Construction and analysis of fourth order, energy consistent, family of explicit time discretizations for dissipative linear wave equations.

Juliette Chabassier, Julien Diaz, Sébastien Imperiale

Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

CONSTRUCTION AND ANALYSIS OF FOURTH ORDER, ENERGY CONSISTENT, FAMILY OF EXPLICIT TIME DISCRETIZATIONS FOR DISSIPATIVE LINEAR WAVE EQUATIONS.

Juliette Chabassier[1], Julien Diaz[1] and Sébastien Imperiale[2]

Abstract. This paper deals with the construction of a family of fourth order, energy consistent, explicit time discretizations for dissipative linear wave equations. The schemes are obtained by replacing the inversion of a matrix, that comes naturally after using the technique of the Modified Equation on the second order Leap Frog scheme applied to dissipative linear wave equations, by explicit approximations of its inverse. The stability of the schemes are studied using an energy analysis and a convergence analysis is carried out. Numerical results in 1D illustrate the space/time convergence properties of the schemes and their efficiency is compared to more classical time discretizations.

1991 Mathematics Subject Classification. 00A71, 35L05, 85A20, 33C55, 65M60.

October 12, 2018.

1. Introduction

Many imaging techniques, such as non-destructive testing, seismic probing and medical imaging, rely on the transient simulation of linear wave equations in complex media. The question of an adapted and efficient time discretization of the underlying Partial Differential Equation naturally arises, and turns out to be a bottleneck in terms of computational burden. In this work, we focus on the formulation of the wave equation that involves a second order time derivative, and we investigate the case of dissipative media. Without being exhaustive, one can distinguish, among existing explicit methods, multistep (as the Leap Frog scheme [1] or Adam Bashforths schemes [15]) and multistage (as Explicit Runge Kutta schemes [14], or fourth order Modified Equation schemes [10]) methods. They exhibit different stability properties, different costs in terms of numbers of matrix/vector products and therefore different efficiencies. On the one hand, the stability conditions of Runge Kutta or Adam Bashforth methods are obtained by quantifying the extent of the imaginary axis which belongs to the stability region of the method formulated at the first order in the complex

Keywords and phrases: Radiation boundary condition, Helmholtz equation, Atmosphere

[1] Magique 3D team – Inria Bordeaux Sud Ouest, 200 avenue de la vieille tour, 33405 Talence Cedex
Université de Pau et des Pays de l'Adour, avenue de l'université, 64013 Pau Cedex

[2] Inria and Paris-Saclay University, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau

	LF	AB3	RK4	ME
cost	1	1	4	2
efficiency	2	≈ 0.65	0.7	1.7
order	2	3	4	4

TABLE 1. Comparison of methods efficiencies for non-dissipative media. LF: Leap Frog, AB3 : Adam Bashforth order 3, RK4 : Runge Kutta order 4, ME : fourth-order Modified Equation scheme.

plane. The associated efficiency can be defined as the ratio between this value and the number of matrix/vector products needed by the method at each time step. On the other hand, methods that preserve a discrete energy, which is consistent with the physical energy, such as Leap Frog or the fourth order Modified Equation scheme, do exhibit explicit stability conditions based on a CFL-condition. The efficiency of these methods can be defined in the same way. In the context of non-dissipative media, the explicit fourth order Modified Equation turns out to be the most efficient scheme, as shown in table 1 where we present the efficiency of Leap-Frop, third order Adams-Bashforth, fourth order Runge Kutta and fourth order Modified Equation.

In the presence of dissipative terms, the stability conditions of Runge Kutta or Adam Bashforth methods are given by an implicit formula, which is not easy to fulfill a priori. This is not the case of the energy-preserving methods that we present here, which still exhibit an explicit CFL-condition. However, the direct application of the Modified Equation technique on the dissipative wave equation does not lead to an explicit scheme (even if finite elements with mass lumping or discontinuous Galerkin methods are used), which greatly hampers the efficiency of the method.

In this article, we aim at circumventing this difficulty. We design a family of explicit fourth order scheme, based on the Modified Equation technique, which can account for physical attenuation in the medium while preserving a discrete energy identity. Our schemes are based upon the use of the first terms of an adequate Neumann series, in order to deal with the implicit part of the obtained equations. To be able to evaluate the efficiency of our approach, we compare the obtained algorithm, first, with results obtained by the standard modified equation scheme, second, with a fourth order time discretization using the explicit Runge-Kutta method. We show that the solution of the modified equation is well approximated and that, in practical applications, our scheme is roughly ten times more accurate for the same computation cost. The paper is organized as follows.

- Section 2 is dedicated to the presentation of the scheme and its formal derivation.
- In Section 3 we study the stability of the scheme by energy techniques.
- In Section 4 we provide a space-time convergence results of our schemes towards the solution obtained by the modified equation.
- Section 5 is devoted to the presentation of one-dimensional space-time convergence results and cost efficiency analysis.
- In Appendix A we discuss an extension of the method to dispersive and dissipative materials in electromagnetism such as Lorentz's materials.
- Finally, in Appendix B, we provide a more dedicated stability analysis for one of the newly derived scheme. It is done using eigenvalue analysis.

In the following, the method we develop is applied to dissipative linear wave equations in a bounded domain Ω. An example of such equations is the following viscous acoustic wave equation, where $u : \mathbb{R}^+ \times \Omega \to \mathbb{R}$ is a pressure, $f : \mathbb{R}^+ \times \Omega \to \mathbb{R}$ is an acoustic source and $R : \Omega \to \mathbb{R}$ is a

damping function.

$$\frac{\partial^2 u}{\partial t^2} + R(x)\frac{\partial u}{\partial t} - \Delta u = f, \quad x \in \Omega, \tag{1}$$

completed by boundary conditions $u(x) = 0$ for $x \in \partial\Omega$. Any solution to this equation satisfies the so-called energy identity

$$\frac{d\mathcal{E}}{dt} = -\int_\Omega R(x)\left(\frac{\partial u}{\partial t}\right)^2 + \int_\Omega f\,\frac{\partial u}{\partial t} \quad \text{where} \quad \mathcal{E}(t) = \frac{1}{2}\int_\Omega \left(\frac{\partial u}{\partial t}\right)^2 + \frac{1}{2}\int_\Omega |\nabla u|^2 . \tag{2}$$

2. The Explicit Modified Equation

This part is devoted to the introduction of a new explicit fourth order time discretization, for dissipative linear wave equations of the form of (but not restricted to) (1).

In order to approach the complexity of the propagating medium and its geometry, the space discretization is assumed to be done, for instance, with high order finite elements, based on a small parameter h devoted to tend to zero, which parametrizes a sequence of finite dimensional spaces $\{V_h\}_h$. In the sequel we identify any element $u_h \in V_h$ and its vectorial representation in a well chosen basis of V_h that we still call u_h. Once the spatial discretization is fixed, we get a differential equation of the kind: Find $u_h(t,\cdot) \in V_h$ such that

$$M_h\frac{d^2 u_h}{dt^2} + B_h\frac{du_h}{dt} + A_h u_h = f_h, \tag{3}$$

where M_h is the mass matrix, B_h the dissipation matrix and A_h is the stiffness matrix. We assume here that the chosen space discretization method is such that M_h is easily invertible, i.e., diagonal or block diagonal. This can be achieved for instance thanks to Finite Difference methods, Finite Element methods with mass lumping (in particular Spectral Element methods [7], [8], [9]) or Discontinuous Galerkin Methods [16]. These types of methods also guarantee that B_h is diagonal or block-diagonal as well. Moreover, we require M_h to be positive symmetric and A_h to be non-negative symmetric. For dissipative problems, it is very likely that B_h is also non-negative symmetric, but we shall not use this property until the numerical analysis.

Since M_h is positive symmetric, Eq. (3) can be rewritten as

$$\frac{d^2 M_h^{\frac{1}{2}} u_h}{dt^2} + M_h^{-\frac{1}{2}} B_h M_h^{-\frac{1}{2}} \frac{dM_h^{\frac{1}{2}} u_h}{dt} + M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}} M_h^{\frac{1}{2}} u_h = M_h^{-\frac{1}{2}} f_h, \tag{4}$$

Obviously, matrix $M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}$ (resp. $M_h^{-\frac{1}{2}} B_h M_h^{-\frac{1}{2}}$) possesses the same properties of symmetry and of positiveness as A_h (resp. B_h). Hence, replacing $M_h^{\frac{1}{2}} u_h$ by u_h, $M_h^{-\frac{1}{2}} f_h$ by f_h, $M_h^{-\frac{1}{2}} B_h M_h^{-\frac{1}{2}}$ by B_h and $M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}$ by A_h, we can consider the simpler formulation

$$\frac{d^2 u_h}{dt^2} + B_h\frac{du_h}{dt} + A_h u_h = f_h, \tag{5}$$

without any loss of generality.

A very robust and efficient time discretization for this equation is the centered and second order, explicit, finite difference scheme known as the Leap-Frog scheme (see [1])

$$\frac{u_h^{n+1}-2u_h^n+u_h^{n-1}}{\Delta t^2}+B_h\frac{u_h^{n+1}-u_h^{n-1}}{2\Delta t}+A_hu_h^n=f_h^n \tag{LF}$$

where $f_h^n=f_h(t^n)$ with $t^n=n\Delta t$. In order to preserve the precision obtained with high order finite elements in space, we wish to design higher order time discretizations, while preserving some interesting mathematical properties as the dissipation of a discrete energy and an efficiency close to the one observed for the second order scheme. More precisely, if B_h is diagonal, scheme (LF) only requires the inversion of a diagonal matrix at each time step.

We propose to design these high-order time schemes thanks to the technique of the Modified Equation for linear equations [10]. It is based on the evaluation of the truncation error of a scheme, and the use of the semi-discrete equation in order to replace some chosen terms. Let us write the truncation error $\mathcal{L}_h$ of the scheme (LF), for the solution u_h to the semi-discrete equation (5), which is supposed as regular in time as needed (the source term is also supposed as regular as needed)

$$\begin{aligned}\mathcal{L}_h&=\frac{u_h(t^{n+1})-2u_h(t^n)+u_h(t^{n-1})}{\Delta t^2}+B_h\frac{u_h(t^{n+1})-u_h(t^{n-1})}{2\Delta t}+A_hu_h(t^n)-f_h^n \\ &=\left[d_t^2u_h(t^n)+B_hd_tu_h(t^n)+A_hu_h(t^n)-f_h(t^n)\right] \\ &\qquad+\frac{\Delta t^2}{12}\left[d_t^2f_h(t^n)+B_hd_t^3u_h(t^n)-A_hd_t^2u_h(t^n)\right]+\mathcal{O}(\Delta t^4)\end{aligned} \tag{6}$$

The first bracket vanishes because u_h is solution to (5). The remaining terms are of order Δt^2, which is the order of the scheme. Equation (5) can be differentiated with respect to time once and twice, in order to replace the terms involving derivatives of u_h. This gives

$$\begin{aligned}\mathcal{L}_h&=\varepsilon_h^2(u_h)+\mathcal{O}(\Delta t^4) \\ \varepsilon_h^2(u_h)&=\frac{\Delta t^2}{12}\left[-B_h^2d_t^2u_h(t^n)-(B_hA_h-A_hB_h)\,d_tu_h(t^n)+A_h^2u_h(t^n)\right. \\ &\qquad\left.+d_t^2f_h(t^n)-A_hf_h(t^n)+B_hd_tf_h(t^n)\right]\end{aligned} \tag{7}$$

Out of linearity, it is possible to substract to scheme (LF) a term consistent with $\varepsilon_h^2(u_h)$. This approach leads to the following scheme, fourth order accurate in time, referred to as "Modified Equation Scheme" (ME) in the literature.

$$\begin{aligned}&\left(I_h+\frac{\Delta t^2}{12}B_h^2\right)\frac{u_h^{n+1}-2u_h^n+u_h^{n-1}}{\Delta t^2}+\left[B_h+\frac{\Delta t^2}{12}\left(B_hA_h-A_hB_h\right)\right]\frac{u_h^{n+1}-u_h^{n-1}}{2\Delta t} \\ &\qquad+A_h\left(I_h-\frac{\Delta t^2}{12}A_h\right)u_h^n=\widehat{f}_h^n\end{aligned} \tag{ME}$$

where the source term needs to be modified as follows,

$$\widehat{f}_h^n=f_h(t^n)+\frac{\Delta t^2}{12}\left[d_t^2f_h(t^n)+B_hd_tf_h(t^n)-A_hf_h(t^n)\right]. \tag{8}$$

The above scheme is implicit. Indeed, it can be rewritten as

$$\left(I_h + \frac{\Delta t}{2}B_h + \frac{\Delta t^2}{12}B_h^2 + \frac{\Delta t^3}{24}\left(B_hA_h - A_hB_h\right)\right)\frac{u_h^{n+1} - u_h^{n-1}}{\Delta t^2} = \widehat{f}_h^n - A_h\left(I_h - \frac{\Delta t^2}{12}A_h\right)u_h^n + 2\left(I_h + \frac{\Delta t^2}{12}B_h^2\right)\frac{u_h^n - u_h^{n-1}}{\Delta t^2}. \quad (9)$$

Even if B_h is diagonal, A_h and B_h do not commute in general, unless the eigen sub-spaces of B_h are invariant by A_h, which is a strong condition on A_h, generally not satisfied. Therefore, the matrix to invert,

$$I_h + \frac{\Delta t}{2}B_h + \frac{\Delta t^2}{12}B_h^2 + \frac{\Delta t^3}{24}\left(B_hA_h - A_hB_h\right), \quad (10)$$

is not diagonal, nor even block diagonal, hence the resulting scheme is implicit. Of course, a natural method would be to use an iterative solver to inverse this non-symmetric matrix. Such solvers rely on an arbitrary stopping criterion, related to some norm of some residual, that allows to stop the algorithm after at a finite number of iterations. Eventually, the algorithm provides an approximate solution that depends non linearly on the solution to the linear problem, which makes it difficult to analyze mathematically.

To deepen the mathematical analysis, the main idea of this paper is to approximate the inverse of the matrix (10) by a truncated Neumann series. The resulting algorithm will require a given number of matrix-vector multiplications, which leads to an explicit algorithm, and can be seen as a linear version of an iterative inversion process with an a priori given number of iterations. The main difficulty will be to prove that this approach does not deteriorate the consistency and stability properties of the resulted scheme.

Let us introduce

$$\widetilde{M}_h = I_h + \frac{\Delta t}{2}B_h + \frac{\Delta t^2}{12}B_h^2 \quad \text{and} \quad C_h = B_hA_h - A_hB_h, \quad (11)$$

such that we have

$$I_h + \frac{\Delta t}{2}B_h + \frac{\Delta t^2}{12}B_h^2 + \frac{\Delta t^3}{24}\left(B_hA_h - A_hB_h\right) = \widetilde{M}_h\left(I_h + \frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right). \quad (12)$$

Denoting $\|\cdot\|_2$ the induced euclidian matrix norm, we assume that $\Delta t^3||\widetilde{M}_h^{-1}C_h||_2 < 24$ for Δt small enough (a rigorous formalization of this statement will be done in the next section of the article), the matrix $I_h + \Delta t^3\widetilde{M}_h^{-1}C_h/24$ is invertible, and its inverse can be written as a convergent Neumann series,

$$\left(I_h + \frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^{-1} = \sum_{k=0}^{+\infty}\left(-\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k. \quad (13)$$

Note that, since $\widetilde{M}_h$ is diagonal or block diagonal, its inverse satisfies the same property. Therefore, truncating the series up to order M, defines a family of "Explicit Modified Equation schemes" (EME-M) that are given by

$$u_h^{n+1} = u_h^{n-1} + \sum_{k=0}^{M} \left(-\frac{\Delta t^3}{24} \widetilde{M}_h^{-1} C_h \right)^k \times$$

$$\widetilde{M}_h^{-1} \left[\Delta t^2 \left(\widehat{f}_h^n - A_h \left(I_h - \frac{\Delta t^2}{12} A_h \right) u_h^n \right) + 2 \left(I_h + \frac{\Delta t^2}{12} B_h^2 \right) (u_h^n - u_h^{n-1}) \right] \quad \text{(EME-M)}$$

3. Analysis

Intuitively, the higher M is, the more accurate the scheme should be, but the higher the computational cost is. A compromise must therefore be found and another criterion for us is the possibility to conduct a mathematical analysis for stability and convergence. For this purpose we perform the following assumption throughout the paper unless specified.

Assumption 1. *The matrix A_h is positive definite and the matrix B_h is symmetric, non-negative.*

Wave equations analysis partially relies on the derivation of an energy identity that shows how the energy, which is generally a semi-norm for the solution, varies in time. It is somewhat natural, in terms of consistency, to construct a discretization that preserves a consistent version of this energy identity. It turns out that the discrete energy identity also provides a so-called CFL condition on the discretization stability, but is also a tool to show the space-time convergence of the fully discrete scheme. In our case, since B_h is assumed symmetric, the symmetry of A_h implies that $C_h = B_h A_h - A_h B_h$ is skew-symmetric, hence, it is possible to show that the classical implicit fourth order centered scheme (ME) satisfies the following energy relation ,in the absence of source.

$$\frac{\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2}}{\Delta t} = -B_h \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t}, \tag{14}$$

where

$$\mathcal{E}_h^{n+1/2} = \frac{1}{2} \left(I_h + \frac{\Delta t^2}{12} B_h^2 - \frac{\Delta t^2}{4} A_h \left[I_h - \frac{\Delta t^2}{24} A_h \right] \right) \frac{u_h^{n+1} - u_h^n}{\Delta t} \cdot \frac{u_h^{n+1} - u_h^n}{\Delta t}$$

$$+ \frac{1}{2} A_h \left(I_h - \frac{\Delta t^2}{12} A_h \right) \frac{u_h^{n+1} + u_h^n}{2} \cdot \frac{u_h^{n+1} + u_h^n}{2}. \tag{15}$$

The energy terms $\mathcal{E}_h^{n+1/2}$ are positive as soon as the CFL condition is satisfied, which is given in this case (the Modified Equation, see) [5]

$$\alpha := \Delta t \left(\frac{\|A_h\|_2}{12} \right)^{\frac{1}{2}} = \frac{\Delta t}{\sqrt{12}} \|A_h^{\frac{1}{2}}\|_2, \quad \alpha \leq 1. \tag{16}$$

Since B_h is non-negative, the right-hand-side of Eq. (14) is non-positive and the energy $\mathcal{E}_h^{n+1/2}$ decays with n. The stability of the scheme in L^2 norm can be deduced (see [6]) from the decay of the energy, which is not necessarily a norm for the solution.

