(Received 16 January 2024, Accepted 12 April 2024, Scheduled 23 April 2024)

Dispersion and Eigenvector Error Analysis of Simplicial Cubic Hermite Elements for 1-D and 2-D Wave Propagation Problems

William A. Mulder[1, 2, *] and Ranjani Shamasundar[2, 3]

[1] *Shell Global Solutions International B.V., The Hague, The Netherlands*
[2] *Faculty of Civil Engineering and Geosciences, Department of Geoscience & Engineering*
Delft University of Technology, Delft, The Netherlands
[3] *Presently at ASML, Veldhoven, The Netherlands*

ABSTRACT: Dispersion error analysis can help to assess the performance of finite-element discretizations of the wave equation. Although less general than the convergence estimates offered by standard finite-element error analysis, it can provide more detailed insight as well as practical guidelines in terms of the number of elements per wavelength needed for acceptable results. We present eigenvalue and eigenvector error estimates for cubic Hermite elements on an equidistant 1-D mesh and on a regular structured 2-D triangular mesh consisting of squares cut in half. The results show that in 1D, the spectrum consists of 2 modes. If these are unwrapped, the spectrum is effectively doubled. The eigenvalue or dispersion error stays below 7% across the entire spectrum. The error in the corresponding eigenvectors, however, increases rapidly once the number of elements per wavelength decreases to one. In terms of element size, the dispersion error is of order 6 and the eigenvector error of order 4. The latter is consistent with the classic finite-element error estimate. In 2D, we provide eigenvalue and eigenvector errors as a series expansion in the element size and obtain the same orders. 2-D numerical tests in the time- and frequency-domain are included.

1. INTRODUCTION

Finite elements are attractive for modeling wave propagation in complex geometries. Elements formulated on the simplex allow for more flexibility in meshing complex geometrical shapes than quadrilaterals and hexahedra. Higher-order elements tend to be computationally more efficient than lower-order elements [1–6]. For the linear element of lowest order, the numerical error mainly shows up as a dispersion error. For that reason, dispersion error analysis is often used as a tool to compare various discretization schemes. With finite-difference schemes, which are translation-invariant on an equidistant grid, this is easily done. With finite elements, translation invariance only occurs at the element level on highly regular uniform meshes for homogeneous problems. For higher-order elements, this leads to a small system instead of a scalar problem in the Fourier analysis of the error. The eigenvalues and eigenvectors of that system describe different modes. Historically, these were classified as 'physical' and 'spurious'. However, proper unwrapping shows that the modes belong to different parts of an enlarged spectrum, each with there own numerical approximation error [7]. Apart from the eigenvalue error in each mode, which is related to the dispersion error, we also have to consider the errors in the corresponding eigenvectors. Because a given 'physical' or exact mode has to be projected on the numerical eigenvectors, each with its error, some energy may end up in the wrong modes. This cross talk is responsible for the 'spurious' behavior.

Here, we apply this eigenvalue and eigenvector analysis to elements based on cubic Hermite interpolating polynomials [8]. These have stronger continuity properties than the classic higher-order elements or their mass-lumped variants [1, 3–5, 9–17], while maintaining a simple structure. In 1D, the cubic polynomials per element are represented by the wavefield and its derivatives on the vertices or nodes. In 2D on the triangle, the element is defined by cubic polynomials with wavefield and its two derivatives on the vertices. To obtain the ten degrees of freedom required to represent a cubic polynomial, a bubble function for the wavefield is added to the interior of the triangle and represented by a wavefield value at its centroid. In 3D on tetrahedra, the element is defined by the wavefield and its three derivatives on the vertices and bubble functions for the wavefield on each of the four faces, providing the 20 degrees of freedom that determine a 3-D cubic polynomial.

Felippa [18] presented a 1-D application of cubic Hermite polynomials to bending elements and included dispersion curves that include a physical and a spurious mode [19]. The latter is also called 'optical' [20, 21]. The eigenvector error analysis is lacking. Here, we will fill that gap.

An application of Hermite elements can be found in [22].

In the next section, we will consider the 1-D case and analyze the dispersion and eigenvector errors. Then, a number of 2-D time- and frequency-domain examples will be presented. The last section summarizes the conclusions.

A preliminary, shorter version of this work appeared in [23].

[*] Corresponding author: William Alexander Mulder (w.a.mulder@tudelft.nl).

doi:10.2528/PIERM24011601

Published by THE ELECTROMAGNETIC ACADEMY

2. ONE DIMENSION

2.1. Finite-Element Discretization

The wave equation in one space dimension reads

$$-k^2u-\frac{\mathrm{d}^2u}{\mathrm{d}x^2}=f. \tag{1}$$

The solution $u(x)$ depends on position x, whereas $f(x)$ is the source term or forcing function. The wavenumber is $k=\omega/c$ for an angular frequency ω and a phase velocity c, which generally depends on ω and x, but will be assumed constant for the dispersion error analysis.

For the finite-element discretization, we choose vertices x_k, $k=0,\dots,N$, that define elements of size $h_\ell=x_\ell-x_{\ell-1}$, $\ell=1,\dots,N$. The element size should typically scale with the local value of the phase velocity to obtain a constant number of elements per wavelength. The basis functions $\phi_i(\xi)$ and $\psi_i(\xi)$, with normalized coordinate $\xi\in[0,1]$ inside the element, should obey

$$\phi_i(j)=\delta_{ij},\quad \frac{\mathrm{d}\phi_i}{\mathrm{d}\xi}(j)=0,\quad \text{for } i,j=0,1, \tag{2}$$

and

$$\psi_i(j)=0,\quad \frac{1}{h}\frac{\mathrm{d}\psi_i}{\mathrm{d}\xi}(j)=\delta_{ij},\quad \text{for } i,j=0,1. \tag{3}$$

In the cubic case, this leads to

$$\begin{aligned}\phi_0(\xi)&=(1-\xi)^2(1+2\xi), & \psi_0(\xi)&=h(1-\xi)^2\xi,\\ \phi_1(\xi)&=\xi^2(3-2\xi), & \psi_1(\xi)&=-h(1-\xi)\xi^2.\end{aligned} \tag{4}$$

Here, h is the length of the element. Note that $\phi_1(\xi)=\phi_0(1-\xi)$ and $\psi_1(\xi)=-\psi_0(1-\xi)$.

The discretization is straight-forward if the phase velocity is constant per element. If the set of basis functions is ordered as $\{\phi_0(\xi),\psi_0(\xi),\phi_1(\xi),\psi_1(\xi)\}$, with the subscript 0 for the left node and 1 for the right node of the element, then the contribution to the mass matrix per element is

$$A=hQ\bar{A}Q,\quad \bar{A}=\frac{1}{420}\begin{pmatrix}156&22&54&13\\22&4&13&3\\54&13&156&22\\13&3&22&4\end{pmatrix}, \tag{5}$$

where

$$Q=\mathrm{diag}\{1,h,1,-h\}. \tag{6}$$

Here, the degrees of freedom are paired as u and $\frac{\mathrm{d}u}{\mathrm{d}x}$ on the left and on the right side of the element. Note that h may vary from element to element. The contribution to the stiffness matrix is

$$B=\frac{1}{h}Q\bar{B}Q,\quad \bar{B}=\frac{1}{30}\begin{pmatrix}36&3&-36&-3\\3&4&-3&1\\-36&-3&36&3\\-3&1&3&4\end{pmatrix}. \tag{7}$$

An alternative evaluation of A and B is based on the nodal-to-modal map [24].

