Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 224 (2007) 1293–1306
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

The discrete origin of FETD-Newmark late time instability, and a correction scheme

Ryan A. Chilton *, Robert Lee

Electroscience Laboratory, Ohio State University, 1320 Kinner Road, Columbus, OH 43212, United States

Received 25 July 2006; received in revised form 25 October 2006; accepted 27 November 2006
Available online 1 February 2007

Abstract

The finite element time domain (FETD) method is commonly used for transient simulation of electromagnetic wave phenomena. Most practitioners consider FETD, when time integrated using the Newmark-Beta method, to be unconditionally stable when $\beta \geqslant 0.25$. Unlike the finite difference time domain (FDTD) ''courant criterion'', FETD-Newmark has no limiting timestep above which the method exhibits exponential growth. However, herein the stability properties of FETD-Newmark will be rigorously investigated by deducing the Jordan canonical form of the FETD-Newmark amplification matrix, and it will be demonstrated that the method does exhibit linear growth for certain field configurations. These modes are none other than the pure-gradient fields associated with ''late time instability''. Though many practical simulations are of short duration and will never observe a linearly growing gradient solution, it can be problematic for simulations which require long time periods to be integrated. A correction scheme for eliminating this late time instability shall be suggested, and numerical results will verify its performance.

Keywords: Computational electromagnetics; Finite elements; Transient simulation; Numerical stability; Von-Neumann analysis; Jordan form

1. Introduction

The finite element time domain method [1], iterated using the Newmark method (FETD-Newmark), is useful for analyzing transient electromagnetic wave interactions when geometry conforming discretization is required. A single update can be characterized as a matrix–vector multiply of the state vector by an amplification matrix **A**. If it has any eigenvalues outside the complex plane unit circle, then the state vector may grow exponentially as repeated updates are applied. That is, unit circle boundedness is a necessary criterion for stability. Investigators have often presumed that FETD-Newmark is unconditionally stable by demonstrating that, for a non-active domain, the amplification matrix has all eigenvalues on or within the unit circle [2,3]. Eigenvalues are strictly on the unit circle for non-dissapative domains. Unfortunately, unit circle boundedness

* Corresponding author. Tel.: +1 614 282 8115.
E-mail address: rac@esl.eng.ohio-state.edu (R.A. Chilton).

0021-9991/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2006.11.021

is not a sufficient criterion for unconditional stability. If the amplification matrix is non-diagonalizable, there may be other stability outcomes: in addition to "no growth" (stability) and exponential growth, a non-diagonalizable matrix can permit polynomial growth [4]. It will be shown herein that **A** is non-diagonalizable, lacks a full set of linearly independent eigenvectors and possesses a non-diagonal Jordan canonical form which permits linear growth for pure gradient fields. However, knowledge of the Jordan form of **A** will naturally lead to a correction algorithm for removing gradient modes and preventing instability. The correction algorithm will be demonstrated with a handful of numerical experiments.

2. The transformed semi-discrete wave equation

Consider the vector wave equation (40) for the electric field in a lossless, source free domain Ω.

$$\nabla \times \left(\mu_r^{-1} \cdot \nabla \times \vec{E}\right) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left(\varepsilon_r \vec{E}\right) = 0. \tag{1}$$

When the electric field is expanded with $\mathscr{E}$ distinct Whitney-1 edge elements $\vec{W}_i$ [5,6] and the residual is weighted through Galerkin testing, a spatially-discrete time-continuous wave equation emerges

$$\mathbf{S}\mathbf{e} + \mathbf{T}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{e} = 0, \tag{2}$$

$$\mathbf{S}_{ij} = \int_\Omega \left(\nabla \times \vec{W}_i\right) \cdot \mu_r^{-1}\left(\nabla \times \vec{W}_j\right) \mathrm{d}v, \quad i,j \in [1..\mathscr{E}], \tag{3}$$

$$\mathbf{T}_{ij} = \int_\Omega \vec{W}_i \cdot \varepsilon_r \cdot \vec{W}_j \, \mathrm{d}v, \quad i,j \in [1..\mathscr{E}]. \tag{4}$$

Here, $\mathscr{E}$ denotes the number of free edges in the FE mesh. Edges which lie on perfect electric conductor (PEC) surfaces are not free, they cannot support a tangential electric field.

The boundary integral which emerges through integration by parts has been discarded, implying the surface $\partial\Omega$ which bounds Ω is either PEC or PMC (perfect magnetic conductor). Constitutive parameters are assumed time-invariant and non-dispersive, but are possibly inhomogeneous. The mass matrix **T** (4) is symmetric and possesses a complete orthonormal set of eigenvectors such that $\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}_T\mathbf{Q}^{\mathrm{T}}$ and $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$. It is also positive definite, which permits a rigorous definition of its "half power". Consistently pick the positive branch of the square root function to produce a matrix $\mathbf{T}^{\frac{1}{2}}$ which is symmetric positive definite (SPD), and therefore invertible.

$$\mathbf{T}^{\frac{1}{2}} = \mathbf{Q}(\mathbf{\Lambda}_T)^{\frac{1}{2}}\mathbf{Q}^{\mathrm{T}}. \tag{5}$$

Perform a change of variables through the matrix $\mathbf{T}^{\frac{1}{2}}$, creating a new field quantity $\tilde{\mathbf{e}} = \mathbf{T}^{\frac{1}{2}}\mathbf{e}$. Not surprisingly, $\tilde{\mathbf{e}}$ satisfies its own semi-discrete wave equation involving a transformed stiffness matrix $\tilde{\mathbf{S}} = (\mathbf{T}^{-\frac{1}{2}})\mathbf{S}(\mathbf{T}^{-\frac{1}{2}})$ and transformed mass matrix $\tilde{\mathbf{T}} = (\mathbf{T}^{-\frac{1}{2}})\mathbf{T}(\mathbf{T}^{-\frac{1}{2}}) = \mathbf{I}$. Throughout this document the tilde symbol will be used to denote quantities associated with this transformed wave equation (6).

$$\tilde{\mathbf{S}}\tilde{\mathbf{e}} + \mathbf{I}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\tilde{\mathbf{e}} = 0. \tag{6}$$

The stiffness matrix $\tilde{\mathbf{S}}$ is positive semi-definite. Partition its eigenvectors into two disjoint subsets: electrostatic eigenvectors $\tilde{\mathbf{s}}_n$ which possess zero discrete curl (these reside in the nullspace of $\tilde{\mathbf{S}}$, and are paired to zero eigenvalues) and electrodynamic eigenvectors $\tilde{\mathbf{d}}_m$ (paired with nonzero eigenvalues denoted by λ_m). Here, $\tilde{\mathbf{s}}$ denotes "static" $\tilde{\mathbf{d}}$ denotes "dynamic". Since $\tilde{\mathbf{S}}$ is symmetric, all eigenvectors can be picked orthonormal to one another (even the electrostatic subset of vectors, which all share a common zero eigenvalue) [7].

$$\tilde{\mathbf{s}}_\mathbf{i}^{\mathrm{T}}\tilde{\mathbf{s}}_\mathbf{j} = \delta_{ij}, \quad \tilde{\mathbf{d}}_\mathbf{i}^{\mathrm{T}}\tilde{\mathbf{d}}_\mathbf{j} = \delta_{ij}, \quad \tilde{\mathbf{s}}_\mathbf{n}^{\mathrm{T}}\tilde{\mathbf{d}}_\mathbf{m} = 0. \tag{7}$$

The number of electrostatic or gradient eigenvectors is $\mathscr{N} - 1$, where $\mathscr{N}$ is the number of nodal electric scalar potential degrees-of-freedom in the mesh (equipotential nodes connected by PEC surfaces are not independent). Physically, samples of scalar electric potential at each free node provide a mechanism for generating any electrostatic field via the discrete gradient operator (an idea to be exploited later). However, one node

is presumed a reference or ground potential against which all other potentials are measured, reducing the rank of electrostatic eigenvectors by one.

