Pollution and accuracy of solutions of the Helmholtz equation
A novel perspective from the eigenvalues

Dwarka, V.; Vuik, C.

DOI
10.1016/j.cam.2021.113549

Publication date
2021

Document Version
Final published version

Published in
Journal of Computational and Applied Mathematics

Journal of Computational and Applied Mathematics 395 (2021) 113549

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Pollution and accuracy of solutions of the Helmholtz equation: A novel perspective from the eigenvalues

V. Dwarka [*], C. Vuik

van Mourik Broekmanweg 6, 2628 XE, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 1 August 2020
Received in revised form 27 January 2021

Keywords:
Helmholtz equation
Pollution
Numerical dispersion
Eigenvalues
FDM
FEM

ABSTRACT

In researching the Helmholtz equation, the focus has either been on the accuracy of the numerical solution (pollution) or the acceleration of the convergence of a preconditioned Krylov-based solver (scalability). While it is widely recognized that the convergence properties can be investigated by studying the eigenvalues, information from the eigenvalues is not used in studying the numerical dispersion which drives the pollution error. Our aim is to bring the topics of accuracy and scalability together for the first time; instead of approaching the pollution error in the conventional sense of being the result of a discrepancy between the exact and numerical wavenumber, we show that the dispersion which drives the pollution error can also be decomposed in terms of the eigenvectors and eigenvalues. Using these novel insights, we construct sharper upper bounds for the total error independent of the grid resolution. While the pollution error can be minimized in one-dimension by introducing a dispersion correction, the latter is not possible in higher dimensions, even for very simple model problems. For our model problem, a correction on the eigenvalues enables us to remove the pollution error and study it in full detail, both in one- and two-dimensions.

1. Introduction

The Helmholtz equation is widely used in applications ranging from geophysics to bio medical physics. Many researches have contributed to the broad range of literature on this topic. In particular, the pollution effect deserved a lot of attention due to its far ranging consequences. In essence, the pollution effect is directly related to numerical dispersion errors due to differences between the actual and numerical wavenumber [1–4]. This error grows with the wavenumber as in the high-frequency range the solutions become very oscillatory.

As a result of this discrepancy, there may be large errors between the actual solution and the obtained numerical solution. Therefore, the solution obtained using fast and efficient solvers, may therefore be severely inaccurate. The fact that the pollution effect for finite element and finite difference methods cannot be avoided in higher-dimensions adds to the problem [2]. No simple solution exists, as it has been shown that for a certain accuracy, the number of grid points needed to retain that accuracy grows along with the wavenumber. However, it grows slower than the order of accuracy of the schemes. In particular, if we let k denote the wavenumber, n the problem size in one-dimension and p the order of a finite difference of finite element scheme, then

$$n = Ck^{\left(\frac{p+1}{p}\right)},$$

[*] Corresponding author.
E-mail addresses: v.n.s.r.dwarka@tudelft.nl (V. Dwarka), c.vuik@tudelft.nl (C. Vuik).
URL: https://homepage.tudelft.nl/d2b4e/ (C. Vuik).

https://doi.org/10.1016/j.cam.2021.113549

where C is a constant that only depends on the accuracy achieved [5]. Therefore, if we wish to increase k while keeping the accuracy of the same order, we need to increase n as well, which leads to larger linear systems.

The literature has proposed several ways to mitigate this persisting issue. One branch has focused on formulating new higher-order discretization schemes. Among the first were a rotated 9-point finite difference scheme [6]. This method was extended by including a 'perfectly matched layer' (PML) [7]. In both works, optimal parameters for the difference scheme were computed in order to improve the accuracy of the numerical solution. A similar strategy was used for the three-dimensional Helmholtz problem, where the 9-point stencil was extended to a 27-point stencil [8]. Furthermore, some line of work developed accurate higher order schemes for the one- and two-dimensional Helmholtz equation, under the assumption that separation of variables can be used [9–12].

In line of this strategy lies the use of compact finite difference schemes [5,13–15]. One advantage of the compact scheme is that no additional boundary conditions are required due to having a larger stencil. While both compact fourth- and sixth-order schemes were developed in the literature, it has been shown that at best sixth-order accuracy can be achieved using compact stencils for the Poisson, and thus inherently, the Helmholtz equation [16]. Apart from using compact higher-order finite difference schemes, others have incorporated wave-ray theory to obtain more accurate solutions [17] or have constructed a modified wavenumber which is closer to the exact wavenumber in order to reduce the numerical dispersion [18]. When using such strategies, all methods depend on a pre-specified propagation angle to provide an accurate solution, as the exact propagation angle is unknown. As a result, for specified angles an accurate solution can be obtained by either incorporating a modified wavenumber or by switching to a higher-order dispersion corrected discretization. A combination of both has been studied by Cocquet et al. [18], where the standard 5-point stencil is replaced by a parametrized 9-point difference scheme including a modified wavenumber. Very recently, using an asymptotic dispersion correction for two-dimensional constant wavenumber problems, these methods have shown to provide up to sixth order accuracy for plane waves given an angle of propagation [19]. In this paper, we aim to provide a theoretical contribution to this field of research by introducing a novel perspective on the pollution error as a result of the numerical dispersion; we study the effect not explicitly by reducing it to a difference between the exact and numerical wavenumber, but we analyse the differences between the analytical and numerical eigenvalues. As we will be using the analytical and numerical eigenvalues, we will use a simple model problem from the literature with Dirichlet boundary conditions. Using this configuration, some new aspects come to light which are of paramount importance. First of all, we will be able to obtain the true characteristics and propagation of the pollution error as k grows due to numerical dispersion instead of a linear dependence on k. As a result, we are able to construct novel yet sharp error estimates which reflect these characteristics. Second of all, we will be able to study the exact eigenmodes in higher-dimensions which reflect the numerical dispersion. Consequently, for the first time, we can pinpoint the pollution effect for particular eigenmodes in one- and two-dimensions and minimize the pollution effect without keeping n uneconomically large. Moreover, for the two-dimensional model problem, the dispersion correction can be obtained for all angles simultaneously.

The paper is organized as follows. We start by defining the model problem in Section 2. Here we also provide the analytical solution, which we need in order to study the pollution error. Section 3 derives the pollution error in the conventional sense by looking at the difference between the analytical and numerical wavenumber. In Section 4 we start by constructing our new upper bound using the eigenvalues as our source for information. We finally present numerical results in Section 5. The experiments are twofold; we show that the bound holds and we provide a way to apply a dispersion correction. We conclude by summarizing our findings in Section 6.

2. Problem definition

In this section we start by defining two model problems. Following a similar approach in the literature, we use the constant wavenumber model with Dirichlet conditions, such that the analytical solution and eigenvalues can be derived [9–12,14,19–22]. We therefore start by focusing on the following model problem

$$\begin{aligned} -\frac{d^2u}{dx^2} - k^2\, u &= \delta(x-x'), x \in \Omega = [0, L] \subset \mathbb{R}, \\ u(0) &= 0, u(L) = 0, \;\; k \in \mathbb{R} \setminus \{0\}. \end{aligned} \tag{1}$$

We will refer to this model problem as MP 1. Working on the unit-domain ($L = 1$), the second order difference scheme with step-size $h = \frac{1}{n}$ leads to

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} - k^2 u_j = f_j, j = 1, 2, 3, \ldots, n, \;\; x_j = jh.$$

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 system and eigenvalues

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

In order to investigate the pollution error in higher dimensions, we define MP 2 to be the two-dimensional version of the original model problem. Therefore, on the standard two-dimensional square unit domain $\Omega = [0, 1] \times [0, 1]$ with constant wavenumber k we consider

$$\begin{aligned} -\Delta u(x,y) - k^2 u(x,y) &= \delta(x-\frac{1}{2}, y-\frac{1}{2}),\ (x,y) \in \Omega \setminus \partial\Omega \subset \mathbb{R}^2, \\ u(x,y) &= 0,\ (x,y) \in \partial\Omega, \end{aligned} \tag{3}$$

2.1. Analytical solution

The general solution to our one-dimensional model problem is given by

$$u(x) = e^{ikx}.$$

However, apart from the general exponential form, we can also express the exact solution to MP 1 in terms of the Green's function $G(x, x')$ given that this contains the eigenvalues. We need to use the Green's function given that we are working with the non-homogeneous Helmholtz equation. We therefore seek a solution of the form

$$u(x) = \int_0^L G(x,x')f(x)dx', \tag{4}$$

where the Green's function satisfies

$$\left(\frac{d^2}{dx^2} - k^2\right) G(x,x') = \delta(x-x').$$

To obtain the Green's function, we need to by rewriting the differential operator from MP 1 in the Sturm–Liouville form [23]. Let $\mathcal{L}(x)$ be the general Sturm–Liouville operator

$$\mathcal{L}(x) = \frac{d}{dx}\left[p\left(x\right)\frac{d}{dx}\right] + q(x) \tag{5}$$

Setting $p(x) = -1$ and $q(x) = -k^2$, we obtain the Sturm–Liouville operator for the Helmholtz boundary value problem, which we will continue to denote by $\mathcal{L}(x)$. Using the Sturm–Liouville operator for the Helmholtz problem, we can rewrite the problem as

$$\mathcal{L}(x)u(x) = f(x).$$

The related eigenvalue problem is

$$\mathcal{L}(x)u(x) = \lambda u(x).$$

Using the eigenfunction expansion, we can rewrite MP 1 (1) as

$$\begin{aligned} \left(\frac{d^2}{dx^2} + \lambda_j\right) u_j(x) &= 0, \\ u_j(0) = u_j(L) &= 0. \end{aligned}$$

Normalizing with a factor $\sqrt{\frac{2}{L}}$, gives the following solution

$$u_j(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{j\pi x}{L}\right) \text{ with } \lambda_j = \left(\frac{j\pi}{L}\right)^2,\ j = 1, 2, 3, \ldots.$$

Integrating over the eigenfunctions for the eigenvalue problem gives

$$\frac{2}{L}\int_0^L \sin\left(\frac{j\pi x}{L}\right)\sin\left(\frac{i\pi x}{L}\right) dx = \delta_{ij}. \tag{6}$$

The Green's function for Eq. (6) is given by

$$G(x,x') = \frac{2}{L}\sum_{j=1}^{\infty} \frac{\sin\left(\frac{j\pi x}{L}\right)\sin\left(\frac{j\pi x'}{L}\right)}{\lambda_j}, k^2 \neq j^2\pi^2, j = 1, 2, 3, \ldots. \tag{7}$$

Consequently on the unit interval, $G(x, x')$ satisfies

$$\begin{aligned} &\mathcal{L}(x)G(x,x') = \delta(x-x'), x \in \Omega = [0, 1] \subset \mathbb{R}, \\ &G(0,x') = G(1,x') = 0,\ x \in \partial\Omega. \end{aligned} \tag{8}$$

In the event that $k^2 = j^2\pi^2$, the eigenfunction expansion would become defective as this would imply resonance and unbounded oscillations in the absence of dissipation. Therefore, we explicitly need to warrant for the latter case and impose the extra condition $k^2 \neq j^2\pi^2$ asserting that our Green's function exists.

