

Introduction and study of fourth order theta schemes for linear wave equations

Juliette Chabassier, Sébastien Imperiale

▶ **To cite this version:**

Juliette Chabassier, Sébastien Imperiale. Introduction and study of fourth order theta schemes for linear wave equations. [Research Report] RR-8090, INRIA. 2012, pp.31. hal-00738324v2

HAL Id: hal-00738324

https://inria.hal.science/hal-00738324v2

Submitted on 4 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

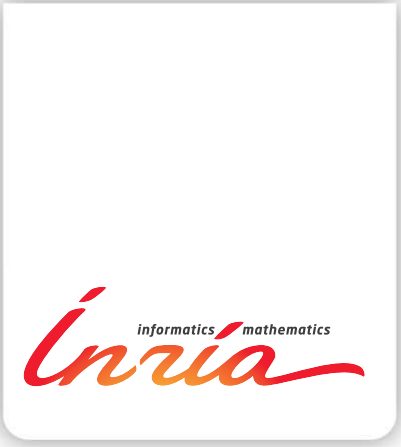


Introduction and study of fourth order theta schemes for linear wave equations.

Juliette Chabassier, Sébastien Imperiale

RESEARCH REPORT
N° 8090
October 2012
Project-Teams Poems and Magique3d

ISSN 0249-6399 ISRN INRIA/RR--8090--FR+ENG



Introduction and study of fourth order theta schemes for linear wave equations.

Juliette Chabassier*†, Sébastien Imperiale‡ †

Project-Teams Poems and Magique 3d

Research Report n° 8090 — October 2012 — 26 pages

Abstract: A new class of high order, implicit, three time step schemes for semi-discretized wave equations is introduced and studied. These schemes are constructed using the modified equation approach, generalizing the θ-scheme. Their stability properties are investigated via an energy analysis, which enables us to design super convergent schemes and also optimal stable schemes in terms of consistency errors. Specific numerical algorithms for the fully discrete problem are tested and discussed, showing the efficiency of our approach compared to second order θ-schemes.

Key-words: Wave equations , High order numerical methods , Time discretization , Theta-scheme, Modified equation

* Magique 3d team, Inria Sud Ouest, 200 Avenue de la Vieille Tour, 33 405 TALENCE, France.
† Poems team, Inria Rocquencourt, Le Chesnay, France.
‡ Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, 10027, USA.

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 Avenue de la Vieille Tour,
33405 Talence Cedex

Introduction et analyse de theta schémas d'ordre quatre pour les équations d'ondes linéaires.

Résumé : Nous introduisons et étudions une nouvelle classe de schémas d'ordre élevé, implicites et à trois pas de temps pour les équations d'ondes semi-discrètes. Ces schémas sont construits sur le principe de l'équation modifiée et généralisent le theta-schéma. Nous étudions leurs propriétés de stabilité via des techniques d'énergie, ce qui nous permet de concevoir des schémas super convergents ainsi que des schémas optimaux en terme d'erreur de consistance. Des algorithmes numériques de résolution pour le problème totalement discrétisé sont testés et critiqués, montrant la supériorité de notre approche comparée aux theta shémas classiques du second ordre.

Mots-clés : Équations d'ondes, méthodes numériques d'ordre élevé, discrétisation en temps, theta schéma, équation modifiée

Contents

1 Introduction **4**

2 Classical results **5**
2.1 Preliminary notations 5
2.2 Classical θ-schemes 6

3 Construction of a family of fourth order implicit schemes **7**
3.1 Modified equation technique for the leap frog scheme 7
3.2 Modified equation technique for the θ-scheme 8
3.3 New fourth order implicit schemes 8

4 Computation of the discrete solution **9**

5 Energy preservation and stability **10**

6 Peculiar (θ, φ)-schemes **12**
6.1 A class of optimal (θ, φ)-schemes 12
6.1.1 Sixth and eighth order stable schemes. 14
6.1.2 Fourth order optimal stable scheme 16
6.1.3 Fourth order unconditionally stable scheme. 17
6.1.4 Conclusions 17
6.2 "Multiple roots" fourth order schemes 18

7 Numerical results **18**
7.1 1D convergence test cases 18
7.2 Propagation of acoustic waves in a strongly heterogeneous 2D domain 20

8 Conclusions and prospects **22**

A Proof of theorem (5.1) **23**

B Optimal order 4 stable scheme **24**

1 Introduction

Linear wave equations play a great role in scientific modeling and are present in many fields of physics. For instance, they arise in the Maxwell equations, the acoustic equation and the elastodynamic equation. A discrete approximation of their solutions can be found with numerical simulations. Spatial discretization of the above equations using classical finite elements methods often leads to a semi-discretized problem of the form: find $u_h \in C^2(t, \mathbb{R}^N)$ such that

$$M_h \frac{d^2}{dt^2} u_h + K_h u_h = 0, \quad u_h(0) = u_{0,h}, \quad \frac{du_h}{dt}(0) = u_{1,h}, \tag{1}$$

where $u_h(t)$ is a vector-unknown in $\mathbb{R}^N$, M_h a symmetric positive definite matrix and K_h a symmetric positive semi-definite matrix.

Several approaches can be adopted to tackle the time discretization of problem (1). The so called "conservative methods" (as for instance the leap frog scheme) preserve a discrete energy which is consistent with the physical energy. They can be shown to be stable as soon as some positivity properties of the discrete energy are satisfied, which generally imposes a restriction, known as the CFL condition, on the time step depending on the matrices M_h and K_h. The leap frog scheme enters a more general class of three points time step, energy preserving, implicit schemes called θ-schemes, which are parametrized by a real number θ. The over cost of these implicit schemes compared to explicit ones is balanced by the fact that stability conditions allow for a bigger time step.

For simple configurations with simple finite elements methods (such as P_1 triangular elements), explicit schemes show good performances. However they have two major drawbacks in complex configurations that have not yet been completely solved:

- If the mesh has different scales of elements, or if the equations involve variable coefficients with strong contracts, the time step must be adapted to the worst situation (for instance the smallest element) because of the CFL condition. A natural way to avoid this restriction is to use local time stepping techniques which divide into two categories. The locally implicit technique, as developed in [CFJ03b], [CFJ03a], [KDE08] and [BJR05], is optimal in term of CFL restriction but "only" second order accurate in time, and requires the inversion of interface matrices. The fully explicit local time stepping, as developed in [DG09], achieves higher order time stepping but without (up to now) a full control over the CFL condition.
- If the mass matrix is non diagonal or non block-diagonal, its inversion (at least one time per iteration) can lead to a dramatic over cost of the explicit schemes, whereas no over cost is observed with implicit schemes (see remark 4.1).

The extension of conservative time discretization schemes to higher orders of accuracy is a natural question. A popular way to design explicit high order three points schemes is the modified equation approach. In this article we extend this approach to design new high order implicit schemes which are stable and present some optimal properties.

The paper is organized as follows. Section 2 recalls some well-known results concerning the leap frog and θ-schemes. In a conclusive remark we present, in the very simple case of the θ-scheme, the non standard approach that we will choose to follow in the rest of the paper. In section 3 we construct a family of energy preserving implicit fourth order schemes, parametrized by two real numbers (θ, φ). In section 4 we discuss the existence of their discrete solution and some practical aspects of computation which can reduce numerical cost. Section 5 is devoted to the study of the stability of the newly introduced schemes via energy techniques. The search for "optimal"

schemes is presented in section 6, it is done by adjusting (θ, φ) to increase accuracy. Finally, numerical results compare theses schemes to classical schemes in section 7.

In the following we will consider the semi-discretized problem

$$\frac{d^2}{dt^2}u_h + A_h u_h = 0, \quad u_h(0) = u_{0,h}, \quad \frac{du_h}{dt}(0) = u_{1,h}, \tag{2}$$

with A_h a symmetric positive semi-definite matrix. With no loss of generality: the analysis done below is valid if $A_h = M_h^{-1} K_h$.

Extensive use will be made of the spectral radius of the matrix A_h, defined as $\rho(A_h) = \sup_{\|v\|=1} A_h\, v \cdot v$ and coinciding with the greatest eigenvalue of A_h. Its exact expression can be given in simple cases, as for instance the 1D wave equation with constant speed c, using finite differences on a regular mesh of size h, for which $\rho(A_h) = 4c^2/h^2$. In other cases, the cost of its numerical evaluation (for example with the power iteration method) is negligible compared to the numerical resolution of the equation.

2 Classical results

In this section we recall the definitions and some properties of the classical leap-frog and θ-schemes, which are widely used for the time discretization of wave equations. In the following we denote $\Delta t > 0$ the time step of the numerical method.

2.1 Preliminary notations

The centered second order approximation of the second order derivative in time of any function $t \mapsto f(t)$ will be denoted

$$D^2_{\Delta t} f(t) = \frac{f(t + \Delta t) - 2\, f(t) + f(t - \Delta t)}{\Delta t^2}. \tag{3}$$

Assuming infinite smoothness on f, let us use a Taylor expansion to write the truncation error of the previous quantity:

$$D^2_{\Delta t} f(t) = \frac{d^2}{dt^2} f(t) + 2 \sum_{m=1}^{\infty} \frac{\Delta t^{2m}}{(2m+2)!} \frac{d^{2m+2}}{dt^{2m+2}} f(t). \tag{4}$$

Classical θ-schemes are based upon the use of a three points centered approximation of $f(t)$ which, for $\theta \in \mathbb{R}$, is defined by

$$\{f(t)\}_\theta = \theta\, f(t + \Delta t) + (1 - 2\theta)\, f(t) + \theta\, f(t - \Delta t). \tag{5}$$

Assuming again infinite smoothness on f, we can write the truncation error of this new quantity:

$$\{f(t)\}_\theta = f(t) + 2\theta \sum_{m=1}^{\infty} \frac{\Delta t^{2m}}{(2m)!} \frac{d^{2m}}{dt^{2m}} f(t). \tag{6}$$

Both the leap frog scheme and the θ-scheme use finite differences to discretize time in order to compute an approximation of the semi discrete solution u_h of (2). Consequently, the unknowns

of those schemes stand for the values of u_h at time $t^n = n\,\Delta t$: $u_h^n \simeq u_h(t^n)$. The discrete versions of (3) and (5), using the same symbols, are

$$D^2_{\Delta t} u_h^n = \frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2}, \qquad \{u_h^n\}_\theta = \theta\, u_h^{n+1} + (1-2\theta) u_h^n + \theta\, u_h^{n-1}. \tag{7}$$

The following algebraic relations will be useful: for all $\theta \in \mathbb{R}$,

$$\{u_h^n\}_\theta = \frac{\theta}{\theta'}\{u_h^n\}_{\theta'} + \frac{\theta' - \theta}{\theta'} u_h^n, \quad \forall\, \theta' \neq 0, \tag{8}$$

$$\{u_h^n\}_\theta = u_h^n + \theta \Delta t^2 D^2_{\Delta t} u_h^n, \tag{9}$$

$$\{u_h^n\}_\theta = (\theta - \frac{1}{4})\Delta t^2 D^2_{\Delta t} u_h^n + \{u_h^n\}_{1/4}. \tag{10}$$

2.2 Classical θ-schemes

The second order accurate leap-frog scheme reads

$$D^2_{\Delta t} u_h^n + A_h u_h^n = 0, \tag{11}$$

which is stable under a restriction on the time step called CFL condition. More precisely, it is possible to show using energy techniques that stability requires the following inequality to hold:

$$\Delta t^2 \leq \frac{4}{\rho(A_h)}, \tag{12}$$

where $\rho(A_h)$ is the spectral radius of the matrix A_h. This scheme enters a more general class of schemes called θ-schemes (convergence proofs are given in [Kar11] and some alternatives are studied in [BM78],[Ryl02] and [LLL10]) :

$$D^2_{\Delta t} u_h^n + A_h \{u_h^n\}_\theta = 0. \tag{13}$$

The leap-frog scheme corresponds to the choice $\theta = 0$. Other choices lead to implicit schemes. When $\theta < 1/4$, these schemes are stable under the CFL condition

$$\Delta t^2 \leq \frac{4}{(1-4\theta)\rho(A_h)}. \tag{14}$$

As long as $\theta \geq 1/4$ they are unconditionally stable.

