin{aligned}
[Z^T \phi_{l_{\min},h}]_i &= \frac{1}{2}\left(\sin((2i-1)h\pi l_{\min,h}) + 2\sin(2ih\pi l_{\min,h}) + \sin((2i+1)h\pi l_{\min,h})\right) \\
&= \frac{1}{2}\left(2\sin(2ih\pi l_{\min,h}) + 2\cos(2ih\pi l_{\min,h})\right)\sin(2ih\pi l_{\min,h}) \\
&= (1+\cos(l_{\min,h}\pi h))\sin(2ih\pi l_{\min,h}) \\
&= C_1(h)\phi_{l_{\min,2h}}.
\end{aligned}$$

Now taking the limit as h goes to zero of the coefficient C_h gives $\lim_{h\to 0} C_1(h) = 2$. For $i = 1, 2, \ldots, n$ we distinguish two cases: i is odd and i is even. We start with the first case

$$\begin{aligned}
[Z\phi_{l_{\min},2h}]_i &= \frac{1}{2}\left(\sin\left(\frac{(i-1)h\pi l_{\min,h}}{2}\right) + \sin\left(\frac{(i+1)h\pi l_{\min,h}}{2}\right)\right) \\
&= \frac{1}{2}\left(\sin((i-1)h\pi l_{\min,h}) + \sin((i+1)h\pi l_{\min,h})\right) \\
&= \cos(l_{\min,h}\phi h)\sin(ih\pi l_{\min,h}) \\
&= C_2(h)\phi_{l_{\min},h}.
\end{aligned}$$

Again, taking the limit as h goes to zero of the coefficient $C_2(h)$ gives $\lim_{h\to 0} C_2(h) = 1$. For i is even, we obtain $Z\phi_{l_{\min},2h} = \sin(\frac{ih\phi l_{\min,h}}{2}) = \sin(ih\pi l_{\min,h}) = \phi_{l_{\min},h}$. We can combine both results to obtain $B\phi_{l_{\min},2h} = Z^T Z\phi_{l_{\min},2h} = Z^T(C_2(h)\phi_{l_{\min},h}) = C_1(h)C_2(h)\phi_{l_{\min,2h}} = \hat{\lambda}_{l_{\min}}(B)\phi_{l_{\min,2h}}$, where $\hat{\lambda}_{l_{\min}}(B)$ represents the perturbed eigenvalue of B at index l due to the approximation error. Taking the limit as h goes to zero provides $\lim_{h\to 0}\hat{\lambda}_{l_{\min}}(B) = \lim_{h\to 0} C_1(h)C_2(h) = 2 = \lambda_{l_{\min,h}}(B)$. ∎

COROLLARY 4.2 (coarse-grid kernel). *Let A_{2h} be the $\frac{n}{2} \times \frac{n}{2}$ matrix given by $A_{2h} = Z^T AZ$, where $Z = I_{2h}^h$ is the prolongation matrix and let $l_{\min}$ be the index of*

the smallest eigenvalue of A in terms of magnitude. Then

$$A_{2h}\phi_{l_{\min,2h}} = C_h\lambda_{l_{\min,h}}(A)\phi_{l_{\min,2h}}, \text{ and } \lim_{h\to 0} C_h = \lambda_{l_{\min},h}(B),$$

where $\phi_{j,2h}$ is the lth eigenvector on the coarse grid of A_{2h} and $\lambda_j(A_{2h})$ is the lth eigenvalue of A_{2h}.

Proof. Using Lemma 4.1 and its proof, we have

$$\begin{aligned} A_{2h}\phi_{l_{\min},2h} &= \left(Z^T AZ\right)\phi_{l_{\min,2h}} \\ &= Z^T A\left(Z\phi_{l_{\min,2h}}\right) \\ &= Z^T A(C_2(h)\phi_{l_{\min},h}) \\ &= C_1(h)Z^T A\phi_{l_{\min},h} \\ &= C_1(h)Z^T\lambda_{l_{\min,h}}(A)\phi_{l_{\min},h} \\ &= \lambda_{l_{\min,h}}(A)C_1(h)\left(Z^T\phi_{l_{\min},h}\right) \\ &= \lambda_{l_{\min,h}}(A)C_1(h)C_2(h)\phi_{l_{\min},2h}. \end{aligned}$$

Using Lemma 4.1 it is easy to see that after taking the limit the eigenvalues of A_{2h} can be written as a product of the eigenvalues of A and the eigenvalues of B. ∎

From Lemma 4.1 and Corollary 4.2 it is clear that for $l_{\min}$, which is within the smooth-frequency range, the near-kernel coarse-grid eigenvalues $\lambda_{l_{\min},h}(A_{2h})$ are equal to the product of $\lambda_{l_{\min},h}(A)$ and $\lambda_{l_{\min,h}}(B)$ when h goes to zero. Consequently, in the limiting case, the coarse-grid kernel and the fine-grid kernel will be aligned proportionally and both A and A_{2h} will reach its smallest absolute eigenvalues at the same index $l_{\min}$.

Recall the behavior of the eigenvalues of PA can be represented by

$$\hat{\beta}^l = \left|\frac{\lambda^l(A)}{\lambda^l(A_{2h})}\right| \text{ for } l = 1, 2, \ldots, \frac{n}{2},$$

where we found that this ratio becomes very small by a mismatch of the smallest absolute eigenvalue of A and A_{2h}, respectively. As in the limit, we can write $\lambda_{l_{\min},h}(A_{2h}) = \lambda_{l_{\min},h}(B)\lambda_{l_{\min},h}(A_h)$, perturbations up to $\lambda_{l_{\min},h}(B)$ will propagate throughout the low-frequency part of the spectrum for $l \in \{1, 2, \ldots, l_{\min,h}\}$, eventually resulting in the errors related to $\lambda^l(A_{2h})$ for $l = l_{\min,h}$. To measure to what extent these perturbations to $\lambda(B)$ lead to errors, we examine the projection error to quantify the error we make when projecting the eigenvector onto the subspace spanned by the column of Z.

THEOREM 4.3 (projection error I). *Let X be the DEF space spanned by column vectors of Z and let the eigenvector corresponding to the smallest eigenvalue of A be denoted by $\phi_{l_{\min,h}} \notin X$. Let $P = ZB^{-1}Z^T$ with $B = Z^T Z$ be the orthogonal projector onto X. Then the projection error E is given by*

$$E = \|(I-P)\phi_{l_{\min},h}\|^2 = {\phi_{l_{\min},h}}^T\phi_{l_{\min},h} - {\phi_{l_{\min},h}}^T ZB^{-1}Z^T\phi_{l_{\min},h}.$$

Proof. By idempotency of the orthogonal projector, we have

$$\begin{aligned} \|(I-P)\phi_{l_{\min},h}\|^2 &= {\phi_{l_{\min},h}}^T(I-P)(I-P)\phi_{l_{\min},h} \\ &= {\phi_{l_{\min},h}}^T(I-P)\phi_{l_{\min},h} \\ &= {\phi_{l_{\min},h}}^T\phi_{l_{\min},h} - {\phi_{l_{\min},h}}^T ZB^{-1}Z^T\phi_{l_{\min},h}. \end{aligned}$$ ∎

We proceed by rewriting the projection error in terms of a perturbation to the eigenvalues of the operator B.

COROLLARY 4.4 (projection error II). *Let X be the DEF space spanned by the column vectors of Z and let the eigenvector corresponding to the smallest eigenvalue of A be denoted by $\phi_{l_{\min},h} \notin X$. Let $P = ZB^{-1}Z^T$ with $B = Z^TZ$ be the orthogonal projector onto X. Then the projection error E is given by*

$$E = \|(I-P)\phi_{l_{\min},h}\|^2 = \left(1 - \frac{\lambda_{l_{\min},h}(B) - \delta_1}{\lambda_{l_{\min},h}(B) - \delta_2}\right) {\phi_{l_{\min},h}}^T \phi_{l_{\min},h},$$

where $\delta_1 = \lambda_{l_{\min},h}(B) - \frac{{\phi_{l_{\min},h}}^T \hat{B}\phi_{l_{\min},h}}{{\phi_{l_{\min},h}}^T \phi_{l_{\min},h}}$ *and* $\delta_2 = \lambda_{l_{\min},h}(B) - \frac{{\phi_{l_{\min},h}}^T \hat{B}\phi_{l_{\min},h}}{{\phi_{l_{\min},h}}^T Z(B^{-1}Z^T\phi_{l_{\min},h})}$.