Eq. (7) immediately provides us with an expression for the analytical eigenvalues. It is apparent that within the bounded domain [0, 1] there are an infinite number of eigenpairs. We employ this expression for the eigenvalues in upcoming sections, where we compare them with the numerical eigenvalues for the linear system of equations. We have expressed the exact solution to MP 1 as an eigenfunction expansion using Green's function. A similar approach will allow us to obtain the exact solution for the two-dimensional MP 2, which is given by

$$\begin{aligned} u(x,y) &= \int_{\Omega} f(x,y)G(x,y,x',y')dx'dy', && (9)\\ &= \int_{\Omega} \delta(x-x',y-x')G(x,y,x',y')dx' && (10)\\ &= G(x,y,x',y'). && (11)\end{aligned}$$

The Green's function $G(x,y,x',y')$ on the unit square becomes

$$\begin{aligned} G(x,y,x',y') = \frac{4}{L}\sum_{j=1}^{\infty}\sum_{j=1}^{\infty} \frac{\sin\left(\frac{j\pi x}{L}\right)\sin\left(\frac{j\pi x'}{L}\right)\sin\left(\frac{j\pi y}{L}\right)\sin\left(\frac{j\pi y'}{L}\right)}{\frac{i^2\pi^2+j^2\pi^2}{L^2}-k^2}&, \\ k^2 \neq i^2\pi^2 + j^2\pi^2, i,j = 1,2,3,\ldots.& \end{aligned} \qquad (12)$$

and satisfies

$$\begin{aligned} &\mathcal{L}(x,y)G(x,y,x',y') = \delta(x-x',y-y')\\ &G(x,0,x',y') = G(x,1,x',y') = 0, \;\; y \in \partial\Omega\\ &G(0,y,x',y') = G(1,y,x',y') = 0, \;\; x \in \partial\Omega\\ &(x,y) \in \Omega = [0,1]\times[0,1] \subset \mathbb{R}^2, \end{aligned} \qquad (13)$$

where $\mathcal{L}(x,y)$ is the two-dimensional Sturm–Liouville operator corresponding to the Helmholtz equation from MP 2.

3. Error bounds

We now briefly explain the classical error bound for the pollution error. It was mentioned, that in order to keep the pollution error at bay, the grid should be refined such that $k^3h^2 < 1$ [1,24]. Such a severe restriction on the step-size is necessary, as the accuracy of the numerical solution deteriorates rapidly when the wavenumber increases. In fact, the numerical wave has dispersive properties, which are not present in the analytical wave. Consequently, a phase shift occurs which forms the primary source of error in the pollution term. Thus, in the case FEM and FDM solutions, a phase lag between the computed and the exact wave is directly related to the dispersive character of the discrete medium (i.e. the computed wave does not propagate at the speed of sound), which causes a difference between the exact and numerical wavenumber. This effect accumulates into the pollution term as k increases.

3.1. Numerical dispersion

To understand how the pollution error depends on the numerical dispersion and consequently on the wavenumber k, note that the dimensionless wavenumber is represented by

$$k = \frac{2\pi f}{\lambda},$$

where $2\pi f$ denotes the angular frequency and λ denotes the phase velocity. Discretizing the one-dimensional Helmholtz equation leads to

$$\frac{u_{j+1}-2u_j+u_{j-1}}{h^2} - k^2u_j = 0. \qquad (14)$$

Moreover, a general continuous solution is given by

$$u\,(x) = e^{ikx}. \qquad (15)$$

Evaluation of expression (15) in the discrete points gives

$$u_j = e^{i\tilde{k}x_j}. \qquad (16)$$

Here i denotes the imaginary unit and $\tilde{k}$ represents the perturbed wavenumber due to having a velocity which is different than the speed of sound. Substituting Eq. (16) into (14) results in

$$u_{j+1} - 2u_j + u_{j-1} = e^{i\tilde{k}x_j}\left(e^{i\tilde{k}h} - 2 + e^{-i\tilde{k}h}\right) = 2\left(\cos\left(\tilde{k}h\right) - 1\right)e^{i\tilde{k}x_j}. \tag{17}$$

Eq. (17) is a good approximation of the exact solution if $\tilde{k}$ solves

$$\frac{2\left(\cos\left(\tilde{k}h\right) - 1\right)}{h^2} - k^2 = 0. \tag{18}$$

Applying Taylor's expansion on the cosine term and substituting into Eq. (18) gives

$$k - \tilde{k} = \mathcal{O}\left(k^2h^2\right)$$

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

$$\text{error}_{\text{pollution}} = \left|e^{ikx_j} - e^{i\tilde{k}x_j}\right| = \left|1 - e^{i\left(\tilde{k}-k\right)x_j}\right| \leq Ck\left|\tilde{k} - k\right| \leq Ck^3h^2. \tag{19}$$

The factor Ck^3h^2 can be decomposed as follows. $\mathcal{O}(k^2h^2)$ provides the error in the numerical wave speed for a wave travelling one period. The extra factor k is called the *pollution error* and corrects the total pollution error by scaling the error over one wave length by the total number of wave lengths travelled over the entire numerical domain [4,24].

In [1] it was noted that the error given in Eq. (19) mainly relates to the dispersion caused by the differing wavenumbers. The total error for the discretized one-dimensional Helmholtz operator is given by

$$\text{error}_{\text{total}} = \frac{\|u - \hat{u}\|}{\|u\|} \leq C_1kh + C_2k^3h^2,\ kh < 1. \tag{20}$$

While applying the rule of thumb $kh \leq 0.625$ is sufficient for keeping the first term under control, it does not harbour properly against the propagation of the pollution error which grows with k, even if kh is kept small enough. Thus, it has been advocated to set the grid resolution to $k^3h^2 \leq \epsilon$ instead of $kh \leq 0.625$. Deraemaeker et al. [1] and Ainsworth [4] have proved that while it is possible to eliminate the pollution effect in one-dimensional Helmholtz problems by implementing a modified wavenumber, a similar conclusion cannot be extended to higher dimensional problems, see Section 3.2 for more details. As a result, much research has been conducted towards minimizing the pollution error. Note that the bound in Eq. (20) also holds in higher dimensions, as long as the second order finite difference method is used. For any general pth order scheme, we obtain the following error bound:

$$\text{error}_{\text{total}} = \frac{\|u - \hat{u}\|}{\|u\|} \leq C_1kh + C_2k(k^ph^p),\ kh < 1. \tag{21}$$

3.2. Dispersion correction

As mentioned earlier, it is possible to eliminate the pollution error for the one-dimensional MP 1. Recall from the previous section that the discretization of MP 1 using second order finite-differences was given by

$$\frac{-u_{j-1h} + 2u_{jh} - u_{j+1h}}{h^2} - k^2u_{jh} = 0,\ \ 1 \leq j \leq n-1, \tag{22}$$

with general solution

$$u\left(x\right) = e^{ikx}. \tag{23}$$

Evaluation of expression (23) in the discrete points led to

$$u_j = e^{i\tilde{k}x_j},\ 1 \leq j \leq n-1, \tag{24}$$

which can be considered as plane-wave solutions of the discrete homogeneous Helmholtz equation, where $\tilde{k}$ represents the numerical wavenumber. Substituting (24) into (22) and using Euler's trigonometric identity to decompose the exponential function, leads to

$$\begin{aligned}
&-2\cos(\tilde{k}h) + 2 - k^2h^2 = 0,\\
&2\cos(\tilde{k}h) = 2 - k^2h^2,\\
&\tilde{k}h = \arccos(1 - \frac{k^2h^2}{2}),\\
&\tilde{k} = \frac{1}{h}\arccos(1 - \frac{k^2h^2}{2}) = k - \frac{k^3h^2}{24} + \mathcal{O}(k^5h^4).
\end{aligned}$$

If we want to eliminate the discretization error introduced into the scheme, we need to set $\tilde{k} = k$, i.e.

$$\tilde{k} = \frac{1}{h}\arccos\left(1 - \frac{k^2h^2}{2}\right) = k \Rightarrow \tilde{k} = \sqrt{\frac{2(1-\cos(kh)}{h^2}}. \tag{25}$$

Unfortunately, this approach only works for one-dimensional problems. To see this, we look at the two-dimensional second order finite difference scheme

$$\frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} - k^2u_{i,j} = 0, \;\; 1 \leq i,j \leq n-1. \tag{26}$$

Again, using plane-wave solutions, we can write $u(x,y) = e^{i(k_1x+k_2y_j)}$, with $(k_1,k_2) = (k\cos\theta, k\sin\theta)$. Evaluating the solution in the discrete grid points (x_i,y_j) gives $u(x_i,y_j) = e^{i(\tilde{k}_1x+\tilde{k}_2y_j)}$, where $(\tilde{k}_1,\tilde{k}_2) = (\tilde{k}\cos\theta, \tilde{k}\sin\theta)$ denotes the numerical wavenumber. Substituting these expressions into the difference scheme (26), the problem becomes

$$-2\cos(\tilde{k}\cos(\theta)h) - 2\cos(\tilde{k}\sin(\theta)h) + 4 - k^2h^2 = 0. \tag{27}$$

Generally the direction of the plane waves θ is unavailable. This is due to the fact that plane waves propagate in an infinite number of directions. Even if there are directionally prevalent components in this decomposition they are not necessarily known a priori [3,25]. Therefore, in order to solve for $\tilde{k}$ to obtain a two-dimensional dispersion correction, Eq. (27) needs to be minimized over all angles θ, which remains problematic.

4. Pollution and spectral properties

The vast majority of works regarding the pollution error focuses on developing numerical discretization schemes to mitigate the pollution effect. Note that in order to study the pollution error, the analytical solution must be known, which limits the scope of potential test problems. Moreover, the a priori upper bound from expression (20) shows that the pollution error can be bounded from above by a term which grows linearly with k. This bound is known to be sharp, but provides little detail as regards the underlying characteristics with respect to its dependence on the numerical dispersion. As we have seen in Section 3.2, this becomes even more problematic in higher-dimensions.

Thus, in order to investigate the explicit translation of the numerical dispersion effect into the pollution error, we will use the information from the eigenvalues. To our current knowledge, this provides a novel theoretical perspective on the pollution error. How the pollution effect influences spectral properties and vice versa has remained an unconventional approach in researching the pollution error. Thus, in order to research these properties, we will start by looking at the differences between the exact and numerical solution of MP 1. The explicit use of the eigenvalues requires that we use a model problem with Dirichlet boundary conditions. The latter model problem has also been researched using the conventional method [11,14,20,21].

4.1. General properties

Recall from Section 2.1 that the one-dimensional MP 1 is given by

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

We also showed that the analytical solution $u(x,x')$ can be expressed in terms of the Green's function by

$$u(x,x') = 2\sum_{j=1}^{\infty}\frac{\sin(j\pi x')}{j^2\pi^2 - k^2}\sin(j\pi x), \;\; k \neq j\pi \text{ for } j = 1,2,3,\ldots. \tag{28}$$

If we define $u_j = u(x_j), j = 1,2,\ldots,n$, where u is evaluated at the discrete grid points, we can represent the n-th term finite solution as a vector $u(\bar{x})$ by

$$u(\bar{x}) = 2\sum_{j=1}^{n}\frac{\sin(j\pi x')}{\lambda_j}v_j(\bar{x}), \;\; k \neq \pi \text{ for } j = 1,2,3,\ldots,n, \tag{29}$$

where $\bar{x} = [x_1,x_2,\ldots,x_n]^T$ and $v_j(\bar{x}) = \frac{\sin(j\pi\bar{x})}{\|\sin(j\pi\bar{x})\|}$ is now the jth orthonormal eigenvector corresponding to the jth eigenvalue. The eigenvectors are exact discretizations of the continuous eigenfunctions. Note that the denominator of each term in the sum consists of the analytical eigenvalues. The right-hand side function $f(\bar{x})$ of MP 1 is known and can also be represented using the same basis of orthonormal eigenvectors

$$f(\bar{x}) = 2\sum_{j=1}^{n}\sin(j\pi x')v_j(\bar{x}). \tag{30}$$

Similarly, we can write the numerical solution vector $\hat{u}$ as follows

$$\begin{aligned}\hat{u} = A^{-1}f(\bar{x}) &= A^{-1}2\sum_{j=1}^{n}\sin(j\pi x')v_j(\bar{x})\\ &= 2\sum_{j=1}^{n}\frac{\sin(j\pi x')}{\hat{\lambda}_j}v_j(\bar{x}),\end{aligned} \tag{31}$$

where $\hat{\lambda}_j$ are the numerical eigenvalues. We will proceed by using the notation u, $\hat{u}$ and f respectively.