We prove in the next section that the schemes (EME-M) ensure a discrete energy relation of the form

$$\frac{\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2}}{\Delta t} = -B_{h,M} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \tag{17}$$

where $\mathcal{E}_h^{n+1/2}$ is defined as in (15) and is therefore positive with the same CFL condition as for the standard modified equation scheme (ME). In the following, we focus on the first four schemes which present the following properties:

- (EME-0). This is is the lowest order of approximation, its cost is close to twice the cost of the Leap-Frog scheme and $B_{h,0} = B_h$, therefore the discrete energy is dissipated.
- (EME-1). In that case, we will show that the scheme is a fourth order scheme under some conditions on the matrices B_h and A_h. We will show that $B_{h,1}$ has no sign and therefore exponential blow up may appear. We will prove that these blow-ups, which are illustrated numerically in appendix B, are of the form $\exp(t\,C\Delta t)$ with $C > 0$ independent of h and Δt and are therefore not observable on usual time scales.
- (EME-2). A similar behavior to the one of scheme (EME-1) can be observed except that fourth order convergence may be obtained in more situations and the exponential stability estimate is more favorable. We will not study this scheme in detail here.
- (EME-3). This scheme provides all the good properties one can expect: it has the maximum order of accuracy and is dissipative. It is however the most expensive in terms of computational cost, among the schemes we analyze, yet its efficiency is better than the scheme RK4.

Remark 1. *Because of Assumption 1, the matrix* $\widetilde{M}_h$ *defined by*

$$\widetilde{M}_h = I_h + \frac{\Delta t}{2} B_h + \frac{\Delta t^2}{12} B_h^2$$

is symmetric and invertible, and its inverse is sparse (diagonal or block diagonal). Moreover $\widetilde{M}_h$ *is a positive symmetric perturbation of the identity and therefore* $\|\widetilde{M}_h^{-1}\|_2 \leq 1$ *and also* $\|\widetilde{M}_h^{-1/2}\|_2 \leq 1$. *We also have,*

$$\|\widetilde{M}_h\|_2 \leq 1 + \frac{\Delta t\,\|B_h\|_2}{2} + \frac{\Delta t^2\,\|B_h\|_2^2}{12}.$$

3.1. **Energy relation**

In order to prove that the Explicit Modified Equation (EME-M) schemes described above ensure an energy relation, we want to rewrite them similarly to the modified equation scheme (ME). One way to obtain such a centered formulation is to compute the inverse of

$$\left[\sum_{k=0}^{M} \left(-\frac{\Delta t^3}{24} \widetilde{M}_h^{-1} C_h\right)^k\right] \widetilde{M}_h^{-1}. \tag{18}$$

Notice that if

$$\frac{\Delta t^3}{24}\|\widetilde{M}_h^{-1}C_h\|_2 < 1, \tag{19}$$

then the inverse of the matrix given in (18) can be expressed thanks to a Neumann series. This motivates us to introduce a second time step restriction. To do so we define

$$\beta := \Delta t\,\|B_h\|_2$$

and give the following lemma

Lemma 3.1. *Assume that $\alpha \leq 1$ (i.e. the CFL condition* (16) *holds). If $\beta < 1$, then* (19) *holds.*

Proof. We have, because of Remark 1,

$$\|\widetilde{M}_h^{-1}C_h\|_2 \leq \|\widetilde{M}_h^{-1}\|_2\,\|C_h\|_2 \leq \|C_h\|_2.$$

Moreover, since $C_h = B_hA_h - A_hB_h$ we have

$$\|C_h\|_2 \leq 2\,\|A_h\|_2\,\|B_h\|_2 \qquad \Rightarrow \qquad \frac{\Delta t^3}{24}\|\widetilde{M}_h^{-1}C_h\|_2 \leq \frac{\Delta t^3}{12}\|A_h\|_2\,\|B_h\|_2.$$

Since we assumed that (16) holds, $\|\Delta t^2 A_h\|_2/12 \leq 1$ and

$$\frac{\Delta t^3}{24}\|\widetilde{M}_h^{-1}C_h\|_2 \leq \Delta t\,\|B_h\|_2 = \beta,$$

which concludes the proof. □

As a consequence of Lemma 3.1 we have the following equivalence theorem.

Theorem 3.2. *Assume that $\alpha \leq 1$ and $\beta < 1$. Then, for $M \in \{0,1,3\}$, the scheme* (EME-M) *is equivalent to*

$$\left(I_h + \frac{\Delta t^2}{12}B_h^2\right)\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \left[B_{h,M} + \frac{\Delta t^2}{12}C_{h,M}\right]\frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} + A_h\left(I_h - \frac{\Delta t^2}{12}A_h\right)u_h^n = \widehat{f}_h^n, \tag{20}$$

where

$$\begin{cases} B_{h,0} = B_h, \qquad C_{h,0} = 0. \\ B_{h,1} = B_h + \dfrac{\Delta t^2}{12}\displaystyle\sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\right)^{2k+1}\left(C_h\widetilde{M}_h^{-1}\right)^{2k+1}C_h, \quad C_{h,1} = \sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\right)^{2k}\left(C_h\widetilde{M}_h^{-1}\right)^{2k}C_h. \\ B_{h,3} = B_h + \dfrac{\Delta t^2}{12}\displaystyle\sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\right)^{4k+3}\left(C_h\widetilde{M}_h^{-1}\right)^{4k+3}C_h, \quad C_{h,3} = \sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\right)^{4k}\left(C_h\widetilde{M}_h^{-1}\right)^{4k}C_h. \end{cases}$$

Proof. Algorithm (EME-M) can be rewritten as

$$\widetilde{M}_h\left[\sum_{k=0}^{M}\left(-\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k\right]^{-1}\frac{u_h^{n+1}-u_h^{n-1}}{\Delta t^2}-2\left(I_h+\frac{\Delta t^2}{12}B_h^2\right)\frac{u_h^n-u_h^{n-1}}{\Delta t^2}$$
$$+A_h\left(I_h-\frac{\Delta t^2}{12}A_h\right)u_h^n=\widehat{f}_h^n.$$

Then, writing

$$\begin{aligned}\widetilde{M}_h\left[\sum_{k=0}^{M}\left(-\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k\right]^{-1} &= I_h+\frac{\Delta t^2}{12}B_h^2+\left(\widetilde{M}_h\left[\sum_{k=0}^{M}\left(-\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k\right]^{-1}-I_h-\frac{\Delta t^2}{12}B_h^2\right),\\ &=: \widehat{M}_{h,M},\end{aligned}$$

algorithm (EME-M) is actually equivalent to

$$\left(I_h+\frac{\Delta t^2}{12}B_h^2\right)\frac{u_h^{n+1}-2u_h^n+u_h^{n-1}}{\Delta t^2}+\widehat{M}_{h,M}\frac{u_h^{n+1}-u_h^{n-1}}{2\Delta t}+A_h\left(I_h-\frac{\Delta t^2}{12}A_h\right)u_h^n=\widehat{f}_h^n. \quad (22)$$

Let us now consider the different cases $M=0$, 1 and 3.

- When $M=0$, we have
$$\frac{2}{\Delta t}\left(\widetilde{M}_h-I_h-\frac{\Delta t^2}{12}B_h^2\right)=B_h,$$
which implies $B_{h,0}=B_h$ and $C_{h,0}=0$.
- For $M=1$ we need to express the inverse of the second term of (22) as a series. We have
$$\widehat{M}_{h,0}\widetilde{M}_h\left[I_h-\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right]^{-1}=\widetilde{M}_h\sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k, \quad (23)$$
and therefore
$$\widehat{M}_{h,1}=B_h+\frac{\Delta t^2}{12}C_h+\frac{2}{\Delta t}\widetilde{M}_h\sum_{k=2}^{+\infty}\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-1}C_h\right)^k,$$
from which we deduce the expression of $B_{h,1}$ and $C_{h,1}$ given by in the theorem's statement by identifying skew-symmetric and symmetric part in the matrix above.
- For the case $M=3$, we will use the property that, for any matrix D_h such that $\|D_h\|_2<1$ we have

$$\begin{aligned}\left[I_h-D_h+D_h^2-D_h^3\right]^{-1}&=\left[(I_h-D_h)(I_h+D_h^2)\right]^{-1}=(I_h+D_h^2)^{-1}(I_h-D_h)^{-1}\\ &=(I_h+D_h^2)^{-1}(I_h-D_h)^{-1}(I_h+D_h)^{-1}(I_h+D_h)=(I_h-D_h^4)^{-1}(I_h+D_h)\end{aligned}$$

therefore
$$\left[I_h-D_h+D_h^2-D_h^3\right]^{-1}=\left(\sum_{k=0}^{+\infty}D_h^{4k}\right)(I_h+D_h). \quad (24)$$

Using (24) with $D_h = \Delta t^3 \widetilde{M}_h^{-1} C_h/24$ one can show that

$$\widehat{M}_{h,3} = B_h + \frac{\Delta t^2}{12} C_h + \frac{2}{\Delta t}\left(\widetilde{M}_h + \frac{\Delta t^3}{24} C_h\right) \sum_{k=1}^{+\infty} \left(\frac{\Delta t^3}{24} \widetilde{M}_h^{-1} C_h\right)^{4k},$$

we obtain the last result of the theorem by identifying symmetric and skew-symmetric part of the above matrix. □

3.2. **Stability**

This section is devoted to the stability analysis of the modified equations schemes. We first give an energy estimate for a large class of schemes before applying to the specific case of the (ME) and (EME-M) schemes.

3.2.1. *General results*

To continue the analysis, we need an intermediate result showing that the energy norm is a semi-norm for the solution. Such a result is inspired from [6]. We recall here the statement and give a sketch of the proof for the sake of completeness.

Proposition 3.3. *If the CFL condition* (16) *is satisfied, then the functional* $\mathcal{E}_h^{n+1/2}$ *defined by* (15) *satisfies*

$$\frac{1}{4}\left\|\frac{u_h^{n+1}-u_h^n}{\Delta t}\right\|_2^2 + (1-\alpha^2)\left\|A_h^{\frac{1}{2}}\frac{u_h^{n+1}+u_h^n}{2}\right\|_2^2 \leq 2\,\mathcal{E}_h^{n+1/2}$$

Proof. If the CFL condition (16) is satisfied, then, recalling that A_h is symmetric,

$$0 \leq A_h(1-\alpha^2) \leq A_h\left(I_h - \frac{\Delta t^2}{12} A_h\right).$$

Moreover, from Assumption 1, B_h is symmetric and non-negative and we have, by inspection of (15),

$$\left(I_h - \frac{\Delta t^2}{4} A_h + \frac{\Delta t^4}{48} A_h^2\right) \frac{u_h^{n+1}-u_h^n}{\Delta t} \cdot \frac{u_h^{n+1}-u_h^n}{\Delta t} + A_h\left(1-\alpha^2\right) \frac{u_h^{n+1}+u_h^n}{2} \cdot \frac{u_h^{n+1}+u_h^n}{2} \leq 2\,\mathcal{E}_h^{n+1/2}.$$

Since the minimum of the positive polynomial $1 - x/4 + x^2/48$ is reached at $x = 6$ and is equal to $1/4$, we get the desired result. □

We address now one of the main results of this paper, which establishes a general stability result for schemes of the following form: for $1 \leq n \leq N$

$$\left(I_h + \frac{\Delta t^2}{12} B_h^2\right) \frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \left[\widetilde{B}_h + \widetilde{C}_h\right] \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} + A_h\left(I_h - \frac{\Delta t^2}{12} A_h\right) u_h^n = \tilde{f}_h^n + A_h^{\frac{1}{2}} \tilde{g}_h^n, \quad (25)$$

where $\widetilde{B}_h$ and $\widetilde{C}_h$ are respectively symmetric and skew-symmetric matrices that may depend on Δt. The stability analysis of the generic scheme (25) is a preliminary step, not only to the stability proof

of schemes (ME) and (EME-M) that we perform in the next paragraph, but also to the convergence analysis that we detail in the next section and which will give appropriate values to $\tilde{f}_h^n$ and $\tilde{g}_h^n$. We assume that the source terms $\tilde{f}_h^n$ and $\tilde{g}_h^n$ are approximations of smooth functions of time compactly supported in $(0,T)$ (hence $\tilde{f}_h^0 = 0$ and $\tilde{g}_h^0 = 0$). Moreover, initial data are assumed equal to zero:

$$u_h^0 = u_h^1 = 0, \tag{26}$$

Theorem 3.4. *Assume that* (26) *holds, that* $\alpha < 1$ *and that there exists* $0 \le c_B < 1/(4\,\Delta t)$ *such that for all* v_h*, we have*

$$\widetilde{B}_h\, v_h \,\cdot\, v_h \ge -c_B \|v_h\|_2^2, \tag{27}$$

then solutions of (25) *satisfies the following energy estimate, for all* $0 \le n \le N$*,*

$$\sqrt{\mathcal{E}_h^{n+1/2}} \le C\Delta t \sum_{k=1}^{N} \left(8 \left\| \tilde{f}_h^k \right\|_2 + \frac{1-\gamma^{-1}}{\Delta t\sqrt{1-\alpha^2}} \left\| \tilde{g}_h^k \right\|_2 + \frac{1}{\sqrt{1-\alpha^2}} \left\| \frac{\tilde{g}_h^k - \tilde{g}_h^{k-1}}{\Delta t} \right\|_2 \right) + \frac{C}{\sqrt{1-\alpha^2}} \sup_{k\in[1,N]} \left\| \tilde{g}_h^k \right\|_2,$$

where the positive scalar C *is given by*

$$C = 2\sqrt{2}\,\gamma\, e^{N(\gamma-1)} \tag{28}$$

and where the amplification factor $\gamma \ge 1$ *is defined by*

$$\gamma := \frac{1 + 4\,\Delta t\, c_B}{1 - 4\,\Delta t\, c_B}.$$

Proof. Multiplying equation (25) by $(u_h^{n+1} - u_h^{n-1})/2\Delta t$ we obtain the following energy relation

$$\frac{\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2}}{\Delta t} = \left(\tilde{f}_h^n + A_h^{\frac{1}{2}} \tilde{g}_h^n\right) \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} - \widetilde{B}_h \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t}.$$

Using Eq. (27), we have the estimation

$$\frac{\mathcal{E}_h^{n+1/2} - \mathcal{E}_h^{n-1/2}}{\Delta t} \le \left(\tilde{f}_h^n + A_h^{\frac{1}{2}} \tilde{g}_h^n\right) \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} + c_B \left\| \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \right\|_2^2.$$

Writing $u_h^{n+1} - u_h^{n-1} = (u_h^{n+1} - u_h^n) + (u_h^n - u_h^{n-1})$ and using Prop. 3.3, we find that

$$\left\| \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \right\|_2^2 \le \frac{1}{2} \left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|_2^2 + \frac{1}{2} \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t} \right\|_2^2 \le 4 \left(\mathcal{E}_h^{n+1/2} + \mathcal{E}_h^{n-1/2} \right) \tag{29}$$

hence

$$(1 - 4\,\Delta t\, c_B)\, \mathcal{E}_h^{n+1/2} - (1 + 4\,\Delta t\, c_B)\, \mathcal{E}_h^{n-1/2} \le \Delta t \left(\tilde{f}_h^n + A_h^{\frac{1}{2}} \tilde{g}_h^n\right) \cdot \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t}.$$

Using the definition of the amplification factor we have

$$\mathcal{E}_h^{n+1/2} - \gamma\,\mathcal{E}_h^{n-1/2} \leq \frac{\gamma\,\Delta t}{1+4\,\Delta t\,c_B}\big(\tilde{f}_h^n + A_h^{\frac{1}{2}}\tilde{g}_h^n\big)\cdot\frac{u_h^{n+1}-u_h^{n-1}}{2\Delta t}.$$

Therefore, one can show, using the above equation recursively, that, for $1 \leq n \leq N$,

$$\mathcal{E}_h^{n+1/2} \leq \gamma^n\,\mathcal{E}_h^{1/2} + \gamma^n \Delta t \sum_{k=1}^{n+1}\gamma^{-k}\left(\tilde{f}_h^k + A_h^{\frac{1}{2}}\tilde{g}_h^k\right)\cdot\frac{u_h^{k+1}-u_h^{k-1}}{2\Delta t}. \tag{30}$$

Note that, to obtain the above equation, we have used $1/(1+4\,c_B\,\Delta t) \leq 1$. Note also that, because of (26), we have $\mathcal{E}_h^{1/2} = 0$. Moreover, one can show that

$$\Delta t\sum_{k=1}^{n}\gamma^{-k}A_h^{\frac{1}{2}}\tilde{g}_h^k\cdot\frac{u_h^{k+1}-u_h^{k-1}}{2\Delta t} = \gamma^{-n}\tilde{g}_h^n\cdot A_h^{\frac{1}{2}}\frac{u_h^{n+1}+u_h^n}{2} + \Delta t\sum_{k=1}^{n-1}\frac{\gamma^{-k-1}\tilde{g}_h^{k+1}-\gamma^{-k}\tilde{g}_h^k}{\Delta t}\cdot A_h^{\frac{1}{2}}\frac{u_h^{k+1}+u_h^k}{2}.$$