The spatially discrete system (6) is integrated in time using the Newmark-Beta method [8], with $\beta = 0.25$. This process requires an implicit matrix solution every timestep, where the "next" field state $\tilde{\mathbf{e}}_{\mathbf{k+1}}$ is determined from two previous field states, $\tilde{\mathbf{e}}_{\mathbf{k}}$ and $\tilde{\mathbf{e}}_{\mathbf{k-1}}$

$$\tilde{\mathbf{B}}_{\mathbf{k+1}}\tilde{\mathbf{e}}_{\mathbf{k+1}} = -\tilde{\mathbf{B}}_{\mathbf{k}}\tilde{\mathbf{e}}_{\mathbf{k}} - \tilde{\mathbf{B}}_{\mathbf{k-1}}\tilde{\mathbf{e}}_{\mathbf{k-1}}, \tag{8}$$

where the $\tilde{\mathbf{B}}_{\mathbf{i}}$'s are $\mathscr{E} \times \mathscr{E}$ matrices defined by

$$\tilde{\mathbf{B}}_{\mathbf{k+1}} = \tilde{\mathbf{B}}_{\mathbf{k-1}} = \frac{1}{4}(c\Delta t)^2\tilde{\mathbf{S}} + \mathbf{I}, \tag{9}$$

$$\tilde{\mathbf{B}}_{\mathbf{k}} = \frac{1}{2}(c\Delta t)^2\tilde{\mathbf{S}} - 2\mathbf{I}. \tag{10}$$

Equivalently to (8), the entire scheme can be characterized as one matrix–vector multiply, with an amplification matrix denoted $\tilde{\mathbf{A}}$.

$$\begin{bmatrix} \tilde{\mathbf{e}}_{\mathbf{k+1}} \\ \tilde{\mathbf{e}}_{\mathbf{k}} \end{bmatrix} = \begin{bmatrix} -\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\tilde{\mathbf{B}}_{\mathbf{k}} & -\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\tilde{\mathbf{B}}_{\mathbf{k-1}} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_{\mathbf{k}} \\ \tilde{\mathbf{e}}_{\mathbf{k-1}} \end{bmatrix} = \tilde{\mathbf{A}} \begin{bmatrix} \tilde{\mathbf{e}}_{\mathbf{k}} \\ \tilde{\mathbf{e}}_{\mathbf{k-1}} \end{bmatrix}. \tag{11}$$

FETD-Newmark timestepping involves repeated multiplication by $\tilde{\mathbf{A}}$, so the Jordan canonical form of $\tilde{\mathbf{A}}$ will be deduced within the next two sections. Fortunately, the eigenspectrum of $\tilde{\mathbf{A}}$ can be constructed from the eigenspectrum of $\tilde{\mathbf{S}}$ (which is known to have a full complement of linearly independent eigenvectors because it is symmetric). However, these matrices are not the same size, $\tilde{\mathbf{A}}$ has twice as many rows and columns as $\tilde{\mathbf{S}}$. We must use each eigenvector of $\tilde{\mathbf{S}}$ to construct two eigenvectors for $\tilde{\mathbf{A}}$.

3. Stable electrodynamic modes

This proof is inspired by [9], in which the authors use a similar strategy to prove stability of the finite integration technique (FIT). To begin, note that $\tilde{\mathbf{d}}_{\mathbf{m}}$ is an eigenvector of each $\tilde{\mathbf{B}}_{\mathbf{i}}$.

$$(\tilde{\mathbf{B}}_{\mathbf{k-1}})\tilde{\mathbf{d}}_{\mathbf{m}} = \left(\frac{1}{4}(c\Delta t)^2\lambda_m + 1\right)\tilde{\mathbf{d}}_{\mathbf{m}}, \tag{12}$$

$$(\tilde{\mathbf{B}}_{\mathbf{k}})\tilde{\mathbf{d}}_{\mathbf{m}} = \left(\frac{1}{2}(c\Delta t)^2\lambda_m - 2\right)\tilde{\mathbf{d}}_{\mathbf{m}}, \tag{13}$$

$$(\tilde{\mathbf{B}}_{\mathbf{k+1}})^{-1}\tilde{\mathbf{d}}_{\mathbf{m}} = \left(\frac{1}{4}(c\Delta t)^2\lambda_m + 1\right)^{-1}\tilde{\mathbf{d}}_{\mathbf{m}}. \tag{14}$$

This step emphasizes the importance of the similarity transform: it requires that $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{T}}$ are simultaneously diagonalizable by the same set of eigenvectors $\tilde{\mathbf{d}}_{\mathbf{m}}$. The easiest way to achieve this is to map either $\mathbf{S}$ or $\mathbf{T}$ to an identity matrix, which is diagonalizable by any linearly independent basis. We cannot map $\mathbf{S}$ to an identity matrix because it is singular, but we can map $\mathbf{T}$ because it is SPD.

Continuing, construct a $2\mathscr{E} \times 1$ column vector denoted $\tilde{\mathbf{w}}_{\mathbf{m}}$ out of scaled copies of $\tilde{\mathbf{d}}_{\mathbf{m}}$. The scalar parameter $\alpha_m \in \mathbb{C}$, $\alpha_m \neq 0$ is a designable degree of freedom which modifies the direction that $\tilde{\mathbf{w}}_{\mathbf{m}}$ points. Because $\tilde{\mathbf{d}}_{\mathbf{m}}$ is an eigenvector of each of the state matrices $\tilde{\mathbf{B}}_{\mathbf{i}}$, the effect of multiplying $\tilde{\mathbf{w}}_{\mathbf{m}}$ by $\tilde{\mathbf{A}}$ is straightforward.

$$\tilde{\mathbf{w}}_{\mathbf{m}} = \begin{bmatrix} \alpha_m\tilde{\mathbf{d}}_{\mathbf{m}} \\ \tilde{\mathbf{d}}_{\mathbf{m}} \end{bmatrix}, \quad \tilde{\mathbf{A}}\tilde{\mathbf{w}}_{\mathbf{m}} = \begin{bmatrix} \left(\frac{2-\frac{1}{2}(c\Delta t)^2\lambda_m}{1+\frac{1}{4}(c\Delta t)^2\lambda_m} - \frac{1}{\alpha_m}\right)\alpha_m\tilde{\mathbf{d}}_{\mathbf{m}} \\ \alpha_m\tilde{\mathbf{d}}_{\mathbf{m}} \end{bmatrix}. \tag{15}$$

Remarkably, α_m can be picked such that $\tilde{\mathbf{w}}_{\mathbf{m}}$ is an eigenvector of $\tilde{\mathbf{A}}$. Such an α_m would have to satisfy (16).

$$\alpha_m = \frac{2-\frac{1}{2}(c\Delta t)^2\lambda_m}{1+\frac{1}{4}(c\Delta t)^2\lambda_m} - \frac{1}{\alpha_m}. \tag{16}$$

When α_m satisfies (16), then $\tilde{\mathbf{w}}_\mathbf{m}$ is by definition an eigenvector because multiplication by $\tilde{\mathbf{A}}$ did not change its direction (only its magnitude – which was stretched by a factor α_m). This quadratic expression has two roots, retain both of them to construct two eigenpairs for $\tilde{\mathbf{A}}$.

$$\alpha_{m,1} = \frac{2 - \frac{1}{2}(c\Delta t)^2\lambda_m + j2\sqrt{\lambda_m}(c\Delta t)}{2 + \frac{1}{2}(c\Delta t)^2\lambda_m} \quad \tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}} = \begin{bmatrix} \alpha_{m,1}\tilde{\mathbf{d}}_\mathbf{m} \\ \tilde{\mathbf{d}}_\mathbf{m} \end{bmatrix}, \tag{17}$$

$$\alpha_{m,2} = \frac{2 - \frac{1}{2}(c\Delta t)^2\lambda_m - j2\sqrt{\lambda_m}(c\Delta t)}{2 + \frac{1}{2}(c\Delta t)^2\lambda_m} \quad \tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}} = \begin{bmatrix} \alpha_{m,2}\tilde{\mathbf{d}}_\mathbf{m} \\ \tilde{\mathbf{d}}_\mathbf{m} \end{bmatrix}. \tag{18}$$

It is easily verified that both $\alpha_{m,1}$ and $\alpha_{m,2}$ are both unitary phasors, so the necessary criterion of unitary eigenvalues is upheld. For $\tilde{\mathbf{A}}$ to be diagonalizable, linear independence of two eigenvectors $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}$ is required. To show they are, consider the quantity $\eta \in \mathbb{C}$ (19).

$$\eta = \frac{(\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}})^H(\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}})}{|\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}||\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}|} = \frac{1}{2} + \frac{4 + \frac{1}{4}\delta^4 - 6\delta^2 - j4\delta + j2\delta^3}{8 + 4\delta^2 + \frac{1}{2}\delta^4} (= \cos\theta), \tag{19}$$

where $\delta = c\Delta t\sqrt{\lambda_m}$.

If $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}$ are linearly dependent, then this quantity will be ± 1. This is a geometric idea: two vectors are dependent (collinear) if and only if they point in the same direction ($\cos\theta = 1$) or point in opposite directions ($\cos\theta = -1$). It can be proven that η is bounded away from ± 1, and the locus that η traces out in the complex plane as δ is varied continuously from 0.1 to 100 is depicted in Fig. 1. This corresponds to a wide dynamic range in timesteps: roughly analogous to 10% of the FDTD [10] "courant limit" up to 100 times the courant limit. Although $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}$ approach linear dependence in the limits as Δt shrinks to zero or grows large without bound, these are not conditions encountered in practical FETD computations.

