Errors Due to Spatial Discretization and Numerical Precision in the Finite-Element Method

Waymond R. Scott, Jr.

Abstract—**The effects of the spatial discretization and the numerical precision on a plane wave propagating through a finite-element mesh are investigated in this work. The spatial discretization results in dispersion in the amplitude and the phase of the wave and in a non-uniform rate of convergence within an element. The finite precision in the calculations used in a finite-element code results in degraded accuracy. These errors are investigated as a function of the node density, the order of the elements, and the precision of the calculations used in the finite element code. The errors for first- through eighth-order elements are investigated both analytically and numerically.**

I. Introduction

The errors introduced by the spatial discretization of the finite-element (FE) method for the scalar Helmholtz equation are investigated in this work. It is widely understood that a plane wave will propagate along a uniform FE mesh; however, it propagates at the wrong velocity, yielding a progressive phase error in the FE solution. This phase dispersion is particularly troublesome for electrically large problems, since it is cumulative and builds up to larger and larger values the farther the wave propagates [1]–[3]. In many types of problems, the phase dispersion is the dominant source of error. The phase dispersion can be reduced by either increasing the node density

Manuscript received February 10, 1994; revised June 13, 1994.

The author is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA.

IEEE Log Number 9406394.

of the mesh or by increasing the order of the elements. The magnitude of this error must be known in order to make an informed decision on the appropriate node density and order of element to use. The phase dispersion is studied in this work for first- through eighth-order, one-dimensional elements as a function of the node density. The phase dispersion is shown to decrease rapidly with the increasing order of the elements. The one-dimensional results are indicative of what to expect in the two- and three-dimensional cases. The two- and three-dimensional cases will yield the same error as the one-dimensional case when the plane wave propagates in the direction of one of the coordinate axis of a uniform mesh of quadrilateral or hexagonal elements. This error has been studied by several investigators for first- and second-order, one-dimensional elements and for first-order, two-dimensional elements [4]–[7].

Other errors are also introduced by the spatial discretization. At low node densities, the plane wave will attenuate as well as propagate at the wrong velocity though the mesh, and the interior nodes of the elements will exhibit much larger errors than the end nodes. These errors are also investigated in this work.

As the node density is increased, the phase dispersion will decrease until it reaches a minimum value and then will begin to increase because of round-off errors that exist due to the finite precision of the computer used to perform the calculations. The effect of the round-off errors is investigated as a function of the node density and the order of the elements.

II. Dispersion Relations

Consider a plane wave in a linear, homogeneous, and isotropic material that is propagating in the z direction and is polarized so that the electric field is x directed:[1]

$$\vec{E}(x,y,z) = E(z)\hat{x}$$
$$\vec{H}(x,y,z) = \frac{1}{\eta}E(z)\hat{y} \tag{1}$$

where an $e^{j\omega t}$ time dependence is assumed, and $\eta = \sqrt{\mu/\varepsilon}$ is the characteristic impedance of the material. The electric field is a solution to the one-dimensional Helmholtz equation:

$$\frac{d^2E}{dz^2} + k^2E = 0 \tag{2}$$

where $k = \omega\sqrt{\mu\varepsilon}$ is the wavenumber. The exact solution is a plane wave with wavenumber k:

$$E(z) = E_0e^{-jkz} \tag{3}$$

where E_0 is a constant. In order to solve the Helmholtz equation using the FE method, the region of interest is divided up into elements. Consider an infinite mesh of n^{th}-order elements with equally spaced nodes; a portion of an infinite mesh of fourth-order elements is shown in Fig. 1(a). The spacing between the nodes is h, and the length of the elements is $\ell = nh$. Using the standard scalar finite-element formulation and the standard higher-order basis functions, the FE method yields an approximate solution which is a plane wave with the numerical wavenumber $\tilde{k}$:

$$\tilde{E}(z + m\ell) = \tilde{E}(z)e^{-j\tilde{k}m\ell} \tag{4}$$

where m is an integer and ℓ is the length of an element. Note that the field at a certain relative position within an element is related to the field at the corresponding position within another element by the wavenumber $\tilde{k}$; however, the field at a certain relative position within a element is not as simply related to the field at a different relative position within an element. This form of the approximate solution is used below with the FE formulation to derive expressions for $\tilde{k}$.