Let $u_h(t)$ be solution of (2). The truncation error of schemes (13) up to fourth order can be written

$$D^2_{\Delta t} u_h(t^n) + A_h\{u_h(t^n)\}_\theta = \Delta t^2 \Big(\frac{1}{12}\frac{d^2}{dt^2} + \theta A_h\Big)\frac{d^2}{dt^2} u_h(t^n) + \mathcal{O}(\Delta t^4). \tag{15}$$

Knowing that $u_h(t)$ is solution of (2), we can replace the second order time derivative by the matrix $-A_h$, giving

$$D^2_{\Delta t} u_h(t^n) + A_h\{u_h(t^n)\}_\theta = -\Delta t^2 \Big(\theta - \frac{1}{12}\Big)\, A_h^2\, u_h(t^n) + \mathcal{O}(\Delta t^4). \tag{16}$$

This expression shows that we obtain second order accuracy except for $\theta = 1/12$, which cancels out the first consistency error term, giving a fourth order scheme.

In some situations, the time step Δt is related to the physics (for sampling reasons) and the

spectral radius $\rho(A_h)$ (via the spatial discretization) is related to the geometry or the a priori knowledge of the solution (in the case of rapidly varying coefficients or boundary layers). Therefore, they have to be considered as data for the numerical analyst. It is then reasonable to choose a numerical scheme suited to these parameters instead of choosing discretization parameters suited to the problem. A good criterion to compare numerical schemes is the coefficient of the first term of the consistency error, which is given, for the specific case of the θ-schemes, by $(\theta - 1/12)$. As these schemes give a degree of freedom, adapting θ to the product $\Delta t^2 \rho(A_h)$ makes it possible to decrease the consistency error while providing a stable scheme. This "optimization" problem is straightforward here: two cases arise, either $\Delta t^2 \rho(A_h) \leq 6$, in which case the value $\theta = 1/12$ leads to a fourth order stable scheme (relation (14) holds), or if $\Delta t^2 \rho(A_h) > 6$, the choice of θ that provides a stable scheme and minimizes the consistency error is

$$\theta = \frac{1}{4} - \frac{1}{\Delta t^2 \rho(A_h)}. \tag{17}$$

The non standard approach mentioned above will be adopted in the discussion that follows: instead of choosing a time step Δt that provides a good accuracy on a given mesh with a given numerical scheme, we will invert the reasoning and choose the best numerical scheme for a given couple $(\Delta t, \rho(A_h))$.

3 Construction of a family of fourth order implicit schemes

3.1 Modified equation technique for the leap frog scheme

The "modified equation" technique, introduced in [SB87], enables to construct higher order schemes from the second order leap frog scheme. More precisely, the order $2p$ scheme is obtained by adding to the leap frog schemes terms that compensate the $(p-1)$ first terms of the truncation error. This error reads, for $u_h(t)$ solution of (2):

$$D^2_{\Delta t} u_h(t^n) + A_h u_h(t^n) = 2 \sum_{m=1}^{\infty} \frac{\Delta t^{2m}}{(2m+2)!} \frac{d^{2m+2}}{dt^{2m+2}} u_h(t^n). \tag{18}$$

Again, we can replace the second order time derivative by the matrix $-A_h$:

$$D^2_{\Delta t} u_h(t^n) + A_h u_h(t^n) = 2 \sum_{m=1}^{\infty} (-1)^{p+1} \frac{\Delta t^{2m}}{(2m+2)!} A_h^{m+1} u_h(t^n). \tag{19}$$

Truncating the series up to the first term, and approaching $u_h(t^n)$ with u_h^n gives the fourth order scheme:

$$D^2_{\Delta t} u_h^n + A_h u_h^n - \frac{\Delta t^2}{12} A_h^2 u_h^n = 0. \tag{20}$$

The stability condition is deduced from energy arguments and reads

$$\Delta t^2 \leq \frac{B_{00}}{\rho(A_h)}, \tag{21}$$

where $B_{00} = 12$.

Using an Hörner algorithm, this scheme is two times more expensive than the original scheme. This over cost can be compensated by adding stabilization terms that allow to increase the time step as explained in [GJ08] and [JR10].

3.2 Modified equation technique for the θ-scheme

We now use the ideas of the modified equation technique applied to the θ-scheme (instead of the leap frog scheme, i.e $\theta = 0$). Let $u_h(t)$ be solution of (2). The truncation error of the θ-scheme (13) is:

$$D^2_{\Delta t} u_h(t^n) + A_h\{u_h(t^n)\}_\theta = \sum_{m=1}^{\infty} \Delta t^{2m} \Big(\frac{2}{(2m+2)!} \frac{d^2}{dt^2} + \frac{2\theta}{(2m)!} A_h \Big) \frac{d^{2m}}{dt^{2m}} u_h(t^n), \tag{22}$$

in which we replace the time derivatives by $-A_h$, as in (20):

$$D^2_{\Delta t} u_h(t^n) + A_h\{u_h(t^n)\}_\theta + \sum_{m=1}^{\infty} e_m(\theta) \Delta t^{2m} \ A_h^{m+1} \ u_h(t^n) = 0, \tag{23}$$

where the coefficients $e_m(\theta)$ are defined by

$$e_m(\theta) = (-1)^m \big(\frac{2}{(2m+2)!} - \frac{2\theta}{(2m)!} \big). \tag{24}$$

To obtain a fourth order scheme, the natural idea would be to follow the modified equation procedure by keeping the first term of the series while replacing $u_h(t^n)$ with u_h^n. This gives

$$D^2_{\Delta t} u_h^n + A_h\{u_h^n\}_\theta + (\theta - \frac{1}{12}) \Delta t^2 \ A_h^2 \ u_h^n = 0. \tag{25}$$

It is possible to study the stability of this scheme using energy techniques. As this will be a special case of the analysis done below in section 6.1, we just give here the result : the time step restriction is given by

$$\Delta t^2 \leq \frac{B_0(\theta)}{\rho(A_h)}, \tag{26}$$

where

$$B_0(\theta) = \begin{cases} \dfrac{12}{(1-12\theta)} & \text{if } \theta \leq \Big(\dfrac{2-\sqrt{3}}{4\sqrt{3}} \Big), \\ r(\theta, 0) & \text{if } \ \theta > \Big(\dfrac{2-\sqrt{3}}{4\sqrt{3}} \Big), \end{cases} \tag{27}$$

where $r(\theta, 0)$ will be introduced in the next sections and satisfies $r(\theta, 0) \lesssim 10$, for $\theta > -(2 + \sqrt{3})/(4\sqrt{3})$. These restrictions turn out to be quite penalizing since, unlike with the classical θ-scheme, there is no possible choice of θ that leads to an unconditionally stable scheme. Indeed, the condition (26) implies that $\Delta t^2 \rho(A_h)$ must bounded by

$$\sup_{\theta \in \mathbb{R}} B_0(\theta) = B_0\Big(\frac{2-\sqrt{3}}{4\sqrt{3}} \Big) = \frac{6}{2-\sqrt{3}} \simeq 22.4. \tag{28}$$

3.3 New fourth order implicit schemes

Trying to improve the previous time step restriction, we choose to introduce a new real number $\varphi \in \mathbb{R}$ and to approximate $u_h(t^n)$ in (23) by $\{u_h^n\}_\varphi$. This will lead us to consider a class of schemes parametrized by $(\theta, \varphi) \in \mathbb{R}^2$, which includes scheme (25) when choosing $\varphi = 0$. We will call (θ, φ)-schemes the following schemes:

$$D^2_{\Delta t} u_h^n + A_h\{u_h^n\}_\theta + \big(\theta - \frac{1}{12}\big) \Delta t^2 A_h^2 \{u_h^n\}_\varphi = 0. \tag{29}$$

Next sections are dedicated to studying the numerical, stability and consistency properties of this general class of two parameters, at least fourth order, implicit schemes.

Remark 3.1. *The value $\theta = 1/12$ cancels out the new added term, retrieving* (13)*. Therefore the stability condition has already been given by* (14)*, which gives here:*

$$\Delta t^2 \leq \frac{6}{\rho(A_h)}. \tag{30}$$

In the following, we will consider $\theta \neq 1/12$.

4 Computation of the discrete solution

Scheme (29) does not seem easy to invert at first sight because it involves powers of the matrix A_h which can already be difficult to invert. Yet, the computation of the solution is not a secondary question to deal with when developing a numerical scheme devoted to be actually used. We propose in this preliminary section an enhancing algorithm based on the complex factorization of a polynomial function. It will allow us at the same time to introduce notations and mathematical objects that will be widely used hereafter during convergence, energy and stability analysis.

The first step towards implementation is to rewrite the numerical scheme in a way that emphasizes its implicit nature. Assuming that $\varphi \neq 0$, and multiplying (29) by $\varphi\,\Delta t^2$, we can use the algebraic relations (8) and (9) to obtain the equivalent scheme

$$\Big(I_h + \theta\Delta t^2 A_h + \varphi(\theta - \frac{1}{12})\Delta t^4 A_h^2\Big)\{u_h^n\}_\varphi = u_h^n + [\theta - \varphi]\Delta t^2 A_h u_h^n =: b_h^n. \tag{31}$$

The computational algorithm associated to the scheme (31) follows two stages:

- We retrieve $\{u_h^n\}_\varphi$ knowing u_h^n by inverting the matrix polynomial

$$P(\Delta t^2 A_h; \theta, \varphi) = I_h + \theta\Delta t^2 A_h + \varphi(\theta - \frac{1}{12})\Delta t^4 A_h^2. \tag{32}$$

  The invertibility of the matrix $P(\Delta t^2 A_h; \theta, \varphi)$ is equivalent to the existence of a discrete solution.