Proof. Using Lemma 4.1 and its proof we know that in the limit $Z^T\phi_{l_{\min},h}$ is an eigenvector of B. We would thus have

$$\begin{aligned}\|(I-P)\phi_{l_{\min},h}\|^2 &= {\phi_{l_{\min},h}}^T \phi_{l_{\min},h} - {\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right)\\ &= {\phi_{l_{\min},h}}^T \phi_{l_{\min},h} - \frac{{\phi_{l_{\min},h}}^T ZZ^T\phi_{l_{\min},h}}{\lambda_{l_{\min},h}(B)}\\ &= {\phi_{l_{\min},h}}^T \phi_{l_{\min},h} - \frac{{\phi_{l_{\min},h}}^T \left(\hat{B}\phi_{l_{\min},h}\right)}{\lambda_{l_{\min},h}(B)}.\end{aligned}$$

Note that $\hat{B}$ has dimension $n \times n$ and has $\frac{n}{2}$ eigenvalues equal to the eigenvalues of B and $\frac{n}{2}$ zero eigenvalues. By Lemma 4.1 and its proof, we also have that $\phi_{l_{\min},h}$ is an eigenvector of $\hat{B}$, which leads to

$$\|(I-P)\phi_{l_{\min},h}\|^2 = \lim_{h\to 0}\left({\phi_{l_{\min},h}}^T \phi_{l_{\min},h} - \frac{{\phi_{l_{\min},h}}^T \left(\lambda_{l_{\min},h}(\hat{B})\phi_{l_{\min},h}\right)}{\lambda_{l_{\min},h}(B)}\right) = 0. \tag{4.1}$$

Now, in the nonlimiting case, we have two sources of errors; the factor containing $\lambda_{l_{\min},h}(B)$ both in the numerator and denominator will be subjected to perturbations. Starting with the denominator, if we let $\tilde{\lambda}_{l_{\min},h}(B)$ denote the perturbed eigenvalue of B, we have

$${\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right) = {\phi_{l_{\min},h}}^T Z\left(\frac{Z^T\phi_{l_{\min},h}}{\tilde{\lambda}_{l_{\min},h}(B)}\right) \neq {\phi_{l_{\min},h}}^T Z\left(\frac{Z^T\phi_{l_{\min},h}}{\lambda_{l_{\min},h}(B)}\right).$$

Reordering leads to

$$\tilde{\lambda}_{l_{\min},h}(B) = \frac{{\phi_{l_{\min},h}}^T ZZ^T\phi_{l_{\min},h}}{{\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right)} = \frac{{\phi_{l_{\min},h}}^T \hat{B}\phi_{l_{\min},h}}{{\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right)}.$$

The perturbation to $\lambda_{l_{\min},h}(B)$ can now be written as

$$\delta_2 = \lambda_{l_{\min},h}(B) - \tilde{\lambda}_{l_{\min},h}(B) = \lambda_{l_{\min},h}(B) - \frac{{\phi_{l_{\min},h}}^T \hat{B}\phi_{l_{\min},h}}{{\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right)}.$$

For the numerator, if we let η denote the error, i.e., $\eta = \hat{B}\phi_{l_{\min},h} - \lambda_{l_{\min},h}(B)\phi_{l_{\min},h}$, then $\hat{B}\phi_{l_{\min},h} = \lambda_{l_{\min},h}(B)\phi_{l_{\min},h} + \eta$ and substitution gives

$$\begin{aligned}\tilde{\lambda}_{l_{\min},h}(B){\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right) &= {\phi_{l_{\min},h}}^T \hat{B}\phi_{l_{\min},h}\\ &= {\phi_{l_{\min},h}}^T \left(\lambda_{l_{\min},h}(B)\phi_{l_{\min},h} + \eta\right).\end{aligned}$$

Letting $\delta_1 = -\frac{{\phi_{l_{\min,h}}}^T \eta}{{\phi_{l_{\min,h}}}^T \phi_{l_{\min,h}}}$, we obtain

$$\tilde{\lambda}_{l_{\min},h}(B)\phi_{l_{\min},h}{}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right) = \left(\lambda_{l_{\min},h}(B) - \delta_1\right){\phi_{l_{\min},h}}^T\phi_{l_{\min},h}.$$

Finally, we can now rewrite the projection error E in terms of perturbations to the eigenvalues of B:

$$\begin{aligned}\|(I-P)\phi_{l_{\min},h}\|^2 &= {\phi_{l_{\min},h}}^T\phi_{l_{\min},h} - {\phi_{l_{\min},h}}^T Z\left(B^{-1}Z^T\phi_{l_{\min},h}\right)\\ &= \left(1 - \frac{\lambda_{l_{\min},h}(B) - \delta_1}{\lambda_{l_{\min},h}(B) - \delta_2}\right){\phi_{l_{\min},h}}^T\phi_{l_{\min},h},\end{aligned}$$

which gives the statement. ∎

We can prove an additional statement with respect to the pollution error. We know that the pollution error is minimized when we keep the step size $h = k^{-\frac{3}{2}}$. We can study the behavior of the projection error by letting k go to infinity.

COROLLARY 4.5 (pollution error). *Let $h = k^{-\frac{3}{2}}$. Let X be the DEF space spanned by column vectors of Z and let the eigenvector corresponding to the smallest eigenvalue of A be denoted by $\phi_{l_{\min,h}} \notin X$. Let $P = ZB^{-1}Z^T$ with $B = Z^TZ$ be the orthogonal projector onto X. Then the projection error E goes to zero,*

$$E = \lim_{k\to\infty} \|(I-P)\phi_{l_{\min},h}\|^2 = 0.$$

Proof. Using Lemma 4.1 and Corollary 4.2 we have

$$\begin{aligned}\left[Z^T\phi l_{\min,h}\right]_i &= (1 + \cos l_{\min,h}\pi h)\phi_{l_{\min},2h}\\ &= \left(1 + \cos l_{\min,h}\frac{\pi}{k^{\frac{3}{2}}}\right)\phi_{l_{\min},2h}.\end{aligned}$$

Now taking $k \to \infty$ gives $\lim_{k\to\infty}[Z^T\phi_{l_{\min},h,i}] = 2\phi_{l_{\min},2h}$. Similarly,

$$\begin{aligned}\left[Z\phi l_{\min,2h}\right]_i &= \cos(l_{\min,2h}\pi h)\phi_{l_{\min},h}\\ &= \cos\left(l_{\min,h}\frac{\pi}{k^{\frac{3}{2}}}\right)\phi_{l_{\min},h}.\end{aligned}$$

Again, taking $k \to \infty$ gives $\lim_{k\to\infty}[Z\phi_{l_{\min},2h,i}] = \phi_{l_{\min},h}$. Now, substituting these expressions into the projection error E gives

$$\begin{aligned}E = \lim_{k\to\infty}\|(I-P)\phi_{l_{\min},h}\|_2^2 &= \lim_{k\to\infty}\phi^T_{l_{\min},h}\phi_{l_{\min},h} - \phi^T_{l_{\min},h}Z(B^{-1}Z^T\phi_{l_{\min},h})\\ &= \lim_{k\to\infty}\left(\phi^T_{l_{\min},h}\phi_{l_{\min},h} - \phi^T_{l_{\min},h}ZB^{-1}(2\phi_{l_{\min},2h})\right)\\ &= \lim_{k\to\infty}\left(\phi^T_{l_{\min},h}\phi_{l_{\min},h} - 2\phi^T_{l_{\min},h}Z(B^{-1}\phi_{l_{\min},2h})\right)\\ &= \lim_{k\to\infty}\left(\phi^T_{l_{\min},h}\phi_{l_{\min},h} - \frac{2}{\lambda B}\phi^T_{l_{\min},h}(Z\phi_{l_{\min},2h})\right)\\ &= \lim_{k\to\infty}\left(\phi^T_{l_{\min},h}\phi_{l_{\min},h} - \frac{2}{\lambda B}\phi^T_{l_{\min},h}\phi_{l_{\min},h}\right)\\ &= \lim_{k\to\infty}\left(1 - \frac{2}{\lambda B}\right).\end{aligned}$$

We know from Corollary 4.2 that $\lambda B \to 2$ when h goes to zero. And thus we obtain the statement. ∎

TABLE 2
Projection error for $\phi_{l_{\min,h}}$ for various values of k. $j_{\min,h}$ and $l_{\min,2h}$ denote the index for the smallest absolute eigenvalue of A and A_{2h}, respectively.

k	E	$l_{\min,h}$	$l_{\min,2h}$	E	$l_{\min,h}$	$l_{\min,2h}$	E	$l_{\min,h}$	$l_{\min,2h}$
	$kh = 0.625$			$kh = 0.3125$			$k^3h^2 = 1$		
10	0.0672	3	3	0.0077	3	3	0.0077	3	3
50	0.4409	16	15	0.0503	16	16	0.0045	16	16
100	0.8818	32	31	0.0503	32	32	0.0032	32	32
500	4.670	162	155	0.5031	162	158	0.0013	162	162
1000	9.2941	324	310	1.0062	324	316	0.0009	324	324

Corollary 4.4 reveals that the projection error due to the inaccurate approximations of the eigenvectors can be represented by deviations from $\lambda_{l_{\min,h}}(B)$. In Table 2 we present the projection error for various k's. The results illustrate that the projection error increases linearly with k. Along with the projection error, the misalignment between $l_{\min,h}$ and $l_{\min,2h}$ increases, shifting the near-zero eigenvalue of A and A_{2h}. If we let $kh = 0.3125$, the projection error is reduced. However, already for $k = 1000$, the error regains magnitude, which explains why, despite resorting to a finer grid, the near-zero eigenvalues reappear when k increases. The results for $k^3h^2 = 1$ are in line with Corollary 4.5. As the step size h gets smaller, the error of the interpolation and restriction operations from the fine to the coarse grid and vice versa reduces. This explains why the projection error decreases as the wave number k increases. This can also be noticed from the last two columns of Table 2. Note that the location of the smallest eigenvalue in terms of the magnitude of A and A_{2h} are always located at the same index.