4.2. One-dimensional spectral properties

We now have a simple expression which can be decomposed into terms containing the eigenvalues. This allows us to identify the polluting terms of the numerical solution. We start by investigating some general properties of the differences between the analytical and numerical eigenvalues.

Lemma 1 (*Difference Eigenvalues*). *Let λ_j be the analytical eigenvalue and $\hat{\lambda}_j$ be the numerical eigenvalue for $j = 1, \ldots, n$, where $n > \pi$. If the expressions for the eigenvalues are given by*

$$\lambda_j = j^2\pi^2 - k^2, \hat{\lambda}_j = \frac{2}{h^2}\left(1-\cos(j\pi h)\right) - k^2,$$

then the difference between the eigenvalues is bounded from above by

$$\lambda_j - \hat{\lambda}_j < \frac{j^4\pi^4h^2}{12}, \tag{32}$$

and from below by

$$\lambda_j - \hat{\lambda}_j \geq \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}. \tag{33}$$

Proof. We start by showing expression (32). The difference between the eigenvalues is given by

$$\lambda_j - \hat{\lambda}_j = j^2\pi^2 - k^2 - \left(\frac{2}{h^2}\left(1-\cos(j\pi h)\right) - k^2\right).$$

Substituting the power series for the cosine term and letting ζ represent our cut-off point, we obtain

$$\begin{aligned}\lambda_j - \hat{\lambda}_j &= j^2\pi^2 - k^2 - \left(\frac{2}{h^2}\left(1 - \left(\sum_{l=0}^{\infty}\frac{(-1)^l(j\pi h)^{2l}}{(2l)!}\right)\right) - k^2\right),\\ &< j^2\pi^2 - k^2 - \left(\frac{2}{h^2}\left(1 - 1(1 - \frac{j^2\pi^2h^2}{2} + \frac{j^4\pi^4h^4}{24} - \zeta^6)\right) - k^2\right),\\ &< j^2\pi^2 - k^2 - \left(\frac{2}{h^2}\left(1 - 1 + j^2\pi^2\frac{h^2}{2} - j^4\pi^4\frac{h^4}{24}\right) - k^2\right),\\ &= j^2\pi^2 - k^2 - \left(j^2\pi^2 - k^2 - \frac{j^4\pi^4h^2}{12}\right),\\ &= \frac{j^4\pi^4h^2}{12}.\end{aligned}$$

This gives us an upper bound with respect to the difference between the analytical and numerical eigenvalue. Now to construct the lower bound in expression (33), we need to show that

$$\lambda_j - \hat{\lambda}_j \geq \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}. \tag{34}$$

We again substitute the power series for the cosine term in the difference equation of the eigenvalues, which gives

$$\begin{aligned}\lambda_j - \hat{\lambda}_j &= j^2\pi^2 - k^2 - \left(\frac{2}{h^2}\left(1 - \left(\sum_{l=1}^{\infty}\frac{(-1)^l(j\pi h)^{2l}}{(2l)!}\right)\right) - k^2\right),\\ &= j^2\pi^2 - k^2 - \left(j^2\pi^2 - k^2 - \frac{j^4\pi^4h^2}{12} + \frac{2j^6\pi^6h^4}{6!} - \frac{2j^8\pi^8h^6}{8!} + \frac{2j^{10}\pi^{10}h^8}{10!} \ldots - \ldots\right),\\ &= \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!} + \frac{2j^8\pi^8h^6}{8!} - \frac{2j^{10}\pi^{10}h^8}{10!} + \ldots - \ldots.\end{aligned}$$

Substituting the difference expression into (34) and grouping terms on the left hand side leads to a true statement if each of the term in parenthesis is non-negative.

$$\left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right) + \left(\frac{2j^8\pi^8h^6}{8!} - \frac{2j^{10}\pi^{10}h^8}{10!}\right) + \dots \geq \left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right). \tag{35}$$

Thus, in order to show that this holds for all j we need to show that each term in parenthesis is non-negative. We can write expression (35) as

$$\left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right) + \sum_{l=2}^{\infty}\left(\frac{2j^{4l}\pi^{4l}h^{4l-2}}{(4l)!} - \frac{2j^{4l+2}\pi^{4l+2}h^{4l}}{(4l+2)!}\right) \geq \left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right). \tag{36}$$

The sum on the left hand side of expression (36) will be greater than the right hand side if we can prove that each grouped term is non-negative. Thus, we need to show that for each $j = 1, 2, \dots n$

$$\begin{aligned}&\left(\frac{2j^{4l}\pi^{4l}h^{4l-2}}{(4l)!} - \frac{2j^{4l+2}\pi^{4l+2}h^{4l}}{(4l+2)!}\right) \geq 0 \Leftrightarrow,\\ &\frac{2j^{4l}\pi^{4l}h^{4l-2}}{(4l)!}\left(1 - \frac{j^2\pi^2h^2}{(4l+2)(4l+1)}\right) \geq 0.\end{aligned} \tag{37}$$

For a positive integer j and $0 < h < 1$, this boils down to showing that for each $j = 1, 2, \dots n$ and $l \geq 2$

$$1 \geq \frac{j^2\pi^2h^2}{(4l+2)(4l+1)} \Leftrightarrow (4l+2)(4l+1) \geq j^2\pi^2h^2. \tag{38}$$

Given that the right hand side of inequality (38) is strictly increasing with respect to j, we can evaluate the minimum at $j = 1$ and maximum at $j = n$ to evaluate the lower bound.

$$(4l+2)(4l+1) \geq \begin{cases}\pi^2h^2, & \text{if } j = 1\\ \pi^2, & \text{if } j = n,\end{cases} \tag{39}$$

where we used that $h = n^{-1} < 1$, where $n > \pi$. In both cases and already for the smallest value of l ($l = 2$), the statement holds. Consequently, we must have

$$\lambda_j - \hat{\lambda}_j \geq \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}. \quad \square$$

Corollary 2 (*Bound for Analytical Eigenvalue*)**.** *Let λ_j be the analytical eigenvalue and $\hat{\lambda}_j$ be the numerical eigenvalue for $j = 1, \dots, n$, where $n > \pi$. Then for each j, the analytical eigenvalue λ_j is bounded in terms of the numerical eigenvalue $\hat{\lambda}_j$ by*

$$\hat{\lambda}_j + \left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right) \leq \lambda_j < \hat{\lambda}_j + \frac{j^4\pi^4h^2}{12}.$$

Proof. This result follows directly from Lemma 1, where we have

$$\begin{aligned}\lambda_j - \hat{\lambda}_j < \frac{j^4\pi^4h^2}{12} &\Rightarrow \lambda_j < \hat{\lambda}_j + \frac{j^4\pi^4h^2}{12},\\ \lambda_j - \hat{\lambda}_j \geq \left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right) &\Rightarrow \lambda_j < \hat{\lambda}_j + \left(\frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^4}{6!}\right). \quad \square\end{aligned}$$

Note that the upper and lower bound are dependent on the truncation error of the numerical discretization method. We will use Lemma 1 and Corollary 2 to obtain a more detailed understanding of the pollution error and how the numerical dispersion contributes to it. Moreover, we aim to find the eigenmodes which are responsible for this dispersive pattern. By writing the numerical eigenvalue as a function of the discretization error to approximate the analytical eigenvalue, we can propose a dispersion correction depending on the discretization scheme (see Section 4.1).

Corollary 3 (*Sum Eigenvalues*)**.** *Let λ_j be the analytical eigenvalue and $\hat{\lambda}_j$ be the numerical eigenvalue for $j = 1, \dots, n$. Then the sum of the reciprocal of the analytical eigenvalues can be bounded in terms of the numerical eigenvalues by*

$$\sum_{j=1}^{n}\left|\frac{1}{\lambda_j}\right| < \sum_{j=1}^{n}\frac{1}{\tilde{\lambda}_j},$$

where we let $\tilde{\lambda}_j = \min\left\{\left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^2}{6!}\right|, \left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12}\right|\right\}$.

Proof. We use Corollary 2. By taking the minimum, we ensure that the analytical eigenvalue is bounded in terms of magnitude. This is necessary as both the continuous and discrete operator are indefinite, which leads to positive and negative eigenvalues. Taking the reciprocal and summing over all eigenvalues gives the statement. □

Lemma 1 and Corollary 3 provides us with a way to express the analytical eigenvalues in terms of the numerical eigenvalues by adding a correction term. This correction term depends on the truncation error of the discretization method. We can now construct an upper bound for the error term between the exact and numerical solution in the theorem below.

Theorem 4 (*Pollution*)**.** *Let u be the (exact) solution to MP 1 and let $\hat{u}$ be the numerical solution obtained by solving $A\hat{u} = f$, where A is a non-singular matrix obtained by using a pth order finite difference scheme. If kh is kept constant, then the absolute error in the L^2-norm is bounded from above by*

$$\|u - \hat{u}\| < 2\sqrt{\sum_{j=1}^{n}\left(\frac{\frac{j^4\pi^4h^2}{12}}{\hat{\lambda}_j\tilde{\lambda}_j}\right)^2},$$

where $\tilde{\lambda}_j = \min\left\{\left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^2}{6!}\right|, \left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12}\right|\right\}.$

Proof. Using the expansion for the right-hand side function $f(x)$, We can write the numerical solution vector $\hat{u}$ as

$$\begin{aligned}\hat{u} = A^{-1}f(\bar{x}) &= A^{-1}(2\sum_{j=1}^{n}\sin(j\pi x'))v_j(\bar{x})\\ &= 2\sum_{j=1}^{n}\frac{\sin(j\pi x')}{\hat{\lambda}_j}v_j(\bar{x}).\end{aligned} \tag{40}$$

Note that this is based on the eigenfunctions evaluated at the discrete grid points and scaled to yield an orthonormal basis (see Section 4.1). Consequently, we have

$$\begin{aligned}\|u - \hat{u}\| &= \left\|2\sum_{j=1}^{n}\frac{\sin(j\pi x')}{\lambda_j}v_j(\bar{x}) - 2\sum_{j=1}^{n}\frac{\sin(j\pi x')}{\hat{\lambda}_j}v_j(\bar{x})\right\|,\\ &= \left\|2\sum_{j=1}^{n}\left(\frac{\sin(j\pi x')}{\lambda_j} - \frac{\sin(j\pi x')}{\hat{\lambda}_j}\right)v_j(\bar{x})\right\|,\\ &= \left\|2\sum_{j=1}^{n}\sin(j\pi x')\left(\frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j}\right)\right\|,\end{aligned}$$

where we used that the eigenvectors are orthonormal. We can write the error in the 2-norm as

$$\begin{aligned}\|u - \hat{u}\| &= \sqrt{4\sin(\pi x')^2(\frac{1}{\lambda_1} - \frac{1}{\hat{\lambda}_1})^2 + 4\sin(2\pi x')^2(\frac{1}{\lambda_2} - \frac{1}{\hat{\lambda}_2})^2 + \ldots + 4\sin(n\pi x')^2(\frac{1}{\lambda_n} - \frac{1}{\hat{\lambda}_n})^2},\\ &= \sqrt{4\sum_{j=1}^{n}\sin(j\pi x')^2\left(\frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j}\right)^2},\\ &< \sqrt{4\sum_{j=1}^{n}\left(\frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j}\right)^2},\\ &= \sqrt{4\sum_{j=1}^{n}\left(\frac{\hat{\lambda}_j - \lambda_j}{\hat{\lambda}_j\lambda_j}\right)^2},\end{aligned} \tag{41}$$

where we used that the eigenvectors are orthonormal and each sine term containing the location of the source is less than one. We would like to find an upper bound for expression (41). We can use Lemma 1 and Corollary 3, to provide element-wise upper bounds. From Lemma 1 it follows that

$$\sum_{j=1}^{n}(\lambda_j - \hat{\lambda}_j)^2 < \sum_{j=1}^{n}\left(\frac{j^4\pi^4h^2}{12}\right)^2. \tag{42}$$

For the denominator, Corollary 3 provides us with

$$\sum_{j=1}^{n}\left(\frac{1}{\hat{\lambda}_j\lambda_j}\right)^2 \leq \sum_{j=1}^{n}\left(\frac{1}{\hat{\lambda}_j}\right)^2\left(\frac{1}{\tilde{\lambda}_j}\right)^2, \tag{43}$$

where we have $\tilde{\lambda}_j = \min\left\{\left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12} - \frac{2j^6\pi^6h^2}{6!}\right|, \left|\hat{\lambda}_j + \frac{j^4\pi^4h^2}{12}\right|\right\}$. Substituting (42) and (43) into inequality (41) gives

$$\sqrt{4\sum_{j=1}^{n}\left(\frac{\hat{\lambda}_j-\lambda_j}{\hat{\lambda}_j\lambda_j}\right)^2} < 2\sqrt{\sum_{j=1}^{n}\left(\frac{\frac{j^4\pi^4h^2}{12}}{\hat{\lambda}_j\tilde{\lambda}_j}\right)^2}. \quad \square$$

4.3. Two-dimensional spectral properties

In this section we will extend the results from Section 4.2 to the two-dimensional case for MP 2. We start by defining the error estimation for the two-dimensional case.