- We then use the knowledge of $\{u_h^n\}_\varphi$ to compute u_h^{n+1}:

$$u_h^{n+1} = \frac{\{u_h^n\}_\varphi + (2\varphi - 1)u_h^n}{\varphi} - u_h^{n-1}. \tag{33}$$

The first stage of this algorithm is not straightforward because the matrix to invert involves powers of the matrix A_h. Indeed, if A_h has a band structure (as in most classical finite elements methods), the bandwidth of A_h^2 will be even larger, which penalizes the use of direct solvers. In the same way, iterative methods may suffer from the bad conditioning of $P(\Delta t^2 A_h; \theta, \varphi)$ since $\rho(A_h^2) = \rho(A_h)^2$ (see for instance remark 7.1). To overcome such difficulties, we propose an algorithm based on the factorization of $P(\lambda, ; \theta, \varphi)$. Different cases arise according to the nature of its roots (r^+, r^-):

- The roots are real. In this case invertibility is not granted and must be verified. If $P(\Delta t^2 A_h; \theta, \varphi)$ is indeed invertible, $\{u_h^n\}_\varphi$ can be obtained in two steps, solving the linear systems with an intermediary unknown v_h^n:

$$\begin{cases} v_h^n = (r^+ I_h - \Delta t^2 A_h)^{-1} b_h^n, \\ \{u_h^n\}_\varphi = r^+ r^- (r^- I_h - \Delta t^2 A_h)^{-1} v_h^n. \end{cases} \tag{34}$$

- The roots are conjugate complex numbers: $r^+ = \overline{r^-}$. As above, we introduce v_h^n computed as:

$$v_h^n = (r^+ I_h - \Delta t^2 A_h)^{-1} b_h^n. \tag{35}$$

  As b_h^n is a real vector, we know that the solution $\{u_h^n\}_\varphi$ will also be a real vector. Therefore we can identify the imaginary parts of both sides in

$$\frac{1}{|r^+|^2}(\overline{r^+} I_h - \Delta t^2 A_h)\{u_h^n\}_\varphi = v_h^n, \tag{36}$$

  to get, without any additional matrix inversion:

$$\{u_h^n\}_\varphi = -\frac{|r^+|^2}{\Im\mathrm{m}(r^+)}\,\Im\mathrm{m}(v_h^n). \tag{37}$$

  This method only requires one matrix inversion which however happens in the complex domain. Adapted techniques can be used to avoid over cost (see [AK00], [Fre90]).

Remark 4.1. *Using the factorization of the polynomial matrix $P(\Delta t^2 A_h; \theta, \varphi)$, if $A_h = M_h^{-1} K_h$ one needs to compute the solution of linear problems of the form*

$$(r^\pm I_h - \Delta t^2 M_h^{-1} K_h) x_h = b_h \Longleftrightarrow (r^\pm M_h - \Delta t^2 K_h) x_h = M_h b_h. \tag{38}$$

This shows that the mass matrix inversion is included in the regular matrix inversion needed by the implicit scheme. This leads to nearly no over cost even if the mass matrix is non-diagonal.

Remark 4.2. *To take into account the non-homogeneous version of* (2)*, namely*

$$\frac{d^2}{dt^2} u_h + A_h u_h = f_h, \quad u_h(0) = u_{0,h}, \quad \frac{du_h}{dt}(0) = u_{1,h}, \tag{39}$$

where f_h is a source term, the fully discrete (31) *must be modified so that the fourth order accuracy is preserved. Using the same arguments as those used to obtained* (16) *from* (15)*, one can show that, when $\varphi \neq 0$, the correct discrete scheme is*

$$P(\Delta t^2 A_h; \theta, \varphi)\{u_h^n\}_\varphi = b_h^n + \varphi\,\Delta t^2 f_h(t^n) + \varphi\,\frac{\Delta t^4}{12}\frac{d^2}{dt^2} f_h(t^n) + \varphi\,\Delta t^4\big(\theta - \frac{1}{12}\big) A_h f_h(t^n).$$

5 Energy preservation and stability

Using relation (10), our (θ, φ)-schemes (29) are equivalent to:

$$\Big[I_h + (\theta - \frac{1}{4})\Delta t^2 A_h + (\theta - \frac{1}{12})(\varphi - \frac{1}{4})\Delta t^4 A_h^2\Big] D_{\Delta t}^2 u_h^n + \Big[A_h + (\theta - \frac{1}{12})\Delta t^2 A_h^2\Big]\{u_h^n\}_{1/4} = 0. \tag{40}$$

This simplifies to the general form:

$$\widetilde{M}_h D_{\Delta t}^2 u_h^n + \widetilde{K}_h \{u_h^n\}_{1/4} = 0, \tag{41}$$

where $\widetilde{M}_h$ and $\widetilde{K}_h$ are symmetric matrices defined by

$$\widetilde{K}_h = A_h\ Q_2(\Delta t^2 A_h;\theta,\varphi),\ \text{and}\ \widetilde{M}_h = \frac{Q_1(\Delta t^2 A_h;\theta,\varphi)}{4}, \tag{42}$$

with

$$\begin{cases} Q_1(\lambda;\theta,\varphi) = 4 + (4\theta-1)\lambda + (4\varphi-1)(\theta-\frac{1}{12})\lambda^2, \\ Q_2(\lambda;\theta,\varphi) = 1 + (\theta-\frac{1}{12})\lambda. \end{cases} \tag{43}$$

Taking the euclidian scalar product $(\cdot,\cdot)$ of equation (41) with $(u_h^{n+1}-u_h^{n-1})/2\Delta t$ gives the energy preservation

$$\frac{\mathcal{E}^{n+1/2}-\mathcal{E}^{n-1/2}}{\Delta t} = 0, \tag{44}$$

where the discrete energy $\mathcal{E}^{n+1/2}$ is defined by

$$\mathcal{E}^{n+1/2} = \frac{1}{2}\left(\widetilde{M}_h\frac{u_h^{n+1}-u_h^n}{\Delta t},\frac{u_h^{n+1}-u_h^n}{\Delta t}\right) + \frac{1}{2}\left(\widetilde{K}_h\frac{u_h^{n+1}+u_h^n}{2},\frac{u_h^{n+1}+u_h^n}{2}\right). \tag{45}$$

One can prove that the positivity of the matrices $\widetilde{M}_h$ and $\widetilde{K}_h$ leads to the stability of the scheme (41). Since those matrices depend on A_h through the polynomials Q_1 and Q_2, a sufficient condition can be derived to ensure this positivity:

$$Q_1(\lambda;\theta,\varphi) \geq 0 \text{ and } Q_2(\lambda;\theta,\varphi) \geq 0,\ \forall\lambda\in[0,\Delta t^2\rho(A_h)]. \tag{46}$$

These conditions lead to an upper bound on Δt^2 that depends on the values of $\theta \neq 1/12$ and φ as stated in the following theorem:

Theorem 5.1 (CFL condition)**.** *The matrices $Q_1(\Delta t^2 A_h;\theta,\varphi)$ and $Q_2(\Delta t^2 A_h;\theta,\varphi)$ are positive matrices if*

$$\rho(A_h)\,\Delta t^2 \leq B(\theta,\varphi) = \min\left(B_{Q_1}(\theta),B_{Q_2}(\theta,\varphi)\right) \tag{47}$$

where

$$B_{Q_2}(\theta) = \begin{cases} +\infty & \text{if } \theta > \dfrac{1}{12}, \\ \dfrac{12}{1-12\theta} & \text{otherwise,} \end{cases} \tag{48}$$

$$B_{Q_1}(\theta,\varphi) = \begin{cases} +\infty & \text{if } (\theta,\varphi)\in\mathcal{D}_{US}, \\ \dfrac{4}{1-4\theta} & \text{if } (\theta,\varphi)\in\mathcal{I}_{1/4}, \\ r(\theta,\varphi) & \text{otherwise,} \end{cases} \tag{49}$$

and

$$\left|\begin{array}{l} \mathcal{I}_{1/4} = (-\infty, 1/4) \cup \{1/4\}, \\ \mathcal{D}_{US} = [1/4, +\infty) \times [1/4, +\infty) \cup \mathcal{S}^- \cup \mathcal{S}^+, \\ \mathcal{S}^- = \left\{ \varphi \leq \dfrac{1}{4}[1 + \dfrac{(4\theta-1)^2}{16(\theta-1/12)}],\ \theta < 1/12 \right\}, \\ \mathcal{S}^+ = \left\{ \varphi \geq \dfrac{1}{4}[1 + \dfrac{(4\theta-1)^2}{16(\theta-1/12)}],\ \theta > 1/12 \right\}, \\ \Delta(\theta,\varphi) = (4\theta-1)^2 - 16\,(4\varphi-1)(\theta-1/12), \\ r(\theta,\varphi) = \dfrac{1-4\theta-\sqrt{\Delta(\theta,\varphi)}}{2(4\varphi-1)(\theta-1/12)}. \end{array}\right. \tag{50}$$

Proof. The proof of this statement will be given in annex. □

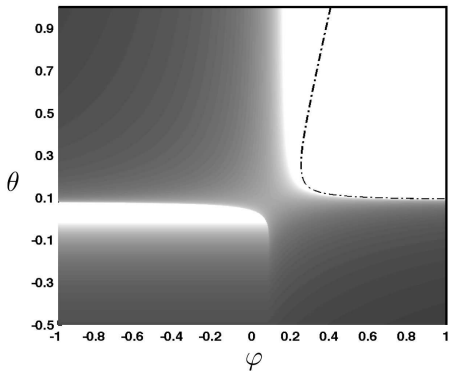


Figure 1: Graphical representation of the CFL condition $B(\theta, \varphi)$. The dotted line represents the demarcation of S^+. The white areas stand for an infinite upper bound (unconditionally stable schemes) whereas the colored scale stands for values of B (the darker being the lower).