[1] Only forward traveling waves are considered in this work: backward traveling waves behave in the same manner. In a region where waves are traveling in both directions, the results for the forward traveling wave can be applied separately to each wave.

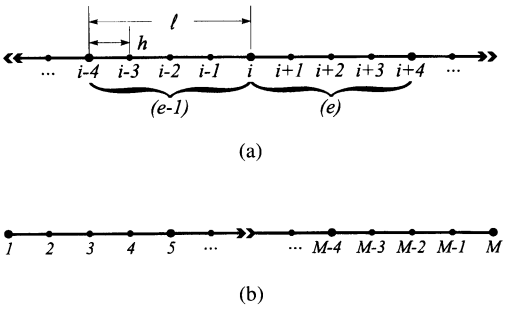


Fig. 1. Sketch of (a) a portion of a an infinite mesh of fourth-order elements, and (b) the mesh of fourth-order elements used in the numerical computations.

Using Galerkin's formulation, residual integrals are computed for each node in the mesh. The contribution of the fourth-order element (e) to the residual integral for node p is

$$R_p^{(e)} = -\int \alpha_p(z)\left[\frac{d^2\tilde{E}(z)}{dz^2} + k^2\tilde{E}(z)\right]dz \tag{5}$$

where $p = i, i+1, i+2, i+3$, and $i+4$; and the α's are the usual finite-element basis functions [8]. The integral is placed in the "weak form" and the approximate solution $\tilde{E}$ is expanded in terms of the basis functions:

$$\tilde{E}(z) = \sum_q E_q\alpha_q(z) \tag{6}$$

where $q = i,\ i+1,\ i+2,\ i+3$, and $i+4$. The residual integrals are then used to generate a set of linear equations for each element, and these equations are assembled to form the global set of equations. The contribution to the residual equations from the fourth-order element (e) is

$$\begin{bmatrix} R_i^{(e)} \\ 0 \\ 0 \\ 0 \\ R_{i+4}^{(e)} \end{bmatrix} = \frac{1}{945\ell} \cdot \begin{bmatrix} 4925 & -6848 & 3048 & -1472 & 347 \\ -6848 & 16640 & -14208 & 5888 & -1472 \\ 3048 & -14208 & 22320 & -14208 & 3048 \\ -1472 & 5888 & -14208 & 16640 & -6848 \\ 347 & -1472 & 3048 & -6848 & 4929 \end{bmatrix} \begin{bmatrix} E_i \\ E_{i+1} \\ E_{i+2} \\ E_{i+3} \\ E_{i+4} \end{bmatrix} - \frac{k^2\ell}{5670} \begin{bmatrix} 292 & 296 & -174 & 56 & -29 \\ 296 & 1792 & -384 & 256 & 56 \\ -174 & -384 & 1872 & -384 & -174 \\ 56 & 256 & -384 & 1792 & 296 \\ -29 & 56 & -174 & 296 & 292 \end{bmatrix} \begin{bmatrix} E_i \\ E_{i+1} \\ E_{i+2} \\ E_{i+3} \\ E_{i+4} \end{bmatrix}. \tag{7}$$

The residual equations for the interior nodes of an element only have a contribution from that element, so the equations for the interior nodes only involve the value of the field at the nodes in that element. Thus, the values of the field at the interior nodes can be expressed in terms of the values at the end nodes (interelement nodes) of the element. The three equations for the interior nodes of element (e) are used to solve for E_{i+1}, E_{i+2}, and E_{i+3} in terms of E_i and E_{i+4}:

$$E_{i+1} = aE_i + bE_{i+4}$$
$$E_{i+2} = c(E_i + E_{i+4})$$
$$E_{i+3} = bE_i + aE_{i+4} \tag{8}$$

TABLE I
NUMERICAL WAVENUMBER $\tilde{k}$ AND THE VALUES OF λ/h FOR WHICH $\mathrm{Im}(\tilde{k}) \neq 0$ FOR FIRST- THROUGH FOURTH-ORDER ELEMENTS