To generate the complete set of the electrodynamic eigenvectors of $\tilde{\mathbf{A}}$, repeat this process for every remaining $\tilde{\mathbf{d}}_\mathbf{j}$ of $\tilde{\mathbf{S}}$. It is important to note that the two vectors $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}$ just constructed will be linearly independent (orthogonal, in fact) from any other $\tilde{\mathbf{w}}_{\mathbf{j},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{j},\mathbf{2}}$ constructed later. Use the orthogonality of the generating vectors $\tilde{\mathbf{d}}_\mathbf{m}$ and $\tilde{\mathbf{d}}_\mathbf{j}$ (7) to show that $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{j},\mathbf{1}}$ are orthogonal (20).

$$(\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}})^H(\tilde{\mathbf{w}}_{\mathbf{j},\mathbf{1}}) = (\alpha^*_{m,1}) \cdot (\alpha_{j,1})(\tilde{\mathbf{d}}_\mathbf{m})^\mathrm{T}(\tilde{\mathbf{d}}_\mathbf{j}) + (\tilde{\mathbf{d}}_\mathbf{m})^\mathrm{T}(\tilde{\mathbf{d}}_\mathbf{j}) = 0. \tag{20}$$

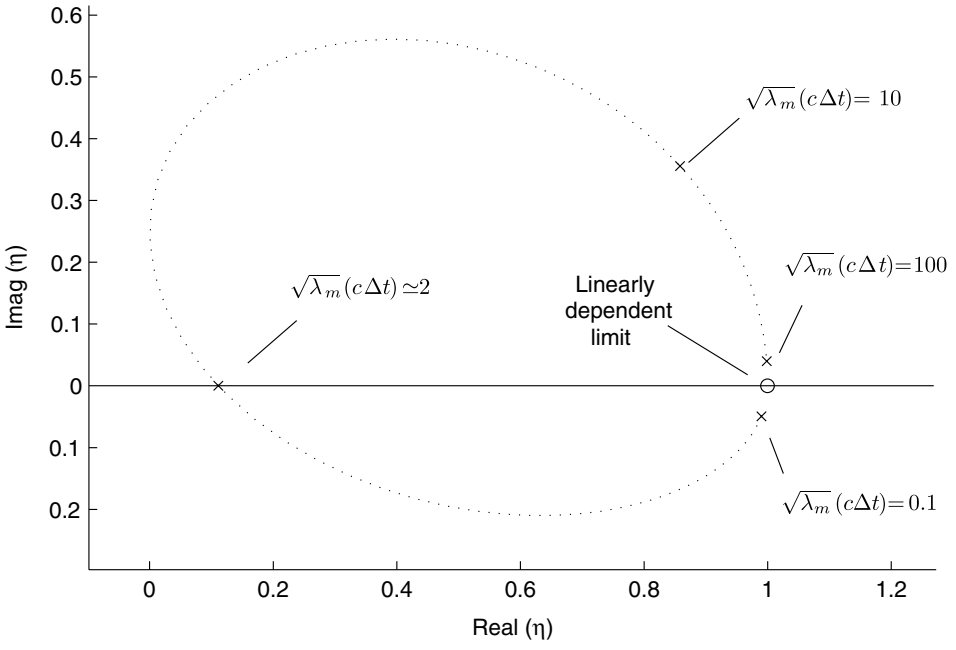

Fig. 1. Linear independence metric for $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{1}}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{2}}$.

4. Electrostatic modes

Encouraged by the success thus far, the same construction process is applied to the electrostatic eigenvectors $\tilde{\mathbf{s}}_\mathbf{n}$ of $\tilde{\mathbf{S}}$. Define the column vector $\tilde{\mathbf{x}}_\mathbf{n}$. As before, the scalar parameter β_n is a designable degree of freedom that modifies the direction of $\tilde{\mathbf{x}}_\mathbf{n}$. Because $\tilde{\mathbf{s}}_\mathbf{n}$ is in the nullspace of $\tilde{\mathbf{S}}$, the effect of multiplying $\tilde{\mathbf{x}}_\mathbf{n}$ by $\tilde{\mathbf{A}}$ is especially simple (21).

$$\tilde{\mathbf{x}}_\mathbf{n} = \begin{bmatrix} \beta_n \tilde{\mathbf{s}}_\mathbf{n} \\ \tilde{\mathbf{s}}_\mathbf{n} \end{bmatrix}, \quad \tilde{\mathbf{A}}\tilde{\mathbf{x}}_\mathbf{n} = \begin{bmatrix} 2\mathbf{I} & -\mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} \tilde{\mathbf{x}}_\mathbf{n} = \begin{bmatrix} \left(2 - \frac{1}{\beta_n}\right)\beta_n \tilde{\mathbf{s}}_\mathbf{n} \\ \beta_n \tilde{\mathbf{s}}_\mathbf{n} \end{bmatrix}. \tag{21}$$

For $\tilde{\mathbf{x}}_\mathbf{n}$ to be an eigenvector of $\tilde{\mathbf{A}}$, β_n should satisfy:

$$\beta_n = 2 - \frac{1}{\beta_n} \quad \text{or} \quad (\beta_n)^2 - 2\beta_n + 1 = 0 \tag{22}$$

The discriminant of Eq. (22) is zero – only $\beta_n = e^{j0} = 1$ is possible. Including the $\tilde{\mathbf{x}}_\mathbf{n}$ eigenvectors indicates the ability of FETD to model any discrete electrostatic field. They are embedded into the amplification matrix with unit eigenvalues: no growth, no phase change, and no decay. Updating does not alter them at all, which is quite satisfying for physical intuition of how a static mode should be treated. However, there is cause for alarm from a numerical viewpoint: we needed to construct two eigenvectors out of $\tilde{\mathbf{s}}_\mathbf{n}$ but have only made one. A complete, linearly independent set of eigenvectors for $\tilde{\mathbf{A}}$ has not been found, so it cannot be claimed that unitary eigenvalues imply unconditional stability. To proceed, hypothesize that $\tilde{\mathbf{A}}$ might not be a diagonalizable matrix, and consider a Jordan canonical form which possesses degenerate eigenvectors algebraic multiplicity 2 but geometric multiplicity 1. Such a matrix can be constructed from 2×2 Jordan sub-blocks $\tilde{\mathbf{j}}$ and "generalized eigenvectors" $\tilde{\mathbf{x}}_\mathbf{n}$ and $\tilde{\mathbf{y}}_\mathbf{n}$.

$$\tilde{\mathbf{j}} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad (\tilde{\mathbf{A}} - \lambda\mathbf{I})\tilde{\mathbf{x}}_\mathbf{n} = 0, \quad (\tilde{\mathbf{A}} - \lambda\mathbf{I})^2 \tilde{\mathbf{y}}_\mathbf{n} = 0. \tag{23}$$

The generalized eigenvector $\tilde{\mathbf{y}}_\mathbf{n}$ must be linearly independent from $\tilde{\mathbf{x}}_\mathbf{n}$ and satisfy the second order eigenpair statement (23) for $\lambda = 1$. For the electrostatic eigenvector, note that $\tilde{\mathbf{A}}$ acted like a "block identity" matrix in Eq. (21). This suggests it may behave like the 2×2 matrix $\tilde{\mathbf{a}}$. This matrix has one repeated eigenvalue and two linearly dependent eigenvectors, and its Jordan canonical form is given (24).

$$\tilde{\mathbf{a}} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \right)^{-1} = \tilde{\mathbf{v}}\tilde{\mathbf{j}}\tilde{\mathbf{v}}^{-1}. \tag{24}$$

The first column of $\tilde{\mathbf{v}}$ resembles $\tilde{\mathbf{x}}_\mathbf{n}$, and the second column suggests that $\tilde{\mathbf{y}}_\mathbf{n}$ might be $[\tilde{\mathbf{s}}_\mathbf{n}; \mathbf{0}]$. Indeed, substitution reveals that this vector satisfies the second order eigenpair statement (23), and clearly $\tilde{\mathbf{y}}_\mathbf{n}$ is linearly independent from $\tilde{\mathbf{x}}_\mathbf{n}$.

$$(\tilde{\mathbf{A}} - \lambda\mathbf{I})^2 \tilde{\mathbf{y}}_\mathbf{n} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_\mathbf{n} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_\mathbf{n} \\ \tilde{\mathbf{s}}_\mathbf{n} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{25}$$

Reapplying this procedure to every electrostatic eigenvector $\tilde{\mathbf{s}}_\mathbf{n}$ of $\tilde{\mathbf{S}}$ gives a set of $\mathscr{N} - 1$ irreducible Jordan blocks. It is straightforward to prove that every $\tilde{\mathbf{x}}_\mathbf{n}$, $\tilde{\mathbf{y}}_\mathbf{n}$ and $\tilde{\mathbf{w}}_{\mathbf{m},\mathbf{i}}$ that has been constructed is linearly independent, the orthogonality of the eigenvectors of $\tilde{\mathbf{S}}$ is the key (7).