Remark 5.1. *The coefficients B_0 of (26) and B_{00} of (21) can be seen as restrictions of B on specific areas of the $\mathbb{R}^2$ plane:*

$$B_0(\theta) = B(\theta, 0), \quad B_{00} = B(0, 0). \tag{51}$$

The following result states that respecting the CFL condition almost always leads to a well-posed discrete problem:

Theorem 5.2 (Existence of the discrete solution)**.** *If $Q_1(\Delta t^2 A_h; \theta, \varphi)$ and $Q_2(\Delta t^2 A_h; \theta, \varphi)$ are positive matrices, then $P(\Delta t^2 A_h; \theta, \varphi)$ is invertible if*

$$(\theta, \varphi) \neq \big(\frac{1}{12} - \frac{1}{\Delta t^2 \rho(A_h)}, \frac{1}{12}\big). \tag{52}$$

Proof. We write the polynomial $P(\lambda;\theta,\varphi)$ as a sum of two positive terms:

$$P(\lambda;\theta,\varphi) = \frac{Q_1(\lambda;\theta,\varphi) + \lambda Q_2(\lambda;\theta,\varphi)}{4}. \tag{53}$$

For $\lambda \in [0, \Delta t^2\rho(A_h)]$, it vanishes if and only if both terms vanish at the same point. It can only happen if $\lambda = \Delta t^2\rho(A_h)$ and if $(\theta,\varphi) = \big(1/12 - (\Delta t^2\rho(A_h))^{-1}, 1/12\big)$. □

6 Peculiar (θ,φ)-schemes

6.1 A class of optimal (θ,φ)-schemes

In the following, we are going to find the "best possible" stable (θ,φ)-schemes, which means to find the optimal values of θ and φ in the $\mathbb{R}^2$ plane that minimize the consistency error of the scheme, under the constraints of scheme stability. Indeed, we know (by construction) that these schemes are fourth order accurate in time, and the consistency errors of order six and eight depend on the values of θ and φ. The consistency error of scheme (29) is obtained, first by evaluating the approximation (6) in the solution $u_h(t^n)$ of (2):

$$\{u_h(t^n)\}_\varphi = u_h(t^n) - \sum_{m=1}^{\infty} c_m(\varphi)\Delta t^{2m} A_h^m u_h(t^n), \tag{54}$$

where the coefficients $c_m(\varphi)$ are defined by

$$c_m(\varphi) = (-1)^{m+1}\frac{2\varphi}{(2m)!}. \tag{55}$$

Then, the choice to approximate $u_h(t^n)$ in (23) by $\{u_h^n\}_\varphi$ leads to

$$\begin{aligned} D_{\Delta t}^2 u_h(t^n) + A_h\{u_h(t^n)\}_\theta + e_1(\theta)\Delta t^2 A_h^2\Big[\{u_h(t^n)\}_\varphi + \sum_{m=1}^{\infty} c_m(\varphi)\Delta t^{2m} A_h^m u_h(t^n)\Big] \\ + e_2(\theta)\Delta t^4 A_h^3 u_h(t^n) + e_3(\theta)\Delta t^6 A_h^4 u_h(t^n) = \mathcal{O}(\Delta t^8), \end{aligned} \tag{56}$$

which gives

$$D_{\Delta t}^2 u_h(t^n) + A_h\{u_h(t^n)\}_\theta + e_1(\theta)\Delta t^2 A_h^2\{u_h(t^n)\}_\varphi = -\varepsilon_3(\theta,\varphi)\Delta t^4 A_h^3 u_h(t^n) - \varepsilon_4(\theta,\varphi)\Delta t^6 A_h^4 u_h(t^n) + \mathcal{O}(\Delta t^8), \tag{57}$$

where the first terms of the consistency error are

$$\begin{cases} \varepsilon_3(\theta,\varphi) = e_2(\theta) + e_1(\theta)\, c_1(\varphi) = \dfrac{1}{360} - \dfrac{\theta}{12} - \dfrac{\varphi}{12} + \theta\,\varphi, & \text{(58a)} \\[2ex] \varepsilon_4(\theta,\varphi) = e_3(\theta) + e_1(\theta)\, c_2(\varphi) = \dfrac{-1}{20160} + \dfrac{\theta}{360} + \dfrac{\varphi}{144} - \dfrac{\theta\,\varphi}{12}. & \text{(58b)} \end{cases}$$

Regarding the stability conditions, we will see that they provide nonlinear constraints on (θ,φ) which depend on $\Delta t^2\rho(A_h)$. This is why we tackle this issue using a non standard point of view : we assume $\Delta t^2\rho(A_h)$ to be known and we solve the corresponding optimization problem:

$$\min_{(\theta,\varphi)\,\in\,\mathbb{R}^2} |\varepsilon_3(\theta,\varphi)|, \quad \begin{cases} Q_1(\lambda;\theta,\varphi) \geq 0, \\ Q_2(\lambda;\theta,\varphi) \geq 0, \end{cases} \quad \forall \lambda \in [0, \Delta t^2\rho(A_h)]. \tag{59}$$

If it is possible to find values of (θ, φ) that make ε_3 vanish in the stability region, then the optimal choices within these values are the one that minimize the absolute value of ε_4.

In section 6.1.1, we will try to mimic the super-convergence phenomenon that we find in the classical θ-scheme when $\theta = 1/12$. Indeed, this second order scheme appears to be fourth order accurate for the peculiar choice of $\theta = 1/12$. In our case, we will see that it will be possible to obtain stable sixth order schemes by restraining the choice of the couple (θ, φ) to a curve in the $\mathbb{R}^2$ plane corresponding to the zeros of (58a), and even stable eighth order schemes by choosing a special couple on this curve that correspond to a zero of (58b). The major drawback of this powerful result is that it can only happen for small values of $\Delta t^2\, \rho(A_h)$.

When this approach is not possible (for $\Delta t^2\, \rho(A_h)$ greater than a certain value), section 6.1.2 investigates which couple (θ, φ) in the $\mathbb{R}^2$ plane leads to the stable numerical scheme that minimizes the absolute value of the consistency error (58a).

Finally in section 6.1.3, we construct "optimal" unconditionally stable schemes by assuming in the optimization process that $\Delta t^2\, \rho(A_h) = +\infty$. This will lead to (θ, φ)-schemes, where θ and φ depend on a small parameter δ, and that minimize the absolute value of (58a) when δ tends to 0.

Remark 6.1. *It is possible to obtain higher order θ-schemes parametrized by more real numbers. For instance, if we extend the previous approach to the next order of approximation, we can introduce $(\theta, \varphi, \psi, \widetilde{\psi}) \in \mathbb{R}^4$ and the obtained schemes read:*

$$D^2_{\Delta t} u^n_h + A_h\{u^n_h\}_\theta + e_1(\theta)\Delta t^2 A_h^2\{u^n_h\}_\varphi + \Delta t^4 A_h^3\Big[e_2(\theta)\{u^n_h\}_\psi + e_1(\theta)c_1(\varphi)\{u^n_h\}_{\widetilde{\psi}}\Big] = 0. \quad (60)$$

The consistency error of these schemes is given by

$$-\Delta t^6 A_h^4\Big[e_3(\theta) + e_1(\theta)c_1(\varphi)c_1(\widetilde{\psi}) + e_1(\theta)c_2(\varphi) + e_2(\theta)c_1(\psi)\Big]u_h(t^n) + \mathcal{O}(\Delta t^8), \quad (61)$$

showing that these schemes are at least sixth order accurate.

6.1.1 Sixth and eighth order stable schemes.

We look for stable (θ, φ)-schemes such that ε_3 vanishes and such that $|\varepsilon_4|$ is minimized. This restricts the choice of θ and φ to a curve in $\mathbb{R}^2$ described by

$$\varphi^\star = (12\,\theta^\star - 1)^{-1}(\theta^\star - \frac{1}{30}). \quad (62)$$

All the values on this curve lead to a sixth order scheme. The stability conditions require that for all $\lambda \in [0, \Delta t^2 \rho(A_h)]$ we must have

$$\begin{cases} Q_1(\lambda; \theta^\star, \varphi^\star) = (4\theta^\star - 1)\lambda + (-\frac{2}{3}\theta^\star + \frac{13}{180})\lambda^2 + 4 \geq 0, \\ Q_2(\lambda; \theta^\star, \varphi^\star) = 1 + (\theta^\star - \frac{1}{12})\lambda \geq 0, \end{cases} \quad (63)$$

where φ has been eliminated using (62). The second inequality is constraining only if $\theta^\star < 1/12$, in which case

$$\Delta t^2 \rho(A_h) \leq \frac{12}{1 - 12\theta^\star} := p(\theta^\star). \quad (64)$$

The first inequality is respected if and only if the polynomial $Q_1(\lambda;\theta^\star,\varphi^\star)$ has no root of multiplicity one in the interval $[0,\Delta t^2\rho(A_h)]$. Let $\Delta(\theta^\star)$ be the discriminant of $Q_1(\lambda;\theta^\star,\varphi^\star)$:

$$\Delta(\theta^\star) = 16\,(\theta^\star)^2 + \frac{8}{3}\theta^\star - \frac{7}{45}. \tag{65}$$

Different cases arise according to the convexity of Q_1:

- Concave when $\theta^\star > \frac{13}{120}$. In this case, there are two simple roots of opposite signs. The positive root must be greater than $\Delta t^2\rho(A_h)$, that is:

$$\Delta t^2\rho(A_h) \le r(\theta^\star) := \frac{1-4\theta^\star-\sqrt{\Delta(\theta^\star)}}{2\,(13/180-2\theta^\star/3)}. \tag{66}$$

- Linear when $\theta_0^\star = \frac{13}{120}$. In this case, Q_1 is a decreasing linear function, which is positive on $[0,\frac{120}{17}]$ to which $\Delta t^2\rho(A_h)$ must belong. This condition extends by continuity relation (66) to its singular point.

- Convex when $\theta^\star < \frac{13}{120}$. In this case, we have $Q_1'(0;\theta^\star,\varphi^\star) < 0$ and two different cases according to the sign of $\Delta(\theta^\star)$ must be considered:

  - $\Delta(\theta^\star) > 0 \Leftrightarrow \theta^\star \notin [-\frac{1}{12}-\frac{\sqrt{15}}{30}, -\frac{1}{12}+\frac{\sqrt{15}}{30}]$: Two simple roots. Therefore the lower root must be greater than $\Delta t^2\rho(A_h)$. This condition turns out to be the same as (66).
  - $\Delta(\theta^\star) \le 0 \Leftrightarrow \theta^\star \in [-\frac{1}{12}-\frac{\sqrt{15}}{30}, -\frac{1}{12}+\frac{\sqrt{15}}{30}]$: No simple root. Therefore the first condition is automatically fulfilled.