The above result is a discrete by-part integration in time, the objective being to exchange discrete time derivative of the solution with multiplication by the square root of A_h of the source term. This is a standard strategy at the continuous level when one studies the stability of the wave equation with source term in dual spaces (for instance $H^{-1}(\Omega)$). Coming back to our estimation, we have, after using Cauchy-Schwarz inequality and the results of Proposition 3.3,

$$\Delta t\sum_{k=1}^{n}\gamma^{-k}A_h^{\frac{1}{2}}\tilde{g}_h^k\cdot\frac{u_h^{k+1}-u_h^{k-1}}{2\Delta t} \leq \frac{\sqrt{2}}{\sqrt{1-\alpha^2}}\,\gamma^{-n}\,\|\tilde{g}_h^n\|_2\,\sqrt{\mathcal{E}_h^{n+1/2}} + \frac{\sqrt{2}\,\Delta t}{\sqrt{1-\alpha^2}}\sum_{k=1}^{n-1}\left\|\frac{\gamma^{-k-1}\tilde{g}_h^{k+1}-\gamma^{-k}\tilde{g}_h^k}{\Delta t}\right\|_2\sqrt{\mathcal{E}_h^{k+1/2}}. \tag{31}$$

Since $u_1 = u_0 = 0$, the term involving $\tilde{f}_h^k$ is first written as follows

$$\Delta t\sum_{k=1}^{n}\gamma^{-k}\,\tilde{f}_h^k\cdot\frac{u_h^{k+1}-u_h^{k-1}}{2\Delta t} = \gamma^{-n}\frac{\Delta t}{2}\tilde{f}_h^n\cdot\frac{u_h^{k+1}-u_h^k}{\Delta t} + \Delta t\sum_{k=1}^{n-1}\frac{\gamma^{-k-1}\tilde{f}_h^{k+1}+\gamma^{-k}\tilde{f}_h^k}{2}\cdot\frac{u_h^{k+1}-u_h^k}{\Delta t},$$

from which we deduce the following estimation, using again Cauchy-Schwarz inequality and Proposition 3.3,

$$\Delta t \sum_{k=1}^{n} \gamma^{-k}\, \tilde{f}_h^k \cdot \frac{u_h^{k+1} - u_h^{k-1}}{2\Delta t}$$

$$\leq \sqrt{2}\,\Delta t\,\gamma^{-n}\,\|\tilde{f}_h^n\|_2 \sqrt{\mathcal{E}_h^{n+1/2}} + 2\sqrt{2}\,\Delta t \sum_{k=1}^{n-1} \left\| \frac{\gamma^{-k-1}\tilde{f}_h^{k+1} + \gamma^{-k}\tilde{f}_h^k}{2} \right\|_2 \sqrt{\mathcal{E}_h^{k+1/2}}. \quad (32)$$

Combining (30), (31), and (32) we obtain

$$\mathcal{E}_h^{n+1/2} \leq \left(\Delta t\,\sqrt{2}\,\gamma \left\| \tilde{f}_h^n \right\|_2 + \frac{\sqrt{2}\,\gamma}{\sqrt{1-\alpha^2}} \|\tilde{g}_h^n\|_2 \right) \sqrt{\mathcal{E}_h^{n+1/2}}$$

$$+ \gamma^{n+1} \Delta t \sum_{k=1}^{n-1} \left(2\sqrt{2} \left\| \frac{\gamma^{-k-1}\tilde{f}_h^{k+1} + \gamma^{-k}\tilde{f}_h^k}{2} \right\|_2 + \frac{\sqrt{2}}{\sqrt{1-\alpha^2}} \left\| \frac{\gamma^{-k-1}\tilde{g}_h^{k+1} - \gamma^{-k}\tilde{g}_h^k}{\Delta t} \right\|_2 \right) \sqrt{\mathcal{E}_h^{k+1/2}}.$$

We define now

$$D := \sup_{k\in[0,N]} \left(\Delta t\,\sqrt{2}\,\gamma \left\| \tilde{f}_h^k \right\|_2 + \frac{\sqrt{2}\,\gamma}{\sqrt{1-\alpha}} \left\| \tilde{g}_h^k \right\|_2 \right).$$

Using Young's inequality $2ab \leq a^2 + b^2$, one can show that for all $1 \leq n \leq N$

$$\frac{1}{2}\mathcal{E}_h^{n+1/2} \leq \frac{1}{2}D^2 + \gamma^n \Delta t \sum_{k=1}^{n-1} d^{k+1/2} \sqrt{\mathcal{E}_h^{k+1/2}},$$

where

$$d^{k+1/2} := 2\sqrt{2}\,\gamma \left\| \frac{\gamma^{-k-1}\tilde{f}_h^{k+1} + \gamma^{-k}\tilde{f}_h^k}{2} \right\|_2 + \frac{\sqrt{2}}{\sqrt{1-\alpha^2}}\,\gamma \left\| \frac{\gamma^{-k-1}\tilde{g}_h^{k+1} - \gamma^{-k}\tilde{g}_h^k}{\Delta t} \right\|_2.$$

Now, remark that $\gamma > 1$ implies, for all $1 \leq n \leq N$,

$$\gamma^n \leq e^{n(\gamma-1)} \leq e^{N(\gamma-1)},$$

we can then deduce for $1 \leq n \leq N$ the inequality

$$\frac{1}{2}\mathcal{E}_h^{n+1/2} \leq \frac{1}{2}D^2 + e^{N(\gamma-1)} \Delta t \sum_{k=1}^{n-1} d^{k+1/2} \sqrt{\mathcal{E}_h^{k+1/2}}, \quad (33)$$

that is still valid for $n = 0$ since $\mathcal{E}_h^{1/2} = 0$. Inspired from the proof of [22], we apply standard arguments to prove a discrete Gronwall's lemma and we introduce, for $1 \leq n \leq N$,

$$\mathcal{E}_h^{n+1/2} \leq \mathcal{F}_h^{n+1/2} := D^2 + 2e^{N(\gamma-1)} \Delta t \sum_{k=1}^{n-1} d^{k+1/2} \sqrt{\mathcal{E}_h^{k+1/2}}.$$

Note that $\mathcal{F}_h^{n+1/2} = 0$ implies $D = 0$, meaning that when the source terms are zero, the solution is also zero. Therefore, without loss of generality we assume that $\mathcal{F}_h^{n+1/2} > 0$. One can see that

$$\begin{aligned}\frac{\mathcal{F}_h^{n+1/2} - \mathcal{F}_h^{n-1/2}}{\Delta t} &= 2\, e^{N(\gamma-1)} d^{n-1/2} \sqrt{\mathcal{E}_h^{n-1/2}} \\ &\leq 2\, e^{N(\gamma-1)} d^{n-1/2} \big(\sqrt{\mathcal{F}_h^{n+1/2}} + \sqrt{\mathcal{F}_h^{n-1/2}}\big).\end{aligned}$$

which implies

$$\frac{\sqrt{\mathcal{F}_h^{n+1/2}} - \sqrt{\mathcal{F}_h^{n-1/2}}}{\Delta t} \leq 2\, e^{N(\gamma-1)} d^{n-1/2}.$$

Hence, for $0 \leq n \leq N$, we have

$$\sqrt{\mathcal{E}_h^{n+1/2}} \leq \sqrt{\mathcal{F}_h^{n+1/2}} \leq \sqrt{\mathcal{F}_h^{1/2}} + 2\, e^{N(\gamma-1)}\, \Delta t \sum_{k=0}^{N-1} d^{k+1/2}. \tag{34}$$

To finish the proof we need to estimate the terms $\sqrt{\mathcal{F}_h^{1/2}} = D$ and $d^{k+1/2}$. First,

$$D \leq \sqrt{2}\, \gamma\, \Delta t \sum_{k=1}^{N} \left\| \tilde{f}_h^k \right\|_2 + \frac{\sqrt{2}\, \gamma}{\sqrt{1-\alpha^2}} \sup_{k \in [1,N]} \left\| \tilde{g}_h^k \right\|_2. \tag{35}$$

Since $\gamma > 1$, one can show, writing $\tilde{g}_h^{k+1} = (\tilde{g}_h^{k+1} - \tilde{g}_h^k) + \tilde{g}_h^k$ and using triangular inequalities that

$$\frac{d^{k+1/2}}{\sqrt{2}\, \gamma} \leq 2 \left\| \tilde{f}_h^k \right\|_2 + 2 \left\| \tilde{f}_h^{k+1} \right\|_2 + \frac{1}{\sqrt{1-\alpha^2}} \left\| \frac{\tilde{g}_h^{k+1} - \tilde{g}_h^k}{\Delta t} \right\|_2 + \frac{1}{\sqrt{1-\alpha^2}} \left(\frac{1-\gamma^{-1}}{\Delta t} \right) \left\| \tilde{g}_h^k \right\|_2.$$

hence

$$\Delta t \sum_{k=0}^{N-1} d^{k+1/2} \leq 2\sqrt{2}\, \gamma \Delta t \sum_{k=1}^{N} \left(4 \left\| \tilde{f}_h^k \right\|_2 + \frac{1}{\sqrt{1-\alpha^2}} \left\| \frac{\tilde{g}_h^k - \tilde{g}_h^{k-1}}{\Delta t} \right\|_2 + \frac{1}{\sqrt{1-\alpha^2}} \left(\frac{1-\gamma^{-1}}{\Delta t} \right) \left\| \tilde{g}_h^k \right\|_2 \right). \tag{36}$$

Finally, combining (35) with (34) and (36), we obtain the result of the theorem using the definition (28) of the scalar C. □

3.2.2. *Stability of the schemes* (ME) *and* (EME-M)

In this section, we specify how Theorem 3.4 can be used to study the scheme (ME) and each scheme (EME-M), $M \in \{0, 1, 3\}$. We first state Proposition 3.5 that gives the value of c_B of Equation (27) for (ME), and (EME-M) with $M = 0$ or $M = 3$. In particular it is shown that $c_B = 0$ and therefore no exponential growth occurs in the stability estimates for schemes (ME), (EME-0) and (EME-3). This is not the case for the scheme (EME-1), for which we establish Proposition 3.6 giving the value of c_B is given in that particular case.

Proposition 3.5. *For all $v_h \in V_h$, $B_{h,0}\, v_h \cdot v_h \geq 0$. Moreover if $\alpha \leq 1$ and $\beta < 1$, then $B_{h,3}\, v_h \cdot v_h \geq 0$. Therefore, Theorem 3.4 can be applied for* (ME), (EME-1) and (EME-3) with $c_B \equiv 0$.

Proof. The first statement comes from the non-negativity of $B_{h,0} = B_h$. The second statement comes from the expression of $B_{h,3}$ given by Theorem 3.2. Indeed, each term of the series is non-negative:

$$\left(C_h \widetilde{M}_h^{-1}\right)^{4k+3} C_h\, v_h \cdot v_h \geq 0,$$

since C_h is skew symmetric. □

Proposition 3.6. *If $\alpha \leq 1$ and $\beta < 1$, then for all $v_h \in V_h$,*

$$B_{h,1}\, v_h \cdot v_h \geq B_h\, v_h \,\cdot\, v_h - \frac{2\,\beta^2}{\Delta t}\widetilde{M}_h\, v_h \,\cdot\, v_h \geq -\frac{4\beta^2}{\Delta t}\|v_h\|_2^2. \tag{37}$$

Proof. Using the expression of $B_{h,1}$ given by Theorem 3.2,

$$B_{h,1} v_h \cdot v_h = B_h\, v_h \,\cdot\, v_h + \frac{\Delta t^2}{12}\left(\sum_{k=0}^{+\infty}\left(\frac{\Delta t^3}{24}\right)^{2k+1}\left(C_h\widetilde{M}_h^{-1}\right)^{2k+1} C_h\right) v_h \,\cdot\, v_h. \tag{38}$$

Since $\widetilde{M}_h$ is symmetric positive, we can define its square root $\widetilde{M}_h^{\frac{1}{2}}$ and write that

$$\begin{aligned}
\frac{\Delta t^2}{12}\left(\frac{\Delta t^3}{24}\right)^{2k+1}\left(C_h\widetilde{M}_h^{-1}\right)^{2k+1} C_h &= \frac{2}{\Delta t}\left(\frac{\Delta t^3}{24}\right)^{2k+2}\widetilde{M}_h^{\frac{1}{2}}\widetilde{M}_h^{-\frac{1}{2}}\left(C_h\widetilde{M}_h^{-1}\right)^{2k+1} C_h\,\widetilde{M}_h^{-\frac{1}{2}}\widetilde{M}_h^{\frac{1}{2}} \\
&= \frac{2}{\Delta t}\left(\frac{\Delta t^3}{24}\right)^{2k+2}\widetilde{M}_h^{\frac{1}{2}}\left(\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^{2k+2}\widetilde{M}_h^{\frac{1}{2}}.
\end{aligned}$$

Denoting by $\mathcal{S}$ the set of eigenvalues of the non positive symmetric matrix

$$\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^2,$$

we have that

$$\begin{aligned}
\sum_{k\geq 0}^{+\infty}\widetilde{M}_h^{\frac{1}{2}}\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^{2k+2}\widetilde{M}_h^{\frac{1}{2}}\, v_h \cdot v_h &= \sum_{k\geq 0}^{+\infty}\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^{2k+2}\widetilde{M}_h^{\frac{1}{2}} v_h \cdot \widetilde{M}_h^{\frac{1}{2}} v_h \\
&\geq \left(\min_{\lambda\in\mathcal{S}}\sum_{k\geq 0}^{+\infty}\lambda^{k+1}\right)\,\widetilde{M}_h^{\frac{1}{2}} v_h \,\cdot\, \widetilde{M}_h^{\frac{1}{2}} v_h,
\end{aligned}$$

and therefore

$$B_h v_h \cdot v_h \geq B_h v_h \cdot v_h + \frac{2}{\Delta t}\left(\min_{\lambda\in\mathcal{S}}\sum_{k\geq 0}^{+\infty}\lambda^{k+1}\right)\,\widetilde{M}_h^{1/2} v_h \,\cdot\, \widetilde{M}_h^{1/2} v_h. \tag{39}$$

We now have to find a lower bound to $\min\limits_{\lambda\in\mathcal{S}}\sum\lambda^{k+1}$. Remark that

$$\rho\left(\left(\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^2\right) \leq \left\|\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right\|_2^2 \leq \|M_h^{-\frac{1}{2}}\|_2^4\, \|C_h\|_2^2 \leq \|C_h\|_2^2$$

since $\|\widetilde{M}_h^{-1/2}\|_2 \leq 1$ from Remark 1. Therefore, following the proof of Lemma 3.1 we have

$$\rho\left(\left(\frac{\Delta t^3}{24}\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\right)^2\right) \leq \beta^2 \quad \text{and} \quad \mathcal{S} \subset \left[-\beta^2, 0\right],$$

which implies, since $\beta < 1$, that

$$\frac{2}{\Delta t}\min_{\lambda\in\mathcal{S}} \sum_{k\geq 0}^{+\infty} \lambda^{k+1} = \frac{2}{\Delta t}\min_{\lambda\in\mathcal{S}} \frac{\lambda}{1-\lambda} \geq -\frac{2}{\Delta t}\max_{\lambda\in\mathcal{S}} |\lambda| \geq -\frac{2\,\beta^2}{\Delta t}.$$

Combining this last inequality with Eq. (39), we obtain the first inequality of the proposition. Morevoer since $\beta < 1$, we have from the definition of $\widetilde{M}_h$ given in Remark 1 that $\|\widetilde{M}_h\|_2 < 19/12 < 2$, therefore omitting the positive term $B_h\, v_h \cdot v_h$ in (37) we obtain the last result of the proposition. □

The result of Proposition 3.6 may seem not sharp enough since, in regards of the stability estimate given in Theorem 3.4, an exponential growth of the energy may occur. The exponential factor is of the form

$$e^{N(\gamma-1)} \quad \text{with} \quad \gamma - 1 = \frac{8\,\Delta t\, c_B}{1 - 4\,\Delta t\, c_B} \quad \text{and so} \quad c_B = \frac{4\,\beta^2}{\Delta t}. \tag{40}$$

We will see however, in the applications we consider, that β is proportional to Δt and therefore, for Δt small enough we have $c_B < 1/(4\Delta t)$, which is a necessary condition to apply Theorem 3.4. Thus,

$$\gamma - 1 \underset{\Delta t\to 0}{\sim} \Delta t^2,$$

meaning that, if $T = N\Delta t$ is the final time of simulation, the exponential growth is of the form $\exp(C\,T\Delta t)$, with C a positive scalar independent of Δt), so that the scheme can still be qualified as convergent and stable. In Appendix B we show that estimate (37) is not optimal in specific situations. However, we illustrate numerically in appendix B that an exponential growth may indeed occur for some discretization parameters.

4. Convergence analysis

In this section we treat "partially" the space and time convergence analysis of our scheme. By partially, we mean that we only prove a space and time convergence result of the solutions of explicit modified equation schemes (EME-M) towards the solutions of the modified equation (ME). We first give an abstract error estimate before tackling the particular case of the wave equation.