Order	Numerical Wavenumber $\tilde{k}$	Values of λ/h for which $\tilde{\alpha} = -\mathrm{Im}(\tilde{k}) \neq 0$.
1	$\frac{1}{h}\cos^{-1}\left[\frac{6-2(kh)^2}{6+(kh)^2}\right]$	None*
2	$\frac{1}{2h}\cos^{-1}\left[\frac{15-26(kh)^2+3(kh)^4}{15+4(kh)^2+(kh)^4}\right]$	$3.62760<\lambda/h<3.97384$
3	$\frac{1}{3h}\cos^{-1}\left[\frac{2800-11520(kh)^2+4860(kh)^4-324(kh)^6}{2800+1080(kh)^2+270(kh)^4+81(kh)^6}\right]$	$2.43347<\lambda/h<2.90855$ or $5.96075<\lambda/h<5.99833$
4	$\frac{1}{4h}\cos^{-1}\left[\frac{19845-148680(kh)^2+134064(kh)^4-28800(kh)^6+1280(kh)^8}{19845+10080(kh)^2+3024(kh)^4+768(kh)^6+256(kh)^8}\right]$	$2.00000^*<\lambda/h<2.48692$, $3.87807<\lambda/h<3.98557$, or $7.99777<\lambda/h<7.99994$

* Only node densities greater than the Nyquist rate ($\lambda/h>2$) are considered in this work.

where

$$a = \frac{1}{16}\left[\frac{-15876+15750(kh)^2-945(kh)^4-312(kh)^6}{-1323+2856(kh)^2-1232(kh)^4+128(kh)^6}\right]$$
$$b = \frac{1}{16}\left[\frac{-5292-1806(kh)^2-497(kh)^4+8(kh)^6}{-1323+2856(kh)^2-1232(kh)^4+128(kh)^6}\right]$$
$$c = \frac{63+14(kh)^2+2(kh)^4}{126-224(kh)^2+32(kh)^4}. \tag{9}$$

The residual integral for node i has contributions from elements (e) and $(e-1)$ and yields the linear equation:

$$\begin{aligned}&[347E_{i-4}-1472E_{i-3}+3048E_{i-2}-6848E_{i-1}+9858E_i\\&\quad-6848E_{i+1}+3048E_{i+2}-1472E_{i+3}+347E_{i+4}]\\&\quad-\frac{8(kh)^2}{3}[-29E_{i-4}+56E_{i-3}-174E_{i-2}+296E_{i-1}+582E_i\\&\quad+296E_{i+1}-174E_{i+2}+56E_{i+3}-29E_{i+4}]=0\end{aligned} \tag{10}$$

where the residual contribution from element $(e-1)$ is computed by shifting the indices in the residual contribution from element (e). Equations (8) and (9) for the interior nodes of element (e) and the corresponding equations for element $(e-1)$ are used to eliminate the values of the field at the interior nodes in (10):

$$\begin{aligned}&[19845+10080(kh)^2+3024(kh)^4\\&\quad+768(kh)^6+256(kh)^8][E_{i-4}+E_{i+4}]\\&\quad+[-39690+297360(kh)^2-268128(kh)^4\\&\quad+57600(kh)^6-25600(kh)^8]E_i=0.\end{aligned} \tag{11}$$

Since we are considering an infinite uniform mesh, each of the residual equations for the interelement nodes is a shifted version of the other equations. Thus, only one of the residual equations is needed to determine $\tilde{k}$. Using (4), E_{i-4} and E_{i+4} can be expressed in terms of E_i:

$$\begin{aligned}E_{i-4} &= E_i e^{j\tilde{k}\ell} = E_i e^{j\tilde{k}4h}\\E_{i+4} &= E_i e^{-j\tilde{k}\ell} = E_i e^{-j\tilde{k}4h}.\end{aligned} \tag{12}$$

Then these equations are used with (11) to solve for $\tilde{k}$. The resulting expression for $\tilde{k}$ for fourth-order elements is in Table I. Expression for $\tilde{k}$ for first- through third-order elements are also included in Table I; they are derived in a similar manner to that used for the fourth-order elements.