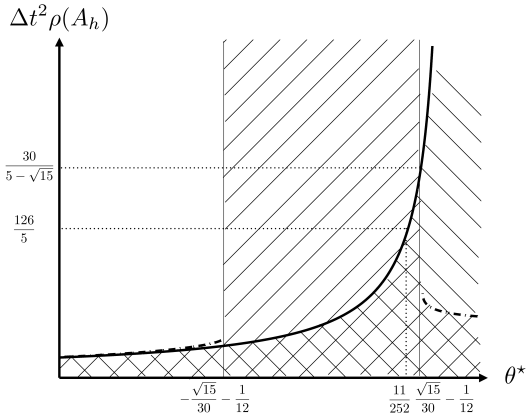


Figure 2: Bounds on $\Delta t^2\rho(A_h)$ for any $\theta^\star$: the scheme is stable if the couple $(\Delta t^2\rho(A_h),\theta^\star)$ lies in both hatched areas. In continuous line, we represented $p(\theta^\star)$ while in dotted line we represented $r(\theta^\star)$ when $\Delta(\theta^\star) > 0$.

To sum up, the upper bound on $\Delta t^2 \rho(A_h)$ is :

$$\begin{cases} \min(p(\theta^\star), r(\theta^\star)) & \text{if } \theta^\star < -\dfrac{1}{12} - \dfrac{\sqrt{15}}{30}, \\ p(\theta^\star) & \text{if } -\dfrac{1}{12} - \dfrac{\sqrt{15}}{30} \leq \theta^\star \leq -\dfrac{1}{12} + \dfrac{\sqrt{15}}{30}, \\ \min(p(\theta^\star), r(\theta^\star)) & \text{if } -\dfrac{1}{12} + \dfrac{\sqrt{15}}{30} < \theta^\star < \dfrac{1}{12}, \\ r(\theta^\star) & \text{if } \dfrac{1}{12} < \theta^\star. \end{cases} \tag{67}$$

In this variety of sixth order schemes, the choice of $\theta^\star$ is guided by the minimization of the absolute value of ε_4, which reads

$$\varepsilon_4(\theta^\star, \varphi^\star) = \frac{-\theta^\star}{240} + \frac{11}{60480}. \tag{68}$$

It is a linear decreasing function of $\theta^\star$, vanishing for the value

$$\theta^{\star\star} = \frac{11}{252} \Rightarrow \varphi^{\star\star} = -\frac{13}{600} \tag{69}$$

- For $\Delta t^2\, \rho \in [0, \frac{126}{5}]$, the couple $(\Delta t^2 \rho(A_h), \theta^{\star\star})$ provides an eighth order accurate stable scheme.
- For $\Delta t^2\, \rho \in [\frac{126}{5}, \frac{30}{5-\sqrt{15}}]$, the couple $(\Delta t^2 \rho(A_h), \theta^{\star\star})$ violates the constraints (67). We can prove that minimizing $|\varepsilon_4(\theta^\star)|$ while satisfying (67) leads to the choice :

$$\theta^\star = \frac{1}{12} - \frac{1}{\Delta t^2 \rho(A_h)}. \tag{70}$$

- For $\Delta t^2 \rho > \frac{30}{5-\sqrt{15}}$, it is not possible to construct sixth order stable schemes anymore. Indeed, all the upper bounds of (67) are lower than $\frac{30}{5-\sqrt{15}}$.

6.1.2 Fourth order optimal stable scheme

We consider now the values of $\Delta t^2 \rho(A_h)$ which could not satisfy the stability conditions of previous section, more precisely we assume that $\Delta t^2 \rho(A_h) > 30/(5 - \sqrt{15})$. As seen before it is not possible to make ε_3 vanish in this case, thus we will construct a stable scheme for a given $\Delta t^2 \rho(A_h)$ that minimizes its absolute value by solving (59).

The second inequality of the constraints in (59) leads to

$$\theta \geq \frac{1}{12} - \frac{1}{\Delta t^2 \rho(A_h)}. \tag{71}$$

For the first inequality, different situations arise according to the convexity and the slope at the origin of $Q_1(\lambda; \theta, \varphi)$. It divides the $\mathbb{R}^2$ plane into six areas delimited by the values $1/12$ and $1/4$ for θ and $1/4$ for φ.

Let us first consider the area $1/12 - 1/(\Delta t^2\, \rho(A_h)) \leq \theta < 1/12,\ \varphi < 1/4$. Q_1 is then a convex parabola with a negative slope at the origin.

- The parabola crosses the y-axis when $\Delta(Q_1) > 0$. In this case, the interesting interval $[0, \Delta t^2\,\rho(A_h)]$ must be before the first root, or in other terms, we must have $Q_1(\Delta t^2\,\rho(A_h)) \geq 0$ and $Q_1'(\Delta t^2\,\rho(A_h)) \leq 0$. But the calculation shows that these two conditions are incompatible because we made the assumption on $\Delta t^2 \rho(A_h) > 8$.
- The parabola never crosses the y-axis when $\Delta(Q_1) \leq 0$ which leads to the condition:

$$\varphi \leq \frac{1}{4}\Big[1 + \frac{(4\theta-1)^2}{16\,(\theta - 1/12)}\Big]. \tag{72}$$

In this part of the plane, the sixth order consistency error is a decreasing function of φ, for any θ:

$$\varepsilon_3(\theta,\varphi) = \frac{1}{360} - \frac{\theta}{12} + \varphi\,(\theta - \frac{1}{12}) \tag{73}$$

and stays positive when (72) is respected. Hence, if θ is fixed, the value of φ that minimizes the absolute value of ε_3 is the greatest value possible. As a consequence, we want to minimize $|\varepsilon_3|$ on the curve

$$\Big(\theta, \varphi = \frac{1}{4}\Big[1 + \frac{(4\theta-1)^2}{16\,(\theta - 1/12)}\Big]\Big), \tag{74}$$

which leads to the 1D minimization of:

$$\varepsilon_3\left(\theta, \frac{1}{4}\Big[1 + \frac{(4\theta-1)^2}{16(\theta - 1/12)}\Big]\right) = \frac{\theta^2}{4} + \frac{\theta}{24} - \frac{7}{2880} \tag{75}$$

This objective function is positive and increasing on the interval $[1/12 - 1/\Delta t^2\,\rho(A_h), 1/12]$ as soon as $\Delta t^2\,\rho(A_h) \geq 4$. Therefore, the optimal values of θ and φ in this area of $\mathbb{R}^2$ are given by:

$$\theta^\sharp = \frac{1}{12} - \frac{1}{\Delta t^2\,\rho(A_h)}, \qquad \varphi^\sharp = \frac{1}{4} - \frac{\Delta t^2\,\rho(A_h)}{64}\big(\frac{2}{3} + \frac{4}{\Delta t^2\,\rho(A_h)}\big)^2. \tag{76}$$

It turns out that in the other areas, either it is not possible to construct stable schemes, or stable schemes give a greater consistency error. We provide in annex a proof of this statement. Notice that if the product $\Delta t^2\,\rho(A_h)$ gets very large, the optimal values $(\theta^\sharp, \varphi^\sharp)$ tend to $(1/12, -\infty)$.

6.1.3 Fourth order unconditionally stable scheme.

When the spectral radius of the operator is not known, we assume it is infinite, which means that the positivity of Q_1 and Q_2 must be ensured on the whole $\mathbb{R}^+$ interval. Unfortunately it is not possible to pass to the limit in the formulae $(\theta^\sharp, \varphi^\sharp)$. We first notice that it is necessary to have

$$\theta > \frac{1}{12}, \quad \varphi > \frac{1}{4}, \tag{77}$$

otherwise Q_2 is a decreasing affine function and Q_1 a concave parabola. Again, we have to distinguish several cases depending on the slope at the origin of Q_1.

- Either $\theta \geq 1/4$ and stability is acquired without any other condition.
- Either $1/12 < \theta < 1/4$ and the parabola Q_1 must not cross the y-axis : its discriminant must be negative, leading to

$$\varphi \geq \frac{1}{4}\Big[1 + \frac{(4\theta-1)^2}{16\,(\theta - 1/12)}\Big]. \tag{78}$$

It both cases, ε_3 is a positive decreasing function of φ, therefore the lowest possible value of φ must be chosen. Again, we tackle a 1D optimization problem. Let us introduce $\delta > 0$ such that

$$\theta^\delta = \frac{1}{12} + \delta \Rightarrow \varphi^\delta = \frac{(-2/3 + 4\delta)^2}{64\,\delta} + \frac{1}{4}. \tag{79}$$

The consistency error parametrized by δ is given by

$$\varepsilon_3(\theta^\delta, \varphi^\delta) = \frac{\delta^2}{4} + \frac{\delta}{12} + \frac{1}{360}. \tag{80}$$

Therefore δ must be chosen as low as possible, even if the lower bound 0 is not achievable. To sum up, we have constructed a family of unconditionally stable (θ, φ)-schemes parametrized by $\delta > 0$ that minimize the consistency error "at the limit".

6.1.4 Conclusions

The results of previous sections are summarized in table 6.1.4. With these choices of (θ, φ)-scheme, the polynomial $P(\Delta t^2\,\rho(A_h); \theta, \varphi)$ defined by (32) has complex roots as soon as

$$\Delta t^2 \rho(A_h) \geq \frac{60}{7} + \frac{24}{7}\sqrt{15} \simeq 21.8502, \tag{81}$$

otherwise the roots are negative. In both cases, the invertibility of the matrix $P(\Delta t^2 A_h, \theta, \varphi)$ is guaranteed.

ρ	$\frac{126}{5}$	$\frac{30}{5-\sqrt{15}}$		$+\infty$
θ	$\frac{11}{252}$	$\frac{\rho-12}{12\rho}$	$\frac{\rho-12}{12\rho}$	$\frac{\delta+12}{12\delta}$
φ	$\frac{-13}{600}$	$\frac{20-\rho}{240}$	$\frac{1}{6} - \frac{\rho}{144} - \frac{1}{4\rho}$	$\frac{1}{6} + \frac{1}{144\delta} + \frac{\delta}{4}$
Order	8	6	4	4

Table 1: Optimal values of θ and φ for a given positive $\rho := \Delta t^2 \rho(A_h)$ and for a positive small parameter δ.