4.1. **Abstract error estimate**

We consider now a family of problems parametrized by a family of positive numbers h that converge to 0, meaning that we consider a family of matrices $\{A_h\}$ and $\{B_h\}$ that may act on larger and larger spaces as h goes to zero. We assume that the time step is given by the CFL condition (16). More precisely we assume $\alpha < 1$ fixed and set

$$\Delta t := \left(\frac{12\,\alpha}{\|A_h\|_2}\right)^{\frac{1}{2}}.$$

With the above equality, the time step Δt can be seen as a function of h. Indeed, in the cases we want to consider, A_h approximates an unbounded operator and $\|A_h\|_2$ blows up when h goes to 0, which implies that Δt goes to 0. Therefore, the space time convergence can be studied be letting h tend to 0. Now we denote by $\{u_h^n\}$ the sequence of iterates obtained by solving the modified equation scheme (ME) and $\{u_{h,M}^n\}$ the sequence of iterates obtained by solving (EME-M), formulated as in (20), and we define

$$e_{h,M}^n := u_h^n - u_{h,M}^n.$$

Our objective is to show that $e_{h,M}^n$ goes to 0 for a given $h-$dependent norm when h goes to 0. By simple computations one can show that the error term $e_{h,M}^n$ satisfies the explicit modified equation scheme (EME-M) with source term depending on u_h^n. More precisely we have, for $1 \leq n \leq N$,

$$\left(I_h + \frac{\Delta t^2}{12} B_h^2\right) \frac{e_{h,M}^{n+1} - 2e_{h,M}^n + e_{h,M}^{n-1}}{\Delta t^2} + \left[B_{h,M} + \frac{\Delta t^2}{12} C_{h,M}\right] \frac{e_{h,M}^{n+1} - e_{h,M}^{n-1}}{2\Delta t} + A_h \left(I_h - \frac{\Delta t^2}{12} A_h\right) e_{h,M}^n = D_{h,M} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t}, \quad (41)$$

with

$$D_{h,M} = B_{h,M} - B_h + \frac{\Delta t^2}{12}(C_{h,M} - C_h).$$

Applying Theorem 3.2 we find

$$D_{h,0} = -\frac{\Delta t^2}{12} C_h, \quad (42)$$

and for $M = 1$ or $M = 3$

$$D_{h,M} = \frac{\Delta t^2}{12} \sum_{k \in \mathcal{N}_M} \left(\frac{\Delta t^3}{24}\right)^k \left(C_h \widetilde{M}_h^{-1}\right)^k C_h, \quad (43)$$

with

$$\mathcal{N}_1 = \mathbb{N}^*, \quad \mathcal{N}_3 = \{4k-1 \,|\, k \in \mathbb{N}^*\} \cup \{4k \,|\, k \in \mathbb{N}^*\}.$$

To continue the analysis, a simple way to proceed would be to apply Theorem 3.4 with

$$\tilde{f}_h^n = D_{h,M} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \quad \text{and} \quad \tilde{g}_h^n = 0. \quad (44)$$

It is easy to obtain an estimation of $(u_h^{n+1} - u_h^{n-1})/2\Delta t$ using the stability of the (ME). However, if the function $R(x)$ is discontinuous, which happens frequently in the context of weve propagation, the term $\|D_{h,M}\|_2$ goes to 0 as h^M. This behavior will be illustrated numerically in the next section. Hence, even with $M = 3$, we will not be able to prove the fourth order convergence of the scheme. An alternative is to write the right hand side of (41) as

$$D_{h,M} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} = A_h^{\frac{1}{2}} \tilde{g}_h^n \quad \text{with} \quad \tilde{g}_h^n = \left(A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}\right) \left(A_h^{\frac{1}{2}} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t}\right), \quad \tilde{g}_h^0 = 0. \quad (45)$$

and to apply Theorem 3.4 with $\tilde{g}^n$ given above and $\tilde{f}^n = 0$. For this, by inspection of the estimation of Theorem 3.4 we need to estimate

$$A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}},\ \tilde{q}_h^n := A_h^{\frac{1}{2}} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \ \text{ and } \ \frac{\tilde{q}_h^n - \tilde{q}_h^{n-1}}{\Delta t} = A_h^{\frac{1}{2}} \frac{u_h^{n+1} - u_h^n - u_h^{n-1} + u_h^{n-2}}{2\Delta t^2}. \quad (46)$$

The next three lemmas aim at estimating these 3 terms successively.

Lemma 4.1. *Assume that $\beta < 1$ and $M = 1$ or $M = 3$. Then, we have,*

$$\|A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}\|_2 \leq \frac{\Delta t^5}{288} \frac{\beta^{M-1}}{1-\beta} \|C_h A_h^{-\frac{1}{2}}\|_2^2.$$

Note that the above Lemma gives an estimation of $A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}$ (for $M = 1$ or $M = 3$) in terms of $C_h A_h^{-\frac{1}{2}}$. We explain in Section 4.2 why we expect that this operator, in the case of the wave equation, has good behavior as h goes to 0 and it is illustrated numerically in Section 5.1.

Proof. Observe that, by definition of $D_{h,M}$ given by (43), we have

$$\frac{12}{\Delta t^2} A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}} = A_h^{-\frac{1}{2}} \left(\sum_{k \in \mathcal{N}_M} \left(\frac{\Delta t^3}{24} \right)^k \left(C_h \widetilde{M}_h^{-1} \right)^k C_h \right) A_h^{-\frac{1}{2}}$$

$$= \frac{\Delta t^3}{24} A_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}} \left(\sum_{k+1 \in \mathcal{N}_M} \left(\frac{\Delta t^3}{24} \right)^k \left(\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}} \right)^k \right) \widetilde{M}_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}.$$

Following the proof of Lemma 3.1 one can show that

$$\left\| \sum_{k+1 \in \mathcal{N}_M} \left(\frac{\Delta t^3}{24} \right)^k \left(\widetilde{M}_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}} \right)^k \right\|_2 \leq \sum_{k=M-1}^{+\infty} \left(\frac{\Delta t^3}{24} \right)^k \|C_h\|_2^k \leq \sum_{k=M-1}^{+\infty} \beta^k = \frac{\beta^{M-1}}{1-\beta}$$

we deduce that

$$\|A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}\|_2 \leq \frac{\Delta t^5}{288} \frac{\beta^{M-1}}{1-\beta} \|A_h^{-\frac{1}{2}} C_h \widetilde{M}_h^{-\frac{1}{2}}\|_2 \, \|\widetilde{M}_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}\|_2$$

$$\leq \frac{\Delta t^5}{288} \frac{\beta^{M-1}}{1-\beta} \|A_h^{-\frac{1}{2}} C_h\|_2 \, \|C_h A_h^{-\frac{1}{2}}\|_2.$$

Finally, since $\|A_h^{-\frac{1}{2}} C_h\|_2 = \|(A_h^{-\frac{1}{2}} C_h)^T\|_2 = \|C_h A_h^{-\frac{1}{2}}\|_2$ we obtain the result of the Lemma. □

Lemma 4.2. *Let u_h^n be the solution of* (ME) *with $\alpha < 1$, $u_h^0 = u_h^1 = 0$, and $\widehat{f}_h^0 = \widehat{f}_h^1 = 0$. Then, for all $1 \leq n \leq N$*

$$\left\| A_h^{\frac{1}{2}} \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} \right\|_2 \leq \frac{36}{\sqrt{1-\alpha^2}} \Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2.$$

Proof. We set

$$v_h^n := \frac{u_h^n - u_h^{n-1}}{\Delta t},$$

then we subtract the scheme (ME) written at time n and at time $n-1$, we obtain, after dividing by Δt, the following equation, valid for all $2 \leq n \leq N$

$$\left(I_h + \frac{\Delta t^2}{12} B_h^2\right) \frac{v_h^{n+1} - 2v_h^n + v_h^{n-1}}{\Delta t^2} + \left[B_h + \frac{\Delta t^2}{12} C_h\right] \frac{v_h^{n+1} - v_h^{n-1}}{2\Delta t} + A_h \left(I_h - \frac{\Delta t^2}{12} A_h\right) v_h^n = \frac{\widehat{f}_h^n - \widehat{f}_h^{n-1}}{\Delta t}. \quad (47)$$

Thanks to the hypothesis $\widehat{f}_h^1 = 0$, we have $u_h^2 = 0$ and thus $v_h^1 = v_h^2 = 0$. We can then apply the stability Theorem 3.4 with $\tilde{g}_h^n = 0$ and $\tilde{f}_h^n \equiv (\widehat{f}_h^n - \widehat{f}_h^{n-1})/\Delta t$ Because B_h is non negative by assumption, we have $c_B = 0$. Then, thanks to Proposition 3.3 we have for all $1 \leq n \leq N$

$$\sqrt{\frac{1-\alpha^2}{2}} \left\| A_h^{\frac{1}{2}} \frac{v_h^{n+1} + v_h^n}{2} \right\|_2 \leq \sqrt{\mathcal{E}_h^{n+1/2}} \leq 9\sqrt{2}\, \Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2,$$

where the energy $\mathcal{E}_h^{n+1/2}$ is given by (15) with v_h^n instead of u_h^n. We obtain the result of the lemma by replacing v_h^n by its value in terms of u_h^n. □

Lemma 4.3. *Let u_h^n be the solution of* (ME) *with $\alpha < 1$ and*

$$u_h^0 = u_h^1 = 0, \quad \widehat{f}_h^0 = \widehat{f}_h^1 = \widehat{f}_h^2 = 0$$

then for all $2 \leq n \leq N$

$$\left\| A_h^{\frac{1}{2}} \frac{u_h^{n+1} - u_h^n - u_h^{n-1} + u_h^{n-2}}{2\Delta t^2} \right\|_2 \leq \frac{36}{\sqrt{1-\alpha^2}} \Delta t \sum_{k=2}^{N} \left\| \frac{\widehat{f}_h^k - 2\widehat{f}_h^{k-1} + \widehat{f}_h^{k-2}}{\Delta t^2} \right\|_2.$$

Proof. We set

$$v_h^n := \frac{u_h^n - u_h^{n-1}}{\Delta t} \text{and} \quad w_h^n := \frac{v_h^n - v_h^{n-1}}{\Delta t},$$

where v_h^n is defined in the proof of corollary 4.2. Subtracting scheme (47) written at time n and at time $n-1$, we obtain, after dividing by Δt, the following equation, valid for all $3 \leq n \leq N$

$$\left(I_h + \frac{\Delta t^2}{12} B_h^2\right) \frac{w_h^{n+1} - 2w_h^n + w_h^{n-1}}{\Delta t^2} + \left[B_h + \frac{\Delta t^2}{12} C_h\right] \frac{w_h^{n+1} - w_h^{n-1}}{2\Delta t} + A_h \left(I_h - \frac{\Delta t^2}{12} A_h\right) w_h^n = \frac{\widehat{f}_h^n - 2\,\widehat{f}_h^{n-1} + \widehat{f}_h^{n-2}}{\Delta t^2}. \quad (48)$$

Using the assumption $\widehat{f}_h^1 = \widehat{f}_h^2 = 0$, we deduce that $v_h^1 = v_h^2 = v_h^3 = 0$ and therefore $w_h^2 = w_h^3 = 0$, so that we can apply the stability Theorem 3.4 with $\tilde{g}_h^n = 0$, $\tilde{f}_h^n \equiv (\widehat{f}_h^n - 2\widehat{f}_h^{n-1} + \widehat{f}_h^{n-1})/\Delta t^2$ and,

once again $c_B = 0$. Then, thanks to Proposition 3.3 we have

$$\sqrt{\frac{1-\alpha^2}{2}}\left\|A_h^{\frac{1}{2}}\frac{w_h^{n+1}+w_h^n}{2}\right\|_2 \le \sqrt{\mathcal{E}_h^{n+1/2}} \le 9\sqrt{2}\,\Delta t\sum_{k=2}^{N}\left\|\frac{\widehat{f}_h^n - 2\,\widehat{f}_h^{n-1}+\widehat{f}_h^{n-2}}{\Delta t^2}\right\|_2,$$

where the energy $\mathcal{E}_h^{n+1/2}$ is given by (15) with w_h^n instead of u_h^n. We obtain the result of the corollary by replacing w_h^n by its value in terms of u_h^n. □

It is now possible to state the following theorem on the estimation of the error term $e_{h,M}^n$ for schemes (EME-M) (with $M \in \{0,1,3\}$).

Theorem 4.4. *Assume $\alpha < 1$, $\beta < 1/6$ and*

$$u_h^0 = u_h^1 = 0, \quad u_{h,M}^0 = u_{h,M}^1 = 0, \quad \widehat{f}_h^0 = \widehat{f}_h^1 = \widehat{f}_h^2 = 0.$$

Then, the error $e_{h,M}^n = u_h^n - h_{h,M}^n$ satisfies, for $1 \le n \le N$,

$$\left(\frac{1}{4}\left\|\frac{e_{h,M}^{n+1}-e_{h,M}^n}{\Delta t}\right\|_2^2 + (1-\alpha^2)\left\|A_h^{\frac{1}{2}}\frac{e_{h,M}^{n+1}+e_{h,M}^n}{2}\right\|_2^2\right)^{\frac{1}{2}}$$

$$\le a_{h,M}\left[b_{h,M}\,\Delta t\sum_{k=1}^{N}\left\|\frac{\widehat{f}_h^k-\widehat{f}_h^{k-1}}{\Delta t}\right\|_2 + N\,\Delta t^2\sum_{k=2}^{N}\left\|\frac{\widehat{f}_h^k-2\widehat{f}_h^{k-1}+\widehat{f}_h^{k-2}}{\Delta t^2}\right\|_2\right],$$

with

$$\begin{cases} a_{h,0} = \Delta t^2\,\dfrac{12}{1-\alpha^2}\|A_h^{-\frac{1}{2}}C_hA_h^{-\frac{1}{2}}\|_2, \\ b_{h,0} = 1, \end{cases}, \quad \begin{cases} a_{h,1} = \dfrac{\sqrt{2}\,\Delta t^5}{1-\alpha^2}\|C_hA_h^{-\frac{1}{2}}\|_2^2\,e^{\,64\,N\,\beta^2}, \\ b_{h,1} = 1+32\,N\,\beta^2. \end{cases}$$

and

$$\begin{cases} a_{h,3} = \dfrac{\Delta t^5}{1-\alpha^2}\,\beta^2\,\|C_hA_h^{-\frac{1}{2}}\|_2^2, \\ b_{h,3} = 1, \end{cases}.$$

Proof. Because of our assumption on the initial data we can apply Theorem 3.4 on scheme (41) with $f_h^n = 0$, $\tilde{g}_h^n$ given by (45) and by $\tilde{g}_h^0 := 0$, and by

$$\widetilde{B}_h = B_{h,M}, \quad \widetilde{C}_h = \frac{\Delta t^2}{12}C_{h,M}.$$

Denoting $\mathcal{E}_h^{n+1/2}$ the energy given by (15) written with $e_{h,M}^n$ instead of u_h^n, we get, for all $0 \le n \le N$,

$$\sqrt{\mathcal{E}_h^{n+1/2}} \le C_M\,\Delta t\sum_{k=1}^{N}\left(\frac{1-\gamma_M^{-1}}{\Delta t\sqrt{1-\alpha^2}}\left\|\tilde{g}_h^k\right\|_2 + \frac{1}{\sqrt{1-\alpha^2}}\left\|\frac{\tilde{g}_h^k-\tilde{g}_h^{k-1}}{\Delta t}\right\|_2\right) + \frac{C_M}{\sqrt{1-\alpha^2}}\sup_{k\in[1,N]}\left\|\tilde{g}_h^k\right\|_2.$$

with $C_M = 2\sqrt{2}\,\gamma_M\, e^{N(\gamma_M-1)}$ and where $\gamma_0 = 1$, $\gamma_3 = 1$ and using (40),

$$\gamma_1 = 1 + \frac{8\,\Delta t\, c_B}{1 - 4\,\Delta t\, c_B} \quad \text{with} \quad c_B = \frac{4\,\beta^2}{\Delta t} \quad \Rightarrow \quad \gamma_1 = 1 + \frac{32\,\beta^2}{1 - 16\,\beta^2}.$$

Note that γ_1 is greater than one and is bounded because of the assumption $\beta < 1/6$. Using the definition of $\tilde{q}_n^h$ given by (46) we can show the following estimation

$$\sqrt{\mathcal{E}_h^{n+1/2}} \leq C_{M,\alpha}\,\Delta t \sum_{k=1}^{N} \left(\frac{1-\gamma_M^{-1}}{\Delta t} \left\|\tilde{q}_h^k\right\|_2 + \left\| \frac{\tilde{q}_h^k - \tilde{q}_h^{k-1}}{\Delta t} \right\|_2 \right) + C_{M,\alpha} \sup_{k\in[1,N]} \left\|\tilde{q}_h^k\right\|_2. \tag{49}$$

where $C_{M,\alpha}$ is given by

$$C_{M,\alpha} = \frac{C_M}{\sqrt{1-\alpha^2}}\, \|A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}\|_2.$$

Moreover, thanks to Lemmas 4.2 and 4.3, we have

$$\begin{cases} \displaystyle\sup_{k\in[1,N]} \left\|\tilde{q}_h^k\right\|_2 \leq \frac{36}{\sqrt{1-\alpha^2}} \Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2, \\ \displaystyle\Delta t \sum_{k=1}^{N} \left\|\tilde{q}_h^k\right\|_2 \leq \frac{36\,N\,\Delta t}{\sqrt{1-\alpha^2}} \Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2, \\ \displaystyle\Delta t \sum_{k=1}^{N} \left\| \frac{\tilde{q}_h^k - \tilde{q}_h^{k-1}}{\Delta t} \right\|_2 \leq \frac{36\,N\,\Delta t}{\sqrt{1-\alpha^2}} \Delta t \sum_{k=2}^{N} \left\| \frac{\widehat{f}_h^k - 2\widehat{f}_h^{k-1} + \widehat{f}_h^{k-2}}{\Delta t^2} \right\|_2. \end{cases}$$

Combining these estimates with the energy estimate (49), one gets

$$\sqrt{\mathcal{E}_h^{n+1/2}} \leq \widetilde{C}_{M,\alpha} \left[\Big(1 + \frac{1-\gamma_M^{-1}}{\Delta t} N\,\Delta t\Big) \Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2 + N\,\Delta t^2 \sum_{k=2}^{N} \left\| \frac{\widehat{f}_h^k - 2\widehat{f}_h^{k-1} + \widehat{f}_h^{k-2}}{\Delta t^2} \right\|_2 \right],$$

where

$$\widetilde{C}_{M,\alpha} = \frac{36\, C_{M,\alpha}}{\sqrt{1-\alpha^2}}.$$

The result of the theorem is obtained, first using Proposition 3.3 that bounds by below the energy with respect to semi-norms of the errors, second using the recursive definition of the constant $\widetilde{C}_{M,\alpha}, C_{M,\alpha}$ and C_M as well as Γ_M, third using the estimation of

$$\|A_h^{-\frac{1}{2}} D_{h,M} A_h^{-\frac{1}{2}}\|_2$$

given by Lemma 4.1 for the cases $M = 1$ and $M = 3$ or Eq. (42) for the case $M = 0$. Finally, note that because of the assumption on β we have

$$\gamma_1 \leq 1 + 64\beta^2 \quad \text{and} \quad 1 - \gamma_1^{-1} \leq 32\,\beta^2.$$