5. Jordan canonical form and linear instability

Combining previous results into the Jordan form $\tilde{\mathbf{A}} = \tilde{\mathbf{V}}\tilde{\mathbf{J}}\tilde{\mathbf{V}}^{-1}$ is straightforward. The Jordan matrix $\tilde{\mathbf{J}}$ has a dynamic block with $2(\mathscr{E} - \mathscr{N} + 1) = 2\mathscr{M}$ unitary phasor eigenvalues, and a static block with $\mathscr{N} - 1$ irreducible Jordan subblocks $\tilde{\mathbf{j}}$.

$$\tilde{J} = \begin{bmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_1^* \end{pmatrix} & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & \begin{pmatrix} \alpha_{\mathscr{M}} & 0 \\ 0 & \alpha_{\mathscr{M}}^* \end{pmatrix} & & & \\ & & & \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} & & \\ & \mathbf{0} & & & \ddots & \\ & & & & & \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{bmatrix}. \tag{26}$$

The transformation matrix $\tilde{\mathbf{V}}$ contains all the conjugate pairs of stable electrodynamic modes ($\tilde{\mathbf{w}}$'s), all the stable electrostatic modes ($\tilde{\mathbf{x}}$'s) and the generalized eigenvectors ($\tilde{\mathbf{y}}$'s).

$$\tilde{V} = \left[\begin{pmatrix} \alpha_1 \tilde{\mathbf{d}}_1 & \alpha_1^* \tilde{\mathbf{d}}_1 & \cdots & \alpha_{\mathscr{M}} \tilde{\mathbf{d}}_{\mathscr{M}} & \alpha_{\mathscr{M}}^* \tilde{\mathbf{d}}_{\mathscr{M}} \\ \tilde{\mathbf{d}}_1 & \tilde{\mathbf{d}}_1 & \cdots & \tilde{\mathbf{d}}_{\mathscr{M}} & \tilde{\mathbf{d}}_{\mathscr{M}} \end{pmatrix} \quad \begin{pmatrix} \tilde{\mathbf{s}}_1 & \tilde{\mathbf{s}}_1 & \cdots & \tilde{\mathbf{s}}_{\mathscr{N}} - 1 & \tilde{\mathbf{s}}_{\mathscr{N}-1} \\ \tilde{\mathbf{s}}_1 & \mathbf{0} & \cdots & \tilde{\mathbf{s}}_{\mathscr{N}-1} & \mathbf{0} \end{pmatrix} \right]. \tag{27}$$

If the initial state is composed only of $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{x}}$ components, clearly no growth of the state vector is permitted as these are all eigenvectors of $\tilde{\mathbf{A}}$ with unit-circle eigenvalues. Unfortunately, if the initial state has a $\tilde{\mathbf{y}}$ component, a linearly growing gradient solution is permitted. This is consistent with practical observations of how "late-time instability" manifests itself in FETD simulations [11].

$$\tilde{\mathbf{A}}\tilde{\mathbf{y}}_\mathbf{n} = \begin{bmatrix} 2\mathbf{I} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}_\mathbf{n}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} 2\tilde{\mathbf{s}_\mathbf{n}} \\ \tilde{\mathbf{s}_\mathbf{n}} \end{bmatrix} = \tilde{\mathbf{y}}_\mathbf{n} + \tilde{\mathbf{x}}_\mathbf{n}, \tag{28}$$

$$\tilde{\mathbf{A}}^2\tilde{\mathbf{y}}_\mathbf{n} = \begin{bmatrix} 2\mathbf{I} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} 2\mathbf{s}_\mathbf{n} \\ \mathbf{s}_\mathbf{n} \end{bmatrix} = \begin{bmatrix} 3\tilde{\mathbf{s}_\mathbf{n}} \\ 2\tilde{\mathbf{s}_\mathbf{n}} \end{bmatrix} = \tilde{\mathbf{y}}_\mathbf{n} + 2\tilde{\mathbf{x}}_\mathbf{n}, \tag{29}$$

$$\tilde{\mathbf{A}}^k\tilde{\mathbf{y}}_\mathbf{n} = \tilde{\mathbf{y}}_\mathbf{n} + k\tilde{\mathbf{x}}_\mathbf{n}. \tag{30}$$

Each multiplication by $\tilde{\mathbf{A}}$ will retain the $\tilde{\mathbf{y}}_\mathbf{n}$ component, and generate an additional $\tilde{\mathbf{x}}_\mathbf{n}$ component. Up to this point, the analysis has been performed for the transformed amplification matrix $\tilde{\mathbf{A}}$, but this is not the update scheme implemented in practice. Rather, the time-integration procedure is applied directly to $\mathbf{S}$ and $\mathbf{T}$ yielding the state matrices $\mathbf{B}_{\mathbf{k}-1}$, $\mathbf{B}_\mathbf{k}$ and $\mathbf{B}_{\mathbf{k}+1}$. They are combined into the conventional amplification matrix $\mathbf{A}$ (31).

$$\begin{bmatrix} \mathbf{e}_{\mathbf{k}+1} \\ \mathbf{e}_\mathbf{k} \end{bmatrix} = \begin{bmatrix} -(\mathbf{B}_{\mathbf{k}+1})^{-1}\mathbf{B}_\mathbf{k} & -(\mathbf{B}_{\mathbf{k}+1})^{-1}\mathbf{B}_{\mathbf{k}-1} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{e}_\mathbf{k} \\ \mathbf{e}_{\mathbf{k}-1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{e}_\mathbf{k} \\ \mathbf{e}_{\mathbf{k}-1} \end{bmatrix}, \tag{31}$$

where

$$\mathbf{B}_{\mathbf{k}-1} = \frac{1}{4}(c\Delta t)^2\mathbf{S} + \mathbf{T} = \left(\mathbf{T}^{\frac{1}{2}}\right)\tilde{\mathbf{B}}_{\mathbf{k}-1}\left(\mathbf{T}^{\frac{1}{2}}\right), \tag{32}$$

$$\mathbf{B}_\mathbf{k} = \frac{1}{2}(c\Delta t)^2\mathbf{S} - 2\mathbf{T} = \left(\mathbf{T}^{\frac{1}{2}}\right)\tilde{\mathbf{B}}_\mathbf{k}\left(\mathbf{T}^{\frac{1}{2}}\right), \tag{33}$$

$$\mathbf{B}_{\mathbf{k}+1} = \frac{1}{4}(c\Delta t)^2\mathbf{S} + \mathbf{T} = \left(\mathbf{T}^{\frac{1}{2}}\right)\tilde{\mathbf{B}}_{\mathbf{k}+1}\left(\mathbf{T}^{\frac{1}{2}}\right). \tag{34}$$

As noted above, each $\mathbf{B}_\mathbf{i}$ matrix is can be equated to its $\tilde{\mathbf{B}}_\mathbf{i}$ counterpart introduced previously (9), (10), through the invertible matrix $\mathbf{T}^{\frac{1}{2}}$. Invert both sides of (34) to find $(\mathbf{B}_{\mathbf{k}+1})^{-1}$ (35).

$$(\mathbf{B}_{\mathbf{k}+1})^{-1} = \left(\mathbf{T}^{-\frac{1}{2}}\right)(\tilde{\mathbf{B}}_{\mathbf{k}+1})^{-1}\left(\mathbf{T}^{-\frac{1}{2}}\right). \tag{35}$$

Using (32), (33) and (35), it is straightforward to show that the matrices $\tilde{\mathbf{A}}$ and $\mathbf{A}$ are similar through a transformation matrix $\mathbf{Z}$.

$$\mathbf{Z}\tilde{\mathbf{A}}\mathbf{Z}^{-1} = \begin{bmatrix} \mathbf{T}^{-\frac{1}{2}} & 0 \\ 0 & \mathbf{T}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} -\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\tilde{\mathbf{B}}_{\mathbf{k}} & -\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\tilde{\mathbf{B}}_{\mathbf{k-1}} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{T}^{\frac{1}{2}} & 0 \\ 0 & \mathbf{T}^{\frac{1}{2}} \end{bmatrix}. \tag{36}$$

Carry out the matrix–matrix multiplies, and insert $\mathbf{I} = \mathbf{T}^{-\frac{1}{2}}\mathbf{T}^{\frac{1}{2}}$ identity matrices where needed:

$$\mathbf{Z}\tilde{\mathbf{A}}\mathbf{Z}^{-1} = \begin{bmatrix} -\mathbf{T}^{-\frac{1}{2}}\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\mathbf{T}^{-\frac{1}{2}}\mathbf{T}^{\frac{1}{2}}\tilde{\mathbf{B}}_{\mathbf{k}}\mathbf{T}^{\frac{1}{2}} & -\mathbf{T}^{-\frac{1}{2}}\left(\tilde{\mathbf{B}}_{\mathbf{k+1}}\right)^{-1}\mathbf{T}^{-\frac{1}{2}}\mathbf{T}^{\frac{1}{2}}\tilde{\mathbf{B}}_{\mathbf{k-1}}\mathbf{T}^{\frac{1}{2}} \\ \mathbf{T}^{-\frac{1}{2}}\mathbf{I}\mathbf{T}^{\frac{1}{2}} & \mathbf{0} \end{bmatrix}. \tag{37}$$

Identify each $\mathbf{B_i}$ from (32), (33) and (35), then identify $\mathbf{A}$ from (31).

$$\mathbf{Z}\tilde{\mathbf{A}}\mathbf{Z}^{-1} = \begin{bmatrix} -\left(\mathbf{B}_{k+1}\right)^{-1}\mathbf{B}_{k} & -\left(\mathbf{B}_{k+1}\right)^{-1}\mathbf{B}_{k-1} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \mathbf{A}. \tag{38}$$

This similarity implies that $\tilde{\mathbf{A}}$ and $\mathbf{A}$ share the same Jordan block and their Jordan transformation matrices (eigenvectors) are related through $\mathbf{V} = \mathbf{Z}\tilde{\mathbf{V}}$. Similarity transformation preserves eigenvector linear independence (unfortunately, it does not preserve orthogonality). Most importantly, if an initial state $\mathbf{y_n} = \mathbf{Z}\tilde{\mathbf{y}}_\mathbf{n}$ is presumed, a linearly growing gradient mode $\mathbf{x_n} = \mathbf{Z}\tilde{\mathbf{x}}_\mathbf{n}$ emerges.

$$\mathbf{A}^k\mathbf{y_n} = \mathbf{y_n} + k\mathbf{x_n}. \tag{39}$$

It must be stressed that this mode is *not* an artifact of the finite element procedure. The underlying continuum vector wave equation supports a linear growth gradient solution, because it resides in the nullspace of both the curl–curl operator and the second time derivative operator (40).

$$\nabla \times \left[\mu_r^{-1} \cdot \nabla \times (t\nabla\Phi)\right] + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\varepsilon_r(t\nabla\Phi) = 0. \tag{40}$$

In the continuum case this mode cannot be excited in a source free domain, it requires a growing electric charge (thus a current) to support. However, in the discrete case exciting this mode is impossible to avoid: the finite precision of floating point arithmetic or the residual error of an iterative solution technique (such as conjugate gradients) will always allow a small spillover of energy into a $\mathbf{y}$ gradient field.