6.2 "Multiple roots" fourth order schemes

The schemes developed in section 6.1 lead to a complex factorization of the polynomial (32) as soon as $\Delta t^2 \rho(A_h)$ is big enough. In this section we build unconditionally stable schemes for which the polynomial $P(\lambda; \theta, \varphi)$ has a double negative root r:

$$P(\Delta t^2 A_h; \theta, \varphi) = r^{-2}(rI_h - \Delta t^2 A_h)^2. \tag{82}$$

Therefore, the inversion of $P(\Delta t^2 A_h;\theta,\varphi)$ can be done by inverting the same matrix $(rI_h - \Delta t^2 A_h)$ twice, which reduces cost when direct solvers are used. This happens when the discriminant of $P(\lambda;\theta,\varphi)$ vanishes, giving the following relation between θ and φ:

$$\varphi = \frac{3\theta^2}{12\theta - 1}. \tag{83}$$

For this choice of φ, the stability conditions read

$$\begin{cases} 1 + (\theta - 1/12)\lambda \geq 0, \\ 4 + (4\theta - 1)\lambda + (\theta^2 - \theta + \frac{1}{12})\lambda^2 \geq 0, \end{cases} \qquad \forall \lambda \geq 0. \tag{84}$$

The first condition implies that $\theta > 1/12$. The second condition cannot be respected unless the leading coefficient is positive, which leads to $\theta \notin [1/2 - 1/\sqrt{6}, 1/2 + 1/\sqrt{6}]$. Since $1/12$ lies in this interval, θ must be greater than $1/2 + 1/\sqrt{6}$, implying that the negative multiple root of $P(\lambda;\theta,\varphi)$ is given by $r = -2/\theta$. To sum up, the (θ,φ)-schemes satisfying (83) and

$$\theta \geq \frac{1}{2} + \frac{1}{\sqrt{6}} \simeq 0.91 \tag{85}$$

are unconditionally stable and we get the factorization

$$P(\Delta t^2 A_h;\theta,\varphi) = (I_h + \frac{\theta\,\Delta t^2}{2} A_h)^2. \tag{86}$$

For these choices of (θ,φ) the absolute value of the consistency error $|\varepsilon_3|$ is a positive and increasing function of θ, so the value $\theta = 1/2 + 1/\sqrt{6}$ (which corresponds to $\varphi = 1/4$) is optimal.

7 Numerical results

7.1 1D convergence test cases

In this section we present an academical test case for which we know the exact solution, allowing us to compute exactly numerical errors. We consider the 1D scalar wave equation with velocity 1 in a domain of length 1 with periodic boundary conditions:

$$\frac{\partial^2}{\partial t^2}u(x,t) - \frac{\partial^2}{\partial x^2}u(x,t) = 0, \quad x \in [0,1], \quad u(0,t) = u(1,t), \quad \frac{\partial}{\partial x}u(0,t) = \frac{\partial}{\partial x}u(1,t). \tag{87}$$

The initial condition is a gaussian pulse centered in the domain. The initial velocity is such that the analytical solution is this gaussian pulse traveling from left to right. More precisely, the initial conditions are

$$u(x,0) = u_0(x) = \begin{cases} e^{(\varepsilon - 1/2)^{-2}\,\ln(\varepsilon)\,(x-1/2)^2} & \text{if } |x-1/2| \leq 1/2 - \varepsilon \\ \varepsilon & \text{if } |x-1/2| > 1/2 - \varepsilon \end{cases}, \qquad \frac{\partial}{\partial t}u(x,0) = -\frac{\partial}{\partial x}u_0(x),$$

where ε is a small parameter (in practice we choose $\varepsilon = 10^{-20}$). The analytic solution is given by

$$u(x,t) = u_0(x - t + n) \text{ where } n \in \mathbb{N} \text{ such that } x - t + n \in [0,1].$$

In order to solve numerically this wave equation, the natural idea is to use the second order centered finite difference scheme. It is even known to provide the exact solution on regular meshes

if the time step is chosen equal to its maximal value allowed by the CFL condition. In this context (regular mesh) it is obvious that our implicit schemes would not compare well. They were originally designed to overcome specific difficulties as distorted meshes or rapidly varying coefficients. In order to pedagogically represent these kinds of difficulties, we propose to introduce a very small element in the mesh. It can seem artificial in this simple 1D example, but is numerically representative of what can happen in 2D or 3D when realistic domains are involved.

Standard explicit and θ-schemes are compared with some of the new (θ, φ)-schemes introduced above, and with the exact solution. The error e is computed as the sup. over time of the relative discrete L^2 error. The estimated cost corresponds to the total number of matrix-vector products with A_h required for the simulation. Non-preconditioned iterative methods (Conjugate Gradient and Minimal Residual, see [Fre90]) have been used to solve the implicit schemes in factorized form as presented in 4.

The mesh is composed of 12 identical elements surrounding a small element of size $\tau = 10^{-4}$, which leads to a spectral radius $\rho(A_h) = 5.62 \times 10^9$ when sixth order spectral finite elements with mass lumping are used to tackle spatial discretization (see [Coh01]). For the explicit scheme, the time step is restricted to 2.7×10^{-5} by the CFL condition, whereas for the implicit schemes, we impose $\Delta t = 0.016$ or $\Delta t = 0.004$. The simulation runs until $T = 10$. Table 2 summarizes the performances of the chosen schemes and figure 3 shows the obtained snapshots at final time.

Scheme	Δt	Error	Cost
Explicit	2.7×10^{-5}	2.4×10^{-5}	374 957
$\theta = 1/4$	1.6×10^{-2}	2.3×10^{-1}	37 158
$(\theta, \varphi) = (1/4, 1/4)$	1.6×10^{-2}	2.8×10^{-2}	67 526
$(\theta, \varphi) = (\theta^\sharp, \varphi^\sharp)$	1.6×10^{-2}	4.1×10^{-3}	47 004
$(\theta, \varphi) = (\theta^\sharp, \varphi^\sharp)$	4.0×10^{-3}	3.1×10^{-5}	73 214

Table 2: Comparison between several schemes: leap-frog explicit ($\theta = 0$), θ-scheme with $\theta = 1/4$, naive (θ, φ)-scheme with $(\theta, \varphi) = (1/4, 1/4)$ and optimal (θ, φ)-scheme adapted to the product $\Delta t^2\, \rho(A_h)$ (given by equation (76)).

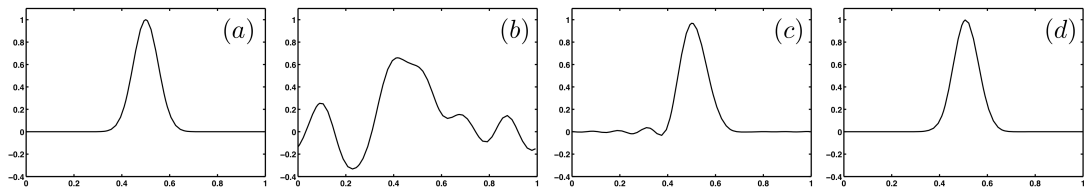


Figure 3: Snapshots of the numerical solutions at final time. Figure (a) corresponds to the explicit scheme, (b) to the θ-scheme, (c) to the naive (θ, φ)-scheme and (d) to the optimal (θ, φ)-scheme. All implicit schemes use the time step $\Delta t = 1.6 \times 10^{-2}$.

As expected, the small element clearly penalizes the explicit scheme: a small step step must be chosen, which increases the cost of the method but also increases its accuracy. This is why

the final snapshot is very close to the analytical solution, and the error is very low, for a very expensive cost. For the three following implicit schemes, we choose a time step of 1.6×10^{-2}. The θ-scheme with $\theta = 1/4$ is very cheap but behaves very poorly, as illustrated by the snapshot and the relative error of about 0.2. The naive (θ, φ)-scheme with $(\theta, \varphi) = (1/4, 1/4)$ gives a much better error and a nicer snapshot, even if numerical dispersion is still visible. The optimal (θ, φ)-scheme clearly overcomes the naive (θ, φ)-scheme, by being more accurate and cheaper. The first observation was to be expected since the criterion of optimality was indeed accuracy, but the second one had not been foreseen and is linked to the correlation between the condition number of the matrix $P(\Delta t^2 A_h; \theta, \varphi)$ and the consistency error ε_3. Finally, by choosing an appropriate time step of 4.0×10^{-3}, we recover with the optimal (θ, φ)-scheme the accuracy of the explicit scheme, while staying about five times cheaper.

Remark 7.1. *To highlight the interest of factorizing the polynom P in this specific 1D case, let us focus on the (θ, φ)-scheme with $(\theta, \varphi) = (1/4, 1/4)$. On the one hand, the natural approach would be to invert the matrix*

$$P(\Delta t^2 A_h; 1/4, 1/4) = I_h + \frac{\Delta t^2}{4} A_h + \frac{\Delta t^4}{24} A_h^2.$$

Its condition number is given by $P(\Delta t^2\, \rho(A_h); 1/4, 1/4) = 8.62 \times 10^{10}$. On the other hand, we can use the factorized form introduced in section 4. The complex roots of the polynomial P read $r^{\pm} = -3 \pm i\sqrt{15}$, hence the matrix to invert is $(-3 + i\sqrt{15}) I_h - \Delta t^2 A_h$ and its condition number is equal to $|r^+ - \Delta t^2\, \rho(A_h)|/|r^+| = 2.94 \times 10^5$. This last matrix is far easier to invert both in terms of computational cost (iterative methods) and numerical precision.

A convergence study has been lead on a regular mesh with sixth order finite elements (or seventh order to obtain an eighth order convergence curve) in space for different (θ, φ)-schemes. The error is computed as the sup norm in time (here $T = 8$) of the discrete L^2 norm (induced by the mass matrix) of difference with the analytical solution. The results are given in figure 4.

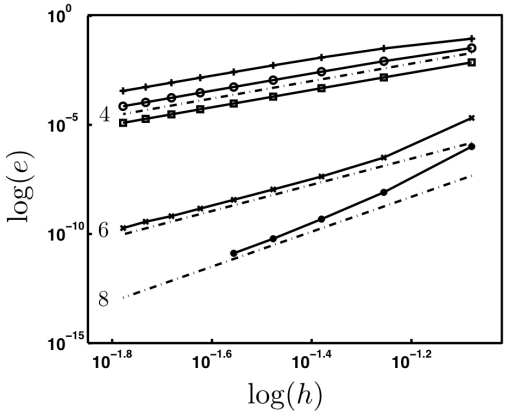


Figure 4: Log of the error w.r.t the log of the mesh size for different (θ, φ)-schemes and discretization parameters. $+$: optimal "multiple roots" scheme with $\Delta t^2 \rho = 120$, $\circ$: optimal (θ, φ)-scheme with $\Delta t^2 \rho = 120$, $\square$: optimal (θ, φ)-scheme with $\Delta t^2 \rho = 60$, $\times$: optimal (θ, φ)-scheme with $\Delta t^2 \rho = 26$, $\bullet$: optimal (θ, φ)-scheme with $\Delta t^2 \rho = 22$.

We obtain the expected rates of convergence for the different (θ, φ)-schemes. All the simulations and matrix inversions have been done with an iterative method without preconditioning. This is clearly non-optimal in term of computational cost but also in term of accuracy since it induced a numerical locking that prevented us to reach better accuracy with the eighth order scheme.