From the expressions for $\tilde{k}$, it is seen that $\tilde{k}$ is purely real for certain values of kh and complex ($\mathrm{Im}(\tilde{k}) \neq 0$) for other values of kh. $\tilde{k}$ is purely real when the argument of the inverse cosine has a magnitude less than one and is complex when the magnitude is greater than one. Note that the length of the elements is an integer multiple of one half of a numerical wavelength when $\tilde{k}$ is complex: $\ell = m\tilde{\lambda}/2 = m\pi/\mathrm{Re}(\tilde{k})$ where m is an integer. Since $\mathrm{Re}(\tilde{k}) \approx k$, the length of the elements is also approximately an integer multiple of $\lambda/2$ when $\tilde{k}$ is complex: $\ell \approx m\lambda/2$. This is because the real part of an inverse cosine of a real number with a magnitude greater than one is an integer multiple of π: $\mathrm{Re}\ (\cos^{-1} x) = mx$ where m is an even integer when $x > 1$ and is an odd integer when $x < -1$. The values of λ/h for which $\tilde{k}$ is complex are presented in Table I. The node density is expressed here in nodes per wavelength ($\lambda/h = 2\pi/kh$) instead of kh. Note that these values are in $n-1$ contiguous regions where n is the order of the elements. The regions correspond to the lengths $\ell = m\tilde{\lambda}/2$ where $m = 1, 2, 3, \cdots, n-1$. For example, the fourth-order elements have three regions for which $\tilde{k}$ is complex: $\ell = \tilde{\lambda}/2, \tilde{\lambda}$, and $3\ \tilde{\lambda}/2$.

Since $\mathrm{Re}(\tilde{k}) \neq k$, a wave propagating along the mesh experiences phase dispersion, which when expressed as a phase error in degrees