2.2. Dispersion Analysis

For the dispersion analysis, we consider a homogeneous problem and assume that the mesh is equidistant and periodic. Then, the mass matrix $\mathcal{M}$ and stiffness matrix $\mathcal{K}$ become

$$\mathcal{M}=\frac{h}{420}\begin{pmatrix}6[52+9(T+T^{-1})] & -13h(T-T^{-1})\\ 13h(T-T^{-1}) & h^2[2+3(2-T-T^{-1})]\end{pmatrix}, \tag{8}$$

and

$$\mathcal{K}=\begin{pmatrix}\frac{6}{5h}(2-T-T^{-1}) & \frac{1}{10}(T-T^{-1})\\ -\frac{1}{10}(T-T^{-1}) & \frac{h}{30}[6+2-T-T^{-1})\end{pmatrix}. \tag{9}$$

The shift operator T is defined by $T^nu_m=u_{m+n}$, where u_m approximates the wavefield at x_m. We also have $T^nu'_m=u'_{m+n}$ for the derivative u'_m that approximates $\frac{\mathrm{d}u}{\mathrm{d}x}(x_m)$. The matrices operate on vectors $(u_m,u'_m)^\mathsf{T}$, where $(\cdot)^\mathsf{T}$ denotes the transpose.

The Fourier symbol of T is $\hat{T}=e^{\mathrm{i}kh}$ for wavenumber k. The dispersion curve follows from the eigenvalues of $\hat{L}=\hat{\mathcal{M}}^{-1}\hat{\mathcal{K}}$, which is now a 2×2 system. Here, $\hat{\mathcal{M}}$ represents the Fourier symbol of the mass matrix and $\hat{\mathcal{K}}$ that of the stiffness matrix. The eigenvalues are

$$\kappa_\pm^2=\frac{6[141-4\zeta(8+\zeta)\pm w]}{h^2[65+\zeta(\zeta-36)]}, \tag{10}$$

with

$$w=\sqrt{13056+\zeta\,[3856+\zeta\,\{-7524+\zeta(1656-19\zeta)\}]}, \tag{11}$$

where $\zeta=\cos(kh)$. For small k,

$$(\kappa_-/k)^2\simeq 1+\frac{(kh)^6}{30240}, \tag{12}$$

demonstrating sixth-order behavior of the dispersion error, relative to the exact wavenumber k.

Figure 1(a) plots the eigenvalues $\kappa_\pm$ as a function of the normalized wavenumber $\eta=kh/(2\pi)$. Note that Nyquist-Shannon sampling theorem requires $|kh|\le\pi$ in the scalar case. Here, with both u and $\frac{\mathrm{d}u}{\mathrm{d}x}$, we have $|kh|\le 2\pi$. The results for negative wavenumbers follow by symmetry and are not plotted.

Had we only shown the results for $|kh|\le\pi$, then one eigenvalue, κ_- would be physical and the other spurious or 'optical' [18, 21]. By enlarging the domain to $|kh|\le 2\pi$, we can unwrap the two eigenvalues: $\kappa_-/(2\pi\eta)$ for $\eta\in[0,\frac{1}{2}]$ and $\kappa^+/(2\pi\eta)$ for $\eta\in[\frac{1}{2},1]$. The other ones, $\kappa^+/(2\pi\eta)$ for $\eta\in[0,\frac{1}{2}]$ and $\kappa_-/(2\pi\eta)$ for $\eta\in[\frac{1}{2},1]$ then remain as spurious modes. The symmetry $\kappa_\pm^2(1-\eta)=\kappa_\pm^2(\eta)$ follows from the dependence of the eigenvalues on $\zeta=\cos(2\pi\eta)$.

Figure 1(b) shows the physical eigenvalues after scaling by the exact eigenvalue. The two values near the discontinuity at $\eta=1/2$ are $\kappa h=\sqrt{168/17}$ and $\sqrt{10}$, both close to π. If all the energy could be restricted to these modes, there would be no spurious modes. For instance, if k is small, all wave energy should be confined to the eigenvector of κ_-. In practice, some energy may end up in the eigenvector of κ_+ for small k. We

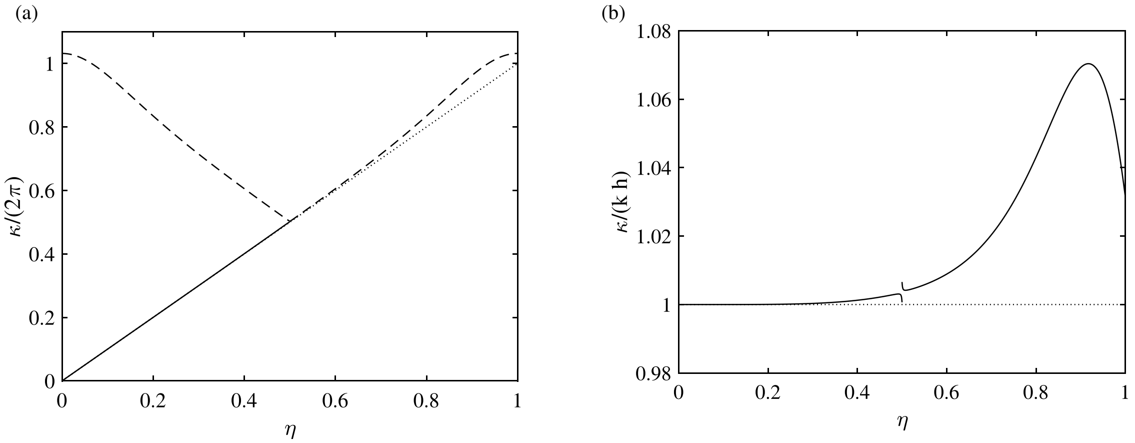

FIGURE 1. (a) The positive square-roots of the two eigenvalues, scaled by 2π, as a function of the normalized wavenumber η. The spurious modes are shown as dashed lines. (b) Unwrapped normalized dispersion curve for the 1-D element based on cubic Hermite polynomials, showing the numerical approximation κ of the wavenumber normalized by the exact one, $kh = 2\pi\eta$. The dotted line is the exact result, the drawn and dashed lines mark the two eigenvalue branches.

will study this in more detail by considering the error in the eigenvectors.

To determine the error in the eigenvectors, we follow [7] and express the Fourier symbol of the spatial operator as $\hat{L} = Q\Lambda Q^{-1}$, where the columns of Q are the eigenvectors of $\hat{L}$ and the diagonal matrix Λ contains the eigenvalues $\kappa_{\pm}^2$ on its diagonal.