7.2 Propagation of acoustic waves in a strongly heterogeneous 2D domain

To conclude the numerical results, we present a more complex configuration. We are interested in the propagation of acoustic pressure $p(\mathbf{x}, t)$ in a square domain $\Omega = [-1, 1] \times [-1, 1]$ in which we assume that the propagation is governed by the non-homogeneous equations

$$\frac{\partial^2}{\partial t^2} p(\mathbf{x}, t) - \mathbf{div}\, c(\mathbf{x})\, \nabla p(\mathbf{x}, t) = f(\mathbf{x}, t), \quad \mathbf{x} \in \Omega, \quad \nabla p \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \partial\Omega, \quad \frac{\partial}{\partial t} p(\mathbf{x}, 0) = p(\mathbf{x}, 0) = 0,$$

where $\mathbf{n}$ is the outward normal, $c(\mathbf{x})$ the material properties of the domain of propagation and $f(\mathbf{x}, t) := g(\mathbf{x})\, h(t)$ a source term. The parameters $c(\mathbf{x})$ and $h(t)$ are represented in figure 5 whereas $g(\mathbf{x})$ is a 2D gaussian pulse centered around $\mathbf{x} = 0$. The explicit expressions of these coefficients are

$$c(x) = 1 + 10^4 e^{-800\,(|\mathbf{x}|^2 - 0.35)^2}, \quad h(t) = \big(2\pi^2 (16\,(t-0.1)-1)^2 - 1\big) e^{-\pi^2\,(16\,(t-0.1)-1)^2}, \quad g(x) = e^{-800\,|\mathbf{x}|^2}.$$

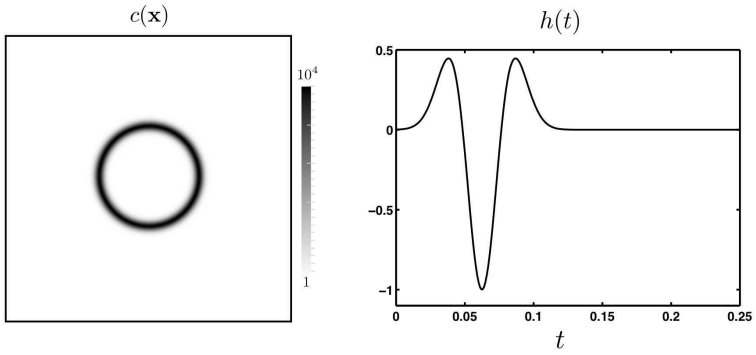


Figure 5: Representation of the coefficients $c(\mathbf{x})$ and $h(t)$.

Explicit computation of the discrete solution of this problem may not be adapted if one wants to minimize the computational cost while reaching a target, but not necessarily very small, accuracy. Indeed, the upper bound for the time step in explicit methods is proportional to $1/\sqrt{\max_{\mathbf{x}} c(\mathbf{x})} = 10^{-2}$. This is a typical situation where implicit and unconditionally stables schemes are often used, since in this case the time step is no longer related to $c(\mathbf{x})$ and can be tuned to reach the desired accuracy and obtain a reasonable computational cost.

We choose to compare three different types of discretization in time to a precomputed refined reference solution:

- **Schemes A & B** Two standard leap frog schemes, of second and fourth order, which are explicit. Hence they must be used with a very small time step. They should yield a very good accuracy yet being more costly than the implicit schemes. Note that the fourth order

leap frog scheme corresponds to the choice $(\theta, \varphi) = (0, 0)$ also called modified equation scheme.

- **Scheme C** The second order θ-scheme with $\theta = 1/4$. It is the most classical scheme that can be used to tackle the implicit discretization of our problem.
- **Scheme D** Our fourth order discretization (θ, φ)-scheme with $(\theta, \varphi) = (1/2 + 1/\sqrt{6}, 1/4)$ as computed in subsection 6.2. This is the most easily implantable alternative that we presented and it should outperform the second-order θ-scheme.

The same space discretization is chosen for all the numerical computations since we are only interested in comparing the time discretization of a semi-discrete problem similar to equation (39). The spatial mesh is a fine enough uniform grid of size $h = 0.025$ (1600 squared elements) and we use fifth order spectral finite elements with mass lumping (see [Coh01]) (160801 degrees of freedom and $\rho(A_h) \simeq 1.2 \times 10^{10}$). Note also that the source term is handled as indicated in remark 4.2 to avoid any accuracy loss.

The explicit schemes A and B must respect their stability condition, leading to a maximal time step allowed of respectively $\Delta t \approx 1.8 \times 10^{-5}$ and $\Delta t \approx \sqrt{3} \times 1.8 \times 10^{-5}$. The last two schemes C and D are unconditionally stable. The time step is chosen to be $\Delta t = 1.8 \times 10^{-3}$, which is close to the maximal time step allowed by the explicit scheme if the coefficient $c(\mathbf{x})$ was constant (equal to 1) over all of the domain. These schemes require one (respectively two) inversions of a real matrix at each time step. The choice of the peculiar fourth order scheme D described in subsection 6.2 has been motivated by the fact that the same (symmetric positive definite) matrix must be inverted twice at each time step, which implies that in practice only one LU factorization is performed before the iterations start. Some snapshots of the solution are presented in figure 6 and the results of the numerical experiments are summed-up in table 3.

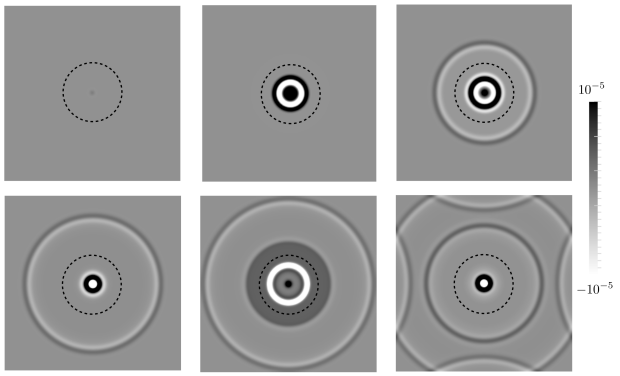


Figure 6: From left to right, top to bottom: snapshots of the pressure field computed with the explicit scheme A, for $T = 0.105$, $T = 0.3$, $T = 0.5$ and $T = 0.7$, $T = 0.9$, $T = 1.1$. The dashed circle is parametrized by $|\mathbf{x}|^2 = 0.35$ and represents the region where $c(\mathbf{x})$ reaches his maximum value.

Type	Scheme	Δt	Relative L^2 error at $T = 1.0$	Relative cost
B (ref)	Explicit Order 4: $(\theta, \varphi) = (0, 0)$	3.11×10^{-6}	Reference	–
A	Explicit Order 2	1.8×10^{-5}	2.33×10^{-6}	100
B	Explicit Order 4: $(\theta, \varphi) = (0, 0)$	3.11×10^{-5}	1.94×10^{-8}	112.7
C	$\theta = 1/4$	1.25×10^{-3}	8.90×10^{-2}	$3.2 + 11.2$
D	$(\theta, \varphi) = (1/2 + 1/\sqrt{6}, 1/4)$	2.5×10^{-3}	4.05×10^{-2}	$2.8 + 8.0$
C	$\theta = 1/4$	2.5×10^{-3}	2.24×10^{-2}	$3.4 + 20.7$
D	$(\theta, \varphi) = (1/2 + 1/\sqrt{6}, 1/4)$	5×10^{-3}	2.87×10^{-3}	$3.0 + 15.8$

Table 3: Comparison between several schemes: L^2 norm relative error at $T = 1.0$. The cost corresponds to the duration of simulation divided by the amount of time needed to run the explicit scheme A (the cost of LU factorization is the first term in the sum for the implicit schemes).

The reference solution is computed with the fourth order classical θ-scheme B, with a time step of $\Delta t = 3.11 \times 10^{-6}$, which is ten times less than the maximum value allowed. Since the time steps taken for the explicit schemes A and B are very low, we see that as expected, they give a very accurate solution for a higher computational cost. More importantly we see that the implicit $\theta-$schemes C give a solution accurate from around 9% to 2% for a cost around 7 to 4 time less, however for nearly the same computational cost $(\theta, \varphi)-$schemes D give solutions accurate from around 4% to 0.3%, outperforming the classical θ-scheme.

8 Conclusions and prospects

We have introduced a new family of fourth order accurate in time, implicit, energy preserving schemes, denoted as the (θ, φ)-schemes, that generalize the famous second order accurate θ-schemes. Their stability properties have been investigated via an energy analysis, leading to a general CFL condition. We have also provided in this family schemes that minimize the consistency error. We have found schemes with super-convergence properties that can be used if $\Delta t^2\, \rho(A_h)$ is small enough, or optimal schemes if $\Delta t^2\, \rho(A_h)$ is large. We have also provided unconditionally stable schemes and finally schemes that require the same numerical treatment as the usual θ-scheme but are fourth order accurate. Numerical results show the interest of such schemes as well as their convergence behavior.

However, we have sometimes noted the need to use preconditioning strategies when a direct factorization of the finite elements matrices was not possible. Several authors have already suggested solutions to deal with the inversion of the complex linear system occurring with the optimal (θ, φ)-scheme ([AK00],[Fre90],[BB08]). These approaches still need to be tested and adapted to the (θ, φ)-scheme. A natural extension of this work would be to consider and analyze the sixth order accurate schemes (see remark 6.1), which offer even more parameters and are therefore expected to give a very low consistency error after the optimization process. Finally, significant extension of this work will be to take into account dissipative terms and non trivial boundary conditions in the equations, which will be the subject of forthcoming work.

A Proof of theorem (5.1)

The positivity of the energy is granted as soon as $Q_1(\lambda;\theta,\varphi)$ and $Q_2(\lambda;\theta,\varphi)$ are positive for any $\lambda \in [0, \Delta t^2\rho(A_h)]$:

$$\begin{cases} Q_1(\lambda;\theta,\varphi) = 4 + (4\theta-1)\lambda + (4\varphi-1)(\theta - \frac{1}{12})\lambda^2 \geq 0, & \text{(88a)} \\ Q_2(\lambda;\theta,\varphi) = 1 + (\theta - \frac{1}{12})\lambda \geq 0. & \text{(88b)} \end{cases}$$

The positivity of each polynomial leads to upper bounds on $\Delta t^2\rho(A_h)$, therefore the positivity of the energy is fulfilled if $\Delta t^2\rho(A_h)$ is lower than the minimum of both bounds.

Positivity of Q_2 This polynomial is an affine function, being 1 at the origin. If its leading coefficient is positive (if $\theta > 1/12$), then (88b) holds. In the opposite case, (88b) holds if and only if $Q_2(\Delta t^2\rho(A_h);\theta,\varphi) \geq 0$. These results lead to the first upper bound (48).