In subsection 3.2.1 we have shown that the spectrum of PA and $PM^{-1}A$ is (apart from a scaling factor) equivalent to

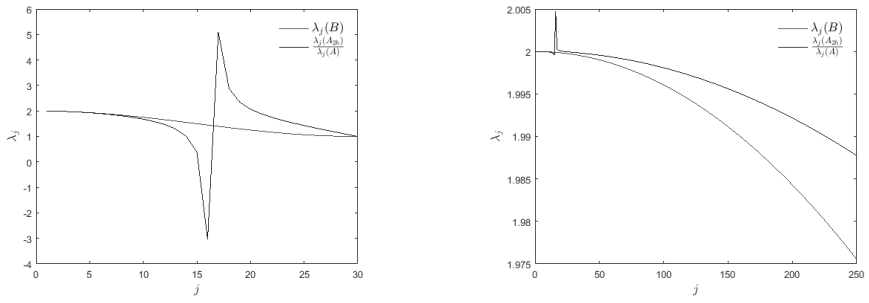
$$\hat{\beta}^l = \left| \frac{\lambda^l(A)}{\lambda^l(A_{2h})} \right|, l = 1, 2, \ldots, \frac{n}{2}.$$

From Lemma 4.1 and Corollary 4.2 we additionally found that in the limit near $l = l_{\min,h}$ we can express the eigenvalues of the coars -grid operator A_{2h} in terms of $\lambda_{l_{\min,h}}(B)$ by $\lambda^{l_{\min,h}}(A_{2h}) = \lambda^{l_{\min,h}}(A)\lambda_{l_{\min,h}}(B)$. Thus in the vicinity of the smallest eigenvalue, we can write

$$\hat{\beta}^l = \left| \frac{\lambda^l(A)}{\lambda^l(A_{2h})} \right| = \frac{1}{\lambda^l(B)}. \tag{4.2}$$

Corollary 4.4 reflects that errors in projecting the eigenvectors onto the coarse grid lead to errors in the eigenvalues of the operator B. These errors accumulate and increase as index l increases due to the eigenvectors becoming more oscillatory. If we account for these errors, then (4.2) becomes $\hat{\beta}^l = |\frac{\lambda^l(A)}{\lambda^l(A_{2h})}| = \frac{1}{\hat{\lambda}^l(B)}$, for some perturbed $\hat{\lambda}^l(B)$. These perturbations to the eigenvalues of B cause inaccurate scaling of the eigenvalues of A, eventually leading to the smallest eigenvalue of A_{2h} being located at a different index $l_{\min,2h} \neq l_{\min,h}$.

In Figures 2(a) and 2(b) we have plotted the eigenvalues of B and the ratio between the eigenvalues of A_{2h} and A according to (4.2). Note that the latter essentially represents the perturbed $\lambda^l(B)$ due to errors accumulated during prolongating and restricting the eigenvectors of A. It can be noted that as h becomes smaller, the ratio slowly converges to $\lambda^l(B)$. This observation is also in line with the projection error decreasing.



(a) $kh = 0.825$ (b) $kh = 0.01$

FIG. 2. $k = 50$. *Plot of the ratio between the fine-grid and coarse-grid eigenvalues* (4.2) *and the eigenvalues of* B. $l_{\min,h} = 16$ *and* $l_{\min,2h} = 15$ *for* $kh = 0.825$. *For* $kh = 0.01$, $l_{\min,h} = l_{\min,2h} = 16$.

5. Higher-order DEF.

5.1. Quadratic approximation. Recall that the grid transfer functions $u_2h = [u_{2h_1}, \ldots, u_{2h_n}]$ from Ω_{2h} to the fine grid Ω_h using standard linear interpolation are given by

$$I_{2h}^{h} : \Omega_{2h} \to \Omega_h, \quad u_{2h} \to I_{2h}^{h}\, u_{2h} \tag{5.1}$$

such that

$$\begin{cases} [u_{2h}]_{i/2} & \text{if } i \text{ is even,} \\ \frac{1}{2}\left([u_{2h}]_{(i-1)/2} + [u_{2h}]_{(i-1)/2}\right) & \text{if } i \text{ is odd,} \end{cases} \quad i = 1, \ldots, n-1. \tag{5.2}$$

A closer look reveals that the current transfer functions are only reinforced at the odd components, leaving the even components unchanged. In fact, these components are mapped to linear combination of their fine-grid counterparts ϕ_{h_l} and a complementary mode $\phi_{h_{n+1-l}}$ with first-order accuracy [11]. A more general representation of the linear interpolation operator for the even components can be given by using rational Bézier curves, which are defined in Definitions 5.1, 5.2, and 5.3. The use of these curves within the context of multigrid methods has been studied in [4, 24]. Using these vectors as vectors for the input of the prolongation and restriction matrices in a multigrid setting is referred to as a *monotone multigrid method*. The monotonicity comes from the construction of the coarse-grid approximations, which ensures that the coarse-grid functions approximate the fine-grid functions monotonically [24, 33]. The higher-order approximation schemes are defined in Definition 5.4.

DEFINITION 5.1 (Bézier curve). *A Bézier curve of degree* n *is a parametric curve defined by*

$$B(t) = \sum_{j=0}^{n} b_{j,n}(t) P_j, \quad 0 \le t \le 1, \quad \textit{where the polynomials}$$
$$b_{j,n}(t) = (n, j)\, t^j (1-t)^{n-j}, \quad j = 0, 1, \ldots, n,$$

are known as the Bernstein basis polynomials of order n. *The points* P_j *are called control points for the Bézier curve.*

DEFINITION 5.2 (rational Bézier curve). *A rational Bézier curve of degree n with control points $P_0, P_1, \ldots, P_n$ and scalar weights $w_0, w_1, \ldots, w_n$ is defined as*

$$C(t) = \frac{\sum_{j=0}^{n} w_j b_{j,n}(t) P_j}{\sum_{j=0}^{n} w_j b_{j,n}(t)}.$$

DEFINITION 5.3 (linear interpolation). *Let $[u_{2h}]_{(j-1)/2}$ and $[u_{2h}]_{(j+1)/2}$, be the endpoints within a component span defined on the coarse grid. Then the prolongation scheme for the even nodes can be characterized by a rational Bézier curve of degree* 1 *with polynomials*

$$b_{0,1}(t) = 1 - t,$$
$$b_{1,1}(t) = t,$$

whenever j is odd by taking the weights $w_0 = w_1 = 1$ and $t = \frac{1}{2}$. Note that in the case $w_0 = w_1$ and is nonrational we obtain the original Bézier curve:

$$C\left(\frac{1}{2}\right) = \frac{\frac{1}{2}[u_{2h}]_{(j-1)/2} + (1 - \frac{1}{2})[u_{2h}]_{(j+1)/2}}{\frac{1}{2} + (1 - \frac{1}{2})} \tag{5.3}$$

$$= \frac{1}{2}\left([u_{2h}]_{(j-1)/2} + [u_{2h}]_{(j+1)/2}\right). \tag{5.4}$$

When j is even, we take the middle component $[u_{2h}]_{j/2}$, which itself gets mapped onto the fine grid.

For large k, the prolongation operator working on the even components is not sufficiently accurate to map the near kernels to adjacent modes on Ω_{2h} and Ω_h. Consequently, we wish to find a higher-order approximation scheme, which takes the even components into account. We thus consider a quadratic rational Bézier curve in order to find appropriate coefficients to yield a higher-order approximation of the fine-grid functions by the coarse-grid functions.