The exact eigenvector corresponding to the mode $e^{\mathrm{i}kx}$ in the Fourier domain is $\hat{\mathbf{e}}_0 = (1, \mathrm{i}k)^\mathsf{T}$. Minus the second spatial derivative turns this into $k^2\hat{\mathbf{e}}_0$, whereas the numerical approximation produces $\hat{L}\hat{\mathbf{e}}_0$. The error in the eigenvector is then something like $k^{-2}\hat{L}\hat{\mathbf{e}}_0 - \hat{\mathbf{e}}_0$. To separate the dispersion error from the error in the eigenvectors, we can replace the numerical eigenvalues κ^2 in Λ by the exact k^2, evaluate the effect of the modified operator $\hat{L}$ on the exact eigenvector $\hat{\mathbf{e}}_0$, divide by k^2 afterwards, and compare the result to the same exact eigenvector. We can also do that for each of the eigenvectors separately by setting the eigenvalues to zero except for the one of interest. Assuming that the first eigenvector corresponds to κ_- and the second to κ_+, we can focus on κ_- for small k. We define vectors

$$\hat{\mathbf{s}}_1 = k^{-2} Q \operatorname{diag}\{k^2, 0\} Q^{-1}\hat{\mathbf{e}}_0 \tag{13}$$

and

$$\hat{\mathbf{s}}_2 = k^{-2} Q \operatorname{diag}\{0, k^2\} Q^{-1}\hat{\mathbf{e}}_0. \tag{14}$$

These describe the following steps: project the exact eigenvector on the numerical ones, propagate with the exact wavenumber, project back, rescale by the squared the wavenumber, and compare to the input. The matrix $\hat{S} = (\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2)^\mathsf{T}$ has these vectors as its first and second column. Then,

$$\hat{S} \simeq \begin{pmatrix} 1 - \frac{2}{4725}(kh)^6 & \frac{2}{4725}(kh)^6 \\ \mathrm{i}k\left[1 + \frac{2}{315}(kh)^4\right] & \mathrm{i}k\left[-\frac{2}{315}(kh)^4\right] \end{pmatrix}. \tag{15}$$

The first column approximates the exact eigenvector $\hat{\mathbf{e}}_0$, and the second column describes how much of it ends up in the other mode and is usually classified as spurious energy. This column has the opposite sign of the error, $\mathbf{s}_1 - \hat{\mathbf{e}}_0$, in the first column, that is, $\mathbf{s}_1 + \mathbf{s}_2 = \hat{\mathbf{e}}_0$. The matrix shows that the first row, corresponding to u, has a sixth-order error, and the second row, corresponding to the derivative of u, has a fourth-order error. The last determines the overall error behavior of the scheme.

To study the eigenvector error over the whole domain, we first rescale the eigenvector to obtain relative errors, by dividing out the factor ik. Let $D = \operatorname{diag}\{1, (\mathrm{i}k)^{-1}\}$. The normalized exact eigenvector becomes $\hat{\mathbf{e}}_1 = D\hat{\mathbf{e}}_0 = (1, 1)^\mathsf{T}$ and the numerical ones the columns of $\tilde{Q} = DQ$:

$$\tilde{Q} = \begin{pmatrix} \frac{a+w}{d} & \frac{a-w}{d} \\ \frac{\sin\xi}{\xi} & \frac{\sin\xi}{\xi} \end{pmatrix}, \tag{16}$$

with $\xi = kh = 2\pi\eta$, $\zeta = \cos(\xi)$, $d = 6(52 - 17\zeta)$, $a = 80 + \zeta(52 - 27\zeta)$ and w as in Equation (11). We then consider the vectors

$$\hat{\mathbf{r}}_1 = \tilde{Q} \operatorname{diag}\{1, 0\} \tilde{Q}^{-1}\hat{\mathbf{e}}_1, \tag{17}$$
$$\hat{\mathbf{r}}_2 = \tilde{Q} \operatorname{diag}\{0, 1\} \tilde{Q}^{-1}\hat{\mathbf{e}}_1. \tag{18}$$

Note that $\hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 = \hat{\mathbf{e}}_1$. The vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ contain 4 components that describe the eigenvector error. The drawn line in Figure 2 consists in $\hat{r}_{1,1} - 1$ for $\eta < \frac{1}{2}$ and $\hat{r}_{2,2} - 1$ for $\eta > \frac{1}{2}$, with 0 at $\eta \to \frac{1}{2}$. The dashed line follows $\hat{r}_{1,2} - 1$ for $\eta < \frac{1}{2}$ and $\hat{r}_{2,1} - 1$ for $\eta > \frac{1}{2}$, with -1 at $\eta \to \frac{1}{2}$. These represent the relative difference between the approximate and exact eigenvectors. The missing components represent the spurious modes and just have the opposite sign, because $\hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 = 1$, and are therefore not shown.

Figure 2 seems to suggest that we should stay at some distance below the Nyquist limit of $\eta = \frac{1}{2}$, since one if the branches shoots off to -1 around $\eta = \frac{1}{2}$. This may be too pessimistic as around $\eta = \frac{1}{2}$, the two eigenvalues $\kappa_{\pm}^2$ are nearly equal. However, the amplitude of the dashed curve rapidly increases for η above $\frac{1}{2}$, so having $|\eta| \leq \eta_{\max}$ with $\eta_{\max}$ just below $\frac{1}{2}$ is advisable. This means that the number of elements per wavelength should be somewhat above 1, or larger for higher accuracy.

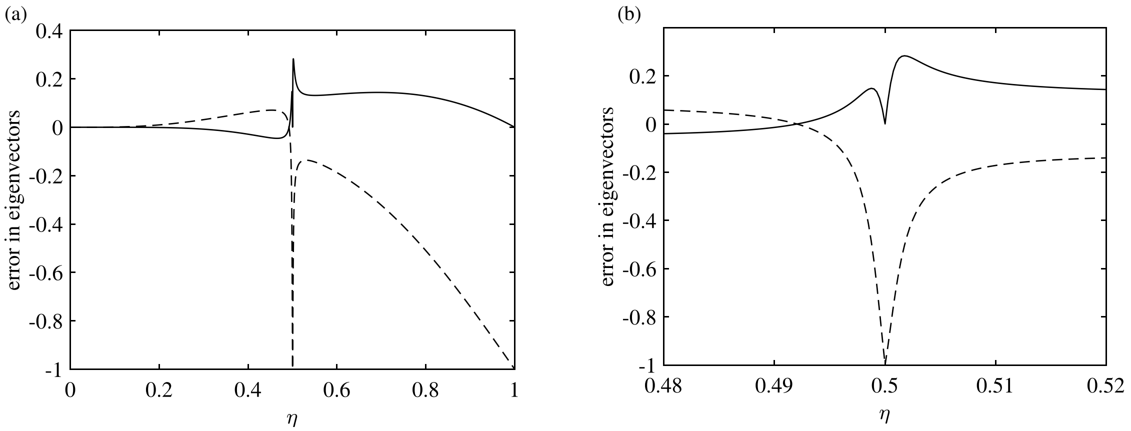

FIGURE 2. Error in the eigenvectors. Only two of the four components are shown, since the other two just have the opposite sign. (b) Detail of (a).