□

4.2. **Application to wave equations**

Our objective is now to apply the results we have obtained to the specific case mentioned in introduction, namely the wave equation (1) with zero initial data. We consider the family of semi-discrete problem of the form (3) and if one consider the total discretization by the modified equation technique (ME) it is natural to expect - if sought solutions are regular enough - a space-time convergence of order 4 of the solution given by (ME) to the solution of (3). Finally, the convergence of (3) to the continuous solution of (1) is an application of the finite-element convergence theory and to restrict the scope of our analysis we assume convergence at the right order is guaranteed, i.e.,

$$\sup_{t^n\in[0,T]} \|u_h^n - u_h(t^n)\|_2^2 \leq h^4\, C_u.$$

for some $C_u > 0$ independent of h. Such results was proved in [6] in the case $R(x) = 0$ for a family of implicit or explicit fourth order schemes. We assume here that adding dissipation does not deteriorate the convergence if the scheme (ME) is used (such result is suggested by the computation of the truncation error (6) and (7)). In the light of the previous equation we seek for fourth order estimation of the error term $e_{h,M}^n$. The proof given in [6] relies, naturally, on regularity properties of the source term. In our context we need equivalent regularity properties and the corresponding assumption reads

Assumption 2. *For any $T > 0$ and Δt given by* (16) *with $\alpha < 1$ independent of h and $N = \lfloor T/\Delta t \rfloor$, there exists a scalar C_f independent of $h > 0$ such that*

$$\Delta t \sum_{k=1}^{N} \left\| \frac{\widehat{f}_h^k - \widehat{f}_h^{k-1}}{\Delta t} \right\|_2 + \sum_{k=2}^{N} \left\| \frac{\widehat{f}_h^k - 2\widehat{f}_h^{k-1} + \widehat{f}_h^{k-2}}{\Delta t} \right\|_2 \leq C_f.$$

In ordee to obtain more specific results in relation with our applications, we also need to assume a given behavior for the norm of the matrices A_h and B_h. It is clear that $\|A_h\|_2$ is equivalent to h^{-2} when h tends to 0, since it corresponds to the discretization of a second order differential operator. Moreover $\|B_h\|_2$ should be bounded independently of h because it corresponds to the discretization of a zero order differential operator. These observations are turn into assumption in what follows

Assumption 3. *We assume that there exist three scalars c_A, C_A and C_B, independent of h, such that*

$$\frac{c_A}{h^2} \leq \|A_h\|_2 \leq \frac{C_A}{h^2} \text{ and } \|B_h\|_2 \leq C_B.$$

Note that, because of the CFL condition (16), the maximum allowed time step Δt is proportional to h. As a direct consequence, we have that

$$\beta = \beta(h) = \Delta t\, \|B_h\|_2 \leq \alpha \left(\frac{12}{\|A_h\|_2}\right)^{\frac{1}{2}} C_B \leq \alpha\, C_B \left(\frac{12}{c_A}\right)^{\frac{1}{2}} h, \quad \text{hence} \quad \beta(h) \underset{h\to 0}{\longrightarrow} 0.$$

Therefore, for h sufficiently small, $\beta < 1$, and the schemes (EME-M) will be well defined (recall that we have assumed that $\alpha < 1$).

Discussions on the norm of $C_h A_h^{-\frac{1}{2}}$. An estimation of the norm of $C_h A_h^{-\frac{1}{2}}$ is required to deduce a space-time convergence result from Theorem 4.4, since the norm of the operator may

blow-up, which is the case in practice as we show numerically in Section 5. A dedicated estimation of the norm of this operator is out of the scope of this article. Instead, in what follows, we define reasonable assumptions on the behavior of the operator with respect to h, by analogy with the continuous setting. We have the following equality

$$\|C_h A_h^{-\frac{1}{2}}\|_2 = \sup_{u_h, v_h} \frac{|v_h^T C_h A_h^{-\frac{1}{2}} u_h|}{\|v_h\|_2 \|u_h\|_2} = \sup_{w_h, v_h} \frac{|v_h^T C_h w_h|}{\|v_h\|_2 \|A_h^{\frac{1}{2}} w_h\|_2}. \tag{50}$$

Note that $\|A_h^{\frac{1}{2}} w_h\|_2$ can be regarded as some approximation of the H^1-norm of the function represented by w_h. If one considers the wave equation (1), we have that C_h is some approximation of the operator $\mathcal{C} : H_0^1(\Omega) \longrightarrow H^{-1}(\Omega)$ defined, for $R \in W^{1,\infty}$ by

$$\begin{aligned}\langle \mathcal{C}w, v\rangle := \langle R\,\Delta w - \Delta(Rw), v\rangle &= -\langle \nabla\cdot(w\nabla R), v\rangle - \langle \nabla R\cdot\nabla w, v\rangle \\ &= (w, \nabla R\cdot\nabla v)_{L^2(\Omega)} - (\nabla R\cdot\nabla w, v)_{L^2(\Omega)} \end{aligned} \tag{51}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality product in $H_0^1(\Omega)$ and $(\cdot,\cdot)_{L^2(\Omega)}$ is the standard L^2 scalar product. Then, in light of equation (50) one can expect that an estimation of the term

$$\frac{|\langle \mathcal{C}w, v\rangle|}{\|v\|_{L^2(\Omega)}\|w\|_{H^1(\Omega)}} \tag{52}$$

will give a precise idea on how to obtain an estimation of $\|C_h A_h^{-\frac{1}{2}}\|_2$ as soon as accurate enough finite elements approximations are used. We consider below a piecewise smooth function R. To do so, we introduce a partition of Ω into L bounded Lipschitz sub-domains

$$\overline{\Omega} = \bigcup_{\ell=1}^{L} \overline{\Omega}_\ell, \quad \Omega_\ell \cap \Omega_k = \emptyset, \;\; k \neq \ell,$$

and define the following subspace of $W^{1,\infty}$ of piecewise regular functions

$$\widetilde{W}^{1,\infty}(\Omega) = \{u \in W^{1,\infty}(\Omega), u|_{\Omega_\ell} \in C^1(\overline{\Omega}_\ell) \cap W^{2,\infty}(\Omega_\ell), \quad l = 1..L\}.$$

Theorem 4.5. *There exists $C_\Omega > 0$, depending on Ω and on the Ω_ℓ only, such that, if $R \in W^{2,\infty}(\Omega)$, then, for all $(v, w) \in H_0^1(\Omega)^2$*

$$\frac{|\langle \mathcal{C}w, v\rangle|}{\|v\|_{L^2(\Omega)}\|w\|_{H^1(\Omega)}} \leq C_\Omega\big(\|\Delta R\|_{L^\infty(\Omega)} + \|\nabla R\|_{L^\infty(\Omega)}\big), \tag{53}$$

and if $R \in \widetilde{W}^{1,\infty}(\Omega)$

$$\frac{|\langle \mathcal{C}w, v\rangle|}{\|v\|_{L^2(\Omega)}\|w\|_{H^1(\Omega)}} \leq C_\Omega\Big(\|\nabla R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\Delta R\|_{L^\infty(\Omega_\ell)}\Big)\Big(1 + \frac{\|v\|_{H^1(\Omega)}^{\frac{1}{2}}}{\|v\|_{L^2(\Omega)}^{\frac{1}{2}}}\Big). \tag{54}$$

Proof. To prove inequality (53) observe that, if $R \in W^{2,\infty}(\Omega)$, we have

$$\langle \mathcal{C}w, v\rangle = -(\Delta R\, w, v)_{L^2(\Omega)} - 2(\nabla R \cdot \nabla w, v).$$

Two applications of Cauchy-Schwartz inequality lead to the estimation (53). To prove (54), observe that, from the right hand side of (51) we have, using Green's formula on each Ω_ℓ,

$$\langle \mathcal{C}w, v\rangle = -(\nabla R \cdot \nabla w, v)_{L^2(\Omega)} - \sum_{\ell=1}^{L} \big((\nabla R \cdot \nabla w, v)_{L^2(\Omega_\ell)} + (\Delta R\, w, v)_{L^2(\Omega_\ell)}\big) + \sum_{\ell=1}^{L} (\nabla R \cdot n\, w, v)_{L^2(\partial\Omega_\ell)},$$

where n on $\partial\Omega_\ell$ denotes the outward unitary normal to Ω_ℓ. Remark that the function $\nabla R \cdot n$ is well defined on $\partial\Omega_\ell$ and

$$\|\nabla R \cdot n\|_{L^\infty(\partial\Omega_\ell)} \le \|\nabla R\|_{L^\infty(\Omega_\ell)}.$$

We can then deduce the following estimation

$$|\langle \mathcal{C}w, v\rangle| \le 2\|\nabla R\|_{L^\infty(\Omega)}\|\nabla w\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|w\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \sum_{\ell=1}^{L} \|\Delta R\|_{L^\infty(\Omega_\ell)} + \sum_{\ell=1}^{L} \|\nabla R\|_{L^\infty(\Omega_\ell)}\|w\|_{L^2(\partial\Omega_\ell)}\|v\|_{L^2(\partial\Omega_\ell)}. \quad (55)$$

The last term of this estimation can be bounded using the following inequality, which is a direct consequence of Theorem 1.5.1.10 of [21].

There exists C_ℓ depending on Ω_ℓ only such that for all $u \in H^1(\Omega_\ell)$ we have

$$\|u\|_{L^2(\partial\Omega_\ell)} \le C_\ell \|u\|_{H^1(\Omega_\ell)} \quad \text{and} \quad \|u\|_{L^2(\partial\Omega_\ell)} \le C_\ell\, \|u\|^{\frac{1}{2}}_{H^1(\Omega_\ell)} \|u\|^{\frac{1}{2}}_{L^2(\Omega_\ell)}.$$

Using these inequalities into (55), one can show that there exists a constant C_Ω, independent of v, w and R, such that

$$|\langle \mathcal{C}w, v\rangle| \le C\Big(\|\nabla R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\Delta R\|_{L^\infty(\Omega_\ell)}\Big)\|w\|_{H^1(\Omega)}\big(\|v\|_{L^2(\Omega)} + \|v\|^{\frac{1}{2}}_{H^1(\Omega)}\|v\|^{\frac{1}{2}}_{L^2(\Omega)}\big).$$

Estimation (54) follows then easily. □

Theorem 4.5 shows that one can expect $\|C_h A_h^{-\frac{1}{2}}\|_2$ to be bounded by a term proportional to

$$\Big(\|\nabla R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\Delta R\|_{L^\infty(\Omega_\ell)}\Big)\left(1 + \sup_{\|v_h\|_2=1} \|A_h^{\frac{1}{2}} v_h\|_2^{\frac{1}{2}}\right), \quad (56)$$

which leads to a conclusion remarking that, thanks to Assumption 3 we have,

$$\|A_h^{\frac{1}{2}} v_h\|_2 \leq \frac{\sqrt{C_A}}{h}\|v_h\|_2.$$

Moreover, if we introduce the space

$$\widetilde{L}^\infty(\Omega) = \{u \in L^\infty(\Omega), u|_{\Omega_\ell} \in C^0(\overline{\Omega}_\ell) \cap W^{1,\infty}(\Omega_\ell), \quad l = 1..L\},$$

then, in view of (56), it appears that one can extend the results obtained in the case $R \in \widetilde{L}^\infty(\Omega)$. Considering a well chosen smooth function $R_h \in \widetilde{W}^{1,\infty}(\Omega)$ that converges towards R in some well chosen norm, then, by inverse inequality (see Section 4.5 of [3]) one can expect to "lose" a power of h. To sum-up, it seems reasonable, to perform the following assumption, which will be confirmed by numerical results in section 5.1).

Assumption 4. *We assume that there exists a constant C_{BA}, independent of h, and $0 \leq r \leq 3/2$, such that*

$$\|C_h A_h^{-\frac{1}{2}}\|_2 \leq \frac{C_{BA}}{h^r}.$$

Remark 2. *Note that it is expected that if R belongs to $W^{2,\infty}(\Omega)$, the Assumption 4 holds with $r = 0$, if $R \in \widetilde{W}^{1,\infty}(\Omega)$, then Assumption 4 holds with $r = 1/2$ and finally if $R \in \widetilde{L}^\infty(\Omega)$, then Assumption 4 holds with $r = 3/2$.*

Discussions on the norm of $A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}$. As previously the norm of $A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}$ must be estimated in order to deduce a space-time convergence result from Theorem 4.4 in the case $M = 0$. Following the above discussion, we end-up to the following assumption by analogy with the continuous setting.

Assumption 5. *We assume that there exists a constant C_{BA} independent of h and $0 \leq r \leq 1/2$ such that*

$$\|A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}\|_2 \leq \frac{C_{BA}}{h^r}.$$

Indeed, with the same arguments as before, one can see that an estimation of the norm $A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}$ should be obtained from an estimation of

$$\frac{|\langle \mathcal{C} w, v\rangle|}{\|w\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}} \tag{57}$$

and we have the following theorem.

Theorem 4.6. *There exists $C_\Omega > 0$, depending on Ω and on the Ω_ℓ only, such that, if $R \in W^{1,\infty}(\Omega)$, then, for all $(v, w) \in H_0^1(\Omega)^2$*

$$\frac{|\langle \mathcal{C} w, v\rangle|}{\|v\|_{L^2(\Omega)}\|w\|_{H^1(\Omega)}} \leq C_\Omega \|\nabla R\|_{L^\infty(\Omega)}, \tag{58}$$

and, if $R \in \widetilde{L}^\infty(\Omega)$ *and* $(v,w) \in \left(H_0^1(\Omega)\cap H^2(\Omega)\right)^2$,

$$\frac{|\langle \mathcal{C}w, v\rangle|}{\|w\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}} \leq C_\Omega \left(\|R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\nabla R\|_{L^\infty(\Omega_\ell)}\right)\left(1 + \frac{\|w\|^{\frac{1}{2}}_{H^2(\Omega)}}{\|w\|^{\frac{1}{2}}_{H^1(\Omega)}} + \frac{\|v\|^{\frac{1}{2}}_{H^2(\Omega)}}{\|v\|^{\frac{1}{2}}_{H^1(\Omega)}}\right). \quad (59)$$

where $\langle\cdot,\cdot\rangle$ *denotes here the duality product in* $H_0^1(\Omega)\cap H^2(\Omega)$.

Proof. From (51), one can see that if $R \in W^{1,\infty}(\Omega)$, then estimation (58) is a consequence of Cauchy-Schwartz inequality. Moreover, assuming $R \in L^\infty(\Omega)$, $(v,w) \in \left(H_0^1(\Omega)\cap H^2(\Omega)\right)^2$ we have

$$R\,\Delta w - \Delta(Rw) \in \left(H_0^1(\Omega)\cap H^2(\Omega)\right)'$$

and one can deduce that

$$\langle \mathcal{C}w, v\rangle = -(R\Delta v, w)_{L^2(\Omega)} + (R\Delta w, v)_{L^2(\Omega)}.$$

Then, if $R \in \widetilde{L}^\infty(\Omega)$, introducing the partition of Ω into sub-domains Ω_ℓ, one can use Green's formulae on each sub-domains Ω_ℓ to obtain

$$\langle \mathcal{C}w, v\rangle = \sum_{\ell=1}^{L} (\nabla R\cdot\nabla v, w)_{L^2(\Omega_\ell)} - (\nabla R\cdot\nabla w, v)_{L^2(\Omega_\ell)} + \sum_{\ell=1}^{L} (R\,\nabla w\cdot n, v)_{L^2(\Omega_\ell)} - (R\,\nabla v\cdot n, w)_{L^2(\Omega_\ell)}. \quad (60)$$

Then, using standard trace inequalities (see again Theorem 1.5.1.10 of [21]), one can show that

$$|(R\,\nabla w\cdot n, v)_{L^2(\Omega_\ell)}| \leq C_\ell \,\|R\|_{L^\infty(\Omega_\ell)}\,\|w\|^{\frac{1}{2}}_{H^2(\Omega)}\,\|w\|^{\frac{1}{2}}_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)}.$$

Therefore, using the above estimation in (60) together with standard Cauchy-Schwartz inequality, we have

$$\begin{aligned}|\langle \mathcal{C}w, v\rangle| \leq C_\Omega \left(\|R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\nabla R\|_{L^\infty(\Omega_\ell)}\right) \\ \cdot \Big(\|\nabla v\|_{L^2(\Omega)}\,\|w\|_{L^2(\Omega)} + \|\nabla w\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} \\ + \|w\|^{\frac{1}{2}}_{H^2(\Omega)}\,\|w\|^{\frac{1}{2}}_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)} + \|v\|^{\frac{1}{2}}_{H^2(\Omega)}\,\|v\|^{\frac{1}{2}}_{H^1(\Omega)}\,\|w\|_{H^1(\Omega)}\Big),\end{aligned}$$

from which we deduce the estimation (59). □

Using the result of Theorem 4.6, we expect the norm of $A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}$ to be proportional to

$$\left(\|R\|_{L^\infty(\Omega)} + \sum_{\ell=1}^{L} \|\nabla R\|_{L^\infty(\Omega_\ell)} \right) \left(2 + \frac{\|A_h w_h\|_2^{\frac{1}{2}}}{\|A_h^{\frac{1}{2}} w_h\|_2^{\frac{1}{2}}} + \frac{\|A_h v_h\|_2^{\frac{1}{2}}}{\|A_h^{\frac{1}{2}} v_h\|_2^{\frac{1}{2}}} \right),$$

if one admits that $\|A_h w_h\|_2$ is an approximation of the $H^2(\Omega)$-norm of the function corresponding to w_h. This justifies Assumption 5 since

$$\frac{\|A_h w_h\|_2^{\frac{1}{2}}}{\|A_h^{\frac{1}{2}} w_h\|_2^{\frac{1}{2}}} \leq \|A_h^{\frac{1}{2}}\|_2^{\frac{1}{2}} \leq \frac{C_A^{\frac{1}{4}}}{h^{\frac{1}{2}}}.$$

Remark 3. *Note that it is expected that if R belongs to $W^{1,\infty}(\Omega)$ then Assumption 5 holds with $r = 0$, if $R \in \widetilde{L}^\infty(\Omega)$ then Assumption 4 holds with $r = 1/2$.*

Main result. Our main result is obtained as a corollary of Theorem 4.4 that takes into account the assumptions 2 to 5.