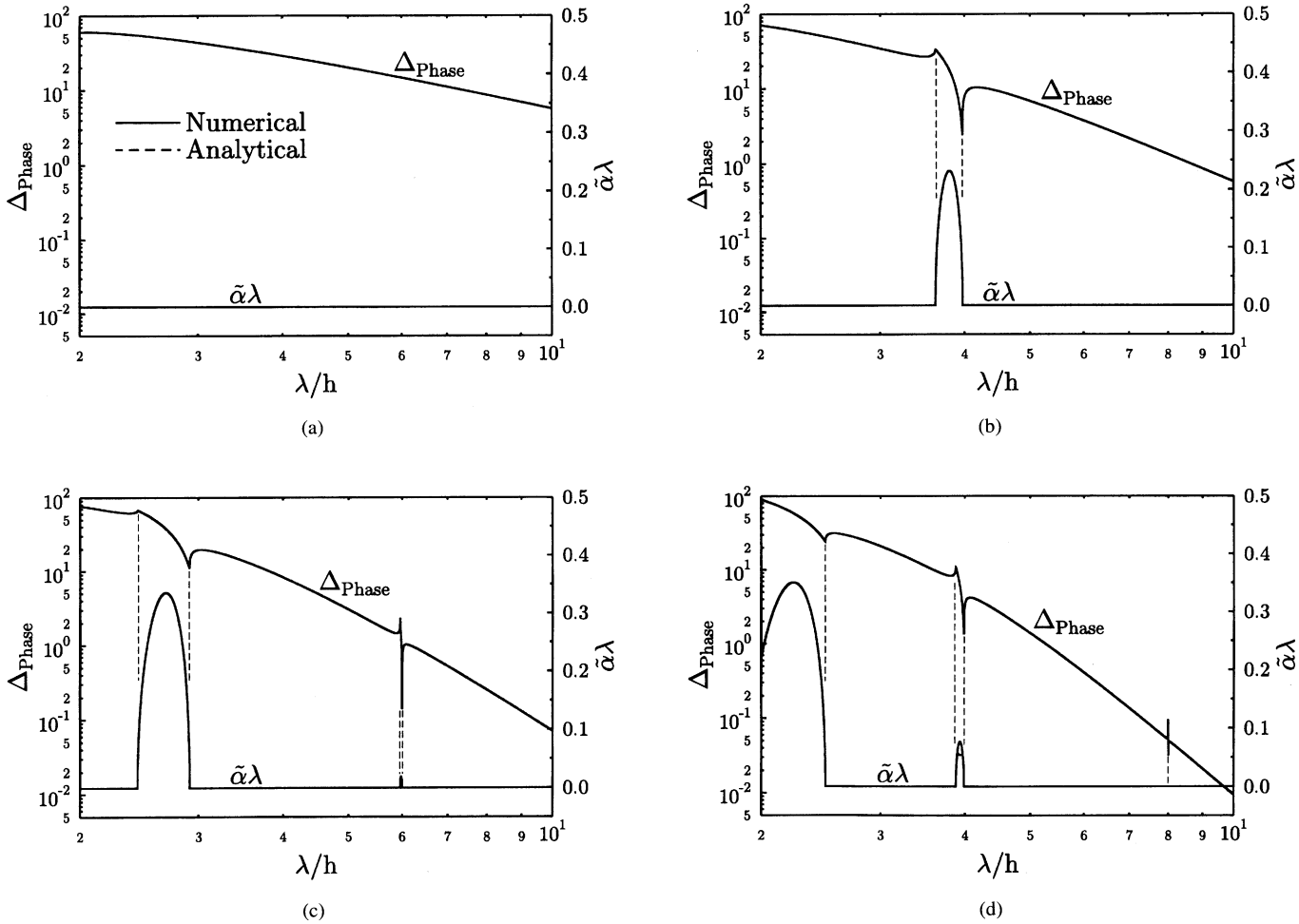


Fig. 2. Graphs of the phase error in degrees per wavelength and of the attenuation constant in nepers per wavelength as a function of the number of nodes per wavelength for (a) first-, (b) second-, (c) third-, and (d) fourth-order elements.

per wavelength is

$$\Delta_{\text{phase}} = 360\left|\frac{\text{Re}(\tilde{k}) - k}{k}\right| \tag{13}$$

When $\text{Im}(\tilde{k}) \neq 0$, a wave propagating along the mesh experiences an amplitude dispersion. The amplitude dispersion yields a exponential decay that is characterized by the attenuation constant $\tilde{\alpha} = -\text{Im}(\tilde{k})$. The attenuation constant is expressed in nepers per wavelength, $\tilde{\alpha}\lambda$.

The phase error and the attenuation constant are graphed in Fig. 2 as a function of node density for first- through fourth-order elements. The attenuation constant is seen to be non-zero for the values of λ/h specified in Table I, and the phase error is seen to change abruptly at the boundaries of these regions.

The phase error is graphed in Fig. 3 for higher node densities than those shown in Fig. 2. These analytical results are denoted by the dashed curves. The phase errors are seen to decrease monotonically with increasing node density when the elements are less that approximately one half of a wavelength long ($\ell \lesssim \lambda/2$ or $\lambda/h \gtrsim 2n$) and to decrease more rapidly for the higher-order elements. The phase errors are seen to proportional to $(h/\lambda)^{2n}$ for large values of λ/h; thus, $\tilde{k}$ converges at the rate $O[(h/\lambda)^{2n}]$ which is a superconvergent rate compared to the ordinary rate of $O[(h/\lambda)^{n+1}]$.

III. Numerical Computations and Roundoff Errors

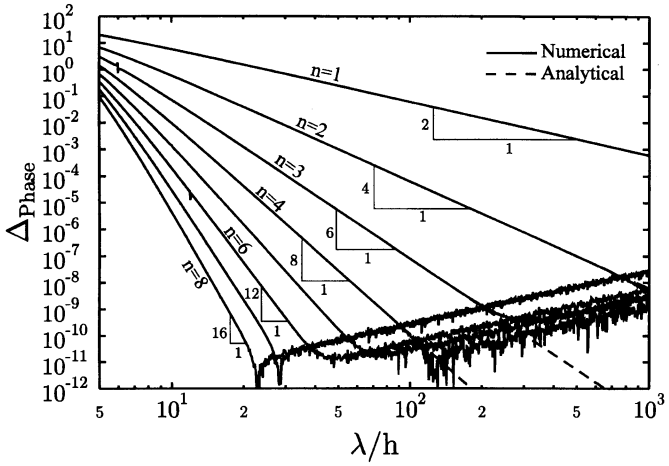
A FE code using n^{th}-order elements was written and used to numerically verify this analysis and to further investigate the errors. The mesh for the code consists of M evenly space nodes divided up in to $(M-1)/n$ elements. A mesh of fourth-order elements is shown in Fig. 1(b). A wave is injected on the left end of the mesh using an essential (Dirichlet) boundary condition. The wave is absorbed on the right end of the mesh by replacing the residual equation for the node on the right end with the constraint equation,