DEFINITION 5.4 (quadratic approximation). *Let $[u_{2h}]_{(j-2)/2}$ and $[u_{2h}]_{(j+2)/2}$ be the endpoints within a component span defined on the coarse grid. Then the prolongation operator can be characterized by a rational Bézier curve of degree* 2 *with polynomials*

$$b_{0,2}(t) = (1 - t)^2,$$
$$b_{1,2}(t) = 2t(1 - t),$$
$$b_{2,2}(t) = t^2,$$

and control point $[u_{2h}]_{j/2}$, whenever j is even. Because we wish to add more weight

to the center value, we take weights $w_0 = w_2 = \frac{1}{2}$, $w_1 = \frac{3}{2}$, *and* $t = \frac{1}{2}$ *to obtain*

$$\begin{aligned} C(t) &= \frac{\frac{1}{2}(1-t)^2[u_{2h}]_{j-1} + \frac{3}{2}2t(1-t)[u_{2h}]_j + \frac{1}{2}(t)^2[u_{2h}]_{j+1}}{\frac{1}{2}(1-t)^2 + \frac{3}{2}2t(1-t) + \frac{1}{2}(t)^2} \\ &= \frac{\frac{1}{2}(1-\frac{1}{2})^2[u_{2h}]_{j-1} + \frac{3}{2}(2)(\frac{1}{2})(1-\frac{1}{2})[u_{2h}]_j + \frac{1}{2}(\frac{1}{2})^2[u_{2h}]_{j+1}}{\frac{1}{2}(1-\frac{1}{2})^2 + \frac{1}{2}(2)(\frac{1}{2})(1-\frac{1}{2}) + \frac{1}{2}(\frac{1}{2})^2} \\ &= \frac{\frac{1}{8}[u_{2h}]_{j-1} + \frac{3}{4}[u_{2h}]_j + \frac{1}{8}[u_{2h}]_{j+1}}{1} \\ &= \frac{1}{8}\left([u_{2h}]_{j-1} + 6[u_{2h}]_j + [u_{2h}]_{j+1}\right). \end{aligned} \tag{5.5}$$

When j *is odd,* $[u_{2h}]_{(j-1)/2}$ *and* $[u_{2h}]_{(j+1)/2}$ *have an even component and we are in the same scenario as is the case with linear interpolation.*

Based on the upper scheme, we redefine the prolongation and restriction operator as follows:

$$I_{2h}^{h}\left[u_{2h}\right]_i = \left\{ \begin{array}{cc} \frac{1}{8}\left(\left[u_{2h}\right]_{(i-2)/2} + 6\left[u_{2h}\right]_{(i)/2} + \left[u_{2h}\right]_{(i+2)/2}\right) & \text{if } i \text{ is even,} \\ \frac{1}{2}\left(\left[u_{2h}\right]_{(i-1)/2} + \left[u_{2h}\right]_{(i+1)/2}\right) & \text{if } i \text{ is odd} \end{array} \right\}, \tag{5.6}$$

for $i = 1, \ldots, n-1$ and

$$I_{h}^{2h}\left[u_h\right]_i = \frac{1}{8}\left(\left[u_h\right]_{(2i-2)} + 4\left[u_h\right]_{(2i+1)} + 6\left[u_h\right]_{(2i)} + 4\left[u_h\right]_{(2i+1)} + \left[u_h\right]_{(2i+2)}\right)$$

for $i = 1, \ldots, \frac{n}{2}$.

Using the new matrices I_{2h}^{h} and I_{h}^{2h}, we can now construct similar analytical expressions for the eigenvalues of A_{2h}, PA, and $P^TM^{-1}A$, where we follow the same approach as [11, 12, 19]. In these works, the basis consisting of eigenvectors is reordered and the projection operator P is block-diagonalized. This allows thorough spectral analysis of each eigenvalue of PA for MP 1-A as each block now contains the nonzero analytical eigenvalues. We therefore start by following a similar approach with respect to the block-diagonalization by reordering the basis consisting of the eigenvectors as follows:

$$V = \left\{\phi_h^1, \phi_h^{(n+1)-1}, \phi_h^2, \phi_h^{(n+1)-2}, \ldots, \phi_h^{\frac{n}{2}}, \phi_h^{(n+1)-\frac{n}{2}}\right\}.$$

Here the fine-grid eigenvectors are given by $\phi_h^l = \sin(l\pi h)$ and the coarse-grid eigenvectors are obtained by substituting $2h$ for h. The prolongation operator maps the coarse-grid eigenvectors for indices $j, l = 1, 2, \ldots \frac{n}{2}$, to

$$\begin{aligned} [I_h^{2h}\phi_{2h}]_j^l &= \frac{1}{8}\left[\sin((j-2)/2)l\pi 2h) + 6\sin((j)/2)l\pi 2h) + \sin((j+2)/2)l\pi 2h)\right] \\ &= \left[\frac{1}{4}\cos(2l\pi h) + \frac{3}{4}\right]\sin(lj\pi h) \end{aligned}$$

for j even and

$$\begin{aligned} [I_h^{2h}\phi_{2h}]_j^l &= \frac{1}{8}\left[4\sin((j-1)/2)l\pi 2h) + 4\sin((j+1)/2)l\pi 2h)\right] \\ &= \left[\cos(l\pi h)\right]\sin(lj\pi h) \end{aligned}$$

for j odd. With respect to the remaining part of the index set containing j, we use that

$$\phi_h^{n+1-l_j} = -(-1)^j \sin(lj\pi h), \quad j = 1, 2, \dots n-1, \text{and } l = 1, 2, \dots \frac{n}{2}. \tag{5.7}$$

Note that (5.7) is only positive when j is odd. Consequently for even j such that $j \in \left\{\frac{n}{2}, \dots, n-1\right\}$ is even, we obtain

$$\begin{aligned}[I_h^{2h}\phi_{2h}]_j^l &= \frac{1}{8}\left[-\sin((j-2)/2)l\pi 2h) - 6\sin((j)/2)l\pi 2h) - \sin((j+2)/2)l\pi 2h)\right] \\ &= \left[-\frac{1}{4}\cos(2l\pi h) - \frac{3}{4}\right]\sin(lj\pi h),\end{aligned}$$

whereas for j odd, we now have

$$\begin{aligned}[I_h^{2h}\phi_{2h}]_j^l &= \frac{1}{8}\left[4\sin((j-1)/2)l\pi 2h) + 4\sin((j+1)/2)l\pi 2h)\right] \\ &= [\cos(l\pi h)]\sin(lj\pi h).\end{aligned}$$

With respect to our basis, we therefore obtain the following 2×1 block for the prolongation operator:

$$[I_{2h}^h]^l = \begin{bmatrix}\cos(l\pi h) + \frac{1}{4}\cos(2l\pi h) + \frac{3}{4} \\ \cos(l\pi h) - \frac{1}{4}\cos(2l\pi h) - \frac{3}{4}\end{bmatrix}.$$

Similarly, the restriction operator is defined by taking $[I_{2h}^h]^{l^T}$ and thus we obtain a 1×2 block. For ease of notation, we now define

$$\begin{aligned}v^l &= \cos(l\pi h) + \frac{1}{4}\cos(2l\pi h) + \frac{3}{4}, \\ v^{n+1-l} &= \cos(l\pi h) - \frac{1}{4}\cos(2l\pi h) - \frac{3}{4}.\end{aligned}$$

Using these expressions, we can now compute the eigenvalue of the Galerkin coarse-grid operator, which is given by the 1×1 diagonal block

$$\lambda^l(A_{2h}) = [I_{2h}^h]^l A^l [I_h^{2h}]^l = \left(v^l\right)^2 \lambda^l(A) + \left(v^{n+1-l}\right)^2 \lambda^{n+1-l}(A). \tag{5.8}$$

In order to obtain the eigenvalues of PA, we have to compute the 2×2 diagonal blocks of the projection operator P first. Recall that P is defined by

$$P^l = I - (I_{2h}^h)^l (A_{2h}^l)^{-1} (I_h^{2h})^l A^l.$$

We thus obtain the following block system:

$$\begin{aligned}P^l &= \begin{bmatrix}1 - \frac{(v^l)^2}{\lambda^l(A_{2h})} & \frac{v^l v^{n+1-l}}{\lambda^l(A_{2h})} \\ \frac{v^{n+1-l}v^l}{\lambda^l(A_{2h})} & 1 - \frac{(v^{n+1-l})^2}{\lambda^l(A_{2h})}\end{bmatrix}\begin{bmatrix}\lambda^l(A) & 0 \\ 0 & \lambda^{n+1-l}(A)\end{bmatrix} \\ &= \begin{bmatrix}\lambda^l(A)\left(1 - \frac{(v^l)^2}{\lambda^l(A_{2h})}\right) & \lambda^{n+1-l}(A)\left(\frac{v^l v^{n+1-l}}{\lambda^l(A_{2h})}\right) \\ \lambda^l(A)\left(\frac{v^{n+1-l}v^l}{\lambda^l(A_{2h})}\right) & \lambda^{n+1-l}(A)\left(1 - \frac{(v^{n+1-l})^2}{\lambda^l(A_{2h})}\right)\end{bmatrix}.\end{aligned} \tag{5.9}$$

From here, we retrieve the eigenvalues of PA by multiplying (5.9) again with the 2×2 diagonal block containing the eigenvalues of A with respect to index l on our defined basis:

$$[PA]^l = \begin{bmatrix} (\lambda^l(A))^2\left(1-\frac{(v^l)^2}{\lambda^l(A_{2h})}\right) & (\lambda^{n+1-l}(A))^2\left(\frac{v^l v^{n+1-l}}{\lambda^l(A_{2h})}\right) \\ (\lambda^l(A))^2\left(\frac{v^{n+1-l}v^l}{\lambda^l(A_{2h})}\right) & (\lambda^{n+1-l}(A))^2\left(1-\frac{(v^{n+1-l})^2}{\lambda^l(A_{2h})}\right) \end{bmatrix}. \tag{5.10}$$