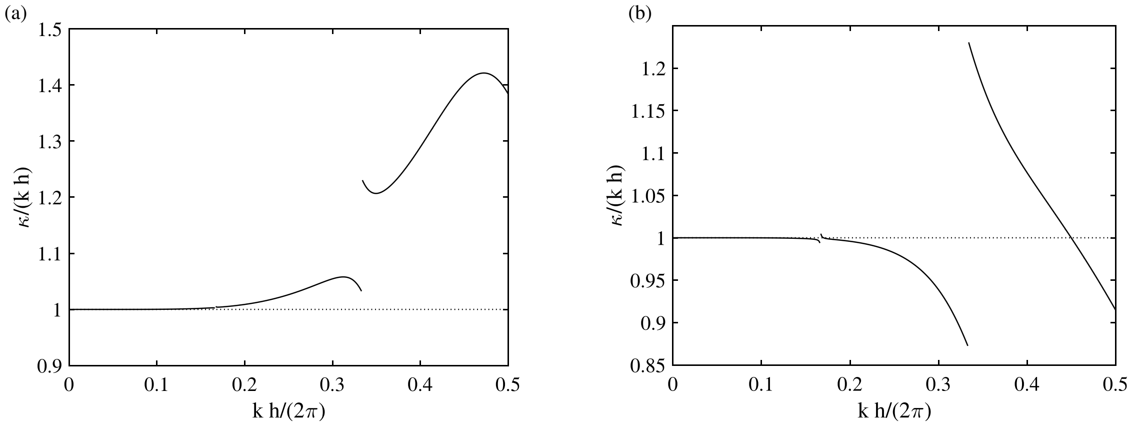

FIGURE 3. Unwrapped normalized dispersion curve for the standard 1-D cubic element with (a) equidistant nodes and a consistent mass matrix and (b) Legendre-Gauss-Lobatto nodes and a lumped mass matrix, showing the numerical approximation κ of the wavenumber normalized by the exact one, $kh/(2\pi)$. The dotted line is the exact result.

We can now compare the above results to the standard cubic element with equidistant internal nodes and to the lumped cubic element with Legendre-Gauss-Lobatto points, both analyzed in [7]. Figure 3(a) shows the unwrapped dispersion curve for the former. The three branches are clearly visible. Note that the spectrum is doubled nor tripled. The errors are larger than with the Hermite cubic element. The rightmost panel of Figure 5 in [7] shows that a significant eigenvector error already appears in the first branch with $\eta \in [0, \frac{1}{6}]$, at considerably smaller values of η than with the cubic Hermite element.

Figure 3(b) shows the unwrapped dispersion curve for the lumped cubic element with Legendre-Gauss-Lobatto points. Again, the errors are larger than with the Hermite cubic element but smaller than for the previous one. The rightmost panel of Figure 7 in [7] again shows a larger eigenvector error at smaller η than with the cubic Hermite element.

3. TWO DIMENSIONS

3.1. Method

The finite-element discretization starts with the reference triangle with barycentric coordinates $\xi_0 = 1 - \xi_1 - \xi_2$, ξ_1 and ξ_2. The true coordinates inside a triangle with vertices $(x, y) = (x_{1,k}, x_{2,k})$, $k = 0, 1, 2$, are $x_j = \sum_{k=0}^{2} \xi_k x_{j,k}$ for $j = 1, 2$. Let $\alpha_j = x_{j,1} - x_{j,0}$ and $\beta_j = x_{j,2} - x_{j,0}$ for $j = 1, 2$. The 10 degrees of freedom are the wavefield values u_k and their derivatives $u_{1,k}$ and $u_{2,k}$ in $x = x_1$ and $y = x_2$, respectively, on the three vertices indexed by $k = 0, 1, 2$, as well as the wavefield p_c at the centroid. We order them as $\{u_0, u_{1,0}, u_{2,0}, u_1, u_{1,1}, u_{2,1}, u_2, u_{1,2}, u_{2,2}, u_c\}$ per element. The corresponding basis functions are

$$\begin{aligned}
\phi_1 &= \xi_0[(3 - 2\xi_0)\xi_0 - 7\xi_1\xi_2],\\
\phi_2 &= \xi_0[\alpha_1\xi_1(\xi_0 - \xi_2) + \beta_1\xi_2(\xi_0 - \xi_1)],\\
\phi_3 &= \xi_0[\alpha_2\xi_1(\xi_0 - \xi_2) + \beta_2\xi_2(\xi_0 - \xi_1)],\\
\phi_4 &= \xi_1[(3 - 2\xi_1)\xi_1 - 7\xi_2\xi_0],\\
\phi_5 &= \xi_1[\gamma_1\xi_2(\xi_1 - \xi_0) + \alpha_1\xi_0(\xi_2 - \xi_1)],\\
\phi_6 &= \xi_1[\gamma_2\xi_2(\xi_1 - \xi_0) + \alpha_2\xi_0(\xi_2 - \xi_1)],\\
\phi_7 &= \xi_2[(3 - 2\xi_2)\xi_2 - 7\xi_0\xi_1],\\
\phi_8 &= \xi_2[\beta_1\xi_0(\xi_1 - \xi_2) + \gamma_1\xi_1(\xi_0 - \xi_2)],\\
\phi_9 &= \xi_2[\beta_2\xi_0(\xi_1 - \xi_2) + \gamma_2\xi_1(\xi_0 - \xi_2)],\\
\phi_{10} &= 27\xi_0\xi_1\xi_2.
\end{aligned} \tag{19}$$

The last is the bubble function. The mass and stiffness matrix per element follow from exact integration over the triangle and serve as input for the global assembly. The differences α_j, β_j and $\gamma_j = \beta_j - \alpha_j$ in the basis functions are related to projections on the edges of the vectors defined by the wavefield gradient. Note that the gradient of the bubble function prevents C^1 continuity across element edges.

Zero Dirichlet boundary conditions are easily implemented by eliminating wavefield values on the boundary from the mass and stiffness matrix. If the boundaries are aligned with the coordinate axes, also their tangential derivatives should be eliminated. Similarly, zero Neumann conditions can be implemented by eliminating the normal wavefield derivatives on the boundary. At corners, this implies that both the horizontal and vertical derivatives are zero and should be removed from the mass and stiffness matrices. If the boundaries are not aligned, a local rotation can be applied for the wavefield derivatives at the vertices that belong to a boundary edge.

To maintain code flexibility, we have chosen to first assemble the global mass matrix $\mathcal{M}$ and stiffness matrix $\mathcal{K}$ without taking the boundary conditions into account. The contributions per element are evaluated from the basis functions (19) in Cartesian coordinates using expressions simply obtained by means of a symbolic algebra package. After global assembly, either zero Dirichlet or Neumann boundary conditions were applied in the test problems we considered. This was implemented by a global rotation matrix R_b. It is an identity operator for the degrees of freedom corresponding to the bubble function and for the degrees of freedom in the interior. For vertices on the boundary, the normal and tangential vectors of the connected boundary edges are computed. If they are the same, we define a local 2×2 rotation matrix for the derivative components with the edge normal on the first row and the tangential component on the second. After rotation, either the first is zero with a Neumann condition or the second if a zero Dirichlet condition holds. This small matrix is inserted in the global one, R_b. If they are different, we have a corner point and the two components of the wavefield derivative should be zero, both for the Dirichlet and the Neumann boundary conditions. In that case, we just use the identity matrix for the local 2×2 rotation matrix.