Lemma 5 (*Difference Eigenvalues*)**.** *Let $\lambda_{i,j}$ be the analytical eigenvalue and $\hat{\lambda}_{i,j}$ be the numerical eigenvalue for $i, j = 1, \ldots, n$, where $n > \pi$. If the expressions for the eigenvalues are given by*

$$\lambda_j = (i^2+j^2)\pi^2 - k^2,$$
$$\hat{\lambda}_j = \frac{1}{h^2}\left(4 - 2\cos(i\pi h) - 2\cos(j\pi h)\right) - k^2,$$

then the difference between the eigenvalues is bounded from above by

$$\lambda_{i,j} - \hat{\lambda}_{i,j} < \frac{(i^4+j^4)\pi^4h^2}{12}, \tag{44}$$

and from below by

$$\lambda_{i,j} - \hat{\lambda}_{i,j} \geq \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!}. \tag{45}$$

Proof. Similar to the one-dimensional case, substituting the power series for both the ith and jth cosine term and letting ζ represent our cut-off point, we obtain

$$\begin{aligned}
\lambda_{i,j} - \hat{\lambda}_{i,j} &= i^2\pi^2 + j^2\pi^2 - k^2 - \left(\frac{1}{h^2}\left(4 - 2\left(\sum_{l=0}^{\infty}\frac{(-1)^l(i\pi h)^{2l}}{(2l)!}\right) - 2\left(\sum_{l=0}^{\infty}\frac{(-1)^l(j\pi h)^{2l}}{(2l)!}\right)\right) - k^2\right),\\
&< i^2\pi^2 + j^2\pi^2 - k^2 - \left(\frac{1}{h^2}\left(4 - 2 + i^2\pi^2\frac{h^2}{2} - 2 + j^2\pi^2\frac{h^2}{2} - i^4\pi^4\frac{h^4}{24} - j^4\pi^4\frac{h^4}{24} + \zeta^6\right) - k^2\right),\\
&= i^2\pi^2 + j^2\pi^2 - k^2 - \left(i^2\pi^2 + j^2\pi^2 - k^2 - \frac{i^4\pi^4h^2}{12} - \frac{j^4\pi^4h^2}{12}\right),\\
&= \frac{(i^4+j^4)\pi^4h^2}{12}.
\end{aligned}$$

To construct the lower bound, we again substitute the power series for the cosine terms in the difference equation, which gives

$$\begin{aligned}
\lambda_{i,j} - \hat{\lambda}_{i,j} &= i^2\pi^2 + j^2\pi^2 - k^2 - \left(\frac{1}{h^2}\left(4 - 2\left(\sum_{l=0}^{\infty}\frac{(-1)^l(i\pi h)^{2l}}{(2l)!}\right) - 2\left(\sum_{l=0}^{\infty}\frac{(-1)^l(j\pi h)^{2l}}{(2l)!}\right)\right) - k^2\right),\\
&= i^2\pi^2 + j^2\pi^2 - k^2 - \left(i^2\pi^2 + j^2\pi^2 - k^2 - \frac{i^4\pi^4h^2}{12} - \frac{j^4\pi^4h^2}{12} + \frac{2i^6\pi^6h^4}{6!} + \frac{2j^6\pi^6h^4}{6!} - \ldots + \ldots\right),\\
&= \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!} + \frac{2(i^8+j^8)\pi^8h^6}{8!} - \frac{2(i^{10}+j^{10})\pi^{10}h^8}{10!} + \ldots - \ldots.
\end{aligned}$$

Substituting the difference expression into (45) and grouping terms on the left hand side only leads to

$$\begin{aligned}
&\left(\frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!}\right) + \left(\frac{2(i^8+j^8)\pi^8h^6}{8!} - \frac{2(i^{10}+j^{10})\pi^{10}h^8}{10!}\right) + \ldots\\
&\geq \left(\frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!}\right).
\end{aligned}$$

We can write this as

$$\left(\frac{(i^4+j^4)\pi^4h^2}{12}-\frac{2(i^6+j^6)\pi^6h^4}{6!}\right)+\sum_{l=2}^{\infty}\left(\frac{2(i^{4l}+j^{4l})\pi^{4l}h^{4l-2}}{(4l)!}-\frac{2(i^{4l+2}+j^{4l+2})\pi^{4l+2}h^{4l}}{(4l+2)!}\right)$$
$$\geq\left(\frac{(i^4+j^4)\pi^4h^2}{12}-\frac{2(i^6+j^6)\pi^6h^4}{6!}\right). \tag{46}$$

The sum on the left hand side of expression (46) will be greater than the right hand side if we can prove that each grouped term is non-negative. Thus, we need to show that for each $i, j = 1, 2, \dots n$

$$\left(\frac{2(i^{4l}+j^{4l})\pi^{4l}h^{4l-2}}{(4l)!}-\frac{2(i^{4l+2}+j^{4l+2})\pi^{4l+2}h^{4l}}{(4l+2)!}\right)\geq 0 \Leftrightarrow,$$
$$\frac{2(i^{4l}+j^{4l})\pi^{4l}h^{4l-2}}{(4l)!}\left(1-\frac{(j^2+i^2)\pi^2h^2}{(4l+1)(4l+2)}\right)\geq 0. \tag{47}$$

For positive integers i, j and $0 < h < 1$, this boils down to showing that for each $i, j = 1, 2, \dots n$ and $l \geq 2$

$$1 > \frac{(j^2+i^2)\pi^2h^2}{(4l+1)(4l+2)} \Leftrightarrow (4l+2)(4l+1) \geq i^2\pi^2h^2 + j^2\pi^2h^2. \tag{48}$$

Given that the right hand side of inequality (48) is strictly increasing with respect to i and j, we can evaluate the minimum at $i, j = 1$ and maximum at $i, j = n$ to evaluate the lower bound.

$$(4l+2)(4l+1) \geq \begin{cases} 2\pi^2h^2, & \text{if } i, j = 1 \\ 2\pi^2, & \text{if } i, j = n, \end{cases} \tag{49}$$

where we used that $h = n^{-1} < 1$ such that $nh = 1$ and $n > \pi$. In both cases and already for the smallest value of l ($l = 2$), the statement holds. Consequently, we must have

$$\lambda_{i,j} - \hat{\lambda}_{i,j} \geq \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!}. \quad \square$$

Similar to the one-dimensional case, we can now bound the analytical eigenvalues in terms of the numerical eigenvalues by using the lower bound.

Corollary 6 (*Sum Eigenvalues*). *Let $\lambda_{i,j}$ be the analytical eigenvalue and $\hat{\lambda}_{i,j}$ be the numerical eigenvalue for $i, j = 1, \dots, n$. Then the sum of the reciprocal of the analytical eigenvalues can be bounded in terms of the numerical eigenvalues by*

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{1}{\lambda_{i,j}}\right| < \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\tilde{\lambda}_{i,j}},$$

where we let $\tilde{\lambda}_{i,j} = \min\left\{\left|\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^2}{6!}\right|, \left|\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12}\right|\right\}$

Proof. The proof is exactly the same as in the one-dimensional case. Using the lower bound and taking the reciprocal of each respective term will give the statement after summing over all i and j. □

We can use Lemma 5 and Corollary 6 to find a similar upper bound for the two-dimensional pollution error. We proceed by extending Theorem 4 to the two-dimensional case.

Corollary 7 (*Pollution 2D*). *Let u be the (exact) solution to MP 2 and let $\hat{u}$ be the numerical solution obtained by solving $A\hat{u} = f$, where A is a non-singular matrix obtained by using a pth order finite difference scheme. If kh is kept constant, then the absolute error in the L^2-norm is bounded from above by*

$$\|u - \hat{u}\| < 4\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\frac{(i^4+j^4)\pi^4h^2}{12}}{\hat{\lambda}_{i,j}\tilde{\lambda}_{i,j}}\right)^2},$$

where $\tilde{\lambda}_{i,j} = \min\left\{\left|\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^2}{6!}\right|, \left|\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12}\right|\right\}.$

Proof. See proof of Theorem 4 for the one-dimensional case and extend it to the case where the index i also goes from 1 to n. □

We now have an upper bound for the total error in terms of the numerical eigenvalues. If we compare this to the conventional pollution term,

$$\text{error}_{\text{pollution}} = \left\| u - \hat{u} \right\| \leq Ck(k^2h^2),$$

we observe that the explicit linear dependence on k has been replaced by the explicit dependence on a superposition of the numerical eigenvalues. One advantage of writing the upper bound in this way is that we can immediately observe that the pollution error can be minimized in both one- and two-dimensions for this model problem. Even for this simple model problem, the latter was deemed impossible due to the wave travelling in infinite directions for the two-dimensional model problem, see Section 4.4.1. It is easy to see that if we can minimize the largest term of the sum, then all other terms, which are by definition smaller, will allow the total sum to be minimized as well.