Similarly, the eigenvalues of $P^TM^{-1}A$ are obtained by simply multiplying (5.9) with the 2×2 block containing the eigenvalues of $M^{-1}A$ instead of A and computing the trace. This operation leads to the following analytical expressions for the eigenvalues of $P^TM^{-1}A$ for $l = 1, 2, \ldots, \frac{n}{2}$:

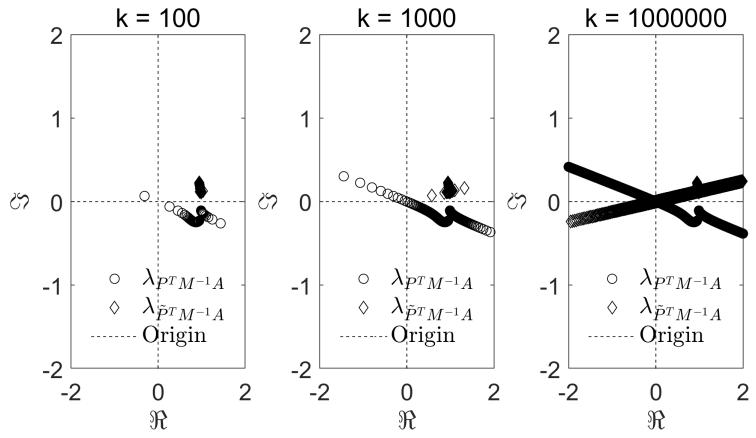
$$\lambda^l(P^TM^{-1}A) = \frac{(\lambda^l(A))^2}{\lambda^l(M)}\left(1-\frac{(v^l)^2}{\lambda^l(A_{2h})}\right) + \frac{(\lambda^{n+1-l}(A))^2}{\lambda^l(M)}\left(1-\frac{(v^{n+1-l})^2}{\lambda^l(A_{2h})}\right). \tag{5.11}$$

Using (5.11), we can proceed with the spectral analysis of the DEF preconditioner for MP 1-A.

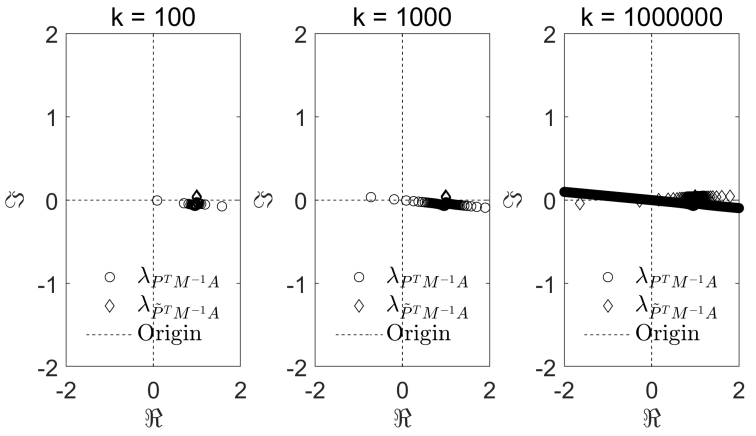
5.2. Spectral analysis. In order to keep track of both (original and adapted) DEF-based preconditioned systems, we will use the $\sim$ notation to denote the adapted system. We will now compare the spectrum of the DEF + CSLP preconditioned matrix ($P^TM^{-1}A$), with the adapted deflation + CSLP precondtioned matrix ($\tilde{P}^TM^{-1}A$) for MP 1-A. In Figure 3 we have plotted the spectrum of both $P^TM^{-1}A$ (dot marker) and $\tilde{P}^TM^{-1}A$ (diamond marker) for very large wave numbers. Starting with the results for $kh = 0.625$, we note that incorporating the new DEF scheme leads to a remarkable reduction in the near-zero eigenvalues. Compared to the original DEF scheme, the spectrum of the adapted scheme is more densely located near the point $(1, 0)$. As a result, the spectrum of the adapted scheme has shorter tails. For example, for $k = 10^3$, there are almost no near-zero eigenvalues. However, as k increases to 10^6, we see the near-zero eigenvalues reappearing. If we switch to a finer grid using $kh = 0.3125$ in Figure 3(b), we observe an even greater improvement. For $k = 10^6$ a few eigenvalues are slightly moving towards the origin; however, these results are negligible compared to the magnitude of the wave number. Table 3 contains the projection error according to Corollary 4.4 for both schemes. The projection error for the new scheme is reduced significantly. However, as k increases we observe that the projection error increases accordingly, which is in line with the spectral analysis.

5.2.1. Parameter sensitivity. We have seen that for very large k such as $k = 10^6$, the adapted scheme using $\tilde{P}$ still has a small number of near-zero eigenvalues. This result is supported by the increasing projection error for $kh = 0.625$ (see Table 3). One explanation is that for these large wave numbers, the low-frequency eigenmodes corresponding to $l^h_{\min}$ for A and $l^{2h}_{\min}$ for $\tilde{A}_{2h}$ are still very oscillatory vectors. Furthermore, apart from these eigenmodes themselves being relatively oscillatory, the high-frequency modes which get activated are again a source for approximation errors when prolonging the coarse-grid eigenvectors. Necessarily, at some point, the scheme based on the adapted DEF vectors will again suffer from accumulation errors as their approximation power reduces when k increases.

One of the characteristics of Bézier curves implies that at systematic intervals some discontinuities appear as sharp corners at certain points [34]. If the eigenvectors become oscillatory due to the wave number being very large, then keeping the grid



(a) $kh = 0.625$



(b) $kh = 0.3125$

FIG. 3. *Eigenvalues of $P^TM^{-1}A$ and $\tilde{P}^TM^{-1}A$. The top row contains the spectrum of $P^TM^{-1}A$ and $\tilde{P}^TM^{-1}A$ for $kh = 0.625$. The bottom row contains the eigenvalues for $kh = 0.3125$.*

TABLE 3
Projection error for the old scheme E and the adapted scheme $\tilde{E}$.

k	E	$\tilde{E}$	E	$\tilde{E}$
	$kh = 0.625$		$kh = 0.3125$	
10^1	0.0672	0.0049	0.0077	0.0006
10^2	0.8818	0.0154	0.1006	0.0008
10^3	9.2941	0.1163	1.0062	0.0014
10^4	92.5772	1.1021	10.0113	0.007
10^5	926.135	10.9784	100.1382	0.0635
10^6	9261.7129	109.7413	1001.3818	0.6282

resolution constant, these discontinuities become a source of approximation error. Instead of diverting to higher-order approximation schemes, the use of rational Bézier curves allow simple modifications which can alter the shape and movement of the utilized curve segments. In fact, the weights of the rational Bézier curve are shape

parameters, which allow control over the curve segments. For example, increasing the weight corresponding to a control point forces the curvature to move more closely and sharply to that control point. Decreasing the weight of a control point, on the other hand, results in the curve flattening and expanding more towards its endpoints. In our case, the quadratic approximation using the rational Bézier curve has one control point per segment. This would lead to the following redefinition:

$$I_{2h}^{h}\left[u_{2h}\right]_i = \begin{cases} \left(\frac{1}{8}\left[u_{2h}\right]_{(i-2)/2} + \left(\frac{3}{4}-\varepsilon\right)\left[u_{2h}\right]_{(i)/2} + \frac{1}{8}\left[u_{2h}\right]_{(i+2)/2}\right) & \text{if } i \text{ is even,} \\ \frac{1}{2}\left(\left[u_{2h}\right]_{(i-1)/2} + \left[u_{2h}\right]_{(i+1)/2}\right) & \text{if } i \text{ is odd} \end{cases}$$

for $i = 1, \ldots, n-1$, and $\varepsilon > 0$. The new scheme alters the expressions for the eigenvalues of $\tilde{P}^T M^{-1} A$ according to

$$\tilde{v}^l = \cos(l\pi h) + \frac{1}{4}\cos(2l\pi h) + \left(\frac{3}{4} - \varepsilon\right),$$
$$\tilde{v}^{n+1-l} = \cos(l\pi h) - \frac{1}{4}\cos(2l\pi h) - \left(\frac{3}{4} - \varepsilon\right).$$

Straightforward substitutions of the altered expressions for $\tilde{v}^l$ and $\tilde{v}^{n+1-l}$ into (5.10) render the analytical expressions for the eigenvalues of $\tilde{P}^T M^{-1} A$. The next question which needs to be answered is, given a fixed kh, how do we find ε? ε should be chosen such that the projection error E is minimized. In order to find this value, we can use two approaches. The first approach is straightforward; our ultimate aim is to have the eigenvalue of $\lambda^l(\tilde{P}^T M^{-1} A)$ at index $l_{\min,h}$ to be equal to 1. Recall from the proof of Corollary 4.2 that in the absence of errors the eigenvalues of A_{2h} can be written as a product of the eigenvalues of A and the eigenvalues of B. Thus, using (5.8), we can write

$$\begin{aligned} \lambda^l(A_{2h}) &= [I_{2h}^h]^l A^l [I_h^{2h}]^l, \\ &= \left(v^l\right)^2 \lambda^l(A) + \left(v^{n+1-l}\right)^2 \lambda^{n+1-l}(A) = \lambda^l(A)\lambda^l(B). \end{aligned} \tag{5.12}$$

Note that the sum of $(v^l)^2$ and $(v^{n+1-l})^2$ in expression (5.12) are exactly equal to $\lambda^l(B)$. If we want (5.12) to hold at index $l_{\min,h}$ in the presence of errors, we need to pick ε such that $(v^{n+1-l})^2 = 0$, which is equivalent to

$$\varepsilon = 0.75 - \left(\cos(l\pi h) - \frac{1}{4}\cos(2l\pi h)\right). \tag{5.13}$$

This way the near-zero eigenvalue of A_{2h} will always be proportional to the near-zero eigenvalue of A. Fortunately, the eigenvalues of B containing the term ε are independent of the eigenvalues of A. Therefore, finding ε primarily depends on the approximation scheme which determines the eigenvalues of B. An interesting observation is that ε is completely determined by the step size h and therefore by the grid resolution kh.