Next, the solution vector is transformed into $R_b\mathbf{u}$, the mass matrix into $R_b\mathcal{M}R_b$, and the stiffness matrix into $R_b\mathcal{K}R_b$. Note that the inverse, or actually the pseudo-inverse, of R_b is R_b. Then, we reduce the size of the vector $R_b\mathbf{u}$ by eliminating the zero boundary values. These are the wavefield values and tangential components of the derivatives for zero Dirichlet boundary conditions or both components at corners points. For the zero Neumann boundary conditions, only the normal components are zero and both in corners points. The result of taking this subset is denoted by $\overline{\mathbf{u}} = S_b R_b \mathbf{u}$. Here, S_b is a non-square matrix with ones and zeros that only selects the non-zero values. Likewise, we obtain $\overline{\mathcal{M}} = S_b R_b \mathcal{M} R_b S_b^{\mathsf{T}}$ and $\overline{\mathcal{K}} = S_b R_b \mathcal{K} R_b S_b^{\mathsf{T}}$, where $(\cdot)^{\mathsf{T}}$ denotes the transpose. If, in the frequency domain and in 2D, we have to solve

$$-k^2 u - \Delta u = f, \tag{20}$$

with a source term $f(\omega, x, y)$, this would require the solution of

$$\left(-\overline{\mathcal{M}} + \overline{\mathcal{K}}\right) \overline{\mathbf{u}} = S_b R_b \mathbf{f}. \tag{21}$$

Note that the mass matrix $\overline{\mathcal{M}}$ contains the factor k^2. We can insert the result $\overline{\mathbf{u}}$ back into a vector of the original size and perform the inverse rotation to obtain $\mathbf{u} = R_b S_b^{\mathsf{T}} \overline{\mathbf{u}}$. In the code, the action of S_b and S_b^{T} is of course implemented with an array of pointers.

A point source $f(\omega, x, y) = w(\omega)\delta(x - x_s)\delta(y - y_s)$ can be represented by Dirac delta functions and a frequency-dependent amplitude $w(\omega)$. Its discrete representation involves

$$\sum_j \int_{\mathcal{T}_j} \phi_{j,k} \delta(x - x_s)\delta(y - y_s)\, \delta x \delta y = \phi_{j_s,k}(x_s, y_s), \tag{22}$$

where the domain Ω is partitioned into elements $\mathcal{T}_j$ and $\phi_{j,k}$ for $k = 1, \ldots, 10$ is one of the basis functions of (19) on that element. Element j_s contains the source. Note that there may be multiple elements containing the source if it happens to lie on an edge or vertex. Because of the continuity of the basis functions, we only have to include one of those.

Since both the mass and stiffness matrix are symmetric and real, sparse Cholesky decomposition after minimum-degree reordering [25] can be used for matrix inversion. Once a solution has been found, we can sample at receiver positions, using the basis functions in (19) for interpolation.

3.2. Dispersion Analysis

Fourier analysis in two dimensions is straightforward if the mesh consists of squares of size $h \times h$, each divided into two triangles. We will first focus on this regular case. Distorted versions [3, 26, 27], based on a parallelepiped instead of a square, are considered in Appendix A.

The square near the origin consists in one triangle with vertices $(0, 0)$, $(h, 0)$, $(0, h)$ and another with $(h, 0)$, (h, h), $(0, h)$. Shift operators T_x and T_y are defined by $T_x^k u_{i,j} = u_{i+k,j}$ and $T_y^k u_{i,j} = u_{i,j+k}$. Their Fourier symbols are $\hat{T}_x = \exp(\mathrm{i}\xi)$ and $\hat{T}_y = \exp(\mathrm{i}\eta)$ with $\xi = k_x h$ and $\eta = k_y h$, where k_x and k_y are the wavenumbers in the x- and y-directions, respectively.

The degrees of freedom are u, $\partial u/\partial x$ and $\partial u/\partial y$ at the vertices, u at the centroid of the first triangle at $(\frac{1}{3}h, \frac{1}{3}h)$ and of the second triangle at $(\frac{2}{3}h, \frac{2}{3}h)$, and all their translates on a periodic mesh. After a spatial Fourier transform, the 10 degrees of freedom per element can be represented by 5 coupled ones on the periodic mesh. We denote the Fourier symbol of the mass matrix by $\hat{\mathcal{M}}$, that of the stiffness matrix by $\hat{\mathcal{K}}$, and that of the spatial operator by $\hat{L} = \hat{\mathcal{M}}^{-1}\hat{\mathcal{K}}$. Its 5 eigenvalues for zero wavenumbers $\xi = \eta = 0$ are $\{0,\, 42,\, 42\,, 84,\, 560/3\}$, two of which are identical. The series expansion for the smallest is given by

$$\begin{aligned} \lambda_1 \simeq (\xi^2 + \eta^2) + \tfrac{1}{5080320}\big[&159(\xi^8 + \eta^8) - 696\,\xi\eta(\xi^6 + \eta^6) \\ &+3221(\xi\eta)^2(\xi^4 + \eta^4) - 6792(\xi\eta)^3(\xi^2 + \eta^2) \\ &+8734(\xi\eta)^4\big], \end{aligned} \tag{23}$$

showing that the dispersion error is of order 6.

The smallest eigenvalue should have an eigenvector that approximates the exact one, given by

$$\hat{\mathbf{e}}_1 = \Big(1,\ \mathrm{i}\xi,\ \mathrm{i}\eta,\ e^{\frac{1}{3}\mathrm{i}(\xi+\eta)},\ e^{\frac{2}{3}\mathrm{i}(\xi+\eta)}\Big)^{\mathsf{T}}. \tag{24}$$

The superscript $(\cdot)^\intercal$ denotes the transpose. The leading error in eigenvector, ignoring terms of order 5 in x and y, is $(a,0,0,0,0)^\intercal$ or, with another scaling by $1/(1+a) \simeq 1-a$, $(0,0,0,-a,-a)^\intercal$, where

$$a = \frac{47(x^4+y^4) - 124\,xy(x^2+y^2) + 192(xy)^2}{27216}. \qquad (25)$$

Another quick way to estimate the cross talk between eigenvectors is provided by the discrete operator acting on the exact first eigenvector. From

$$\hat{L}\mathbf{e}_1 - (\xi^2+\eta^2)\mathbf{e}_1 \simeq \left(-84a, d_2, d_3, \frac{308}{3}\,a, \frac{308}{3}\,a\right)^\intercal, \qquad (26)$$

where

$$\begin{aligned} d_2 \;&=\; -\frac{\mathrm{i}}{360}\big(8\xi^5 - 1095\xi^4\eta + 2050\xi^3\eta^2 \\ &\quad -1795\xi^2\eta^3 + 850\xi\eta^4 - 28\eta^5\big) \end{aligned} \qquad (27)$$

and

$$\begin{aligned} d_3 \;&=\; \frac{\mathrm{i}}{360}\big(28\xi^5 - 850\xi^4\eta + 1795\xi^3\eta^2 \\ &\quad -2050\xi^2\eta^3 + 1095\xi\eta^4 - 458\eta^5\big). \end{aligned} \qquad (28)$$

This obviously produces an error of order 4, as in the 1-D case.

As already mentioned, distorted triangles based on a periodic pattern of parallelepipeds are considered in Appendix B. The results are qualitatively the same.

3.3. Time Domain

We have tested the method on a 2-D standing-wave problem in a model with constant wave speed. The partial differential equation is

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \qquad (29)$$

The time stepping scheme lets

$$\mathbf{u}^{n+1} = 2\mathbf{u}^n - \mathbf{u}^{n-1} - (\Delta t)^2 \mathcal{M}^{-1}\mathcal{K}\mathbf{u}^n, \qquad (30)$$

where the superscript denotes time $t^n = t^0 + n\Delta t$. The vector $\mathbf{u}$ denotes the degrees of freedom. The time step Δt should be chosen such that $0 \le (\Delta t)^2 L \le 4$, with $L = \mathcal{M}^{-1}\mathcal{K}$.