Corollary 8 (*Minimized Pollution 2D*)**.** *Let u be the (exact) solution to MP 2 given by expression (29) and suppose the L^2-norm of the exact solution is always smaller than 1, i.e. $\|u\| < 1$. Let $(i_{\min}, j_{\min})$ and $(\hat{i}_{\min}, \hat{j}_{\min})$ denote the location of the smallest analytical and numerical eigenvalue respectively and suppose $\left|\lambda_{i_{\min},j_{\min}}\right| \leq \left|\hat{\lambda}_{\hat{i}_{\min},\hat{j}_{\min}}\right|$. Then, if*

$$\left(\frac{4\frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12}}{\hat{\lambda}_{i_{\min},j_{\min}}(\hat{\lambda}_{i_{\min},j_{\min}} + \frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12} - \frac{2(i^6_{\min}+j^6_{\min})\pi^6h^4}{6!})} \right)^2 = \mathcal{O}(h^2),$$

then the relative error is bounded by

$$\frac{\left\| u - \hat{u} \right\|}{\|u\|} \leq 1.$$

Proof. Note that reciprocal of the smallest analytical value in terms of magnitude is the largest term in the set of the reciprocals of both the analytical and numerical eigenvalues. Now, unless $(i_{\min}, j_{\min}) = (\hat{i}_{\min}, \hat{j}_{\min})$, and $\lambda_{i_{\min},j_{\min}} \approx \hat{\lambda}_{i_{\min},j_{\min}}$, the difference between the reciprocals will be largest there and thus it will provide the largest contribution to the sum. As a result, we must have

$$\left(\frac{4\frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12}}{\hat{\lambda}_{i_{\min},j_{\min}}(\hat{\lambda}_{i_{\min},j_{\min}} + \frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12} - \frac{2(i^6_{\min}+j^6_{\min})\pi^6h^4}{6!})} \right)^2 \geq \left(\frac{4\frac{(i^4+j^4)\pi^4h^2}{12}}{\hat{\lambda}_{i,j}(\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!})} \right)^2, \tag{50}$$

for all $i, j = 1, 2, \ldots, n$. Each (i, j)-term can be bounded from above by the left-hand side of inequality (50). Now substituting for each term in the upper bound from Corollary 7, we obtain

$$\left\| u - \hat{u} \right\| < 4\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{\frac{(i^4+j^4)\pi^4h^2}{12}}{\hat{\lambda}_{i,j}(\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!})} \right)^2}, \tag{51}$$

$$\begin{aligned}
&= \sqrt{\left(\frac{4\frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12}}{\hat{\lambda}_{i_{\min},j_{\min}}(\hat{\lambda}_{i_{\min},j_{\min}} + \frac{(i^4_{\min}+j^4_{\min})\pi^4h^2}{12} - \frac{2(i^6_{\min}+j^6_{\min})\pi^6h^4}{6!})} \right)^2 + \sum_{\substack{i=1 \\ i\neq i_{\min}}}^{n-1}\sum_{\substack{j=1 \\ j\neq j_{\min}}}^{n-1} \left(\frac{4\frac{(i^4+j^4)\pi^4h^2}{12}}{\hat{\lambda}_{i,j}(\hat{\lambda}_{i,j} + \frac{(i^4+j^4)\pi^4h^2}{12} - \frac{2(i^6+j^6)\pi^6h^4}{6!})} \right)^2}, \\
&= \sqrt{\mathcal{O}(h^2) + (n-1)\mathcal{O}(h^2)}, \\
&= 1.
\end{aligned}$$

The proof for the case $\left|\lambda_{i_{\min},j_{\min}}\right| \geq \left|\hat{\lambda}_{\hat{i}_{\min},\hat{j}_{\min}}\right|$ is exactly the same. □

The upper corollary reveals the paramount importance of the accuracy of the near-zero eigenvalues and eigenvectors. These dictate the upper bound for the remaining terms in the sum. If the near-zero eigenmodes are approximated with high accuracy, then the dispersion part of the pollution error can be minimized. This also means that if we need a rough estimate which is in the ball park of the true error, we can simply take the reciprocal of the smallest eigenvalue in magnitude due to its largest contribution to the entire sum. In the next section we will use the results from this section to construct a dispersion correction for the one- and two-dimensional model problems.

4.4. Eigenvalue based dispersion correction

Using this novel perspective, we can construct a dispersion correction using the eigenvalues and eigenvectors. Note that for the one-dimensional MP 1, this can easily be constructed and produces similar results compared to using the modified wavenumber, see Section 4.4.1. However, one advantage we now have is that we can use the same method in the higher-dimensional problem MP 2 to explicitly study how the numerical dispersion translates into the pollution error. In the next section, we will provide numerical evidence for the accuracy ranging from fine to very coarse grids ($kh \geq 1$). The latter will allow to solve and study the current model problem very intricately, while keeping the problem size economically feasible compared to determining the step-size according to $k^3h^2 \leq 1$.

4.4.1. One-dimensional dispersion correction

We start by rewriting our original system as follows. Note that for our matrix A, if $\hat{\lambda}_j$ is an eigenvalue of A corresponding to eigenvector v_j, then

$$Av_j = \hat{\lambda}_j v_j \Longrightarrow (A + cI)v_j = (\hat{\lambda}_j + c)v_j,$$

and thus $\hat{\lambda}_j + c$ is an eigenvalue of $(A + cI)$. Consequently, if the analytical solution is known, a very simple remedy to obtain better accuracy according to our proposition, would be to let

$$c = -\hat{\lambda}_{j_{\min}} + \lambda_{j_{\min}}. \tag{52}$$

This alleviates the mismatch between the exact near zero eigenvalue and the numerical eigenvalue at index $j_{\min}$. Recall from Section 3.2 that the pollution error for MP 1 can be eliminated by incorporating a modified wavenumber $\tilde{k}$. The latter represents an explicit correction of the wavenumber with respect to the dispersion error. Consequently, we can test for the elimination of pollution by comparing the relative error between the exact and numerical solution after solving the following two systems

$$\tilde{A} = A - \tilde{k}I, \text{ where } \tilde{k} = \sqrt{\frac{2(1 - \cos(kh))}{h^2}},$$
$$A_c = A + cI, \text{ where } c = -\hat{\lambda}_{j_{\min}} + \lambda_{j_{\min}}.$$

We furthermore denote

$$\hat{u}_{\tilde{k}} : \tilde{A}\hat{u}_{\tilde{k}} = f \text{ and } \hat{u}_c : A_c\hat{u}_c = f.$$

For the one-dimensional case, our results from Section 4.4.1 suggest that this is often enough to alleviate the adverse effects of numerical dispersion by adding the constant c. However, in some cases, and especially for the two-dimensional model problem, we need a way to shift more smaller eigenvalues while keeping the corresponding eigenvectors unchanged. The reason for this is that in the two-dimensional case there may be a higher algebraic multiplicity and corresponding locations $(i_{\min}, j_{\min})$ where the smallest eigenvalue is located and consequently there may be more than one value for c. In order to circumvent this difficulty, we will make use of some theorems, starting with Brauer's theorem [26].

Theorem 9 (*Brauer*). *Let A be a diagonalizable matrix with $Av_j = \lambda_j v_j$ and suppose r is a vector such that $r^{\intercal} v_j = 1$, then for any scalar $\hat{\lambda}_j$, the eigenvalues of the matrix*

$$\hat{A} = A + (\hat{\lambda}_j - \lambda_j)v_j r^{\intercal},$$

consist of those of A, except that one eigenvalue λ_j of A is replaced with $\hat{\lambda}_j$. Moreover, the eigenvector v_j is unchanged, that is $\hat{A}v_j = \hat{\lambda}_j v_j$.

Proof. For a proof see [26]

Corollary 10. *Let A be a diagonalizable matrix with $Av_j = \lambda_j v_j$ and suppose $r = v_j$ then for any scalar $\hat{\lambda}_j$, the eigenvalues of the matrix*

$$\hat{A} = A + (\hat{\lambda}_j - \lambda_j)v_j v_j^{\intercal},$$

*consist of those of A, except that one eigenvalue λ_j of A is replaced with $\hat{\lambda}_j$. Moreover, **all** the eigenvectors remain unchanged.*

Proof. By the diagonalization property of A, we can write $A = P\Sigma P^{-1}$, where Σ consist of the diagonal matrix containing the eigenvalues of A. Then v_j lies in the j-th column of P. Let e_j be the j-th column of the identity matrix. Then we can take

$$\hat{A} = A + (\hat{\lambda}_j - \lambda_j)P(e_j e_j^{\intercal})P^{-1},$$
$$= A + (\hat{\lambda}_j - \lambda_j)(Pe_j)(e_j^{\intercal}P^{-1}),$$

where $r^{\intercal} = e_j^{\intercal}P^{-1}$, is precisely the jth column of the matrix P^{-1}. □

Using the above theorem and lemma, we can correct each eigenvalue, without shifting the eigenvectors of the previous system. Our dispersion correction for the two-dimensional case will use the above theorem recursively, which is extended into the following lemma.

Lemma 11. *Let A be a diagonalizable matrix such that we can write $A = P^{-1}\Sigma P$, where P is the matrix containing the eigenvectors of A. Then, the same basis can be used for diagonalizing $\hat{A}$, where $\hat{\Sigma}$ is the matrix containing the shifted eigenvalues of A such that $\hat{\Sigma}(j,j) = \hat{\lambda_j}$ and we can write $\hat{A} = P\hat{\Sigma}P^{-1}$.*

Proof. We start by applying Theorem 9 and Corollary 10 recursively. For the first eigenvalue λ_1 we obtain

$$\hat{A} = A + (\hat{\lambda_1} - \lambda_1)(Pe_1)(e_1^{\intercal}P^{-1}),$$

where $\hat{A}$ has exactly the same eigenvectors as the original matrix A, but the first eigenvalue λ_1 is shifted to $\hat{\lambda_1}$. Applying this for all $j = 1, 2, \dots, n$, we finally obtain

$$\hat{A} = A + \sum_{j=1}^{n}(\hat{\lambda_j} - \lambda_1)(Pe_j)(e_j^{\intercal}P^{-1}). \tag{53}$$

We proceed by multiplying Eq. (53) from the left by P^{-1}. If we let I, denote the identity matrix, we obtain

$$\begin{aligned} P^{-1}\hat{A} &= P^{-1}A + \sum_{j=1}^{n}(\hat{\lambda_j} - \lambda_1)(P^{-1}Pe_j)(e_j^{\intercal}P^{-1}), \\ &= P^{-1}A + \sum_{j=1}^{n}(\hat{\lambda_j} - \lambda_1)(Ie_j)(e_j^{\intercal}P^{-1}). \end{aligned} \tag{54}$$

Note that for each j the term $(e_j)(e_j^{\intercal}P^{-1})$ is an all zero matrix apart from the j-th row vector of P^{-1}. Next we multiply Eq. (54) from the right by P, which leads to

$$\begin{aligned} P^{-1}\hat{A}P &= P^{-1}AP + \sum_{j=1}^{n}(\hat{\lambda_j} - \lambda_1)(e_j)(e_j^{\intercal}P^{-1}P), \\ &= \Sigma + \sum_{j=1}^{n}(\hat{\lambda_j} - \lambda_1)(e_j)(e_j^{\intercal}I), \\ &= \Sigma + (\hat{\Sigma} - \Sigma) = \hat{\Sigma}. \quad \square \end{aligned}$$

We can use Lemma 11 to correct the eigenvalues, while keeping the eigenvectors of the original matrix unchanged. We now proceed by constructing the corrected eigenvalues of the new matrix $\hat{A}$. We know that the eigenvalues are bounded from above by a term which is in fact similar to the remainder from the truncation error of the discretization method used. Thus, the method is reminiscent of switching to a higher cut-off point in constructing higher-order discretization stencils. One advantage of this approach is that can now explicitly study the eigenmodes which cause the pollution error as a direct result of numerical dispersion to grow. When constructing higher-order pollution-free discretization schemes, each grid function cannot be tied explicitly to a measure of having numerical dispersion inducing properties. Whereas, the contribution of the particular eigenmodes are now clearly visible in the solution and therefore the error. In our case, we therefore correct the eigenvalues by adding a finite part of the remainder in order to better approximate the analytical eigenvalue. When using Dirichlet boundary conditions, the effect of each eigenmode contributing to the overall pollution term can be studied in one-, two- and three-dimensions.

$$\tilde{\lambda_j} = \hat{\lambda}_j + \sum_{n=2}^{10}\frac{(-1)^n(j\pi)^{2n}h^{2(n-1)}}{(2n)!}.$$

For the one-dimensional case in particular, we need the eigendecomposition and the new matrix containing the corrected eigenvalues to obtain the solution. With respect to the one-dimensional model problem, it is much more efficient to solely correct one eigenvalue, in particular the smallest eigenvalue (see Section 4.4.1). However, for the two-dimensional dispersion correction, we propose a different method, which is based on using the one-dimensional eigendecomposition. As a result, for our model problem, the pollution error can be studied for large wavenumbers in higher-dimensions at reasonable computational costs.