We can take advantage of this k-independence, as it enables us to determine an ε without having to account for the wave number. Also, once we find an ε which works for some kh, then it will work for all k as long as kh remains constant. Thus, especially for practical applications of higher-dimensional problems, instead of computing the exact smallest eigenvalues of the fine- and coarse-grid operator, we can find the ε by

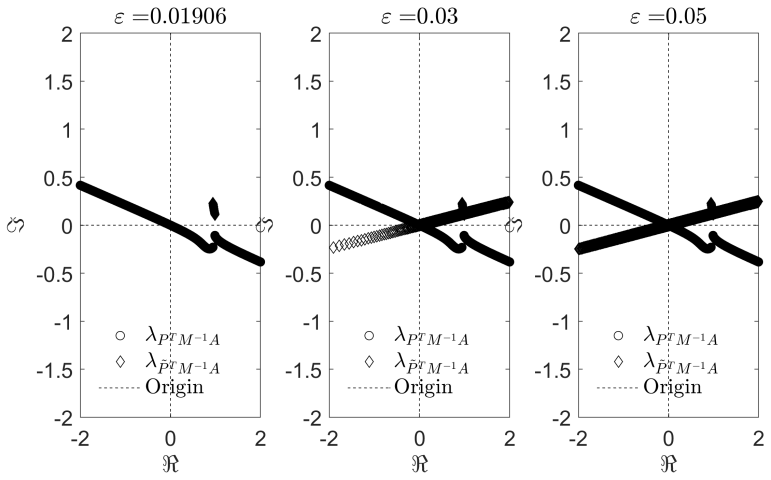


FIG. 4. *Eigenvalues of $P^TM^{-1}A$ and $\tilde{P}^TM^{-1}A$ using $kh = 0.625$ for various weight parameters ε. The wave number k has been set to 10^6.*

performing a grid search for some small k. A similar strategy was used in [16] for the open cavity problem in order to find the optimal parameter for a given k and a given paritition in the context of optimized Schwarz methods (with overlap). There, the best parameter was chosen to be the one which resulted in the smallest GMRES residual. In our case, the best parameter ε is the one which minimizes the projection error for some fixed h. Therefore, for MP 2 and MP 3, we will use the heuristic in Algorithm 5.1. This provides a practical alternative to computing the analytical expressions for the eigenvalues of B.

Algorithm 5.1 Projection error minimizer.

Initialize k small, $\phi_{j_{\min}}$, $\varepsilon_0 = 0.0001$, tol $= 10^{-4}$
for $c = 1, 2, \ldots, m$ **do**
  Compute $E_{c\varepsilon_0}$ using $c\varepsilon_0$ to construct Z
  $y1 = Z^T\phi_{j_{\min}}$, $w = \phi^T_{j_{\min}}Z, B = Z^TZ$
  $By2 = y1$, solve for $y1$ (direct or iteratively)
  $E_{c\varepsilon_0} = \phi^T_{j_{\min}}\phi_{j_{\min}} - wy2$
  while $\bar{E}_{c\varepsilon_0} >$ tol **do**
    Compute $E_{(c+1)\varepsilon_0}$ and repeat until $\bar{E}_{(c+1)\varepsilon} <$ tol
  end while
end for
Set $\varepsilon = \tilde{c}\varepsilon_0$ for some $\tilde{c} \in [1, m]$.

We proceed by reexamining the spectrum of MP 1-A for $k = 10^6$ after introducing the weight parameter. We have plotted the eigenvalues for $kh = 0.625$ for $\varepsilon = 0.01906$ (left), $\varepsilon = 0.03$ (center), and $\varepsilon = 0.05$ (right) in Figure 4. It immediately becomes apparent that using the right parameter to minimize the projection error completely shifts the spectrum. Particularly, the left column contains the results where the optimal ε has been used and it can be noted that the spectrum stays clustered near $(1, 0)$ independent of the wave number k.

In the next section, we provide numerical experiments with these parameters for MP 1-A in order to test whether we obtain a reduced number of iterations as theorized.

TABLE 4

Number of GMRES iterations for MP 1-A *using the adapted preconditioned DEF scheme APD(ε). ε has been determined using* (5.13). *APD*(0) *is the adapted DEF scheme without the projection error minimizer ε. The shift in CSLP has been set to* $(1, 1)$ *and the preconditioner has been inverted inexactly.*

k	APD(0.1250)	APD(0.0575)	APD(0.01906)	**APD(0)**	APD(0.00125)	**APD(0)**
	$kh = 1$	$kh = 0.825$	$kh = 0.625$	$kh = 0.625$	$kh = 0.3125$	$k^3h^2 = 1$
10^1	2	3	4	**4**	3	**4**
10^2	6	5	4	**4**	3	**4**
10^3	6	5	4	**6**	3	**4**
10^4	6	5	4	**12**	3	**4**
10^5	6	5	4	**59**	3	**4**
10^6	6	5	4	**509**	3	**4**

TABLE 5

Projection error for MP 1-A *E(ε) for various kh. ε has been determined using* (5.13).

k	E(0.1250)	E(0.0575)	E(0.01906)	E(0.00125)
	$kh = 1$	$kh = 0.825$	$kh = 0.625$	$kh = 0.3125$
10^1	0.0127	0.0075	0.0031	0.0006
10^2	0.0233	0.0095	0.0036	0.0007
10^3	0.0245	0.0095	0.0038	0.0007
10^4	0.0246	0.0095	0.0038	0.0007
10^5	0.0246	0.0095	0.0038	0.0007
10^6	0.0246	0.0095	0.0368	0.0007

6. Numerical experiments. We start by examining the convergence behavior of the adapted solver using various kh. Unless stated otherwise, we deploy the CSLP preconditioner with $(\beta_1, \beta_2) = (1, 1)$ as we approximate the inverse of M using one $V(1, 1)$-multigrid iteration. The tolerance level for the relative residual has been set to 10^{-7}.

6.1. One-dimensional constant wave number model.

6.1.1. MP 1-A. For MP 1-A the results are presented in Tables 4 and 5. Table 4 gives the number of iterations and Table 5 provides the projection error for increasing k. The numerical results presented are in line with with the theoretical results from section 4 and the spectral analysis from Figure 5. The consistently clustered spectrum near $(1, 0)$ is reflected in a significant reduction in the number of iterations. On coarser levels, the number of iterations is still constant yet higher. In particular, compare the 6 iterations for $kh = 1$ with the 5 iterations for $kh = 0.825$. Even for such a simple model problem as MP 1-A, these results present the first numerical evidence of obtaining true wave number independent convergence for very high wave numbers without having to resort to keeping the shift in the CSLP preconditioner small and inverting the preconditioner exactly.

If we keep the grid resolution economical, $kh = C$, where $C \in [0.3125, 1]$, we observe that, unlike the previous DEF scheme using linear interpolation, the adapted scheme has an almost constant projection error as the wave number increases; see Table 5. With respect to the pollution error, it is necessary to keep the grid resolution $k^3h^2 \approx 1$. The last column of Table 4 contains the number of iterations using $k^3h^2 \approx 1$. These results are in line with the theory from section 4, Corollary 4.5, and corroborate that an increasing wave number in fact leads to a lower projection error (Table 5) and hence a decreasing number of iterations (Table 4). This brings us to the final observation. The use of the weight parameter ε becomes redundant in the case we

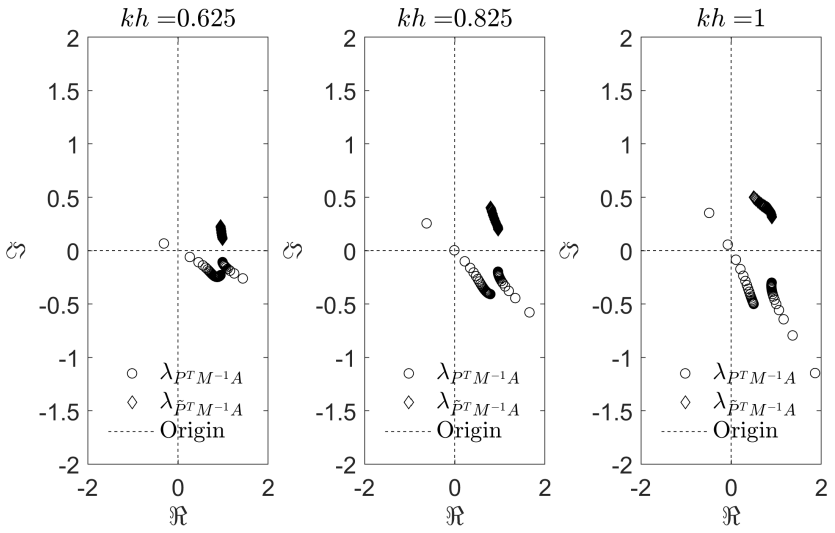


FIG. 5. *Eigenvalues for $k = 10^6$ of $P^TM^{-1}A$ and $\tilde{P}^TM^{-1}A$ using various kh. The weight parameter ε has been determined using* (5.13).