Before the time stepping starts, we apply a sparse Cholesky decomposition on the real symmetric matrix $\mathcal{M}$ after minimum-degree reordering [25]. During the time stepping, the result is used to compute the action of $\mathcal{M}^{-1}$ on the vector $\mathcal{K}\mathbf{u}^n$.

The domain for the test problem has a size $[0,2] \times [0,1]$ in dimensionless units. We choose a unit wave speed c. The exact solution is a standing wave of the form $u = \sin(\alpha_1 x)\sin(\alpha_2 y)\cos(\omega t)$, with $\omega = c(\alpha_1^2 + \alpha_2^2)^{1/2}$, $\alpha_k = 2\pi m_k$ and $m_1 = 4$, $m_2 = 2$. The solution obeys zero Dirichlet boundary conditions. The initial-value problem is started at time zero and runs until $t_{\max} = 2\pi/\omega$. The mesh was generated by taking a uniform background mesh with square cells, perturbing internal vertices randomly by at most 10% to make it more irregular, and applying a Delaunay triangulation. Figure 4 shows a fairly coarse mesh and the initial wavefield. For plotting purposes, the latter was interpolated from the given degrees of freedom on the mesh to a much finer Cartesian grid, using the cubic Hermite polynomial representation.

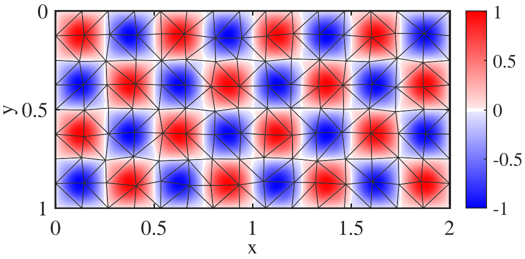

FIGURE 4. Coarse mesh with the exact initial wavefield as backdrop.

The root-mean-square (RMS) errors in the wavefield p at the vertices, u_c at the element centroids, and in the horizontal and vertical derivatives at the vertices was measured at a time $t_{\max}$. Figure 5 plots the results as a function of $n_\lambda = \lambda/\sqrt{|\Omega|/n_{\mathrm{e}}}$, where $|\Omega|$ is the area of the domain and n_{e} the number of elements. This quantity estimates the number of elements per wavelength and is proportional to the inverse of the average element size. Power-law fits provide an error of order 4 for u and u_c and of order 3 for $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$, as expected for a cubic-polynomial representation of the wavefield and a smooth solution. The second-order time-stepping scheme ran at about half the maximum allowable value, which apparently produces time-stepping errors small enough to prevent them from showing up as a second-order trend in the graphs, implying that the spatial errors dominate in this case.

In this example, the RMS error in the solution u at the vertices is about 6% for $n_\lambda = 4$.

3.4. Frequency Domain

The first test has a point source at the origin at a frequency that corresponds to a wavelength of $\lambda = 2\pi/k = 0.66$ m in a square domain of size $[-5,5] \times [-5,5]$ m^2.

This combination of wave propagation speed, frequency, and domain size does not produce resonances. Zero Neumann boundary conditions are imposed all around.

A sequence of meshes was generated with MESH2D version 3.1 [28], and the corresponding solutions were computed in Matlab®. Figure 6 shows the solution on a mesh with $N = 1156893$ vertices. Data were recorded by a line of receivers at $x_r = (0.05\,j)$ m and $y_r = 0$ m, where $j = 1, 2, \ldots, 100$. The results were compared to the exact solution, listed in Appendix B. The maximum and root-mean-square (RMS) errors in Figure 7 as a function of $N^{1/2}$ roughly follow fourth-order convergence, despite the logarithmic singularity at the source.

In this example, the RMS error in the solution u at the vertices is about 7.4% for $n_\lambda = 5$.

To demonstrate the capability of the method, the second example has a non-square domain with zero Neumann boundary conditions and corners at $(0,0)$, $(6,-3)$, $(9,-1)$, $(9,6)$, and $(3,6)$, all in meters. A point source is located at $x_s = 8$ m and

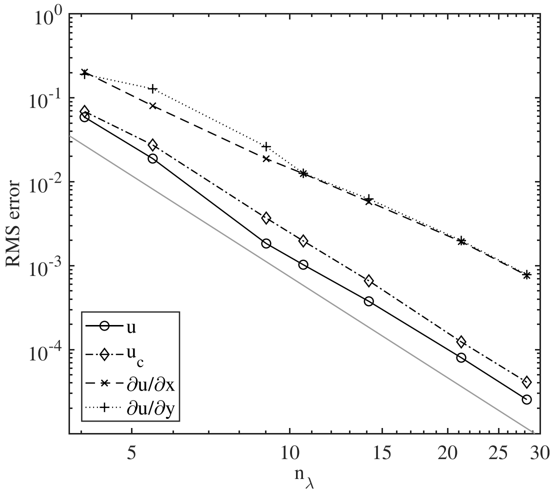

FIGURE 5. Convergence with cubic Hermite polynomials as basis functions for a 2-D test problem. The root-mean-square error as a function of n_λ, the number of elements per wavelength, shows fourth-order convergence, indicated by the gray line, for the wavefield u at the nodes and u_c at the centroids, whereas the horizontal and vertical derivatives have third-order convergence.

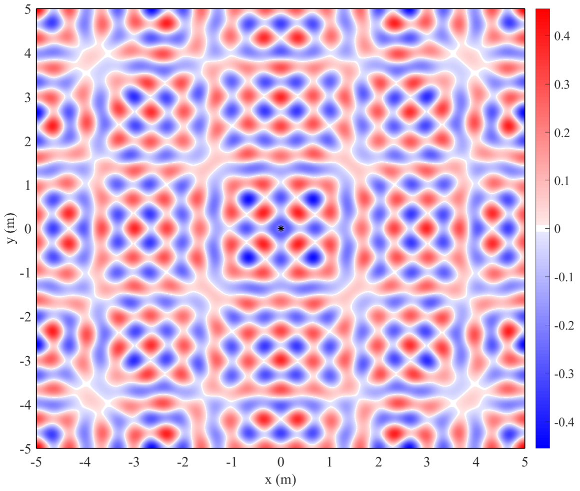

FIGURE 6. Solution for a point source at the center of a square domain with Neumann boundary conditions.

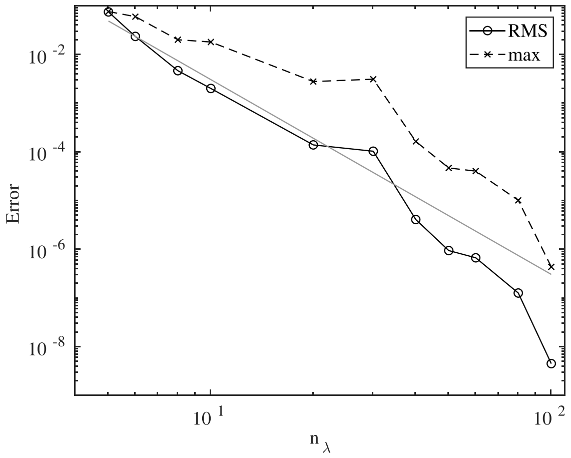

FIGURE 7. Convergence for data recorded at a line of receivers as a function of n_λ, the number of elements per wavelength. Both the maximum and root-mean-square (RMS) errors roughly follow fourth-order convergence, marked by the gray line.