6. Provable stability vs. practical stability

Although the FETD-Newmark amplification matrix $\mathbf{A}$ is non-diagonalizable and permits linearly growing states which are electrostatic (purely gradient) fields, this eigenspectrum analysis shows that multiplication by $\mathbf{A}$ is closed for the stable electrodynamic and electrostatic spaces. That is, any pure-electrodynamic input $\mathbf{w} = \mathbf{Z}\tilde{\mathbf{w}}$ yields a pure-electrodynamic output and the same is true for a stable electrostatic input $\mathbf{x} = \mathbf{Z}\tilde{\mathbf{x}}$. Furthermore, any "defective state" which contains a $\mathbf{y}$ component has no predecessor other than another defective state. Assuming that the simulation starts out non-defective, and no $\mathbf{y}$ component is added deliberately, then all $\mathbf{y}$ components are zero for all time and the system is proven stable. These assumptions are upheld in practical simulation: systems are initialized at rest and carefully implemented current sources do not generate new gradient fields. So why does FETD exhibit instability?

The answer is that updating is inexact, which disrupts the closedness property of $\mathbf{A}$. Most non-trivial sized problems use conjugate gradient (CG) [12,13] for the implicit update (or another fast but approximate linear solver), so the "next state" always has a residual error. Some of this error could be numerical energy delivered to a $\mathbf{y}$ component. It will be kick-started orders of magnitude smaller than the desired electrodynamic solution, but it will grow until it dominates the total energy in late time.

Even direct linear solvers are not immune, because round off error from inexact floating point arithmetic will shift the desired solution out of pure-dynamic space and give it a small $\mathbf{y}$ component. Direct linear solution will probably exhibit stable behavior for a larger number of timesteps than iterative solution, because the energy delivered into $\mathbf{y}$ eigenvectors will be governed by the floating point precision of the underlying hardware, not the user defined CG tolerance (which is typically several orders of magnitude above machine precision). This is consistent with previous observations [14], and the upcoming numerical experiments. Exact arithmetic (representing field samples as rational numbers with integer precision) should maintain stability

indefinitely. However, the computational expense of such an FETD scheme is intractable for non-trivial problems.

For many practical problems of short duration that are stepped with a highly accurate linear solver and smart preconditioner, no gradient term will ever be observed. Using a stricter linear solver tolerance can push the emergence of gradient field growth far "into the future", hence the name "late time instability" is appropriate. There are, however, some specific problems which require very long integration times. For example, a large cavity with two resonant modes closely spaced in frequency would need a long true-time response to distinguish the two modes in the frequency domain.

Thinking more abstractly, if the FETD-Newmark method is not "late time stabilized" then essentially it is not convergent. For example, suppose a non-dissipative cavity problem is investigated and global h-refinement (halving all edge lengths) and t-refinement (halving the timestep) is performed repeatedly to imitate the continuum behavior as $\Delta h \to 0$ and $\Delta t \to 0$. To simulate the same time duration (and thus achieve the same spectral precision), the number of timesteps must be doubled each trial. But here a crime has been committed: without removing the linear gradient solution, it cannot be guaranteed that the simulation remains stable and non-corrupted for the increased number of timesteps. Thus, the convergence of the solution in the $\Delta h,\ \Delta t \to 0$ infinitesimal limit cannot be probed. This is the same justification given for the "tree-cotree" [15] stabilization process used to solve the "low-frequency instability" [16] problem in frequency domain finite elements. Many useful problems can be analyzed without the use of tree-cotree but severe ill-conditioning in the $\Delta h \to 0$ limit will make convergence study impossible. Late-time instability in the time domain and low-frequency instability in the frequency domain are closely related.

The next three sections will introduce and demonstrate a tractable strategy for eliminating late-time instability. It will be shown that removing the linear growth term is inexpensive although it prevents the FETD-Newmark method from modeling stable gradient modes ($\mathbf{x}$).

7. Discrete gradient operator

Although the vectors $\mathbf{s}_n$ provide an orthonormal basis for the nullspace of $\mathbf{S}$, diagonalization is an impractical strategy for computing them. Fortunately, a linearly independent (although not orthogonal) basis for this space can be generated through the $\mathscr{E} \times \mathscr{N}$ gradient incidence matrix, $\mathbf{G}'$ (41).

$$\mathbf{G}'_{\mathbf{i},\mathbf{j}} = \begin{cases} +1 & \text{Edge } i \text{ points into Nodeset } j, \\ -1 & \text{Edge } i \text{ points out of Nodeset } j, \\ 0 & \text{Edge } i \text{ and Nodeset } j \text{ are not incident.} \end{cases} \tag{41}$$

When assembling G', nodes connected by PEC surfaces are renumbered and combined into "Nodesets", which are characterized by a single scalar potential degree-of-freedom. Free standing nodes which are not affixed to PEC surfaces are themselves Nodesets of cardinality one.

This concept is familiar from discrete lattice electromagnetism literature [17] or FEM potential-based $(\vec{A} - V)$ formulations [18]. Notably, $\mathbf{SG}' = \mathbf{0}$, indicating that each column $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{\mathscr{N}}\}$ of $\mathbf{G}'$ is an $\mathscr{E} \times 1$ vector in the nullspace of $\mathbf{S}$. However, recalling that one free node is assigned a reference potential it is clear that $\mathbf{G}'$ has one extra, redundant column vector. Any of them can be discarded, we will define the last node of the domain as a zero potential ground and discard the last column of $\mathbf{G}'$. In "matlab" notation, define $\mathbf{G} = \mathbf{G}'(:, 1 : \mathscr{N} - 1) = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{\mathscr{N}-1}]$.

8. Correcting instability

Previous efforts to remove linear instability by isolating the gradient subspace are present in the literature [19,20]. These efforts have focused on prohibiting the gradient mode from ever existing in the state vector by constraining the Krylov space sequence during CG solution. The computational effort required is two additional sparse matrix vector multiplies per residual computation. The proposed scheme will require an implicit linear solve step, but can be applied infrequently (once every several hundred timesteps) and still recover the correct, stable solution. Samples in between corrections may still possess gradient errors.

First, an important observation [21] must be made about how the similarity transform matrix $\mathbf{T}^{\frac{1}{2}}$ changes the orthogonality properties of the eigenvectors $\mathbf{s_n}$ and $\mathbf{d_m}$. Recalling that $\tilde{\mathbf{s}}_\mathbf{n}$ and $\tilde{\mathbf{d}}_\mathbf{m}$ were orthogonal vectors (7), it is clear that the vectors $\mathbf{s_n} = (\mathbf{T}^{-\frac{1}{2}})\tilde{\mathbf{s}}_\mathbf{n}$ and $\mathbf{d_m} = (\mathbf{T}^{-\frac{1}{2}})\tilde{\mathbf{d}}_\mathbf{m}$ are orthogonal if a new inner product is defined as $\langle \mathbf{a}, \mathbf{b} \rangle_\mathbf{T} = \mathbf{a}^\mathrm{T}\mathbf{T}\mathbf{b}$. Since $\mathbf{T}$ is a positive definite matrix, this is a valid inner product.

$$\langle \mathbf{d_m}, \mathbf{s_n} \rangle_\mathbf{T} = \mathbf{d_m}^\mathrm{T}\mathbf{T}\mathbf{s_n} = \tilde{\mathbf{d}_\mathbf{m}}^\mathrm{T}\left(\mathbf{T}^{-\frac{1}{2}}\right)\mathbf{T}\left(\mathbf{T}^{-\frac{1}{2}}\right)\tilde{\mathbf{s}}_\mathbf{n} = \tilde{\mathbf{d}}_\mathbf{m}^\mathrm{T}\tilde{\mathbf{s}}_\mathbf{n} = 0. \tag{42}$$

Since the columns of the Jordan transformation matrix $\mathbf{V} = \mathbf{Z}\tilde{\mathbf{V}}$ are linearly independent, they form a complete basis for spanning the FETD-Newmark state space. Consider the decomposition of an arbitrary state vector (43).

$$\begin{bmatrix} \mathbf{e_{k+1}} \\ \mathbf{e_k} \end{bmatrix} = \sum_{m=1}^{\mathscr{E}-\mathscr{N}+1} \left(a_m \begin{bmatrix} \alpha_m \mathbf{d_m} \\ \mathbf{d_m} \end{bmatrix} + b_m \begin{bmatrix} \alpha_m^\star \mathbf{d_m} \\ \mathbf{d_m} \end{bmatrix} \right) + \sum_{n=1}^{\mathscr{N}-1} c_n \begin{bmatrix} \mathbf{s_n} \\ \mathbf{s_n} \end{bmatrix} + \sum_{n=1}^{\mathscr{N}-1} d_n \begin{bmatrix} \mathbf{s_n} \\ \mathbf{0} \end{bmatrix}. \tag{43}$$