4.4.2. Two-dimensional dispersion correction

As mentioned previously, we can use the one-dimensional eigendecomposition to construct the new two-dimensional coefficient matrix $\hat{A}$. One important feature we need is that the original partial differential equation can be solved using separation of variables. A similar prerequisite is needed and posed in some methods developed in the literature [9–12]. To construct our new two-dimensional matrix $\hat{A}$, the following pseudo-code from Algorithm 1 can be used.

Algorithm 1 Pollution corrected coefficient matrix $\hat{A}_{2D}$ using A_{1D}

```
procedure $\hat{A}$
    Construct eigendecomposition of 1D coefficient matrix $A_{1D}$ such that $D = P^{-1}A_{1D}P$
    for $j = 1 : n$ do
        $\tilde{\lambda}_j = \hat{\lambda}_j + \sum_{n=2}^{10} \frac{(-1)^n (j\pi)^{2n} h^{2(n-1)}}{(2n)!}$
        Replace $\hat{\lambda}_j$ in $D$ with $\tilde{\lambda}_j$
        $\tilde{D}(j, j) = \tilde{\lambda}_j$
    end for
    Use corrected matrix $\tilde{D}$ to construct $\hat{A}_{1D} = P^{-1}\tilde{D}P$
    Construct 2D coefficient matrix $\hat{A}_{2D} = (\hat{A}_{1D} \otimes I_{1D}) + (I_{1D}\hat{A}_{1D}\otimes)$
end procedure
```

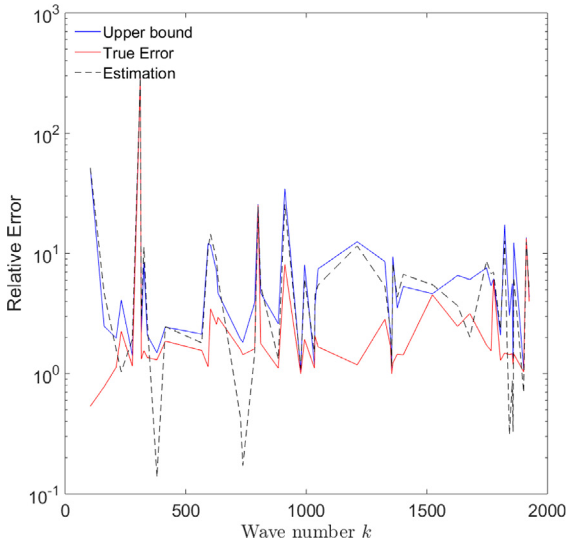


Fig. 1. 1D Relative error and upper bound for various randomly generated k using $kh = 0.625$.

5. Numerical results

We start by examining the error estimates for the pollution error for MP 1 and MP 2. In both cases we evaluate how close our error estimates lie relative to the true error. We then continue by examining the performance of our eigenvalue-based dispersion correction for both model problems. We mentioned that the conventional approach to studying pollution focuses on the notion of a discrepancy between the numerical and exact wavenumber k. In these instances, the exact solution is generally expressed in exponential form, and the eigenvalues are not expressed explicitly. An interesting observation is that this discrepancy between the numerical and exact wavenumber manifests itself through inaccurate near zero eigenvalues. Thus, if the numerical eigenvalues were better approximations of their continuous counterparts, then we expect the relative error to decrease. Section 4.4.1 contains the results for MP 1, while Section 4.4.2 covers MP 2. All one-dimensional systems are solved using a direct method in Matlab R2018a. For the two-dimensional model problems with large k ($k > 300$), we use a standard preconditioned GMRES-solver to obtain the numerical solution, due to the increasing density of the coefficient matrix.

5.1. One-dimensional constant wavenumber model

5.1.1. Error estimation

In Fig. 1 we plot the relative error (red) for random values of k between 100 and 2000 and the upper bound (blue) based on Theorem 4. Additionally, the dashed line is the reciprocal of the smallest numerical eigenvalue in magnitude. This allows us to assess how well this estimate is in the ballpark of the true relative error.

From Fig. 1 we can see that the upper bound always holds, as the blue line is always either above or exactly on the red line. The lines for the error (red) and upper bound (blue) never intersect, and the bound is sharp. Moreover, it shows that the true error behaves more erratically and has a more oscillatory nature which is in direct relation to the smallest eigenvalue in magnitude (dotted line). In particular for example, $k = 1000$ yields a true relative error of 1.493. If we use the bound where the error grows linearly with k, then we have that the pollution term is estimated to be bounded by $k^3h^2 = 390.625$. Using the information from the eigenvalues, our upper bound gives 3.238. Note that the true error (red) follows an oscillatory pattern with peaks appearing for certain k. These are instances where one of the eigenmodes are close to resonant modes and the numerical approximation is poor. If $\lambda_{j_{\min}}$ or $\hat{\lambda}_{j_{\min}}$ is closer to zero than its counterpart, the reciprocal becomes very large. As the intrinsic oscillatory behaviour of the actual error become visible, we observe that the proxy based solely on the smallest eigenvalue (dashed black line) provides a close representation of the actual relative error. Thus, a lot of information can be deduced by simply looking at the smallest eigenvalue in terms of magnitude. Note the proxy is meant to perform as an estimate of the true relative error and not as an upper bound. In some cases, the estimation underestimates the actual relative error.

5.1.2. One-dimensional dispersion correction

For the one-dimensional case, we will use the dispersion correction in Eq. (52). It is also possible to correct each eigenvalue in order to obtain very accurate solutions. However, the results we obtained by using the simple correction with respect to the smallest eigenvalue produces comparable results relative to including the modified wavenumber, which is known to eliminate the pollution error to a satisfactory level. Thus, we start by adding the correction term, which is based on adding terms of truncation error, to the coefficient matrix A,

$$c = -\hat{\lambda}_{j_{\min}} + \sum_{n=2}^{10} \frac{(-1)^n (j_{\min}\pi)^{2n} h^{2(n-1)}}{(2n)!}. \tag{55}$$

This alleviates the mismatch between the exact near zero eigenvalue and the numerical eigenvalue at index $j_{\min}$. As mentioned, recall from Section 3.2 that the pollution error for MP 1 can be eliminated by incorporating a modified wavenumber $\tilde{k}$. The latter represents an explicit correction of the wavenumber with respect to the dispersion error. Consequently, we can test for the elimination of pollution by comparing the relative error between the exact and numerical solution after solving the following two systems

$$\tilde{A} = A - \tilde{k}I, \text{ where } \tilde{k} = \sqrt{\frac{2(1-\cos(kh))}{h^2}},$$

$$A_c = A + cI, \text{ where } c = -\hat{\lambda}_{j_{\min}} + \sum_{n=2}^{10} \frac{(-1)^n (j_{\min}\pi)^{2n} h^{2(n-1)}}{(2n)!}.$$

We furthermore denote

$$\hat{u}_{\tilde{k}} : \tilde{A}\hat{u}_{\tilde{k}} = f \text{ and } , \hat{u}_c : A_c\hat{u}_c = f,$$

and

$$e_{\tilde{k}} = \frac{\|u - \hat{u}_{\tilde{k}}\|}{\|u\|}, \; e_c = \frac{\|u - \hat{u}_c\|}{\|u\|}.$$

Table 1 contains the results for randomly chosen wavenumbers k between 100 and 1000 using 10 grid points per wave length ($kh = 0.625$) and approximately 6 grid points per wave length ($kh = 1$). The latter represents the results of applying the dispersion correction on a very coarse grid. The reason we consider a coarse grid is that in absence of dominating pollution, which has been corrected by either $\tilde{k}$ or c, we should be able to obtain accurate results. The results from Table 1 show that using the eigenvalue correction c leads to significant reduction of the relative error. In some instances it provides even better accuracy than using the adjusted wavenumber $\tilde{k}$. Similar conclusions can be drawn from the results when letting $kh = 1$. While e_c exceeds $e_{\tilde{k}}$ occasionally, we see that $e_{\tilde{k}}$ is more much insensitive to changes in the grid resolution. In particular for $\tilde{k}$, the average error for $kh = 0.625$ appears to be fixed around 0.06, and increases to about 0.18 for $kh = 1$, whereas even for $kh = 1$ even further reductions of the error can be obtained by using the eigenvalue correction c.

5.2. Two-dimensional constant wavenumber model

5.2.1. Error analysis

In this section we provide numerical results for MP 2. We start by presenting the error and the upper bound using the eigenvalues in Fig. 2. To put illustrate the pollution effect, we will present the solution and error for various examples in Figs. 3 and 4.

Starting with Fig. 2, we observe that the upper bound always holds. Similar to the one dimensional case, we again observe the oscillatory nature of the actual true error. The spikes in the error provide great insight relative to the linear relation between k and the increasing error. From Fig. 2 we additionally notice that almost for all k, the relative error is always larger than one. While the upper bound is of the same order as the true error, it is often larger than the true error. Yet, it follows the same oscillatory pattern as the true error from which we can deduce how much each eigenmode contributes to the error. For the first time to our knowledge, we are therefore able to break down and study

Table 1
Relative error e before and after dispersion correction using the eigenvalue-correction c and $\tilde{k}$ for $kh = 0.625$ (left) and $kh = 1$ (right).

k	e	$e_{\tilde{k}}$	e_c	k	e	$e_{\tilde{k}}$	e_c
104	0.8305	0.0675	0.0069	168	1.1536	0.1855	0.0927
170	24.7322	0.0688	0.0714	175	1.3693	0.1864	0.2516
175	1.3315	0.0681	0.0880	210	1.0913	0.1869	0.0271
195	6.2046	0.0687	0.0697	222	1.2413	0.1877	0.0615
237	1.4284	0.0679	0.0302	230	1.1270	0.1861	0.0374
245	2.8360	0.0685	0.0686	230	1.1270	0.1861	0.0374
249	0.9651	0.0679	0.0169	263	1.6076	0.1871	0.2472
306	6.9451	0.0681	0.0273	265	1.4997	0.1883	0.2833
336	1.0465	0.0681	0.0030	315	19.6416	0.1884	0.2404
380	4.0850	0.0680	0.0027	315	19.6416	0.1884	0.2404
498	2.5642	0.0685	0.0335	333	21.4829	0.1880	0.1880
505	1.2707	0.0682	0.0163	337	1.0714	0.1868	0.0484
575	0.9914	0.0680	0.0018	415	1.5989	0.1882	0.1950
584	12.1368	0.0683	0.0708	459	21.2139	0.1883	0.1958
641	1.8812	0.0684	0.0687	461	1.1824	0.1879	0.0462
688	1.0008	0.0682	0.0002	488	1.4006	0.1873	0.0613
720	2.5973	0.0680	0.0116	561	13.4295	0.1882	0.0814
773	1.4760	0.0684	0.0690	594	0.9996	0.1876	0.0131
797	1.3180	0.0682	0.0897	621	18.6735	0.1878	0.2712
814	1.0074	0.0681	0.0065	659	1.6383	0.1879	0.2276
835	1.4264	0.0682	0.0781	820	1.0003	0.1879	0.0024
843	6.1061	0.0684	0.0943	867	21.4856	0.1882	0.1887
922	1.3107	0.0681	0.0333	881	1.5010	0.1884	0.3452
965	1.0184	0.0681	0.0107	882	1.1125	0.1881	0.0445
996	0.9955	0.0682	0.0023	919	1.3408	0.1883	0.0920