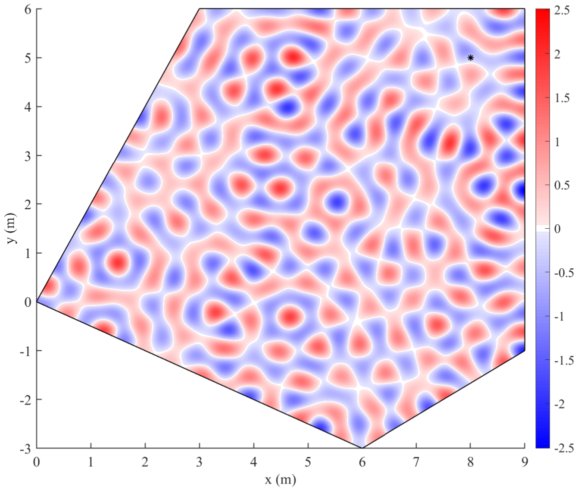

FIGURE 8. Solution for a point source at $(8, 5)$, marked by the asterisk, in a polygonal domain with Neumann boundary conditions.

$y_s = 5\,\mathrm{m}$. The wavenumber is the same as in the previous example. Figure 8 shows the solution obtained on a mesh with 694692 vertices.

4. CONCLUSIONS

We have analyzed the dispersion properties of finite elements based on cubic Hermite polynomials applied to the wave equation with a constant phase velocity. In one space dimension, the dispersion curve has a sixth-order error, whereas the eigenvector error that describes the cross talk with the spurious mode has an error of order six for the wavefield and of order four for the wavefield gradient. This results in an overall order four in terms of the element size, consistent with the classic finite-element error estimate. The accuracy at higher wavenumbers is reasonable up to a value somewhat below the Nyquist limit for the scalar case, implying a bit more than one element or two degrees of freedom per wavelength.

Dispersion analysis in two space dimensions on a regular structured mesh consisting of squares divided into two triangles shows the same orders of accuracy, 6 for the dispersion curve and 4 for the eigenvectors. 2-D numerical tests on two simple problems in the time or frequency domain show fourth-order spatial accuracy for the wavefield and one order less for its gradient. Reasonable results can be obtained already for 5 elements per wavelength.

ACKNOWLEDGEMENT

The second author was funded by the Industrial Partnership Programme 'Computational sciences for energy research' of the Netherlands Organisation for Scientific Research (NWO) and Shell Global Solutions International B.V.

APPENDIX A. DISTORTED 2-D CASE

The Fourier analysis for the homogeneous periodic problem with squares divided into two triangles can be generalized to parallelepipeds. The first triangle then has vertices $(0,0)$, $(h,0)$, (ht_x,hs_y) and the second $(h,0)$, $(h+ht_x,hs_y)$, (ht_x,hs_y), with a relative translation t_x and positive relative height s_y. Squares are recovered for $t_x=0$ and $s_y=1$.

The scaled wavenumbers ξ and η have to be redefined. From $k_xx+k_yy=j_1\xi+j_2\eta$ for vertices $y_{j_1,j_2}=j_2hs_y$ and $x_{j_1,j_2}=j_1h+j_2ht_x$, we obtain $\xi=k_xh$ and $\eta=k_yhs_y+\xi t_x$.

The 5 eigenvalues for zero wavenumbers become $(42/s_y^2)\tilde{\lambda}_{0,k}$ with

$$\tilde{\lambda}_{0,1}=0,\quad \tilde{\lambda}_{0,2}=1,\quad \tilde{\lambda}_{0,3}=t_x^2+s_y^2,$$

$$\tilde{\lambda}_{0,4}=(1-t_x)^2+s_y^2,\ \tilde{\lambda}_{0,5}=\frac{20}{9}\left[1-t_x(1-t_x)+s_y^2\right] \tag{A1}$$

The smallest eigenvalue is $\lambda_1\simeq[\xi^2+((\eta-\xi t_x)/s_y)^2]$ with an error of order 8 in ξ and η. Specifically,

$$\lambda_1\simeq\left[\xi^2+((\eta-\xi t_x)/s_y)^2\right] + \frac{\sum_{k=0}^{8}c_k\xi^{8-k}\eta^k}{2540160\,p[p-1+q(1-q)]}, \tag{A2}$$

where $p=1-t_x(1-t_x)+s_y^2$, $q=1-t_x$ and

$$\begin{aligned}
c_0 &= 3(p-q)(25p+3q),\\
c_1 &= 24(p-q)(5-15p-4q),\\
c_2 &= -4[100+p(5-261p+126q)-q(205-109q)],\\
c_3 &= 24[50-p(15+78p-29q)-q(94-45q)],\\
c_4 &= -2[770-p(245+1098p-270q)-5q(313-154q)],\\
c_5 &= 24[45-p(25+63p-9q)-q(94-50q)],\\
c_6 &= -4[109-p(139+126p)-5q(41-20q)],\\
c_7 &= 24(4-14p-5q),\\
c_8 &= 3(-3+28p),
\end{aligned}$$

The corresponding exact eigenvector is the same as in Equation (24). Its approximation has a leading error term $(0,0,0,-a,-a)^\intercal$, now with

$$a=\frac{19\xi^4-68\xi^3\eta+96\xi^2\eta^2-56\xi\eta^3+28\eta^4+3b/p}{13608}, \tag{A3}$$

where

$$\begin{aligned}
b \;=\;& (\xi^2-\eta^2)(3\xi^2+3\eta^2+4\xi\eta)\\
&+t_x\xi(2\eta-\xi)\left[3\xi^2+10\eta(\eta-\xi)\right].
\end{aligned} \tag{A4}$$

APPENDIX B. EXACT SOLUTION

Consider a two-dimensional square box $\Omega=[-L,L]^2$ with a delta function source at the center and Neumann boundary conditions all around. Helmholtz's equation

$$-k^2u-\Delta u=\delta(x)\delta(y),\quad k=\frac{\omega}{c}, \tag{B1}$$

can be solved with eigenfunctions $\phi_{m_1,m_2}(x,y) = \phi_{m_1}(x)\phi_{m_2}(y)$, where $\phi_m = c_m\cos(\pi mx/L)$ for $m=0,1,\ldots$. Given the orthogonality relation

$$\int_{-L}^{L}\phi_m\phi_n x=\delta_{m,n}(1+\delta_{m,0})Lc_m^2, \tag{B2}$$

substitution of

$$u(x,y)=\sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty}a_{m_1,m_2}\phi_{m_1,m_2}(x,y) \tag{B3}$$

into (B1) and integration against $\phi_{n_1,n_2}(x,y)$ over the domain leads to

$$\begin{aligned}
a_{m_1,m_2} \;=\;& \Big\{(1+\delta_{m_1,0})(1+\delta_{m_2,0})\\
&\left[\pi^2(m_1^2+m_2^2)-(kL)^2\right]\Big\}^{-1}.
\end{aligned} \tag{B4}$$

The implicit assumption is that no resonances occur for the chosen parameters k and L. For receivers on the line $y=0$, the summation over m_2 results in

$$\begin{aligned}
u(x,0) &= \sum_{m=0}^{\infty}\frac{\cos(\pi mx/L)}{2q\tanh(q)(1+\delta_{m,0})},\\
q &= \sqrt{(m\pi)^2-(kL)^2},
\end{aligned} \tag{B5}$$

or

$$\begin{aligned}
u(x,0) &= -\sum_{m=0}^{\infty}\frac{\cos(\pi mx/L)}{2s\tan(s)(1+\delta_{m,0})},\\
s &= \sqrt{(kL)^2-(m\pi)^2}.
\end{aligned} \tag{B6}$$

Converge is slow for large m, but we can subtract

$$\sum_{m=1}^{\infty}\frac{\cos(\pi mx/L)}{2m\pi}=-\frac{1}{4\pi}\log\left[2\left(1-\cos(\pi x/L)\right)\right]. \tag{B7}$$

This leaves terms of $O(m^{-3})$ instead of $O(m^{-1})$ in the summation, still requiring a fairly large range for m, up to the order of 10^5, and better summed from high m to low.