To correct late time instability we seek to isolate the electrostatic components of the $\mathbf{e_{k+1}}$ and $\mathbf{e_k}$ samples and subtract them out, while leaving the electrodynamic components unchanged. While this will prevent the FETD-Newmark method from modeling stable gradient fields (they will be "shorted out" at regular intervals by the correction step), they are of little practical interest in most simulations. Consider the $\mathbf{e_{k+1}}$ sample first (44).

$$\mathbf{e_{k+1}} = \sum_{m=1}^{\mathscr{E}-\mathscr{N}+1} \left(a_m \alpha_m \mathbf{d_m} + b_m \alpha_m^\star \mathbf{d_m} \right) + \mathbf{z}, \quad \mathbf{z} = \sum_{n=1}^{\mathscr{N}-1} (c_n + d_n)\mathbf{s_n} \tag{44}$$

Here all the electrostatic components have been lumped together into an unknown vector $\mathbf{z}$. Next a change of basis is employed. Because $\mathbf{z}$ is a linear combination of $\mathbf{s_n}$ vectors, it is a member of the nullspace of $\mathbf{S}$. We already possess a set of linearly independent vectors which span the nullspace of $\mathbf{S}$ and can be computed efficiently, the set $\{\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g}_{\mathscr{N}-1}\}$. There must exist a linear combination of $\mathbf{g}$ vectors which yields the $\mathbf{z}$ vector (45).

$$\mathbf{e_{k+1}} = \sum_{m=1}^{\mathscr{E}-\mathscr{N}+1} (a_m \alpha_m \mathbf{d_m} + b_m \alpha_m^\star \mathbf{d_m}) + \sum_{i=1}^{\mathscr{N}-1} p_i \mathbf{g_i}. \tag{45}$$

Here, p_i stands for potential. This correction step is trying to deduce the scalar potentials that would have to exist to at each node in order to generate the pure-gradient field $\mathbf{z}$. One constraint for finding the $\mathscr{N}-1$ unknown p_i's may be generated by taking the $\mathbf{T}$-inner product of (45) with another member of the nullspace of $\mathbf{S}$, $\mathbf{g_j}$. This is the key step for isolating the electrostatic and electrodynamic components of $\mathbf{e_{k+1}}$-each quantity $\mathbf{d_m}^\mathrm{T}\mathbf{T}\mathbf{g_j}$ must be zero. Since $\mathbf{g_j}$ is a member of the nullspace of $\mathbf{S}$, it is a summation of $\mathbf{s_n}$ vectors and every of them is $\mathbf{T}$-orthogonal to any $\mathbf{d_m}$ vector.

$$(\mathbf{e_{k+1}})^\mathrm{T}\mathbf{T}\mathbf{g_j} = \sum_{m=1}^{\mathscr{E}-\mathscr{N}+1} \left(a_m \alpha_m \mathbf{d_m}^\mathrm{T}\mathbf{T}\mathbf{g_j} + b_m \alpha_m^\star \mathbf{d_m}^\mathrm{T}\mathbf{T}\mathbf{g_j} \right) + \sum_{i=1}^{\mathscr{N}-1} p_i \mathbf{g_i}^\mathrm{T}\mathbf{T}\mathbf{g_j}. \tag{46}$$

When expanded out, Eq. (46) looks like a single row of a matrix equation for the $\mathscr{N}-1$ unknown p_i's.

$$\left(\mathbf{g}_0^\mathrm{T}\mathbf{T}\mathbf{g}_j\right) \cdot p_0 + \left(\mathbf{g}_1^\mathrm{T}\mathbf{T}\mathbf{g_j}\right) \cdot p_1 + \cdots \left(\mathbf{g}_{\mathscr{N}-1}^\mathrm{T}\mathbf{T}\mathbf{g}_j\right) \cdot p_{\mathscr{N}-1} = (\mathbf{e_{k+1}})^\mathrm{T}\mathbf{T}\mathbf{g}_j. \tag{47}$$

Testing $\mathbf{e_{k+1}}$ with each $\mathbf{g_j}$ will yield an $\mathscr{N}-1$ by $\mathscr{N}-1$ matrix equation.

$$(\mathbf{G}^\mathrm{T}\mathbf{T}\mathbf{G})\mathbf{p} = \mathbf{G}^\mathrm{T}\mathbf{T}\mathbf{e_{k+1}}. \tag{48}$$

It is straightforward to prove that $\mathbf{G}^\mathrm{T}\mathbf{T}\mathbf{G}$ is positive definite and therefore invertible. Consider the quadratic form $\mathscr{W} = \mathbf{p}^\mathrm{T}\mathbf{G}^\mathrm{T}\mathbf{T}\mathbf{G}\mathbf{p}$ – this quantity is a measure of the electrostatic energy associated with the scalar potential values $\mathbf{p}$. Since $\mathbf{T}$ is a positive definite matrix, $\mathscr{W} \geqslant 0$, and only zero if the vector $\mathbf{G}\mathbf{p} = \mathbf{0}$. Recalling that $\mathbf{G}$ computes differences in potential among adjacent nodes it is clear that it will only generate an all zero vector if each node has the same potential and they are all equal to the reference potential, which is assumed zero. That is, $\mathbf{G}\mathbf{p}$ is zero if and only if $\mathbf{p} = \mathbf{0}$. Subsequently $\mathbf{p}^\mathrm{T}\mathbf{G}^\mathrm{T}\mathbf{T}\mathbf{G}\mathbf{p} > 0$ if $\mathbf{p} \neq \mathbf{0}$, indicating positive definiteness. The system is straightforward to solve, either directly or iteratively. Upon solving for the unknown p_i coefficients,

the vector $\mathbf{z} = \mathbf{Gp}$ can be subtracted from $\mathbf{e_{k+1}}$ to create a new state $\mathbf{e_{k+1}}$, valid which has only stable electrodynamic components. The same strategy is used to remove the gradient components from the $\mathbf{e_k}$ sample, completing the correction algorithm.

$$\mathbf{e_{k+1,valid}} = \mathbf{e_{k+1}} - \mathbf{G}(\mathbf{G}^{\mathrm{T}}\mathbf{TG})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{Te_{k+1}}, \tag{49}$$

$$\mathbf{e_{k,valid}} = \mathbf{e_k} - \mathbf{G}(\mathbf{G}^{\mathrm{T}}\mathbf{TG})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{Te_k}. \tag{50}$$

The correction step (49) and (50) involves two linear solves, so it is roughly equal in computational complexity as two FETD-Newmark update cycles. However, since correction removes all the accumulated erroneous gradient fields, it does not need to be applied every update. In practical simulations, the correction step should be applied "often enough" that the contribution of the linear gradient mode to the total field remains at essentially noise level. A fixed number of bits is used to represent the sum of both the true electrodynamic response and spurious gradient mode. If no gradient removal is ever performed, the relative contribution of the gradient solution can grow so large that all significant bits are used to represent it and the desired electrodynamic response is essentially at machine-noise level. Though it would require a staggering number of timesteps, the growing gradient solution can even exhaust the range of floating point exponents and the total field must be represented as the "infinity" bit pattern. At that point, the simulation contains no information.

These are extreme outcomes that can be avoided by applying gradient field removal often enough that it provides negligible contribution to the total field. It is likely that "often enough" is problem specific, because it depends upon the conditioning of the amplification matrix. Floating point precision of underlying hardware and the residual tolerance of the iterative linear solve step will also be factors in determining the optimum correction frequency. In the upcoming numerical experiments, a CG residual tolerance of $\epsilon = 10^{-9}$ and correction frequency of once every 1000 timesteps will be shown to be a good rule of thumb to keep the gradient contribution at noise level for a sample problem.

9. Numerical experiments

To validate the Jordan form analysis presented here, a small but nontrivial PEC spherical cavity problem was investigated using the FETD-Newmark scheme. A cutaway depiction of a radius 1 meter discretized sphere [22] appears in Fig. 2. The mesh contains $\mathscr{E} = 2871$ free edges and $\mathscr{N} = 341$ free nodes. In all experiments, the implicit update was performed using the CG method, the problem is sufficiently small that no preconditioner was used.

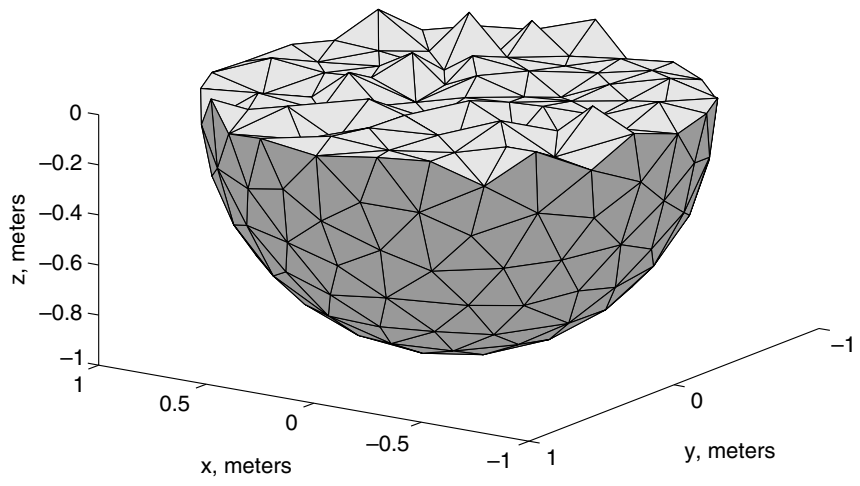

Fig. 2. Test geometry – air-filled sphere with PEC boundary.