Corollary 4.7. *Let $T > 0$ and Δt given by (16) with $\alpha < 1$ independent of h and $N = \lfloor T/\Delta t \rfloor$. Let assumptions 2, 3, 4, and 5 hold. Assume that*

$$u_h^0 = u_h^1 = 0, \quad u_{h,M}^0 = u_{h,M}^1 = 0, \quad \widehat{f}_h^0 = \widehat{f}_h^1 = \widehat{f}_h^2 = 0.$$

Then, there exists $C > 0$ such that for h sufficiently small we have

$$\left\| A_h^{\frac{1}{2}} e_{h,M}^{n+1} \right\|_2^2 \leq C\,(1+T)\, m(h; M, r),$$

where $m(h; M, r)$ are decreasing functions of h given by

$$m(h; 0, r) = h^{2-r} \quad \textit{for} \quad 0 \leq r \leq 1/2$$

and

$$m(h; 1, r) = e^{C\,T\,h}\, h^{5-2r} \quad \textit{and} \quad m(h; 3, r) = h^{7-2r} \quad \textit{for } 0 \leq r \leq 3/2.$$

Proof. Note that, for h sufficiently small, we have $\beta < 1/6$ and we can apply Theorem 4.4. This shows that

$$\left(\frac{1}{4} \left\| \frac{e_{h,M}^{n+1} - e_{h,M}^n}{\Delta t} \right\|_2^2 + (1-\alpha^2) \left\| A_h^{\frac{1}{2}} \frac{e_{h,M}^{n+1} + e_{h,M}^n}{2} \right\|_2^2 \right)^{\frac{1}{2}} \leq a_{h,M} \big(b_{h,M}\, C_f + T\, C_f \big), \tag{61}$$

where we have used Assumption 2 to estimate the source term contribution and where the value of $a_{h,M}$ and $b_{h,M}$ are given by Theorem 4.4. Using Assumptions 4 and 5, the right hand side of (61) can be bounded by $C\,(1+T)\, m(h; M, r)$ for some constant C independent of h. Then, one can observe that

$$A_h^{\frac{1}{2}} e_{h,M}^{n+1} = A_h^{\frac{1}{2}} \frac{e_{h,M}^{n+1} + e_{h,M}^n}{2} - \frac{\Delta t}{2} A_h^{\frac{1}{2}} \frac{e_{h,M}^n - e_{h,M}^{n+1}}{\Delta t}$$

and because of the CFL condition (16) we have $\Delta t \, \|A_h^{\frac{1}{2}}\|_2 < 12$. Therefore, by the triangular inequality,

$$\|A_h^{\frac{1}{2}} e_{h,M}^{n+1}\|_2 \leq \left\| A_h^{\frac{1}{2}} \frac{e_{h,M}^{n+1} + e_{h,M}^{n}}{2} \right\|_2 + 6 \left\| \frac{e_{h,M}^{n} - e_{h,M}^{n+1}}{\Delta t} \right\|_2 .$$

Then, up to the definition of another constant $C > 0$, we can deduce from equation (61) that

$$\|A_h^{\frac{1}{2}} e_{h,M}^{n+1}\|_2 \leq C \, (1+T) \, m(h; M, r)$$

which concludes the proof of the Corollary. □

5. Numerical results

For the numerical investigation of the schemes (EME-M), we use the following 1D dissipative wave equation as a model problem.

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + R(x)\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = f, & x \in (0,1), \quad t \in (0,T), \\ u = 0, & x \in \{0,1\}, \quad t \in (0,T), \\ u = 0, \quad \dfrac{\partial u}{\partial t} = 0, & x \in (0,1), \quad t = 0. \end{cases} \tag{62}$$

The source term f is given by

$$f(x,t) = e^{-\left(\frac{x-x_0}{r_0}\right)^2} e^{\alpha_0 \frac{u(t)}{u(t)-1}}, \qquad u(t) = \left(\frac{t-t_0}{\tau}\right)^2. \tag{63}$$

In practice, we set $x_0 = 0.8$, $r_0 = 0.025$, $t_0 = 0.3$, $\tau = 0.08$ and $\alpha_0 = 100$. In the following, we investigate two different behaviors for the dissipative function $R(x)$: either $R(x) = R_0(x) \in \widetilde{L}^\infty(\Omega)$, $R(x) = R_1(x) \in \widetilde{W}^{1,\infty}(\Omega)$, or $R(x) = R_2(x) \in C^\infty(\Omega)$, with R_0 and R_1 given by

$$R_0(x) = \sigma \, \mathbf{1}_{[0,x_1]}(x), \quad R_1(x) = \sigma \, \mathbf{1}_{[x_1-r_1, x_1+r_1]}(x) \frac{r_1 - |x - x_1|}{r_1}$$

and R_2 given by

$$R_2(x) = \sigma \, \mathbf{1}_{[-1,1]} s(x) \exp\left(\alpha_1 + \frac{\alpha_1}{s(x)^2 - 1}\right), \quad s(x) = \frac{x - x_1}{r_1}.$$

In the above expressions, we set $r_1 = 0.2$, $x_1 = 0.3$ and $\alpha_1 = 10$. The positive scalar σ corresponds to the maximum value of the dissipation profile and is a parameter for the numerical investigation.

The discretization in space is done using fourth order spectral finite elements method [7] leading to a diagonal mass matrix.

5.1. Numerical investigations of $\|C_h A_h^{-\frac{1}{2}}\|_2$ and $\|A_h^{-\frac{1}{2}} C_h A_h^{-\frac{1}{2}}\|_2$

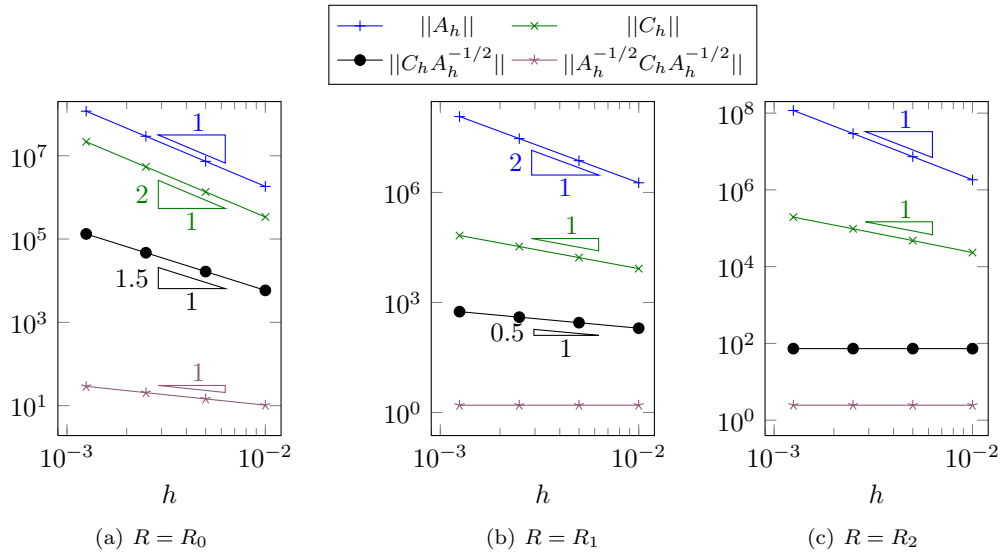
In this section, we investigate numerically the validity of Assumption 4 and Assumption 5. To do so we assemble finite element matrices constructed with fourth order finite elements on a uniform mesh of $[0,1]$. Denoting $h = 1/N$ the space step, we display in Figure 1 the euclidean norm of

$$A_h,\ C_h,\ C_h A_h^{-1/2} \text{ and } A_h^{-1/2}\, C_h\, A_h^{-1/2}$$

for decreasing values of h, with the three different dissipation profiles introduced before (with $\sigma = 1$).

We observe a perfect agreement with Assumption 4 and Remark 2 by looking at the black curves with plain circles $\left\|C_h A_h^{-1/2}\right\|$. If $R \in \widetilde{L}^\infty(\Omega)$, we observe in Fig. 1(a) that Assumption 4 holds with $r = 3/2$; if $R \in \widetilde{W}^{1,\infty}(\Omega)$, we observe in Fig. 1(b) that Assumption 4 holds with $r = 1/2$; and if R belongs to $W^{2,\infty}(\Omega)$, we observe in Fig. 1(c) that Assumption 4 holds with $r = 0$.

Finally, we observe a perfect agreement with Assumption 5 and Remark 3 by looking at the red curves with stars $\left\|A_h^{-1/2} C_h A_h^{-1/2}\right\|$ If $R \in \widetilde{L}^\infty(\Omega)$, we observe in Fig. 1(a) that Assumption 5 holds with $r = 1/2$. and if R belongs to $W^{1,\infty}(\Omega)$ or $W^{2,\infty}(\Omega)$, we observe in Fig. 1(c) and Fig. 1(b) that Assumption 5 holds with $r = 0$.



(a) $R = R_0$ (b) $R = R_1$ (c) $R = R_2$

FIGURE 1. Norms of A_h, C_h, $C_h A_h^{-1/2}$ and $A_h^{-1/2} C_h A_h^{-1/2}$ with respect to h with $\sigma = 1$ on a uniform mesh of $[0,1]$ using fourth order finite elements.

5.2. Convergence of (EME-M) towards (ME)

In this section, we illustrate numerically the results of Theorem 4.4 in the context of wave equations, for which we have investigated the values of r in Assumption 4 and Assumption 5. All

these results are summed up in Corollary 4.7. In Fig. 2 are displayed the H^1-norms of the error between the solution to (ME) and the solution to (EME-0), (EME-1) and (EME-3), as h goes to zero. We keep the CFL-number α equal to 0.9, therefore Δt is asymptotically proportional to h. We choose to test two different values of $\sigma \in \{1, 100\}$ for each dissipation profile R_i, $i \in \{0, 1, 2\}$. The parameters of the simulations are the same as in the previous subsection, and the final time is set to $T = 3$.

First, let us observe the convergence of scheme (EME-0), displayed with + signs in blue in Fig. 2. Corollary 4.7 predicts a convergence in h^{2-r}, with r depending on the regularity of the dissipation profile. The observed convergence are in perfect agreement with the corollary and Remark 3. We observe second order convergence for the profiles R_2 and R_1, which was expected since $r = 0$ for R_2 and R_1, which belong to $W^{1,\infty}(\Omega)$. For the profile R_0, we observe a convergence at order 3/2, which was also expected since $r = 1/2$ for R_0, which belongs to $\widetilde{L}^{\infty}(\Omega)$.

Now, let us focus on the convergence of scheme (EME-1), displayed with $\circ$ signs in red in Fig. 2. Corollary 4.7 predicts a convergence in $e^{CTh}h^{5-2r}$ with r depending on the regularity of the dissipation profile. The observed convergence are again in perfect agreement with the corollary and Remark 2. We observe a convergence at order 5 for the profile R_2, as expected, since $r = 0$ for $R \in W^{2,\infty}(\Omega)$. For the profile R_1, we observe a convergence at order 4, as expected, since $r = 1/2$ for R_1, which belongs to $\widetilde{W}^{1,\infty}(\Omega)$. Finally, for the profile R_0, we observe a convergence at order 2, again as expected, since $r = 3/2$ for R_0, which belongs to $\widetilde{L}^{\infty}(\Omega)$.

Let us then observe the convergence of scheme (EME-3) displayed with $\bullet$ signs in black in Fig. 2. Corollary 4.7 predicts a convergence in h^{7-2r} with r depending on the regularity of the dissipation profile. The computed values of the errors are very small (close to machine precision), which illustrates that this scheme is very accurate. Consequently, for these values of dissipation amplitudes, it is difficult to assess the asymptotical regime, except for the profile R_0 with $\sigma = 100$. In this last case, the observed convergence happens at order 4, which is in perfect agreement with the corollary and Remark 2 since $r = 3/2$ for R_0, which belongs to $\widetilde{L}^{\infty}(\Omega)$.

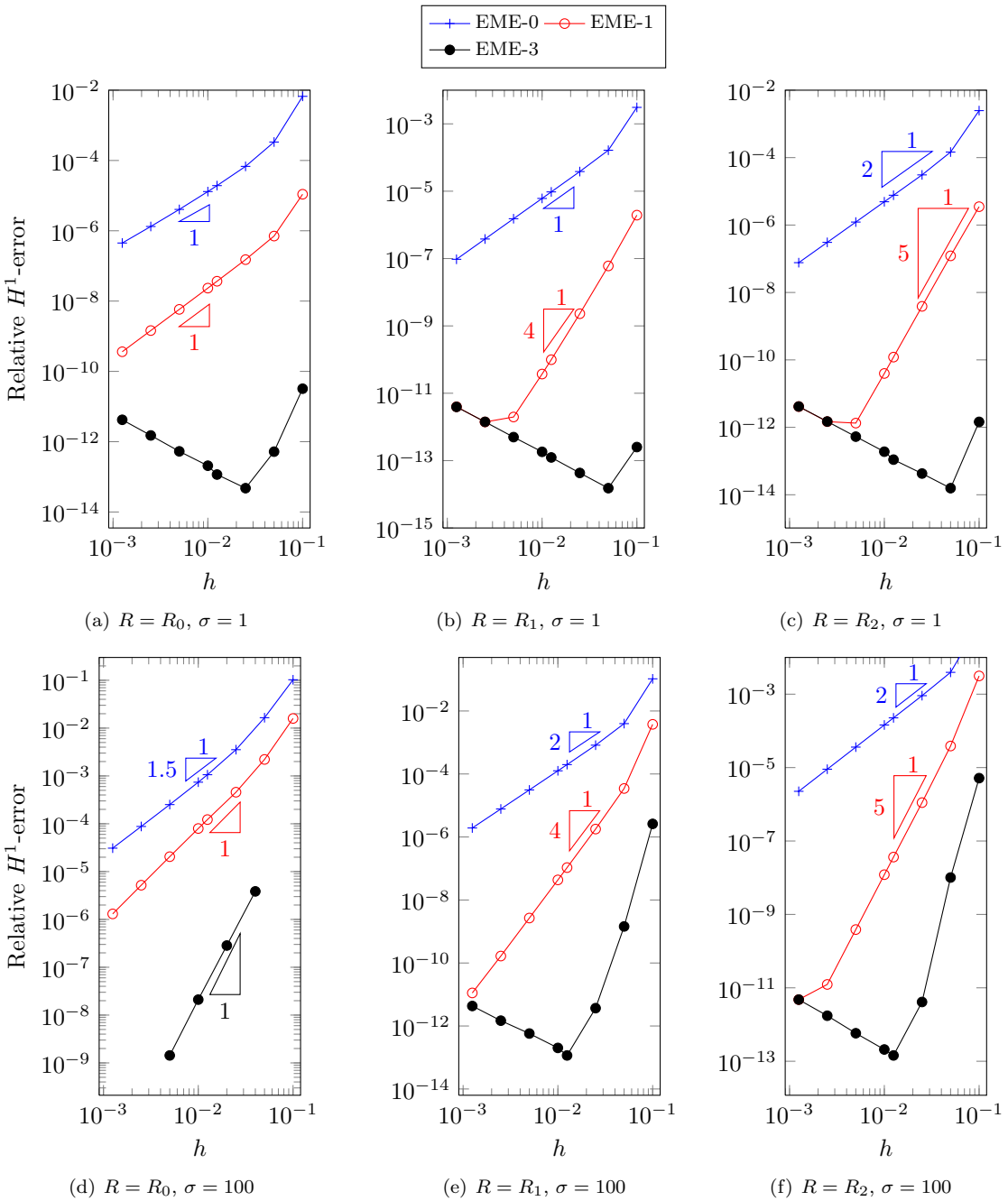
Finally, we have represented in Figure 3 the convergence of scheme (EME-3) for large values of σ with a coarse discretization in the cases $R = R_1$ and $R = R_2$, in order to observe the convergence regime. We see that in both cases the convergence is better than a convergence of order 6 which is in agreement with Corollary 4.7 and Remark 2 when considering dissipation profiles in $\widetilde{W}^{1,\infty}(\Omega)$.

5.3. **Space/Time convergence analysis**

For the space/time convergence analysis, we choose N larger and larger ($N \in [10, 400]$) and we choose 99% of the largest allowed time step, i.e. $\alpha = 0.99$. Four different schemes are compared:

- The Leap Frog scheme (LF). The maximum allowed time step is $2/\sqrt{\rho(A_h)}$.
- The Modified Equation scheme (ME). The maximum allowed time step is $2\sqrt{3}/\sqrt{\rho(A_h)}$.
- The explicit Modified Equation scheme (EME-M). The same time step than for (ME) is used.
- The explicit 4th order Runge-Kutta scheme (RK4). Without dissipation, the maximum allowed time step is $2\sqrt{2}/\sqrt{\rho(A_h)}$. With dissipation, the maximum allowed time step must be implicitly deduced from the stability region and the spectral properties of the semi-discretized equation.