REFERENCES

[1] Mulder, W. A., "A comparison between higher-order finite elements and finite differences for solving the wave equation," in *Proceedings of the Second ECCOMAS Conference on Numerical Methods in Engineering*, 344–350, 1996.

[2] Zhebel, E., S. Minisini, A. Kononov, and W. A. Mulder, "A comparison of continuous mass-lumped finite elements with finite differences for 3-D wave propagation," *Geophysical Prospecting*, Vol. 62, No. 5, 1111–1125, 2014.

[3] Geevers, S., W. A. Mulder, and J. J. W. van der Vegt, "Dispersion properties of explicit finite element methods for wave propagation modelling on tetrahedral meshes," *Journal of Scientific Computing*, Vol. 77, No. 1, 372–396, Oct. 2018.

[4] Geevers, S., W. A. Mulder, and J. J. W. van der Vegt, "New higher-order mass-lumped tetrahedral elements for wave propagation modelling," *SIAM Journal on Scientific Computing*, Vol. 40, No. 5, A2830–A2857, 2018.

[5] Geevers, S., W. A. Mulder, and J. J. W. van der Vegt, "Efficient quadrature rules for computing the stiffness matrices of mass-lumped tetrahedral elements for linear wave problems," *SIAM Journal on Scientific Computing*, Vol. 41, No. 2, A1041–A1065, 2019.

[6] Mulder, W. A., "Performance of old and new mass-lumped triangular finite elements for wavefield modelling," *Geophysical Prospecting*, Vol. 72, No. 3, 885–896, 2024.

[7] Mulder, W. A., "Spurious modes in finite-element discretizations of the wave equation may not be all that bad," *Applied Numerical Mathematics*, Vol. 30, No. 4, 425–445, 1999.

[8] Ciarlet, P. G. and P.-A. Raviart, "General Lagrange and Hermite interpolation in R^n with applications to finite element methods," *Archive for Rational Mechanics and Analysis*, Vol. 46, No. 3, 177–199, Jan. 1972.

[9] Tordjman, N., "Eléments finis d'ordre élevé avec condensation de masse pour l'équation des ondes," Ph.D. dissertation, L'Université Paris IX Dauphine, France, 1995.

[10] Cohen, G., P. Joly, J. E. Roberts, and N. Tordjman, "Higher order triangular finite elements with mass lumping for the wave equation," *SIAM Journal on Numerical Analysis*, Vol. 38, No. 6, 2047–2078, 2001.

[11] Chin-Joe-Kong, M. J. S., W. A. Mulder, and M. van Veldhuizen, "Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation," *Journal of Engineering Mathematics*, Vol. 35, 405–426, 1999.

[12] Cohen, G., P. Joly, J. E. Roberts, and N. Tordjman, "Higher order triangular finite elements with mass lumping for the wave equation," *SIAM Journal on Numerical Analysis*, Vol. 38, No. 6, 2047–2078, 2001.

[13] Mulder, W. A., "Higher-order mass-lumped finite elements for the wave equation," *Journal of Computational Acoustics*, Vol. 9, No. 2, 671–680, 2001.

[14] Mulder, W. A., "New triangular mass-lumped finite elements of degree six for wave propagation," *Progress In Electromagnetics Research*, Vol. 141, 671–692, 2013.

[15] Cui, T., W. Leng, D. Lin, S. Ma, and L. Zhang, "High order mass-lumping finite elements on simplexes," *Numerical Mathematics: Theory, Methods and Applications*, Vol. 10, No. 2, 331–350, 2017.

[16] Liu, Y., J. Teng, T. Xu, and J. Badal, "Higher-order triangular spectral element method with optimized cubature points for seismic wavefield modeling," *Journal of Computational Physics*, Vol. 336, 458–480, 2017.

[17] Mulder, W. A., "More continuous mass-lumped triangular finite elements," *Journal of Scientific Computing*, Vol. 92, No. 2, 38, 2022.

[18] Felippa, C. A., "Customizing high performance elements by Fourier methods," in *Trends in Computational Structural Mechanics*, 283–296, W. A. Wall, K.-U. Bletzinger, and K. Schweizerhof (Eds.), CIMNE, Barcelona, Spain, 2001.

[19] Park, K. C. and D. L. Flaggs, "A Fourier analysis of spurious mechanisms and locking in the finite element method," *Computer Methods in Applied Mechanics and Engineering*, Vol. 46, No. 1, 65–81, 1984.

[20] Mullen, R. and T. Belytschko, "Dispersion analysis of finite element semidiscretizations of the two-dimensional wave equation," *International Journal for Numerical Methods in Engineering*, Vol. 18, No. 1, 11–29, 1982.

[21] Cottrell, J. A., A. Reali, Y. Bazilevs, and T. J. R. Hughes, "Isogeometric analysis of structural vibrations," *Computer Methods in Applied Mechanics and Engineering*, Vol. 195, No. 41, 5257–5296, 2006.

[22] Boucher, C. R., Z. Li, C. I. Ahheng, J. D. Albrecht, and L. R. Ram-Mohan, "Hermite finite elements for high accuracy electromagnetic field calculations: A case study of homogeneous and inhomogeneous waveguides," *Journal of Applied Physics*, Vol. 119, No. 14, 143106, 2016.

[23] Shamasundar, R., "Finite element methods for seismic imaging: Cost reduction through mass matrix preconditioning by defect correction," Ph.D. dissertation, Delft University of Technology, The Netherlands, 2019.

[24] Strang, G. and G. J. Fix, *An Analysis of the Finite Element Method*, Series in Automatic Computation, XIV, Prentice-Hall, Inc., Englewood Cliffs, NY, 1973.

[25] Davis, T. A., *Direct Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, 2006.

[26] Wu, J.-Y. and R. Lee, "The advantages of triangular and tetrahedral edge elements for electromagnetic modeling with the finite-element method," *IEEE Transactions on Antennas and Propagation*, Vol. 45, No. 9, 1431–1437, 1997.

[27] Mulder, W. A., E. Zhebel, and S. Minisini, "Time-stepping stability of continuous and discontinuous finite-element methods for 3-D wave propagation," *Geophysical Journal International*, Vol. 196, No. 2, 1123–1133, 2014.

[28] Engwirda, D., "Locally optimal Delaunay-refinement and optimisation-based mesh generation," Ph.D. dissertation, The University of Sydney, Australia, 2014.