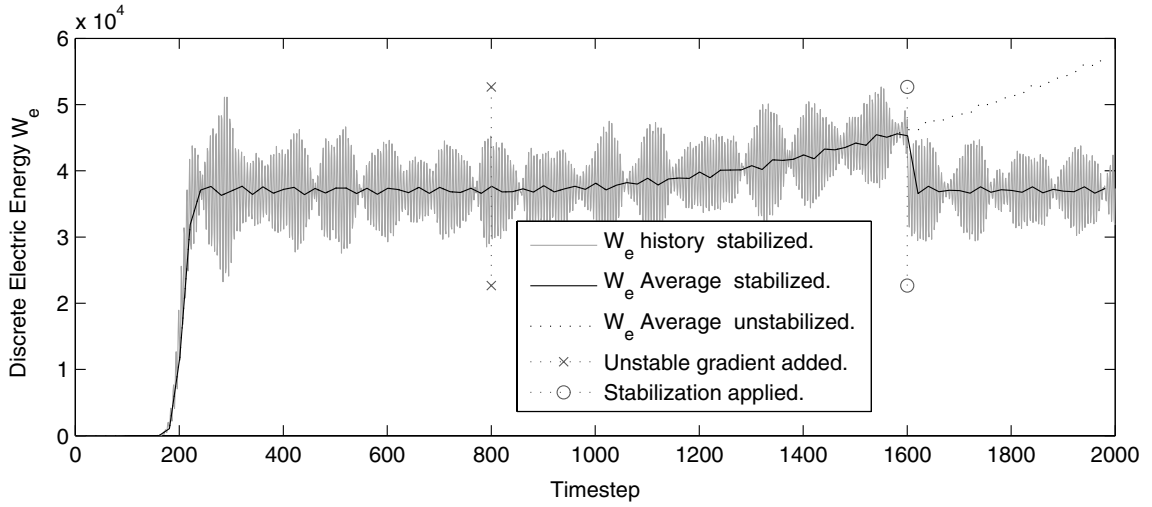

Fig. 3. Generation and cancellation of linear field growth gradient instability.

9.1. Experiment 1 – generation and cancellation of instability

This experiment verifies that late-time stability exists and the correction scheme of Section 8 eliminates it. Although the former is not in doubt among practitioners, it must be shown that the Jordan analysis has identified the correct form of a state vector which can exhibit growth. The sphere is initialized at rest and energized with pure electrodynamic energy, almost all of which is above the cutoff frequency of the cavity. The system is advanced in time with a tight relative residual tolerance $\epsilon = 10^{-9}$. At each timestep, a discrete electric energy quantity $\mathscr{W}_e$ is measured and plotted in Fig. 3

$$\mathscr{W}_e = \mathbf{e}_{\mathbf{k+1}}^{\mathrm{T}}\mathbf{T}\mathbf{e}_{\mathbf{k+1}} + \mathbf{e}_{\mathbf{k}}^{\mathrm{T}}\mathbf{T}\mathbf{e}_{\mathbf{k}}. \tag{51}$$

Since this quantity is not explicitly conserved from one timestep to the next, there is "jitter" in the plot and a need to low pass filter or average the energy history.

At $t = 800$ steps, a defective state is generated by perturbing $\mathbf{e}_{\mathbf{k+1}}$ with a random gradient field. This introduces a difference between the gradient components among $\mathbf{e}_{\mathbf{k+1}}$ and $\mathbf{e}_{\mathbf{k}}$, so the complete state contains unstable gradient-producing modes ($\mathbf{y}$'s). The energy added is minuscule (no significant jump in energy norm appears at $t = 800$ steps), but growth behavior soon emerges. Although the spurious gradent electric field is expected to grow linearly, the chosen energy metric is proportional to field intensity squared and thus exhibits quadratic growth. At $t = 1600$ steps, the correction algorithm is applied, which halts the growth behavior and restores the energy norm to the correct pure-electrodynamic level by removing the accumulated gradient field components. If no correction is applied, fields and energy continue to diverge.

9.2. Experiment 2 – cancellation of stable gradient modes

This experiment emphasizes that not all gradient fields are inherently unstable. Once more a sphere at rest is energized with electrodynamic energy and perturbed with gradient fields at $t = 800$ steps (Fig. 4). However, for this experiment both $\mathbf{e}_{\mathbf{k}}$ and $\mathbf{e}_{\mathbf{k+1}}$ are perturbed with the same random gradient field. Although a large field is added (note the jump in energy norm) the complete state is not defective, it only has stable gradient components ($\mathbf{x}$'s). No growth occurs, but when correction is applied at $t = 1600$ steps the gradient fields are removed and only electrodynamic energy remains. It is unfortunate that stable gradient modes are removed, because they *are* Maxwellian solutions to the vector wave equation. However, they are of little practical interest in most scattering and radiation simulations. Feed structures which support TEM modes are a notable exception (coaxial cable, for instance). When exciting such a structure with a transient pulse which has nonzero average (DC) value, the field distribution will be incorrect after the

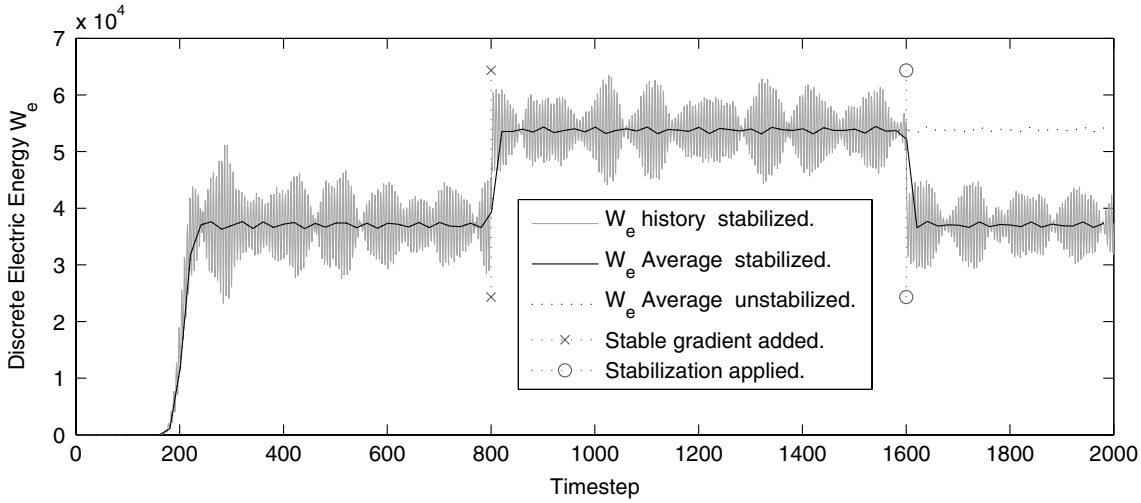

Fig. 4. Cancellation of stable gradient modes.

stabilization process is applied. The correct response possesses stable gradient mode **x**-components, but they are removed.

However, if it is known that a stable gradient mode (**x**) should be present in the solution, it can always be added back in after removing the unstable gradient modes (**y**). The operator of the simulation knows what stable gradient field should exist (either the domain was initialized with a gradient field state or it was added through an impressed current waveform with nonzero DC component, both of these are under direct control of the operator) and can reintroduce it after the correction method is applied.

9.3. Experiment 3 – instability from linear solution inaccuracy

Next it shall be shown that inaccurate linear system solution can introduce instability and the suggested correction scheme can fix this problem. The initialization step is the same as prior experiments, except the tolerance of the CG solve step is relaxed to $\epsilon = 10^{-5}$. This loss of accuracy results in growing gradient solutions after very few timesteps (Fig. 5). The correction scheme (49), (50) is applied every 1000 timesteps. In this exper-

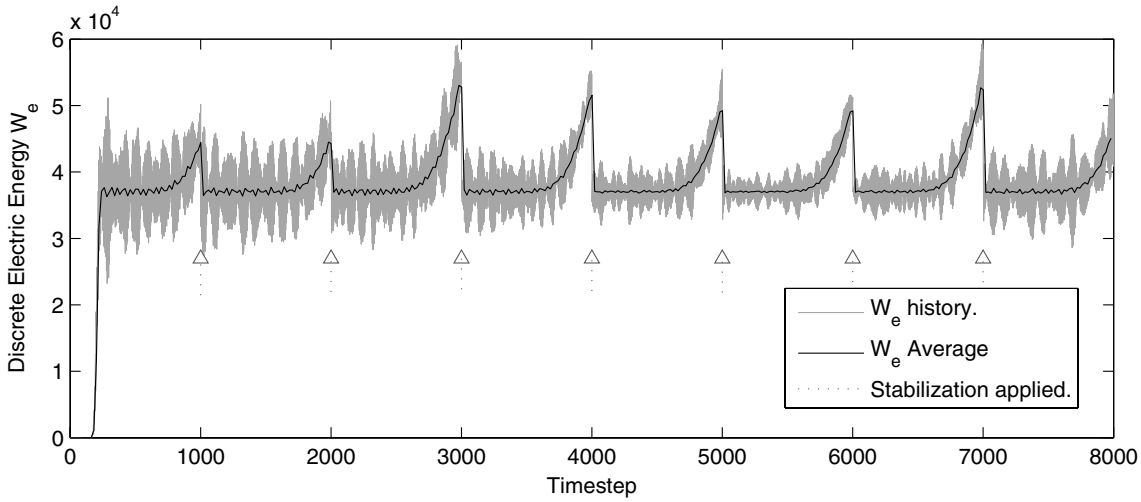

Fig. 5. Stability behavior for loose CG residual tolerance.