Remark 4. *The spectral radius of symmetric real matrices can be efficiently computed using the power iteration algorithm. No such simple algorithm exists to estimate if the eigenvalues of non*



(a) $R = R_0$, $\sigma = 1$ (b) $R = R_1$, $\sigma = 1$ (c) $R = R_2$, $\sigma = 1$

(d) $R = R_0$, $\sigma = 100$ (e) $R = R_1$, $\sigma = 100$ (f) $R = R_2$, $\sigma = 100$

FIGURE 2. Convergence of schemes EME-0,1 and 3 to (ME) w.r.t. the space step h.

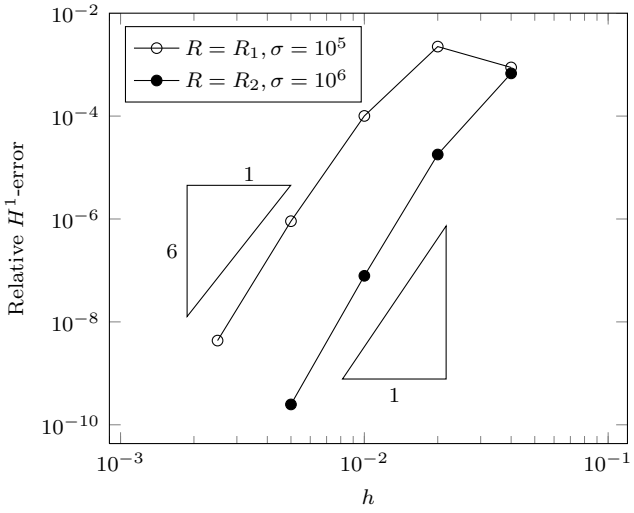


FIGURE 3. Convergence of schemes EME-3 to (ME) w.r.t. the space step h for $R = R_1$ or $R = R_2$ with σ large.

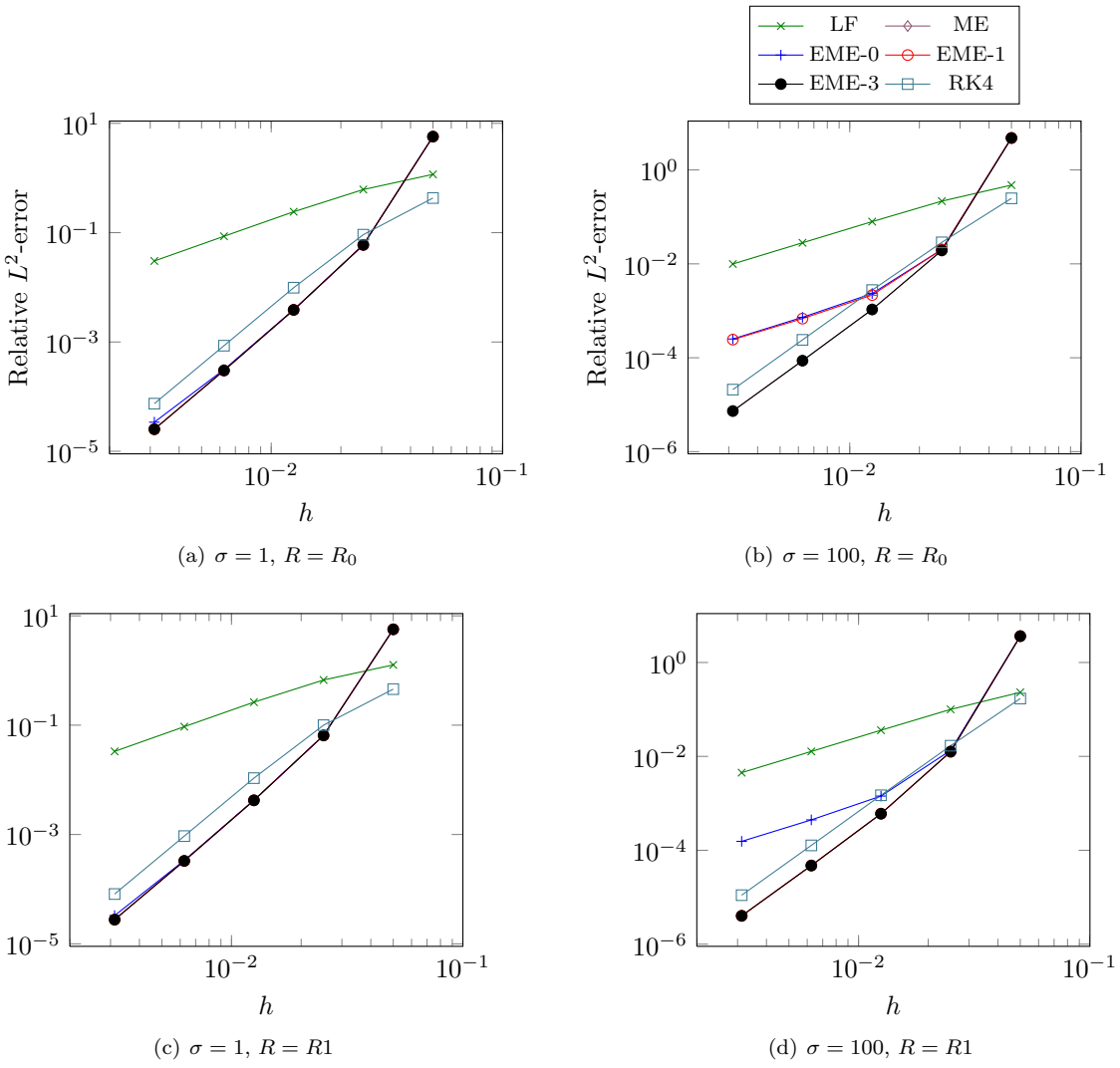
symmetric matrices lie in a given region of the complex plane. This is why, in practice, for Runge-Kutta methods, one tries a time step and reduces it in case of numerical instability. Note however that, in the examples presented below, Runge Kutta schemes were stable even in the presence of dissipation when the time step was was set to $2\sqrt{2}/\sqrt{\rho(A_h)}$.

We assume that initial data are zero and that the source term is a compactly supported regular function in space and time. A reference solution is computed using an explicit 4th order Runge-Kutta scheme on a fine grid. We plot in Figure 4(a) and 4(b) the relative L^2 error in space, L^∞ in time obtained when $R = R_0$ and $\sigma \in \{1, 100\}$. In Figure 4(c) and 4, we plot the error obtained when $R = R_1$ and $\sigma \in \{1, 100\}$.

We observe that, if the damping profile is smooth enough, then the explicit modified equation scheme (EME-1) gives results similar to the modified equation scheme (ME). This nice property is lost if the damping profile is discontinuous and we expect that a loss of space/time convergence could occur in this case. Note however that the method, in the tested parameter range, gives relatively accurate results. Since, in realistic applications, we expect the damping amplitude to be small, we believe that even in the case of discontinuous damping profiles the scheme (EME-1) is interesting. To assess more accurately the efficiency of our scheme, we plot on Figure 5(b) and 4(b) the error obtained with respect to the complexity of the algorithm. We compare the Leap Frog scheme, the explicit Modified Equation scheme and the Runge-Kutta 4 scheme applying the following rule: for each iteration in time, we count the number of multiplications by the sparse finite element matrices (A_h, C_h, B_h and M_h). For each scheme we obtain

(LF) Only one operation is counted: 1 multiplication by A_h.

(EME-M) 2+M operations are counted: 2 multiplications by A_h and M by C_h. We have assumed that a multiplication by C_h is as costly as a multiplication by A_h which is a pessimistic estimation in regards of the presented numerical results and for the implementation choice we made.



(a) $\sigma = 1$, $R = R_0$

(b) $\sigma = 100$, $R = R_0$

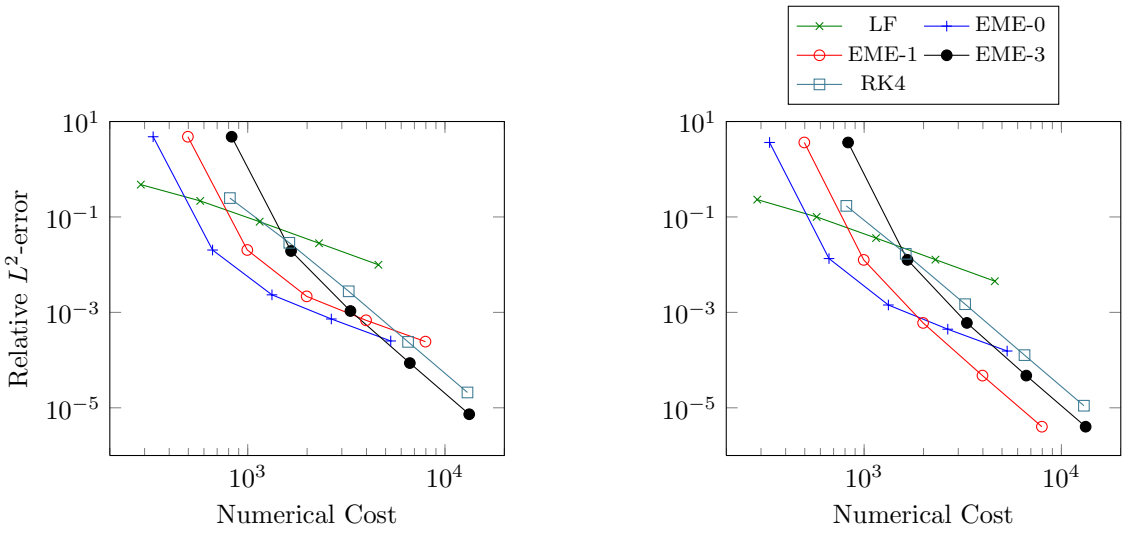
(c) $\sigma = 1$, $R = R1$

(d) $\sigma = 100$, $R = R1$

FIGURE 4. Relative error for LF, ME, EME, RK4, w.r.t. the space step h.

(RK4) At each stage of the 4 stages method, a multiplication by A_h needs to be computed: 4 multiplications by A_h are counted in total.

Note that the Modified Equation is not compared since it requires the inversion of a matrix, that could be done efficiently by iterative methods, but whose complexity is hard to assess.



(a) $R = R_0$, $\sigma = 100$. (b) $R = R_1$, $\sigma = 100$

FIGURE 5. Relative error for LF, EME, RK4, w.r.t. the numerical cost.

APPENDIX A. EXTENSION OF THE METHOD: APPLICATION TO LORENTZ'S MATERIALS

We are interested in the propagation of electromagnetic waves in Lorentz's material that are governed by the following system of equations (see [19] section 4.4) set in a domain Ω

$$\begin{cases} \varepsilon\, \partial_t^2 E + \varepsilon\, \Omega_e^2 \partial_t^2 P + \mathrm{rot}\, \mu^{-1}\, \mathrm{rot} E = 0 & \text{(64a)} \\ \partial_t^2 P + \alpha_e \partial_t P - E = 0 & \text{(64b)} \end{cases}$$

where ε and μ are the permittivity and the permeability constants, and the vector field P is an additional variable that accounts for non-local dissipative and dispersive effects in the material. These effects are governed by the non negative parameters Ω_e (responsible for dispersion) and α_e (responsible for losses). Assuming adequate boundary conditions along $\partial\Omega$ ($E \times n = 0$ for instance), it is standard to show that the following energy identity holds

$$\frac{1}{2}\frac{d}{dt}\left(\int_\Omega \varepsilon |\partial_t E|^2 d\mathbf{x} + \int_\Omega \mu^{-1} |\mathrm{rot}\, E|^2 d\mathbf{x} + \int_\Omega \varepsilon \Omega_e^2 |\partial_t^2 P|^2 d\mathbf{x},\right) = -\int_\Omega \varepsilon \alpha_e \Omega_e^2 |\partial_t^2 P|^2 d\mathbf{x}.$$

Setting $Q = \Omega_e\, \partial_t P$, one obtains

$$\begin{cases} \varepsilon\, \partial_t^2 E + \varepsilon\, \Omega_e \partial_t Q + \mathrm{rot}\, \mu^{-1}\, \mathrm{rot} E = 0, \\ \varepsilon\, \partial_t^2 Q - \varepsilon\, \Omega_e \partial_t E + \varepsilon\, \alpha_e \partial_t Q = 0. \end{cases} \tag{65}$$

Note that a discretization of system (65) would lead to an algebraic system of ODEs of the form given by (5), but the matrix B_h would be not symmetric. Even if our analysis was done for the

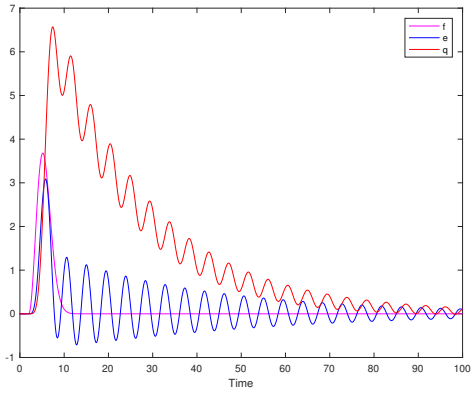


FIGURE 6. Source term and solution of system (66) (with zero initial condition) w.r.t. time.

case where B_h is symmetric, algorithm (EME-0) does not rely on the positivity properties of B_h. We test our scheme on the corresponding simple ODE problems

$$\begin{cases} \varepsilon\,\partial_t^2 e + \varepsilon\,\Omega_e \partial_t q + \lambda^2 \mu^{-1} e = f, \\ \varepsilon\,\partial_t^2 q - \varepsilon\,\Omega_e \partial_t e + \varepsilon\,\alpha_e \partial_t q = 0, \end{cases} \tag{66}$$

where $e(t)$ and $q(t)$ are the scalar unknowns, $f(t)$ is the source term (see Figure 6) and we choose zero initial condition for both variables. In the following, we represent the obtained results, as well as the relative error with respect to time for different values of parameters: following [19], we choose $\varepsilon = \mu = 1$, $\Omega_e = 1$, $\alpha_e = 0.1$ and $\lambda = 1$. For the discretization parameters, we set $\Delta t = 0.05$ and $T = 100$. The reference solution is obtained using the Leap Frog scheme with a time step 1000 times smaller. We compare the solution obtained with the Leap Frog scheme, the Modified Equation scheme and the explicit Modified Equation scheme (EME-1). The relative errors on $e(t)$ in the supremum norm over $t \in [0, T]$ are 8.6×10^{-4} for the (LF) scheme, 2.5×10^{-7} for the (ME) scheme and 1.3×10^{-7} for the (EME-1) scheme. A plot of the solution and the source term are given in Figure 6.
The obtained numerical results show the applicability of the method for dispersive equations like the propagation of electromagnetic waves in Lorentz Material. However, the stability analysis is more complicated since the energy relation can not be easily obtained for the (ME) scheme or the (EME-1) scheme applied to system (66).

APPENDIX B. STABILITY OF (EME-1) BY EIGENVALUE ANALYSIS

The previously obtained result shows that one may obtain exponential increasing behavior of the solution in the energy norm. Our estimate ensures that if β is proportional to Δt, this exponential instability in time is of the form $\exp(C\,T\Delta t)$ (with C a positive scalar independent of Δt) . Although it can be satisfactory for some applications, we show in what follows that this result can be improved asymptotically for small Δt or if $C_B = \|B_h\|_2$ is small enough. For the following

analysis we introduce the notations

$$\widehat{B}_h = \frac{B_h}{C_B}, \quad \varepsilon = C_B \Delta t.$$

Our result is based upon an asymptotic analysis in the parameter ε and is inspired from the work of [17] and of [18]. We consider the homogeneous version of the algorithm (EME-1) (i.e. without any source term) written as a first order system. To do so we introduce the variable $v_h^{n+1} = u_h^n$. One can see that the unknowns (u_h^{n+1}, v_h^{n+1}) are computed using (u_h^n, v_h^n). More precisely,

$$\begin{pmatrix} u_h^{n+1} \\ v_h^{n+1} \end{pmatrix} = \mathcal{A}_h(\varepsilon) \begin{pmatrix} u_h^n \\ v_h^n \end{pmatrix}. \tag{67}$$

where we get, from the scheme (EME-1), that

$$\mathcal{A}_h(\varepsilon) = \begin{pmatrix} D_h(\varepsilon) & 0 \\ 0 & I_h \end{pmatrix} \begin{pmatrix} 2\Big(I_h + \dfrac{\varepsilon^2}{12}\widehat{B}_h^2\Big) - \Delta t^2 \widetilde{A}_h & -2\Big(I_h + \dfrac{\varepsilon^2}{12}\widehat{B}_h^2\Big) \\ I_h & 0 \end{pmatrix} + \begin{pmatrix} 0 & I_h \\ 0 & 0 \end{pmatrix},$$

with

$$\begin{cases} D_h(\varepsilon) = \widetilde{M}_h^{-1}(\varepsilon) - \varepsilon \dfrac{\Delta t^2}{24} \widetilde{M}_h^{-1}(\varepsilon)\, \widehat{C}_h\, \widetilde{M}_h^{-1}(\varepsilon), \\ \widetilde{M}_h(\varepsilon) = I_h + \dfrac{\varepsilon}{2}\widehat{B}_h + \dfrac{\varepsilon^2}{12}\widehat{B}_h^2, \end{cases} \quad \text{and} \quad \begin{cases} \widehat{C}_h = \Big(\widehat{B}_h A_h - A_h \widehat{B}_h\Big), \\ \widetilde{A}_h = A_h \left(I_h - \dfrac{\Delta t^2}{12} A_h\right). \end{cases}$$

Note that we have

$$\mathcal{A}_h(0) = \begin{pmatrix} 2\, I_h - \Delta t^2\, \widetilde{A}_h & -I_h \\ I_h & 0 \end{pmatrix}.$$

The eigenvalues of $\mathcal{A}_h(0)$ can be deduced from the eigenvalues of A_h. Let us denote by λ_i^h a positive eigenvalue of A_h and by ϕ_i^h a corresponding eigenvector. Since the matrix A_h is real and symmetric, the vectors $\{\phi_i^h\}$ are real and can be chosen as an orthonormal basis of $\mathbb{R}^{N_h}$ for the euclidian norm (N_h is the number of degrees of freedom of the underlying finite element space V_h). In particular, this implies that all eigenvalues are semi-simple (their algebraic multiplicity corresponds to their geometric multiplicity). Now, we first remark that we have, for each $i \in \{1, \cdots, N_h\}$,

$$\Delta t^2\, \widetilde{A}_h\, \phi_i^h = \mu_i^h\, \phi_i^h, \quad \text{with} \quad \mu_i^h(\Delta t) = \Delta t^2\, \lambda_i^h \big(1 - \frac{\Delta t^2\, \lambda_i^h}{12}\big).$$

Thus, the set $\{\phi_i^h\}$ is also a basis of eigenvectors of $\Delta t^2\, \widetilde{A}_h$ associated to the eigenvalues $\{\mu_i^h\}$. Then, it can be shown that, for each $i \in \{1, \cdots, N_h\}$, such that $\mu_i^h \neq 0$ and $\mu_i^h \neq 4$, we have

$$\mathcal{A}_h(0) \begin{pmatrix} \phi_i^h \\ \dfrac{1}{\eta_i^h}\phi_i^h \end{pmatrix} = \eta_i^h \begin{pmatrix} \phi_i^h \\ \dfrac{1}{\eta_i^h}\phi_i^h \end{pmatrix} \text{ with } \eta_i^h = \eta_{i,+}^h \text{ or } \eta_i^h = \eta_{i,-}^h$$

and

$$\eta^h_{i,\pm} = 1 - \frac{\mu^h_i}{2} \pm \frac{i}{2}\sqrt{\mu^h_i(4-\mu^h_i)}. \tag{68}$$

Theorem B.1. *If for all $i \in \{1, \cdots, N_h\}$ we have $\mu^h_i \neq 0$ and $\mu^h_i \neq 4$ then the spectrum of $\mathcal{A}_h(0)$ is the set $\{\eta^h_{i,\pm}\}$ and all these eigenvalues are semi-simple.*

If the CFL condition is satisfied (i.e. $\alpha \leq 1$), then $0 \leq \mu^h_i \leq 4$. In that case we find that $|\eta^h_{i,\pm}(\Delta t)|^2 = 1$. Moreover, according to the previous theorem, if $0 < \mu^h_i < 4$ the scheme (67) is stable for $\varepsilon = 0$ since the amplification matrix $\mathcal{A}_h(0)$ is diagonalizable (all its eigenvalues are semi-simple) and its spectral radius is exactly equal to 1. Such a situation occurs if the CFL condition is satisfied strictly (i.e. $\alpha < 1$) and thanks to Assumption 1.