$$E_M = E_{M-n}e^{-j\tilde{k}\ell}. \tag{14}$$

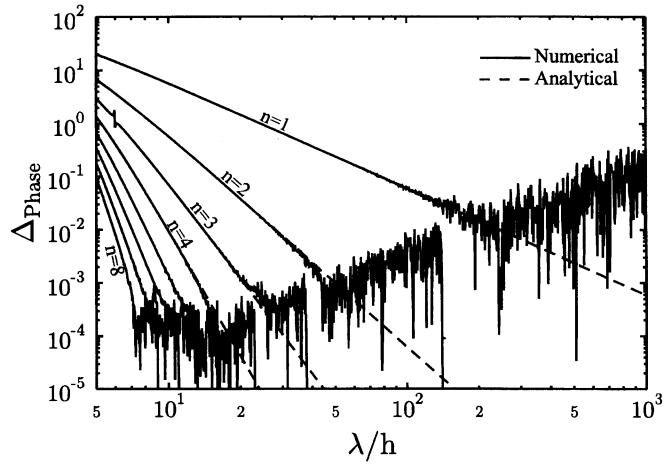
This absorbing boundary condition is exact since it relates two interelement nodes using $\tilde{k}$ from Table I. Since expressions for $\tilde{k}$ are not available when $n > 4$, k is used in (14) for these cases; then, the boundary condition is not exact.

The numerical wavenumber is determined from the field obtained with the FE code. The real part of $\tilde{k}$ is determined from the unwrapped phase of the field and the imaginary part of $\tilde{k}$ is determined from the exponential decay of the amplitude of the field:

$$\tilde{k} = \frac{j\ln(E_M/E_1)}{(M-1)h} = \frac{-\text{Phase}(E_M/E_1) + j\ln(|E_M/E_1|)}{(M-1)h}. \tag{15}$$



(a)



(b)

Fig. 3. Graphs of the phase error in degrees per wavelength as a function of the number of nodes per wavelength when (a) 64-bit arithmetic and (b) 32-bit arithmetic are used in the FE code.

The phase error and the amplitude constant from the FE code are compared to those from the analytical formulas in Fig. 2 for first- through fourth-order elements. The results are seen to be essentially identical. For this comparison, $M = 41$.

The phase error from the FE code for first- through eighth-order elements is graphed in Fig. 3(a) for 64-bit arithmetic and in Fig. 3(b) for 32-bit arithmetic.[2] The phase error from the FE code is seen to decrease until it reaches a minimum and then begins to increase; this increase is caused by the round-off errors in the calculations. For the lower node densities where the round-off errors are insignificant, the phase errors determined from the code and the analytical dispersion relations are seen to be essentially identical; however, for the higher node densities where the round-off errors are significant, the phase errors from the code and the analytical relations are seen to differ. The minimum phase error and the node density at which the minimum phase error occurs are seen to be functions of the precision of the arithmetic and the order of the elements. The minimum phase errors

[2]A longer mesh ($M = 5041$) was used for these results because the absorbing boundary condition for the fifth- through eight-order elements is not exact. The relative error introduced by the inexact boundary condition decreases when the length of the mesh is increased; this error is insignificant for a mesh of this length.

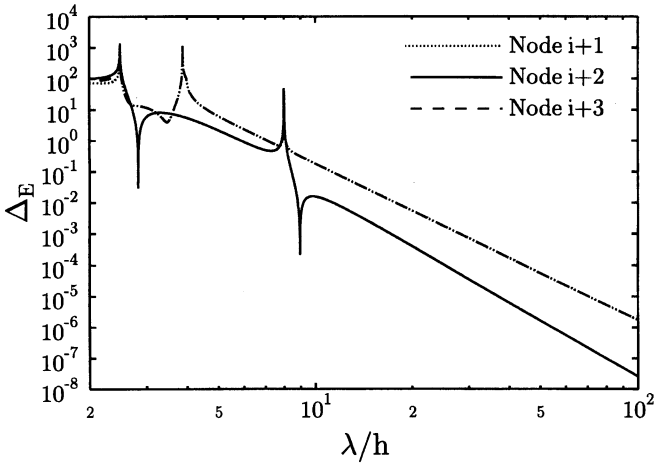


Fig. 4. Graphs of the relative difference Δ_E as a function of the number of nodes per wavelength for the three interior nodes of a fourth-order element.