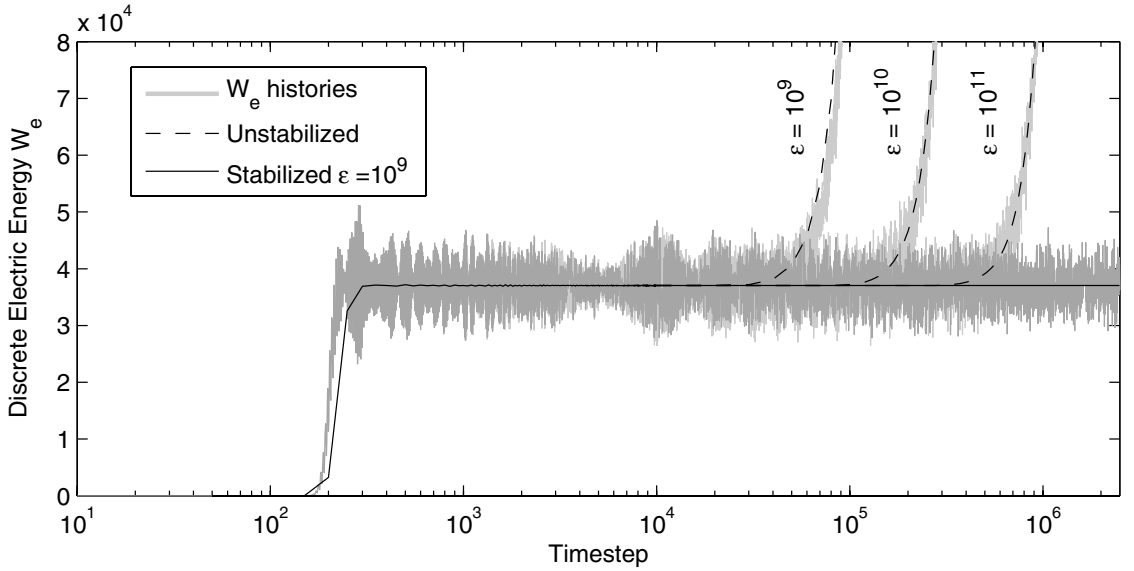

Fig. 6. For properly selected CG tolerance and correction frequency the energy response is perfectly flat. Tighter residual tolerance does not eliminate late-time instability, but stabilization via gradient removal does.

iment, the implicit solve for the correction step $(\mathbf{G}^{\mathrm{T}}\mathbf{T}\mathbf{G})\mathbf{p} = \mathbf{G}^{\mathrm{T}}\mathbf{T}\mathbf{e}$ remains $\epsilon = 10^{-9}$. This is hardly additional computational expense because the correction step is executed so seldom. After every correction the energy metric returns to the same pure-electrodynamic level, indicating that the gradient modes are reliably removed.

9.4. Experiment 4 – preventing late time instability in practical simulations

The previous experiment demonstrated that loose tolerance for an iterative solver causes growing gradient fields to emerge after very few timesteps. Applying the correction scheme removes the gradient field, but the energy history still exhibits an erroneous "sawtooth" shaped artifact that corrupts the flat electrodynamic response. This is unacceptable in practical computations. To remove the sawtooth artifact, the tolerance for the implicit update solve and the frequency of gradient removal should be chosen such that the gradient solution is not given sufficient time to grow above a noise level. Here we perform the same cavity simulation, using a residual tolerance of $\epsilon = 10^{-9}$. When iterated without gradient correction, the simulation diverges at approximately 50,000 timesteps, but when gradient correction is applied every 1000 timesteps, the average energy response is indistinguishable from a flat line. A sawtooth component is certainly present in this waveform, but the frequent gradient correction prevents it from growing to any significant size before being removed. Its contribution to the total energy is negligible. Fig. 6 also demonstrates that using tighter residual tolerance only pushes the instability problem further into future, and cannot remove it completely. Unstabilized simulations with tolerances of $\epsilon = 10^{-10}$ and $\epsilon = 10^{-11}$ are shown to diverge after 80,000 and 300,000 timesteps, respectively. The stabilized $\epsilon = 10^{-9}$ simulation maintains stability up to 2.5 million timesteps, at which point the experiment is halted.

10. Conclusions

In this paper, the amplification matrix of the FETD-Newmark algorithm has been thoroughly dissected by deducing its Jordan canonical form. It has been demonstrated that the FETD-Newmark supports linearly growing pure-gradient fields which are non-physical despite the fact that they satisfy the underlying continuum vector wave equation. The Jordan form aided the derivation of a correction scheme for removing gradient fields by exploiting the **T**-orthogonality of the electrostatic and electrodynamic eigenspaces. Removing the gradient field is not needed for many practical problems of short duration, but simulations which require long

times to be integrated can eventually become overwhelmed with growing gradient modes causing a loss of precision in the desired electrodynamic response. The correction scheme was validated via numerical experimentation on a small but non-trivial PEC spherical cavity problem.

References

[1] Jin-Fa Lee, Zachary Sacks, Whitney elements time domain methods, IEEE Transactions on Magnetics 31 (3) (1995).
[2] Stephen D. Gedney, Umesh Navsariwala, An unconditionally stable finite element time-domain solution of the vector wave equation, IEEE Microwave and Guided Wave Letters 5 (10) (1995) 332–334.
[3] J.-F. Lee, R. Lee, A.C. Cangellaris, Time-domain finite-element methods, IEEE Transactions on Antennas and Propagation 45 (1997) 430–442.
[4] Shumin Wang, Fernando L. Teixeira, Some remarks on the stability of time-domain electromagnetic simulations, IEEE Transactions on Antennas and Propagation 52 (3) (2004) 895–898.
[5] A. Bossavit, I. Mayergoyz, Edge elements for scattering problems, IEEE Transactions on Magnetics 25 (1989) 2816–2821.
[6] A. Bossavit, Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism, IEEE Proceedings 135 (Pt. A, no. 8) (1988).
[7] Gilbert Strang, Linear Algebra and its Applications, Academic Press, New York, NY, 1980.
[8] N.M. Newmark, A method of computation for structural dynamics, ASCE 85 (1959) 67–94.
[9] R. Schuhmann, T. Weiland, Convervation of discrete energy and related laws in the finite integration technique, Progress in Electromagnetics Research 32 (2001) 301–316.
[10] Kane Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, IEEE Transactions on Antennas and Propagation 14 (1966).
[11] Chieh-Tsao Hwang, Ruey-Beei Wu, Treating late-time instability of hybrid finite-element/finite-difference time-domain method, IEEE Transactions on Antennas and Propagation 47 (2) (1999).
[12] Magnus R. Hestenes, Eduard Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards 49 (1952) 409–436.
[13] Jonathan Richard Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.
[14] Shumin Wang, Improved accuracy algorithms for time-domain finite methods in electromagnetics, Dissertation, OSU Electroscience Library, 2003.
[15] R. Albanese, G. Rubinacci, Solution of three dimensional eddy current problems by integral and differential methods, IEEE Transactions on Magnetics 24 (1998) 98–101.
[16] Seung Cheol Lee, Jin-Fa Lee, Robert Lee, Hierarchical vector finite elements for analyzing waveguiding structures, IEEE Transactions on Microwave Theory and Techniques 51 (8) (2003).
[17] T. Weiland, Time domain electromagnetic field computation with finite difference methods, International Journal of Numerical Modelling: Electronic Networks, Devices and Fields 9 (1996) 259–319.
[18] Romanus Dyczij-Edlinger, Guangha Peng, Jin-Fa Lee, Efficient finite element solvers for the Maxwell equations in the frequency domain, Computer Methods in Applied Mechanics and Engineering 169 (3) (1999) 297–309.
[19] Neelakantam V. Venkatarayalu, Marinos N. Vouvakis, Yeow Beng Gan, Jin-Fa Lee, Suppressing linear time growth in edge element based finite element time domain solution using divergence free constraint equation, in: 2005 IEEE APS International Antennas and Propagation Symposium, Washington, DC, USA, 2005.
[20] Neelakantam V. Venkatarayalu, Jin-Fa Lee, Removal of spurious DC modes in edge element solutions for modeling three-dimensional resonators, IEEE Transactions on Microwave Theory and Techniques 54 (7) (2006) 3019–3025.
[21] Daniel A. White, Joseph M. Koning, Computing solenoidal eigenmodes of the vector Helmholtz equation: a novel approach, IEEE Transactions on Magnetics 38 (5) (2002).
[22] Per-Olog Persson, Gilbert Strang, A simple mesh generator in MATLAB, SIAM Review 46 (2) (2004) 329–345.