TABLE 6
Number of GMRES iterations for MP 1-B *using APD(ε) and* Sommerfeld *radiation conditions. ε has been determined using* (5.13). *The shift in CSLP has been set to* $(1, 1)$ *and has been inverted inexactly.*

k	APD(0.1250)	APD(0.0575)	APD(0.01906)	APD(0.00125)	**APD(0)**
	$kh = 1$	$kh = 0.825$	$kh = 0.625$	$kh = 0.3125$	$k^3h^2 = 1$
10^1	2	3	5	4	**5**
10^2	8	6	5	4	**5**
10^3	8	6	5	4	**5**
10^4	8	6	5	4	**5**
10^5	8	6	5	4	**5**
10^6	8	6	5	4	**5**

let $k^3h^2 = 1$. Recall that the weight parameter is necessary in order to capture the perturbations which arise in mapping the eigenvectors as the wave number increases. Corollary 4.5 shows why this becomes unneccessary as the mappings naturally become more accurate as we let h go to zero.

Finally, compared to the CSLP preconditioner which shows h-independent convergence behavior, the use of the APD preconditioner could allow for more accurate solutions while keeping the number of iterations constant and small. For example, one could use a higher-order finite difference scheme, combined with a coarser grid resolution in order to solve large scale problems more accurately without being penalized by an increased number of iterations.

6.1.2. MP 1-B. Table 6 contains the results for MP 1-B. We observe that including Sommerfeld radation conditions does not lead to deviating conclusions. While the results of the rigorous Fourier analysis for MP 1-A are not analogously applicable to the case where we use Sommerfeld radiation conditions, we have used the same values for ε determined for MP 1-A and observe that the convergence behavior is very similar. This provides numerical evidence for the notion that the convergence

TABLE 7

Number of iterations for MP 2 using $kh = 0.625$ using APD(ε). ε has been optimized using Algorithm 5.1. Approximate CSLP inversion using one V(1, 1) cycle. Exact inversion includes the CSLP shift $(1, k^{-1})$. AD contains no CSLP preconditioner.

k	n^2	APD(0)	AD(0.01906)	APD(0.01906)	APD(0)	CSLP
		Approximate inversion			Exact inversion	
50	6400	4	13	5	3	9
100	25600	5	13	6	3	12
250	160000	10	13	6	5	20
500	640000	15	14	8	5	28
750	1440000	37	16	9	7	36
1000	2560000	53	18	9	8	45

behavior for MP 1-A and MP 1-B are very similar and in both cases we obtain pure wave number independent convergence.

6.2. Two-dimensional constant wave number model. In this section perform numerical experiments for the two-dimensional model problem using a constant wave number k and Dirichlet boundary conditions. The weight parameter ε has been optimized using Algorithm 5.1. Table 7 contains the number of iterations for $kh = 0.625$. We start with the case where we use the APD scheme without using the weight parameter ε. In this case, the third column shows that we can solve for $k = 1000$ in 53 iterations. To see the effect of the DEF technique without the influence of the CSLP preconditioner, the fourth column contains the number of iterations for the AD scheme including the weight parameter. Remarkably, we can solve for $k = 1000$ in 18 iterations. Finally, combining both the weight parameter and the approximate inversion of the CSLP preconditioner, it takes 9 iterations to solve for $k = 1000$. If we would have inverted the CSLP preconditioner exactly using a small shift to compensate for the use of no weight parameter, it would take the solver 8 iterations to solve for $k = 1000$. These results are almost similar, but the use of the weight parameter and approximate inversion is less computationally expensive compared to exact inversion of the CSLP preconditioner. This is very promising as this implies that we can include a powerful preconditioner without having to pay the price of exact inversion at the finest level. While we do see a slight increase in the number of iterations throughout Table 7, these are the lowest reported number of iterations for a sequential implementation using such high wave numbers. Without the use of the DEF preconditioner, CSLP preconditioned GMRES would need 45 iterations to converge despite using a small shift of order $k^{-1} = 10^{-3}$.

We now repeat the same analysis for $kh = 0.3125$, with results reported in Table 8. Note that in this case we do not include an adjusted weight coefficient parameter, i.e., we set $\varepsilon = 0$. The inclusion of ε may in particular be more useful when using coarser grids. The reason behind this is that increasing the problem size already results in more accuracy and faster convergence (see Corollary 4.5). We also compare the performance of the adapted scheme with and without the inclusion of the CSLP preconditioner. Results are reported in Table 8. If we compare these results to the ones obtained from Table 7, we note that, with the inclusion of the CSLP preconditioner, increasing the problem size leads to faster convergence as theorized. Two important remarks can be made with respect to letting $kh = 0.3125$. First of all, in the case we set $\varepsilon = 0$, we go from 53 iterations for $kh = 0.625$ to 8 iterations for $kh = 0.3125$ when $k = 1000$. However, once we include the weight parameter (Table 7, column 5), we obtain 9 iterations for $kh = 0.625$ and 8 iterations for $kh = 0.3125$ and the

TABLE 8

Number of iterations for MP 2 *using* $kh = 0.3125$ *using APD*(ε). *Approximate CSLP inversion using one* $V(1,1)$*-cycle. AD contains no CSLP preconditioner.*

k	n^2	AD(0)	APD(0)
		Iterations	Iterations
25	6400	4	4
50	25600	4	4
100	102400	3	4
250	640000	4	4
500	2560000	5	5
750	5760000	5	5
1000	10240000	7	8

TABLE 9

Results for the Marmousi problem using 10 *gpw. All solvers are combined with the inexact inversion of the CSLP preconditioner using shifts* $(1,1)$. *TL denotes two level.*

f	DEF-TL	APD-TL	DEF-TL	APD-TL
	Iterations		Solve time (s)	
1	3	6	1.72	4.08
10	16	5	7.30	3.94
20	31	5	77.34	19.85
40	77	5	1175.99	111.78

TABLE 10

Results for the Marmousi problem using 10 *gpw using no CSLP preconditioner. TL denotes two level.*

f	DEF-TL	APD-TL	DEF-TL	APD-TL
	Iterations		Solve time (s)	
1	10	12	1.41	2.76
10	20	12	2.44	2.80
20	35	12	17.15	15.15
40	82	12	219.39	85.87

convergence behaviors become very similar irrespective of using a finer grid resolution. Second of all, the number of iterations with and without the CSLP preconditioner is almost the same for all reported values of k in Table 8. It may be argued that for fine grid resolutions, some computational time can be saved by excluding the CSLP preconditioner as we need one multigrid iteration to approximate the inverse. The numerical results from the previous and current section show that there are plenty of optimization strategies to exploit when it comes to balancing a small and fixed number of iterations and a fine-grid resolution. The latter is equally important to obtain accurate solutions.

6.3. Two-dimensional nonconstant Marmousi model. In this section we present the numerical results for the industrial two-dimensional Marmousi problem (MP 4) (section 2). Results are reported in Tables 9 and 10. Starting with Table 9 we implement no correction using ε given that the grid for this model problem has been resolved such that $kh \leq 0.39$ on average and the maximum wave number is approximately 400.[2] Table 9 contains the results for frequencies $f = 1, 10, 20$, and 40 using 10 grids points per wavelength (gpw) for the largest wave number k. The results show that even for this challenging problem, the APD scheme leads to very

[2]If we use the dimensionless model we obtain a wave number of $\sqrt{\frac{2\pi 40}{2587.5}^2 \times 2048 \times 8192} \approx 398$.

TABLE 11
Number of iterations for MP 3 using $kh = 0.625$. AD contains no CSLP preconditioner. APD contains the CSLP with shift $(1, 1)$, which has been inverted inexactly.

k	n^3	APD(0)	APD(0.00125)
		Iterations	Iterations
5	512	4	4
10	4096	4	4
25	64000	5	4
50	512000	5	4
75	1728000	6	4

satisfactory results. If we compare the results between DEF-TL and APD-TL, we note an improved performance in terms of both metrics: solve time and iterations. For $f = 1$, the number of iterations for APD-TL is larger than DEF-TL. The latter method takes 6 iterations, while the former takes 3 iterations, which is clearly reflected in the lower solve time. Once we start increasing the frequency, we note that the APD-TL scheme quickly catches up in terms of both iterations and solve time. For example for $f = 40$, we obtain 5 iterations and a total solve time of 111.78 seconds.