are seen to be much smaller for the 64-bit arithmetic than for the 32-bit arithmetic, and the node densities at which the minimum phase error occurs are seen to decrease with the increasing order of the elements and are seen to be much smaller for the 32-bit arithmetic than for the 64-bit arithmetic. Note that the minimum phase error occurs at relatively low node densities for the highest-order elements. For example with eight-order elements, the minimum phase error occurs at approximately 7 nodes per wavelength for 32-bit arithmetic and at approximately 22 nodes per wavelength for 64 bit arithmetic.

IV. Non-Uniform Rate of Convergence

Consider a plane wave propagating along an infinite mesh of n^{th}-order elements. One would expect that the field to be

$$\overline{E}(z) = E_0 e^{-j\tilde{k}z} \tag{16}$$

where E_0 is a constant at any point in the mesh; however, the field is more complicated; it has the form given by (4). In order to investigate the behavior of the field in the interior of the elements, let the value of the field at the interelement nodes be

$$\tilde{E}(m\ell) = E_0 e^{-j\tilde{k}m\ell} \tag{17}$$

where m is a integer. The field in the interior of the element is then determined from the field at the interelement nodes by using the residual equations for the element. For example, the field at the interior nodes of a fourth-order element is given by the expression (8), and the field between the nodes is determined from (6). The relative difference between the FE field $\tilde{E}(z)$ and the expected field $\overline{E}(z)$ is defined as

$$\Delta_E = \left| \frac{\tilde{E}(z) - \overline{E}(z)}{\overline{E}(z)} \right|. \tag{18}$$

The relative difference varies within an element but it does not vary element to element; the relative difference is zero at the interelement nodes. Fig. 4 is a graph of the relative difference as a function of λ/h for the three interior nodes of a fourth-order element. The asymptotic rate of convergence is seen to be $O[(h/\lambda)^{n+2}]$ for the mid node and to be $O[(h/\lambda)^{n+1}]$ for the other two nodes. The relative differences are seen to be very large and nonlinear near regions for which $\tilde{\alpha} \neq 0$; thus, the field near these regions will be very different than that expected, (16). Similar behavior is also exhibited by the other elements with different order.

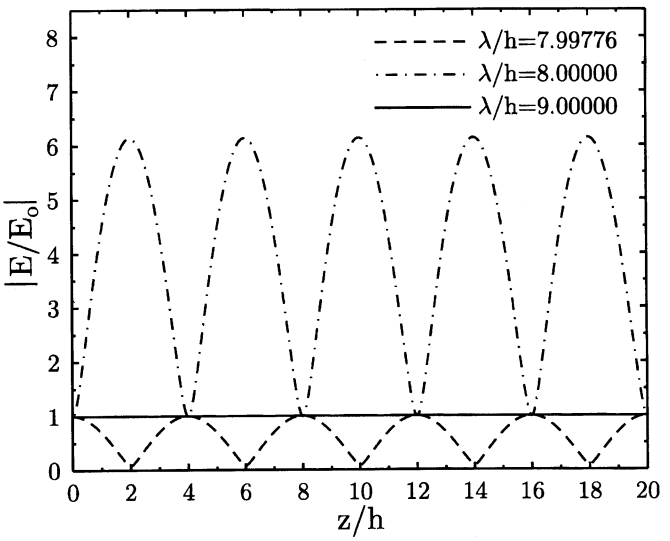


Fig. 5. Graphs of the magnitude of the field as a function of z/h for $\lambda/h = 7.99776$, 8, and 9 for fourth-order elements.

As an example, the field was calculated using the FE code for fourth-order elements with $M = 21$. Fig. 5 is a graph of the magnitude of the field as a function of z/h for three node densities: $\lambda/h = 7.99776$ and $\lambda/h = 8$ which are near a region where $\tilde{\alpha} \neq 0$, and $\lambda/h = 9$ which is not near one of these regions. One would expect these graphs to be constant since $\tilde{k}$ is real for these node densities. The graph for $\lambda/h = 9$ is seen to be relatively constant; however, the graphs for $\lambda/h = 7.99776$ and 8 are seen to have large variations. The size of the variations at the interior nodes are consistent with those predicted in Fig. 4. Note that from Fig. 4 the variations would be even larger near the other regions where $\tilde{\alpha} \neq 0$.