To study the spectrum of $\mathcal{A}_h(\varepsilon)$, we use the perturbation theory of linear operators [13]. More precisely, we use Theorem 5.4 of [13] to compute the derivatives, w.r.t. ε, of each eigenvalue $\eta^h_{i,\pm}$ of $\mathcal{A}_h(\varepsilon)$ at $\varepsilon = 0$. A first step is to remark that $\mathcal{A}_h(\varepsilon)$ is continuous for small ε, and to compute $d\mathcal{A}_h(\varepsilon)/d\varepsilon$ at $\varepsilon = 0$. Observe that

$$\frac{d}{d\varepsilon}\left(I_h + \frac{\varepsilon^2}{12}\widehat{B}^2_h\right)(0) = \frac{d}{d\varepsilon}\widetilde{A}_h = 0, \quad \frac{d}{d\varepsilon}\widetilde{M}^{-1}_h(0) = -\frac{\widehat{B}_h}{2}.$$

Taking into account the fact that $\widetilde{M}_h$ and $\widehat{B}_h$ are diagonal or block diagonal matrices, we deduce that

$$\frac{d}{d\varepsilon}D^{-1}_h(0) = -\frac{\widehat{B}_h}{2} - \frac{\Delta t^2}{24}\widehat{C}_h.$$

Finally, one can compute that

$$\frac{d}{d\varepsilon}\mathcal{A}_h(0) = -\frac{1}{2}\begin{pmatrix} \widehat{B}_h + \dfrac{\Delta t^2}{12}\widehat{C}_h & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 2I_h - \Delta t^2\widetilde{A}_h & -2I_h \\ 0 & 0 \end{pmatrix}.$$

Denote by $\eta = \eta^h_{i,+}$ a complex eigenvalue of $\mathcal{A}_h(0)$ associated to some real $\mu = \mu^h_i$ that satisfies $0 < \mu^h_i < 4$ (the case $\eta = \eta^h_{i,-}$ is treated similarly). As previously shown, η is semi-simple and we assume that its multiplicity may be greater than 1. We denote by $\Sigma(\mu)$ the set of numbers such that $\mu^h_j = \mu$ for all $j \in \Sigma(\mu)$. The dimension of $\Sigma(\mu)$ is equal to the multiplicity of the eigenvalue μ of $\Delta t^2\widetilde{A}_h$. For all $j \in \Sigma(\mu)$, we denote by $\psi^h_{j,R}$ and $\psi^h_{j,L}$ the right and left eigenvectors of $\mathcal{A}_h(0)$ constructed as follows

$$\psi^h_{j,R} = \begin{pmatrix} \phi^h_j \\ \dfrac{1}{\eta}\phi^h_j \end{pmatrix} \text{ and } \psi^h_{j,L} = \begin{pmatrix} \phi^h_j \\ \dfrac{-1}{\eta}\phi^h_j \end{pmatrix}.$$

Remark that, since the eigenvectors ϕ^h_j are orthonormal for all $j \in \Sigma(\mu)$, the same property holds for the eigenvectors ψ^h_j. With all these properties at hand, one can construct the eigenprojection matrix $P(\eta)$ defined as follows

$$P(\eta) = \frac{1}{1-\eta^{-2}}\sum_{j\in\Sigma(\mu)} \psi^h_{j,R}\,\psi^{h,T}_{j,L}.$$

We can show that $P(\eta)$ is a projection operator that commutes with $\mathcal{A}_h(0)$, and

$$P(\eta)\mathcal{A}_h(0) = \mathcal{A}_h(0)P(\eta) = P(\eta)\mathcal{A}_h(0)P(\eta) = \eta P(\eta).$$

We can now apply Theorem 5.4 of [13]: Any eigenvalue $\eta(\varepsilon)$ of $\mathcal{A}_h(\varepsilon)$ is differentiable at $\varepsilon = 0$, and there exist an eigenvalue η of $\mathcal{A}_h(0)$ and an eigenvalue $\tilde{\eta}$ of

$$P(\eta)\left(\frac{d}{d\varepsilon}\mathcal{A}_h(0)\right)P(\eta)$$

in the subspace spanned by the family $\{\psi^h_{j,R}\}_{j\in\Sigma(\mu)}$, such that, for ε sufficiently small, we have

$$\eta(\varepsilon) = \eta + \varepsilon\,\tilde{\eta} + o(\varepsilon).$$

To continue this analysis, note that, since $\widehat{C}_h$ is skew-symmetric, we have

$$P(\eta)\left(\frac{d}{d\varepsilon}\mathcal{A}_h(0)\right)P(\eta) = -\frac{1}{2}\,\frac{2-\mu-2\eta^{-1}}{1-\eta^{-2}}\,\widehat{B}_h(\mu),$$

with

$$\widehat{B}_h(\mu) = \left(\sum_{j\in\Sigma(\mu)}\phi^h_j(\phi^h_j)^T\right)\widehat{B}_h\left(\sum_{j\in\Sigma(\mu)}\phi^h_j(\phi^h_j)^T\right).$$

Since $|\eta| = 1$, we have $\eta^{-1} = \overline{\eta}$ and because of (68) we have $\eta^2 - (2-\mu)\,\eta + 1 = 0$. Therefore,

$$\frac{2-\mu-2\eta^{-1}}{1-\eta^{-2}} = \frac{\eta^2-1}{\eta-\overline{\eta}} = \eta,$$

which implies that any eigenvalue $\eta(\varepsilon)$ behaves, for ε small enough, as

$$\eta(\varepsilon) = \eta\,(1-\varepsilon\alpha_h) + O(\varepsilon^2)$$

where η is an eigenvalue associated to some μ and α_h is one of the non-negative eigenvalues of $\widehat{B}_h(\mu)$. If the eigenvalues of $\widehat{B}_h(\mu)$ are positive for all μ, one can deduce, by continuity of the eigenvalues $\eta(\varepsilon)$ with respect to ε, that $|\eta(\varepsilon)| \le 1$ for all $0 \le \varepsilon \le \varepsilon_0$, for some ε_0 depending on $\mathcal{A}_h(\varepsilon)$. For these values of ε, the scheme is stable. These observations are summed up in the following theorem

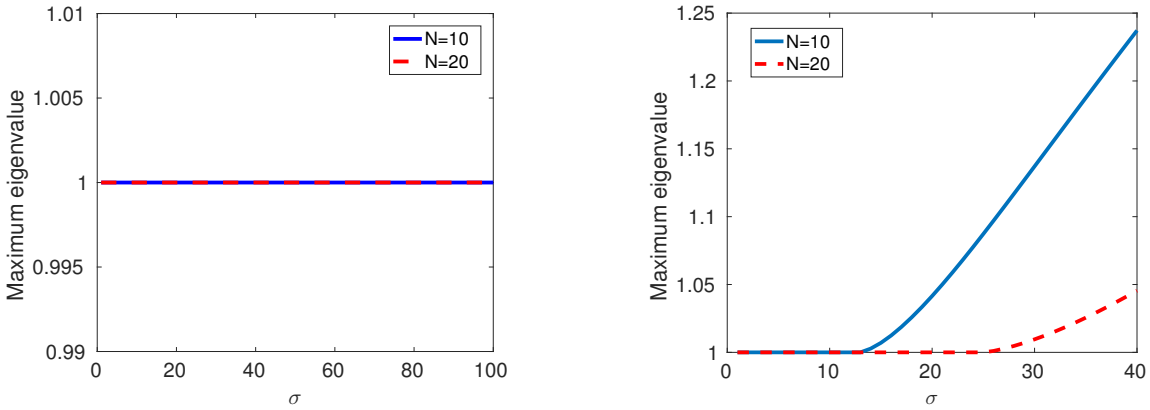
Theorem B.2. *Assume that Assumption 1 and assumptions of Theorem B.1 hold. Then, if*

$$(\phi^h_j)^T\,\widehat{B}_h\phi^h_j > 0$$

for all eigenvectors ϕ^h_j of $\Delta t^2\widetilde{A}_h$, there exists $\varepsilon_0 > 0$ such that the scheme (67) *is stable for all $0 \le \varepsilon \le \varepsilon_0$. Equivalently, all eigenvalues of $\mathcal{A}_h$ are semi-simple and of module lower or equal to* 1.

Theorem B.2 gives a sharper result compared to what was obtained in the previous section thanks to an analysis by energy technique, in the case of a small enough damping or a small time step. Note that a stronger result would be obtained if one could guaranty that $(\phi_j^h)^T \widehat{B}_h \phi_j^h > 0$. Such a property depends on the spatial discretization process and some pathological cases may arise, corresponding to situations where eigenvectors with compact support exist. Note however that, in a one-dimensional setting, the eigenvectors ϕ_j^h are equidistributed in the domain if a uniform mesh is used (see [20]).

We present now some numerical results to illutrate the theoretical results obtained above. We consider the same discretization parameters as in the above section (with either $N = 10$ or $N = 20$). The discrete system of equations for the (EME-1) scheme is written as a first order induction relation as in equation (67) and we compute the eigenvalues when σ varies. The largest eigenvalue amplitude is plotted w.r.t σ in Figs. 7(a) and 7(b). In Fig. 8, the trajectories of the eigenvalues, with respect to σ, are plotted. One can see that the amplitude of the eigenvalues indeed decreases for small values of σ, as predicted by Theorem B.2.



(a) $R = R_1$ (b) $R = R_0$.

FIGURE 7. Evolution of the largest eigenvalue amplitude w.r.t. σ, $N = 10, 20$.

REFERENCES

[1] R. Dautray, J-L. Lions. Mathematical Analysis and Numerical Methods for Science and Technology - Volume 5 and 6 - Evolution Problems I and II. Springer-Verlag Berlin, 2000.

[2] P. Joly. The mathematical model for elastic wave propagation. "Effective computational methods for wave propagation". Numererical Insights, Chapman & Hall/CRC, vol. 5, pp. 247–266, 2008.

[3] S. C. Brenner, L. R. Scott, The Mathematical Theory of Finite Element Methods (Third Edition), Springer, New York, 2008.

[4] J.C. Gilbert, P. Joly. Higher order time stepping for second order hyperbolic problems and optimal CFL conditions. Partial Differential Equations, vol 16, pp. 67–93, 2008.

[5] J. Chabassier, S. Imperiale. Introduction and study of fourth order theta schemes for linear wave equations. Journal of Computational and Applied Mathematics, vol. 245, pp. 194–212, 2013.

[6] J. Chabassier, S. Imperiale. Space/time convergence analysis of a class of conservative schemes for linear wave equations, In Comptes Rendus Mathematique, vol 355(3), pp 282–289, 2017.

[7] G. Cohen. High-order numerical methods for transient wave equations. Springer-Verlag, 2001.

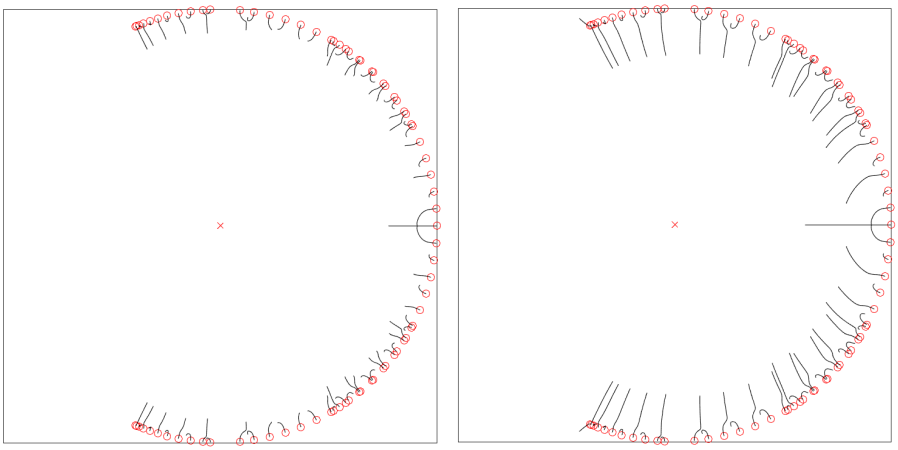


FIGURE 8. Trajectories of the eigenvalues of system (67) in the complex plane (with $N = 10$ and $R = R_0$). Left: σ varies between 0 and 10. Right: σ varies between 0 and 20. Note that for $\sigma = 0$ the eigenvalues are all located on the unit circle and their exact location is represented by circles. For $\sigma \leq 10$ no eigenvalue is of absolute value greater than one. For $\sigma = 20$ two eigenvalues are of absolute value greater than one .

[8] G. Cohen, P. Joly and N. Tordjman. Higher-order finite elements with mass-lumping for the 1D wave equation. Finite Element Analysis and Design, vol. 16(3-4), pp. 329–336, 1994.

[9] G. Cohen, P. Joly, J. E. Roberts and N. Tordjman Higher order triangular finite elements with mass lumping for the wave equation. SIAM Journal of Numererical Analysis, vol. 38(6), pp. 2047–2078, 2001.

[10] G.R. Shubin, J.B. Bell. A modified equation approach to constructing fourth-order methods for acoustic wave propagation. Society for Industrial and Applied Mathematics. Journal on Scientific and Statistical Computing, vol. 8(2), pp.135–151, 1987.

[11] P. Joly. Topics in computational wave propagation. "Variational methods for time-dependent wave propagation problems". Lecture Notes in Computational Science and Engineering, Springer Berlin, vol 31, pp. 201–264, 2003.

[12] J. Chabassier, S. Imperiale. Stability and dispersion analysis of improved time discretization for simply supported prestressed Timoshenko systems. Application to the stiff piano string. Wave Motion, vol. 50(3), pp. 456-480, 2012.

[13] T. Kato, Perturbation Theory for Linear Operators. Classics in Mathematics, Springer-Verlag Berlin Heidelberg, vol 132, 1995.

[14] J. G. Verwer, Runge–Kutta methods and viscous wave equations, Numerische Mathematik, vol. 112(3), pp. 485–507, 2009.

[15] M.J. Grote and T. Mitkova, High-order explicit local time-stepping methods for damped wave equations, Journal of Computationnal and Appplied Mathematics , vol. 239, pp 270–289, 2013.

[16] M. J. Grote, A. Schneebeli and D. Schötzau, Discontinuous Galerkin finite element method for the wave equation, SIAM Journal of Numerical Analysis vol. 44, 2408-–2431, 2006.

[17] P. Freitas, M. Grinfeld and P. A. Knight, Stability of Finite-Dimensional Systems with Indefinite Damping, Advances in Mathematical Sciences and Applications, vol. 17, 435–446, 1997.

[18] P. Freitas, E. Zuazua, Stability Results for the Wave Equation with Indefinite Damping, Journal of Differential Equations , vol. 132(2), pp. 338–352, 1996.

[19] M. Cassier, P. Joly, M. Kachanovska, Mathematical models for dispersive electromagnetic waves: An overview, Computers and Mathematics with Applications, vol. 74(11), pp. 2792–2830, 2017,

[20] S. Ervedoza, A. Marica, E. Zuazua, Numerical meshes ensuring uniform observability of one-dimensional waves: construction and analysis, IMA Journal of Numerical Analysis, vol. 36(2) pp. 503-–542, 2016.

[21] P. Grisvard. Elliptic Problems in Nonsmooth Domains. Pitman, Boston, 1985.
[22] S. S. Dragomir, Some Gronwall type inequalities and applications, Nova Science Pub Incorporated, 2003.