Table 10 repeats the same simulation without the use of the CSLP preconditioner. We observe very similar behavior as compared to the results obtained for the constant wave number problem (Table 7). Excluding the CSLP preconditioner and solely using the DEF preconditioner results in a constant number of iterations and a significant reduction in sequential solve time. If we use the old DEF preconditioner based on the linear interpolation scheme, then a similar effect cannot be observed. For example, for $f = 40$ Hz, we obtain 82 iterations versus 12 for the adapted scheme. These results provide a promising basis for future research where the coarse-grid solve can be optimized and balanced with respect to the number of iterations and time scalability of the overarching solver.

6.4. Three-dimensional constant wave number model. In this section we present some three-dimensional numerical results for MP 3. We have used the same weight parameter ε from the two-dimensional test problem MP 2. From Table 11 we can see that even without the weight parameter ε, the three-dimensional results show promising features for scalability with respect to the number of iterations. These results are in line with the previous results obtained for the one- and two-dimensional constant wave number model. We similarly expect the importance of ε to decrease along with kh.

7. Conclusion. We have shown that the near-zero eigenvalues for DEF-based preconditioners are related to the near-kernel eigenmodes of the fine-grid operator A and coarse-grid operator A_{2h} being misaligned. This effect can be attributed to the interpolation scheme not being able to sufficiently approximate the transferring of the grid functions at very large wave numbers.

We have presented the first scheme to analytically measure the effect of these errors on the construction of the projection preconditioner by means of the projection error. The quality of the DEF vectors determines whether the projection error dominates. To minimize the projection error, we propose the implementation of a higher-order approximation scheme to construct the deflation vectors. Incorporating a weight parameter within the approximation scheme provides sufficient counterbalance to mitigate the reappearance of the near-zero eigenvalues. Two options are available for determining the weight parameter. The first is to use the analytical eigenvalues

of B at the smallest index $l_{\min,h}$ and solve for ε. This approach is straightforward to use as it primarily depends on the eigenvalues of B, which can be computed independently of the eigenvalues of A. The second approach is to use the projection error minimizing algorithm, which finds the ε which minimizes the error on average.

Even without adjusting the weight parameter, the spectrum of our proposed operator is still the most favorable compared to other preconditioning operators based on DEF. We have performed numerical testing and simulation of our model problems ranging from the simple one-dimensional constant wave number problem to the challenging industrial Marmousi problem. The numerical results are in line with the theoretical results as the number of iterations for both the one-, two-, and three-dimensional constant wave number model problems are more or less wave number independent. We furthermore provide numerical evidence supporting the notion that our method also works for non-self-adjoint and heterogeneous problems, even when the CSLP preconditioner is excluded. The latter allows for a substantial speedup.

Acknowledgment. We would like to thank three anonymous referees for their constructive feedback and valuable comments which allowed us to greatly improve an earlier version of this paper.

REFERENCES

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

[2] M. Bollhöfer, M. J. Grote, and O. Schenk, *Algebraic multilevel preconditioner for the Helmholtz equation in heterogeneous media*, SIAM J. Sci. Comput., 31 (2009), pp. 3781–3805.

[3] P.-H. Cocquet and M. J. Gander, *How large a shift is needed in the shifted Helmholtz preconditioner for its effective inversion by multigrid?*, SIAM J. Sci. Comput., 39 (2017), pp. A438–A478.

[4] M. Donatelli, *A Note on Grid Transfer Operators for Multigrid Methods*, preprint, arXiv:0807.2565, 2008.

[5] M. Eiermann, O. G. Ernst, and O. Schneider, *Analysis of acceleration strategies for restarted minimal residual methods*, J. Comput. Appl. Math., 123 (2000), pp. 261–292.

[6] M. Embree, *How descriptive are GMRES convergence bounds?*, manuscript.

[7] Y. Erlangga and E. Turkel, *Iterative schemes for high order compact discretizations to the exterior Helmholtz equation*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 647–660.

[8] Y. A. Erlangga and R. Nabben, *On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian*, Electron. Trans. Numer. Anal., 31 (2008), 3.

[9] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik, *A novel multigrid based preconditioner for heterogeneous Helmholtz problems*, SIAM J. Sci. Comput., 27 (2006), pp. 1471–1492.

[10] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee, *On a class of preconditioners for solving the Helmholtz equation*, Appl. Numer. Math., 50 (2004), pp. 409–425.

[11] O. G. Ernst and M. J. Gander, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, in Numerical Analysis of Multiscale Problems, Springer, Berlin, 2012, pp. 325–363.

[12] O. G. Ernst and M. J. Gander, *Multigrid methods for Helmholtz problems: A convergent scheme in* $1d$ *using standard components*, in Direct and Inverse Problems in Wave Propagation and Applications. De Gruyter, Berlin, 2013, pp. 135–186.

[13] C. Farhat, A. Macedo, and M. Lesoinne, *A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems*, Numer. Math., 85 (2000), pp. 283–308.

[14] M. J. Gander, I. G. Graham, and E. A. Spence, *Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed?*, Numer. Math., 131 (2015), pp. 567–614.

[15] M. J. Gander, F. Magoulés, and F. Nataf, *Optimized Schwarz methods without overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 24 (2002), pp. 38–60.

[16] M. J. Gander and H. Zhang, *Optimized Schwarz methods with overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 38 (2016), pp. A3195–A3219.

[17] M. J. Gander and H. Zhang, *Restrictions on the use of sweeping type preconditioners for Helmholtz problems*, in International Conference on Domain Decomposition Methods, Springer, Cham, Switzerland, 2017, pp. 321–332.
[18] M. J. Gander and H. Zhang, *A class of iterative solvers for the Helmholtz equation: Factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods*, SIAM Rev., 61 (2019), pp. 3–76.
[19] L. G. Ramos and R. Nabben, *On the spectrum of deflated matrices with applications to the deflated shifted Laplace preconditioner for the Helmholtz equation*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 262–286.
[20] I. Graham, E. Spence, and E. Vainikko, *Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption*, Math. Comp., 86 (2017), pp. 2089–2127.
[21] I. Graham, E. Spence, and J. Zou, *Domain Decomposition with Local Impedance Conditions for the Helmholtz Equation*, preprint, arXiv:1806.03731, 2018.
[22] I. G. Graham, E. A. Spence, and E. Vainikko, *Recent results on domain decomposition preconditioning for the high-frequency Helmholtz equation using absorption*, in Modern Solvers for Helmholtz Problems, Springer, 2017, Cham, Switzerland, pp. 3–26.
[23] A. Greenbaum and Z. Strakos, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, Springer, New York, 1994, pp. 95–118.
[24] M. Holtz and A. Kunoth, *B-spline-based monotone multigrid methods*, SIAM J. Numer. Anal., 45 (2007), pp. 1175–1199.
[25] I. C. Ipsen, *Expressions and bounds for the GMRES residual*, BIT, 40 (2000), pp. 524–535.
[26] I. C. Ipsen, *Departure from Normality and Eigenvalue Perturbation Bounds*, Technical report, North Carolina State University. Center for Research in Scientific Computation, 2003.
[27] J. Liesen and Z. Strakos, *Convergence of GMRES for tridiagonal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 233–251.
[28] G. Meurant and J. D. Tebbens, *The role eigenvalues play in forming GMRES residual norms with non-normal matrices*, Numer. Algorithms, 68 (2015), pp. 143–165.
[29] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[30] R. B. Morgan, *GMRES with deflated restarting*, SIAM J. Sci. Comput., 24 (2002), pp. 20–37.
[31] R. B. Morgan, Z. Yang, and B. Zhong, *Pseudoeigenvector bases and deflated GMRES for highly nonnormal matrices*, Numer. Linear Algebra Appl., 23 (2016), pp. 1032–1045.
[32] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.
[33] A. Pinkus, *On L1-Approximation*, Cambridge Tracts in Math. 93, Cambridge, Cambridge University Press, 1989.
[34] D. F. Rogers, *An Introduction to NURBS: With Historical Perspective*, Morgan Kaufmann, San Francisco, 2000.
[35] A. Sheikh, *Development of the Helmholtz Solver Based on a Shifted Laplace Preconditioner and a Multigrid Deflation Technique*, Technical report, Delft University of Technology, Delft, The Netherlands, 2014.
[36] A. Sheikh, D. Lahaye, L. G. Ramos, R. Nabben, and C. Vuik, *Accelerating the shifted Laplace preconditioner for the Helmholtz equation by multilevel deflation*, J. Comput. Phys., 322 (2016), pp. 473–490.
[37] A. Sheikh, D. Lahaye, and C. Vuik, *On the convergence of shifted Laplace preconditioner combined with multilevel deflation*, Numer. Linear Algebra Appl., 20 (2013), pp. 645–662.
[38] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga, *Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods*, J. Sci. Comput., 39 (2009), pp. 340–370.
[39] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.