Clearly, one would not want to use node densities near the regions where $\tilde{\alpha} \neq 0$. Note that for the highest-order elements, these regions where $\tilde{\alpha} \neq 0$ overlap with the regions where round-off errors are significant. For example for eight-order elements, $\tilde{\alpha} \neq 0$ when $\lambda/h \approx 16/m$ with $m = 1, 2, 3, \ldots, 7$; and the round-off errors are significant when $\lambda/h \gtrsim 7$ for 32-bit arithmetic; thus, making it impractical to avoid both the regions where $\tilde{\alpha} \neq 0$ and where round-off errors are significant.

Often boundary conditions that involve a derivative of the field with respect to z at the end of an element are used to inject and absorb waves at the boundaries of a FE mesh. These boundary conditions will not perform properly when the node density is near one of the regions where $\tilde{\alpha} \neq 0$, since the derivatives with respect to z will deviate significantly from what is expected.

V. Conclusion

The numerical dispersion of a plane propagating through a FE mesh has been investigated in this work. Analytical dispersion relationships were derived and used to investigate the numerical dispersion for first- through fourth-order elements, and a FE code was written and used to calculate and investigate the numerical dispersion for first- through eighth-order elements. All of the elements are shown to exhibit a phase dispersion at all node densities, and the elements with order greater than first are shown to exhibit an amplitude dispersion $\tilde{\alpha} \neq 0$ for certain ranges of node densities. The numerical dispersion calculated analytically and numerically were shown to be essentially identical. The phase error was shown to decrease rapidly with increasing order of the elements. For example, at 10 nodes per wavelength, the phase error is 60 degrees per wavelength for first-order elements, while it is only 10^{-2} degrees per wavelength for fourth-order elements, and it is only 5×10^{-6} degrees per wavelength for eighth-order elements.

Roundoff errors were shown to limit the minimum achievable phase error. The minimum phase errors were shown to be much smaller for 64-bit arithmetic than for 32-bit arithmetic. The minimum phase error was also shown to be a function of the order of the elements.

It was also shown that the field calculated with the FE method exhibited large unexpected errors when node densities were near the regions where $\tilde{\alpha} \neq 0$. Thus, in order to obtain an accurate result, one needs to choose a node density that is not near a region where $\tilde{\alpha} \neq 0$ as well as a node density that results in a sufficiently small phase error.

References

[1] A. Bayliss, C. I. Goldstein, and E, Turkel, "On accuracy conditions for the numerical computation of waves," *J. Comput. Phys.*, vol. 59, pp. 396–404, 1995.

[2] A. F. Peterson and R. J. Baca, "Error in the finite element discretization of the scalar Helmholtz equation over electrically large regions," *IEEE Microwave and Guided Wave Lett.*, vol. 1, pp. 219–222, Aug. 1991.

[3] A. C. Cangellaris and R. Lee, "On the accuracy of numerical wave simulations based on finite methods," *J. Electromag. Wav. Appl.*, vol. 6, pp. 1635–1653, 1992.

[4] G. W. Platzman, "Some response characteristics of finite-element tidal modes," *J. Comput. Phys.*, vol. 40, pp. 36–63, 1981.

[5] R. Mullen and T. Belytschko, "Dispersion analysis of finite-element semidiscretization of the two-dimensional wave equation," *Int. J. Num. Methods Eng.*, vol. 18, pp. 11–29, 1982.

[6] D. R. Linch, K. D. Paulsen, and J. W. Strohbehn, "Finite-element solution of Maxwell's equation for hypothermia treatment planning," *J. Comput. Phys.*, vol. 58, pp. 246–269, 1985.

[7] R. Lee and A. C. Cangellaris, "A study of discretization error in the finite-element approximation of wave solutions," *IEEE Trans. Antennas Propagat.*, vol. 40, pp. 542–549, May 1992.

[8] D. S. Burnett, *Finite Element Analysis: From Concepts to Applications*. Reading, MA: Addison-Wesley, 1987.