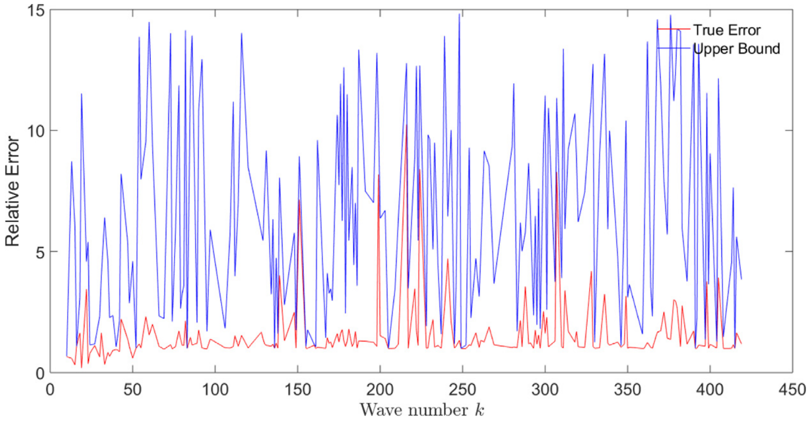
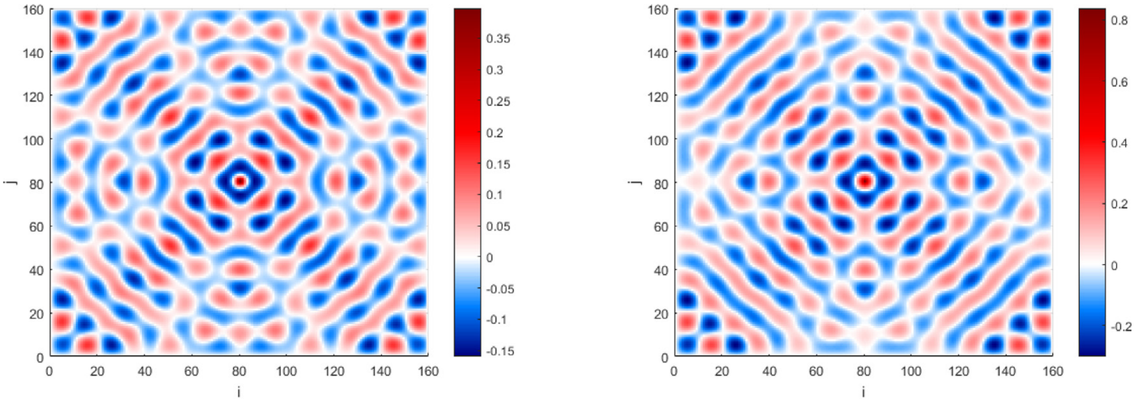


Fig. 2. 2D Relative error for with upper bound for various k between 10 and 425 using $kh = 0.625$.

the dispersive property of the numerical solution in higher-dimensions. The oscillatory error pattern also reveals that the largest contribution in terms of the dispersion can be pointed to the smallest eigenvalues which drive the total sum in Corollaries 6 and 7.

Secondly, as mentioned previously, in some cases the upper bound is much larger than the actual error. This can be understood by noting that in this model problem the source is located at the centre of the numerical domain. Thus, at all even indices j, the sine-term related to the source will be zero and these terms will not be included into the sum. In cases where we see an overshoot, either the smallest numerical or analytical eigenvalue is located at an even index. While it is not part of the actual error, due to being eliminated by the sine-term containing $\frac{\pi}{2}$, it is in fact still included in our upper bound. Note that in creating the upper bound, we do not differentiate between even and odd indices. The reason for this is that we prefer an upper bound which covers the worst case scenario and is not limited to fixing the location of the point source for this model problem.

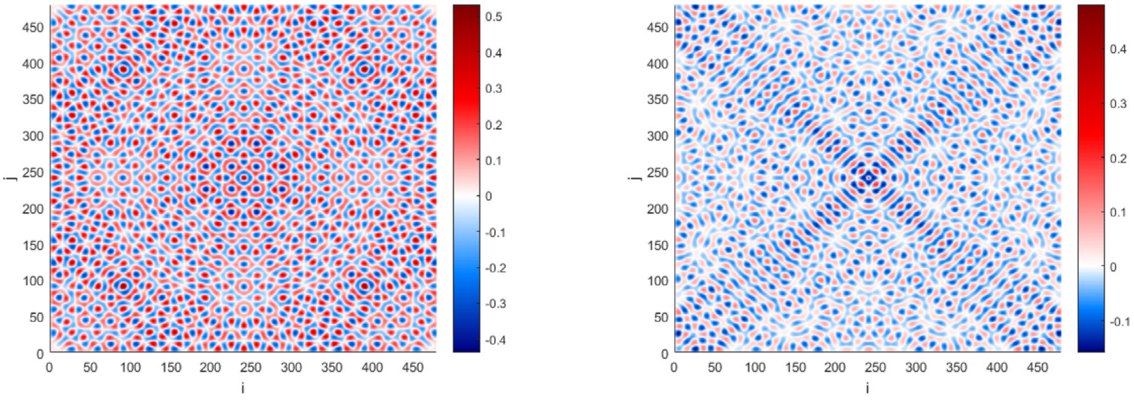
To illustrate the full pollution effect, we continue by plotting some solutions for several values of k. We have plotted the results for $k = 50$ and $k = 150$ in Figs. 3 and 4. Note that here we are using 20 grid points per wave length which results in $kh = 0.3125$. On the x- and y-axis respectively, we have the index i, j corresponding the grid point (x_i, y_j). The colorbar indicates the value of $u(x_i, y_j)$. In all subfigures, blue hues correspond to negative values, whereas red hues correspond to positive values.



(a) Exact solution

(b) Numerical solution

Fig. 3. Exact and numerical solution for MP 2 using second order finite differences and $k = 50$. $kh = 0.3125$, $n^2 = 25\,600$.



(a) Exact solution

(b) Numerical solution

Fig. 4. Exact and numerical solution for MP 2 using second order finite differences and $k = 150$. $kh = 0.3125$, $n^2 = 230\,400$.

We can see from Fig. 3 that for a medium size wavenumber ($k = 50$), the numerical solution is a fair approximation of the exact solution. We can see from the contour of both figures that most of the error does not come from numerical dispersion. If the latter would be the case, the contour of the numerical solution would differ significantly from the exact solution (see Fig. 4 for example).

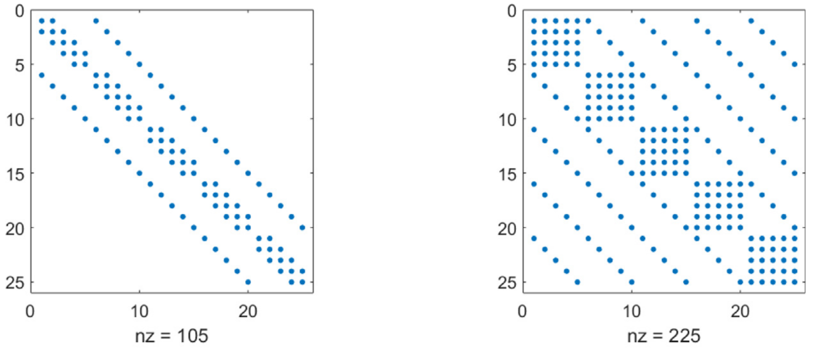
We repeat the analysis for a larger wavenumber; $k = 150$. From Fig. 4(b) we can see that the accuracy deteriorates rapidly as k increases. Fixing the resolution at $kh = 0.3125$ does not suffice in keeping both the phase and amplitude differences under control. We can see from Fig. 4(a) that the exact and numerical solution do not coincide, forcing the conclusion that severe differences between the exact and numerical wavenumber are present. It furthermore supports the observation that increasing the number of grid points mainly results in a substantial resolve of the amplitude differences, rather than the phase differences.

5.2.2. Two-dimensional dispersion correction

We now investigate the effect of applying a dispersion correction using the eigenvalues for the two-dimensional MP 2. Note that for the two-dimensional case it will not suffice to simply add the constant

$$c = -\hat{\lambda}_{i_{\min},j_{\min}} + \sum_{n=2}^{10} \frac{(-1)^n (i_{\min}\pi^{2n} + j_{\min}\pi^{2n}) h^{2(n-1)}}{(2n)!}.$$

There may be multiple locations $(i_{\min}, j_{\min})$ where the smallest eigenvalue is located and thus there may be more than one value for c. If the algebraic multiplicity of the smallest eigenvalue is exactly two, then adding the constant c will still reduce the overall error. However, in the two-dimensional case, the algebraic multiplicity may often be larger than two. Therefore, we will follow the steps described in Algorithm 1. Given that we are solving for the underlying Green's function and general solution, the property that separation of variables can be applied, results in the fact that we can start



(a) Original matrix A

(b) Corrected matrix $\hat{A}$

Fig. 5. Sparsity pattern for $k = 10$ using $kh = 1.5$.

Table 2
Relative (RE) and corrected relative error (CRE) for various k and kh. ⊘ represents the case where the numerical smallest eigenvalue becomes zero.

k	RE	CRE	RE	CRE	RE	CRE	RE	CRE
	$kh = 0.625$		$kh = 1$		$kh = 1.5$		$kh = 2$	
50	0.5997	4.5735e−14	0.9345	8.2702e−12	1.0734	1.6792e−13	2.3670	1.3682e−13
100	2.9899	2.0658e−13	3.4088	3.7729e−13	2.2193	4.9494e−13	⊘	1.3589e−13
150	2.1181	3.8782e−14	4.1974	7.0672e−13	1.5188	3.4133e−12	4.1095	8.2933e−13
200	1.5251	1.0603e−14	6.9603	2.8059e−13	1.2006	2.4750e−13	⊘	8.2361e−13
250	6.0865	1.6424e−13	1.7587	3.0356e−12	1.6301	6.9112e−13	10.1217	1.8701e−13
300	1.6199	9.5418e−13	8.9052	3.5591e−13	1.5293	4.0104e−12	⊘	1.0000e−13
350	1.0057	2.6876e−13	1.0839	7.3402e−13	1.8533	9.1389e−12	2.1767	2.0000e−13
400	1.1679	2.5257e−13	1.0581	1.7699e−13	8.3801	2.3536e−13	⊘	7.0000e−12
450	2.1156	1.9757e−13	3.5765	3.3000e−13	2.0732	3.0631e−13	5.2939	9.1044e−13

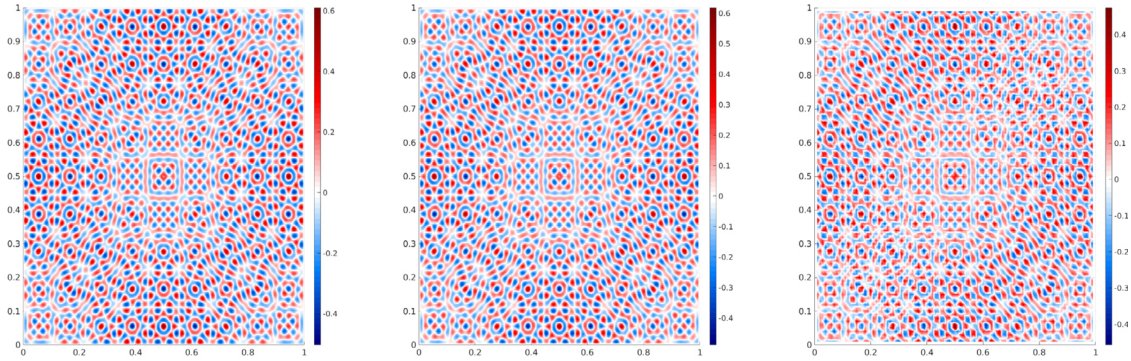
correcting the eigenvalues already in the one-dimensional case and use those to construct the new coefficient matrix $\hat{A}$. As this leads to a correction which is independent of the true analytical wavenumber and pre-specified propagation angles, the resulting coefficient matrix will become more dense and subjected to a different sparsity pattern. In Fig. 5 we have plotted the sparsity pattern of the corrected coefficient matrix $\hat{A}$ for $k = 10$, using $kh = 1.5$. It is apparent that many diagonals are added to the matrix. Additionally, we can see the formation of clear blocks in the centre of the adjusted matrix. For smaller kh, the new coefficient matrix $\hat{A}$ will contain many diagonals and larger blocks, being much more dense yet sparse compared to the original coefficient matrix A.

Before we solve the linear systems explicitly, we verify the two-dimensional dispersion correction. Irrespective of the solution method, we can use the series representation of the discrete solution using the dispersion correction, to establish whether the resulting solution will indeed be dispersion free. Thus, in Table 2 we report the results for various k and kh using the dispersion correction on the numerical eigenvalues which we construct from the one-dimensional case. Note that we do not need to compute the two-dimensional eigenvalues and eigenvectors in Algorithm 1 and proceed until step 6 in the algorithm. We note that in almost all cases the true relative error is always larger than 1 without the dispersion correction. Using the new correction for this model problem, the error is reduced significantly and shows relative independence as regards kh. Even when we move to very coarse grids, which will allow for solving the corresponding linear systems accurately and iteratively, the error stays almost constant despite being in the high-frequency range, which to our current knowledge, is a novel theoretical result. For $kh = 2$, ⊘ represents a case where the numerical smallest eigenvalue without correction becomes zero and we have resonance. This shows the severity of the dispersion causing the pollution, as the actual analytical eigenvalue is still far away from zero.

We now assess the performance in terms of computation time and iterations. In order to make a fair comparison, we solve the linear systems using second-order finite differences using the rule $k^3h^2 = 5$, as this should reduce the pollution error to some extent. We then increase k and report the relative error and number of iterations. From Table 2 we observe that we can use coarser grids to solve for the same wavenumber k and we compare the differences. We will use GMRES as the iterative solver and apply the standard Complex Shifted Laplacian preconditioner (CSLP) with a complex shift set to 1 using multigrid. We use one V-cycle with one pre- and post-smoothing step. Some important remarks are in place. First of all, the accuracy achieved from the iterative solver will depend on the stopping criterion and we set the tolerance at 10^{-6}. Second of all, higher accuracy could have been received of order 10^{-2} by taking k^3h^2 smaller. However, that would lead to large linear systems and thus we report up to $N = 320^2$. Finally, the number of iterations needed to reach convergence for GMRES remains unaffected by the increased accuracy and a detailed study on the convergence behaviour lies beyond the

Table 3
Exact and numerical solutions for $k = 200$. Exact on a fine-grid $kh = 0.625, n^2 = 101761$ and the numerical on coarse-grids using the eigenvalue dispersion correction. For $kh = 2$, we have $n^2 = 9801$.

k	$(A_{2D},\ k^3h^2 \approx 5)$				$(\hat{A}_{2D}\ kh \approx 1)$				$(\hat{A}_{2D}\ kh \approx 2)$			
	n	RE	Its	CPU (s)	n	RE	Its	CPU (s)	n	RE	Its	CPU (s)
10	15	0.046	18	0.094	10	1.571e−08	13	0.066	5	1.010e−09	6	0.052
20	40	0.083	53	0.247	20	4.842e−07	64	0.178	10	5.192e−08	15	0.072
40	114	0.291	111	6.726	40	8.060e−09	225	2.763	20	2.685e−08	217	0.589
60	208	0.522	377	113.888	60	4.991e−07	480	35.072	10	3.981e−07	464	3.653
80	320	1.8612	654	1386.827	80	3.823e−07	712	151.486	40	6.123e−07	901	18.845



(a) Exact solution for $kh = 0.625$ (b) Numerical solution for $kh = 1$ (c) Numerical solution $kh = 2$

Fig. 6. Exact and numerical solutions for $k = 200$. Exact on a fine-grid $kh = 0.625, n^2 = 101\,761$ and the numerical on coarse-grids using the eigenvalue dispersion correction. For $kh = 2$, we have $n^2 = 9801$.

scope of this work. For normal matrices in general, GMRES convergence is governed by the smallest eigenvalues in terms of magnitude. Thus, while the resulting eigenvalues may be more accurate, they may still be small leading to hampered convergence.

Table 3 sheds light on some interesting observations made previously. Using the dispersion correction, we can solve for the same wavenumber k while using coarser grids which lead to smaller linear systems. This is beneficial as this implies that the theoretical study of the pollution error can now be studied from all angles simultaneously in higher-dimensions using coarser systems. If we for example look at $k = 80$, we note that even with $N = 320^2$, which is equivalent to using 27 grid points per wavelength ($k^3h^2 \approx 5$), the error keeps increasing and require even finer grids to obtain accurate solutions. Moreover, the standard iterative solver needs 654 iterations and approximately 1386 s to reach convergence. On the contrary, using $N = 40^2$, which is equivalent to using 3 grid points per wave length ($kh \approx 2$), the error is reduced to the tolerance level of GMRES.

In Fig. 6 we have plotted the exact solution for $k = 200$ on a fine grid and compare it to the numerical solution computed on a very coarse grids using the eigenvalue based dispersion correction. We can see that the accuracy and resolution for such a high wavenumber computed on a very coarse grid ($kh = 2$) are still satisfactory. The figures illustrate what we observed for $k = 200$ in Table 2; the error, after introducing the dispersion correction, at its best is of order 10^{-14} and at its worse of order 10^{-13}. Even for a simple model problem such as ours, achieving an explicit dispersion correction independent of the propagation angle in higher-dimensions is unprecedented.

6. Concluding remarks and summary

In this paper we researched the pollution error due to numerical dispersion for the Helmholtz problem using Dirichlet conditions from an unconventional and novel perspective; the eigenvalues. We have sought to provide the first theoretical basis for defining the pollution error in terms of the eigenvalues. Our work also aims to build a bridge between studying the relation between iterative solvers and the accuracy of numerical solutions now that both have been expressed in terms of a common denominator; the near-zero eigenvalues. This is especially interesting due to the fact that these near-zero eigenvalues, which are generally responsible for hampering the convergence of iterative solvers, are in fact indicators for the pollution effect. Furthermore, by examining the behaviour of the eigenvalues, we proposed an upper bound for the relative error. In particular, we have shown that if the near-zero eigenvalues and eigenvectors are approximated with high accuracy, then the dispersion part of the pollution error can be minimized considerably. Our results also illustrate that the error grows in an oscillatory manner, and our error bound is able to capture and reveal this effect. For higher-dimensional problems, it has been shown theoretically that the pollution error cannot be avoided [2]. However, we have provided a theoretical framework where the pollution error can be brought to approximately zero for very large wavenumbers, irrespective of the grid resolution (kh). The basis of this approach lies in correcting the respective eigenvalues with the remainder, which depends on the order of the truncation error of the finite difference scheme.

Using our new results, we have shown that on a coarse-grid it is possible to obtain pollution-free and therefore accurate one- and two-dimensional solutions. In particular the latter exhibits the novelty of our approach as it is generally considered impossible, even for simple problems to reduce the pollution error to such a degree. The solutions obtained account for all propagation angles simultaneously and do not rely on pre-determined angles for plane-wave propagation, which promotes a detailed study of the pollution effect.

References

[1] A. Deraemaeker, I. Babuška, P. Bouillard, Dispersion and pollution of the fem solution for the helmholtz equation in one, two and three dimensions, Internat. J. Numer. Methods Engrg. 46 (1999) 471–499.

[2] I.M. Babuška, S.A. Sauter, Is the pollution effect of the fem avoidable for the helmholtz equation considering high wave numbers? SIAM J. Numer. Anal. 34 (1997) 2392–2423.

[3] F. Ihlenburg, I. Babuška, Finite element solution of the helmholtz equation with high wave number part ii: the hp version of the fem, SIAM J. Numer. Anal. 34 (1997a) 315–358.

[4] M. Ainsworth, Discrete dispersion relation for hp-version finite element approximation at high wave number, SIAM J. Numer. Anal. 42 (2004) 553–575.

[5] E. Turkel, D. Gordon, R. Gordon, S. Tsynkov, Compact 2d and 3d sixth order schemes for the helmholtz equation with variable wave number, J. Comput. Phys. 232 (2013) 272–287.

[6] C.H. Jo, C. Shin, J.H. Suh, An optimal 9-point, finite-difference, frequency-space, 2-d scalar wave extrapolator, Geophysics 61 (1996) 529–537.

[7] Z. Chen, D. Cheng, W. Feng, T. Wu, An optimal 9-point finite difference scheme for the helmholtz equation with pml, Int. J. Numer. Anal. Model. 10 (2013).

[8] Z. Chen, D. Cheng, T. Wu, A dispersion minimizing finite difference scheme and preconditioned solver for the 3d helmholtz equation, J. Comput. Phys. 231 (2012) 8152–8175.

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

[10] K. Gerdes, F. Ihlenburg, On the pollution effect in fe solutions of the 3d-helmholtz equation, Comput. Methods Appl. Mech. Engrg. 170 (1999) 155–172.

[11] K. Wang, Y. Wong, J. Huang, Analysis of pollution-free approaches for multi-dimensional helmholtz equations, Int. J. Numer. Anal. Model. 16 (2019) 412–435.

[12] K. Wang, Y.S. Wong, Is pollution effect of finite difference schemes avoidable for multi-dimensional helmholtz equations with high wave numbers? Commun. Comput. Phys. 21 (2017) 490–514.

[13] S. Britt, S. Tsynkov, E. Turkel, Numerical simulation of time-harmonic waves in inhomogeneous media using compact high order schemes, Commun. Comput. Phys. 9 (2011) 520–541.

[14] T. Wu, A dispersion minimizing compact finite difference scheme for the 2d helmholtz equation, J. Comput. Appl. Math. 311 (2017) 497–512.

[15] T. Wu, R. Xu, An optimal compact sixth-order finite difference scheme for the helmholtz equation, Comput. Math. Appl. 75 (2018) 2520–2537.

[16] I. Singer, E. Turkel, High-order finite difference methods for the helmholtz equation, Comput. Methods Appl. Mech. Engrg. 163 (1998) 343–358.

[17] C.C. Stolk, A dispersion minimizing scheme for the 3-d helmholtz equation based on ray theory, J. Comput. Phys. 314 (2016) 618–646.

[18] P.H. Cocquet, M.J. Gander, X. Xiang, A finite difference method with optimized dispersion correction for the helmholtz equation, in: International Conference on Domain Decomposition Methods, Springer, 2017, pp. 205–213.

[19] P.H. Cocquet, M.J. Gander, X. Xiang, Closed form dispersion corrections including a real shifted wavenumber for finite difference discretizations of 2d constant coefficient helmholtz problems, SIAM J. Sci. Comput. 43 (2021) A278–A308.

[20] L.L. Thompson, P.M. Pinsky, Complex wavenumber fourier analysis of the p-version finite element method, Comput. Mech. 13 (1994) 255–275.

[21] J. Galkowski, E.H. Müller, E.A. Spence, Wavenumber-explicit analysis for the helmholtz h-bem: error estimates and iteration counts for the dirichlet problem, Numer. Math. 142 (2019) 329–357.

[22] Y. Du, H. Wu, Z. Zhang, Superconvergence analysis of linear fem based on polynomial preserving recovery for helmholtz equation with high wave number, J. Comput. Appl. Math. 372 (2020) 112731.

[23] W. Read, Analytical solutions for a helmholtz equation with dirichlet boundary conditions and arbitrary boundaries, Math. Comput. Model. 24 (1996) 23–34.

[24] F. Ihlenburg, I. Babuška, Dispersion analysis and error estimation of galerkin finite element methods for the helmholtz equation, Internat. J. Numer. Methods Engrg. 38 (1995) 3745–3774.

[25] F. Ihlenburg, I. Babuška, Solution of helmholtz problems by knowledge-based fem, Comput. Assist. Mech. Eng. Sci. 4 (1997b) 397–416.

[26] C.Y. Chiang, M.M. Lin, The eigenvalue shift technique and its eigenstructure analysis of a matrix, J. Comput. Appl. Math. 253 (2013) 235–248.