Positivity of Q_1 This polynomial is an affine function if $\varphi = 1/4$. In this case, either $\theta > 1/12$ and (88a) is always true, either $\theta < 1/12$ and (88a) holds if and only if $Q_1(\Delta t^2\rho(A_h);\theta,\varphi) \geq 0$, which leads to the upper bound $4/(1-4\theta)$. Assume now that $\varphi \neq 1/4$. The polynomial Q_1 being 4 at the origin, several possible situations arise according to the sign of the leading coefficient and the sign of the slope at the origin:

	$\varphi \in (-\infty, 1/4)$	$\varphi \in (1/4, +\infty)$
$\theta \in (1/4, +\infty)$	concave, slope ≥ 0	convex, slope ≥ 0
$\theta \in (1/12, 1/4)$	concave, slope ≤ 0	convex, slope ≤ 0
$\theta \in (-\infty, 1/4)$	convex, slope ≤ 0	concave, slope ≤ 0

The three concave situations lead to the following upper bound: $\Delta t^2\rho(A_h)$ must be lower than the positive root of Q_1. The formula for this positive root $r(\theta,\varphi)$ is given by (50). In the case convex with positive slope, Q_1 is then positive at the origin and increasing, thus the condition (88a) is automatically fulfilled, therefore we set the upper bound to $+\infty$. Finally, the two convex with negative slope situations are a little more complicated. Two cases arise according to the sign of the discriminant $\Delta(\theta,\varphi)$ defined in (50). Either there are no distinct roots when $\Delta(\theta,\varphi) \leq 0$ (if $(\theta,\varphi) \in \mathcal{S}^+$ or $(\theta,\varphi) \in \mathcal{S}^-$), thus (88a) is fulfilled and we set the upper bound to $+\infty$. Or there are two positive roots when $\Delta(\theta,\varphi) > 0$, therefore the condition (88a) is fulfilled if $\Delta t^2\rho(A_h)$ is lower than the first root of Q_1, which is still $r(\theta,\varphi)$ since the sign of the leading coefficient has changed.

B Optimal order 4 stable scheme

In this appendix, we complete the proof given in 6.1.2 which states that the optimal value of $|\varepsilon_3(\theta,\varphi|)$ when $\Delta t^2\rho(A_h) > 30/(5-\sqrt{15})$ is obtained in the quadrant $\{\theta < 1/12, \varphi < 1/4\}$ for

$$\begin{cases} \theta^\sharp = \dfrac{1}{12} - \dfrac{1}{\Delta t^2\,\rho(A_h)}, \\ \varphi^\sharp = \dfrac{1}{4} - \dfrac{\Delta t^2\,\rho(A_h)}{64}\left(\dfrac{2}{3} + \dfrac{4}{\Delta t^2\,\rho(A_h)}\right)^2. \end{cases} \tag{89}$$

It is shown in 6.1.2 that in this quadrant, this choice gives the optimal value of $|\varepsilon_3|$:

$$\varepsilon_3(\theta^\sharp, \varphi^\sharp) = \frac{1}{360} - \frac{1}{\Delta t^2\, \rho(A_h)} + \frac{1}{4(\Delta t^2\, \rho(A_h))^2}, \tag{90}$$

which is lower than $1/360$ because $\Delta t^2\, \rho(A_h) > 30/(5-\sqrt{15})$. We will see that in the other regions of the (θ, φ)-plane, either the schemes cannot be stable, either the stable schemes give a greater consistency error.

The quadrant $\{\theta < 1/12, \varphi \geq 1/4\}$ The positivity of Q_2 on the interval $[0, \Delta t^2\, \rho(A_h)]$ imposes that $\theta > 1/12 - 1/\Delta t^2\, \rho(A_h)$. Moreover, the positivity of the concave polynomial Q_1 on the same interval is respected if $Q_1(\Delta t^2\, \rho(A_h), \theta, \varphi) \geq 0$. These two conditions are incompatible because $\Delta t^2\, \rho(A_h) > 30/(5-\sqrt{15})$. Therefore, there are no stable schemes in this quadrant.

The half space $\{\theta > 1/12\}$ The positivity of Q_2 is granted in this area. As for Q_1, we divide the space into three zones:

- $\varphi < 1/4$: in this area, $Q_1(\lambda; \theta, \varphi)$ is a concave parabola whose positivity on the interval $[0, \Delta t^2\, \rho(A_h)]$ is acquired if and only if $Q_1(\Delta t^2\, \rho(A_h); \theta, \varphi) \geq 0$, which gives the following inequality:

$$\varphi \geq \frac{1}{4}\Big[1 - \frac{4 + (4\theta - 1)\Delta t^2\, \rho(A_h)}{(\theta - 1/12)(\Delta t^2\, \rho(A_h))^2}\Big], \tag{91}$$

  which can be respected only if $\theta > 1/4 - 1/\Delta t^2\, \rho(A_h)$.

- $\varphi \geq 1/4$ and $\theta \geq 1/4$: in this area, $Q_1(\lambda; \theta, \varphi)$ is a convex parabola with a positive slope at the origin, being 4 at the origin. It is then positive for any $\lambda \geq 0$. All schemes are stable in this region.

- $\varphi \geq 1/4$ and $1/12 < \theta < 1/4$: in this area, $Q_1(\lambda; \theta, \varphi)$ is a convex parabola with a negative slope at the origin. Its positivity on $[0, \Delta t^2\, \rho(A_h)]$ can be acquired either if there are no distinct roots ($\Delta(\theta, \varphi) \leq 0$) or if the first root $r(\theta, \varphi)$ is greater than $\Delta t^2\, \rho(A_h)$. The second condition can be written in a way that avoids the inversion of the relation $r(\theta, \varphi)$: $Q_1(\Delta t^2\, \rho(A_h); \theta, \varphi) \geq 0$ and $Q_1'(\Delta t^2\, \rho(A_h); \theta, \varphi) \leq 0$. Introducing

$$\begin{cases} \varphi^\Delta(\lambda, \theta) = \dfrac{1}{4}\Big[1 + \dfrac{(4\theta - 1)^2}{16(\theta - 1/12)}\Big], \\ \varphi^{Q_1}(\lambda, \theta) = \dfrac{1}{4}\Big[1 - \dfrac{4 + (4\theta - 1)\lambda}{(\theta - 1/12)\lambda^2}\Big], \\ \varphi^{Q_1'}(\lambda, \theta) = \dfrac{1}{4}\Big[1 - \dfrac{4\theta - 1}{2(\theta - 1/12)\lambda}\Big], \end{cases} \tag{92}$$

  the positivity of $Q_1(\lambda; \theta, \varphi)$ on $[0, \Delta t^2\, \rho(A_h)]$ is equivalent to

$$\varphi \geq \varphi^\Delta\big(\Delta t^2\, \rho(A_h), \theta\big) \text{ or } \begin{cases} \varphi \geq \varphi^{Q_1}\big(\Delta t^2\, \rho(A_h), \theta\big), \\ \varphi \leq \varphi^{Q_1'}\big(\Delta t^2\, \rho(A_h), \theta\big). \end{cases} \tag{93}$$

It is easy to show that

$$\varphi^{Q_1}\big(\Delta t^2\,\rho(A_h),\theta\big) \le \varphi^{\Delta}\big(\Delta t^2\,\rho(A_h),\theta\big), \qquad \forall \quad \theta > \frac{1}{12}, \tag{94}$$

$$\varphi^{Q_1}\big(\Delta t^2\,\rho(A_h),\theta\big) \le \varphi^{Q_1'}\big(\Delta t^2\,\rho(A_h),\theta\big) \qquad \Leftrightarrow\ \theta \ge \frac{1}{4} - \frac{2}{\Delta t^2\,\rho(A_h)}, \tag{95}$$

$$\varphi^{\Delta}\big(\Delta t^2\,\rho(A_h),\theta\big) \le \varphi^{Q_1'}\big(\Delta t^2\,\rho(A_h),\theta\big) \qquad \Leftrightarrow\ \theta \in [\frac{1}{4} - \frac{2}{\Delta t^2\,\rho(A_h)}, \frac{1}{4}]. \tag{96}$$

Therefore, the lower bound on φ will be $\varphi^{\Delta}\big(\Delta t^2\,\rho(A_h),\theta\big)$ if $\theta \in (1/12, 1/4 - 2/\Delta t^2\,\rho(A_h))$ and $\varphi^{Q_1}\big(\Delta t^2\,\rho(A_h),\theta\big)$ if $\theta \in [1/4 - 2/\Delta t^2\,\rho(A_h), 1/4]$.

To sum up, the stable schemes of the half plane $\theta > 1/12$ are obtained in the region

$$\begin{cases} \varphi \ge \varphi^{\Delta}\big(\Delta t^2\,\rho(A_h),\theta\big), & \text{if } \dfrac{1}{12} < \theta \le \dfrac{1}{4} - \dfrac{2}{\Delta t^2\,\rho(A_h)}, \\ \varphi \ge \varphi^{Q_1}\big(\Delta t^2\,\rho(A_h),\theta\big), & \text{if } \theta \ge \dfrac{1}{4} - \dfrac{2}{\Delta t^2\,\rho(A_h)}. \end{cases} \tag{97}$$

Since the zero level set of $\varepsilon_3(\theta,\varphi)$ is outside this stability zone, the minimum value is achieved on the boundary of the zone, which leads to the following optimization problem:

$$\min_{\theta>1/12} \begin{cases} \dfrac{\theta^2}{4} + \dfrac{\theta}{24} - \dfrac{7}{2880} & \text{when } \dfrac{1}{12} < \theta \le \dfrac{1}{4} - \dfrac{2}{\Delta t^2\,\rho(A_h)}, \\ \theta\,(\dfrac{1}{6} - \dfrac{1}{\Delta t^2\,\rho(A_h)}) - \dfrac{13}{720} + \dfrac{1}{4\Delta t^2\,\rho(A_h)} - \dfrac{1}{(\Delta t^2\,\rho(A_h))^2} & \text{when } \theta \ge \dfrac{1}{4} - \dfrac{2}{\Delta t^2\,\rho(A_h)}. \end{cases} \tag{98}$$

This function is continuous and increasing and can be extended by continuity up to $\theta = 1/12$: the limit value is $1/360$, which is always greater than the optimal value found in the other quadrant. This concludes our proof.

References

[AK00] O Axelsson and A Kucherov. Real valued iterative methods for solving complex symmetric linear systems. *Numerical linear algebra with applications*, 7(4):197–218, 2000.

[BB08] M Benzi and D Bertaccini. Block preconditioning of real-valued iterative algorithms for complex linear systems. *IMA Journal of Numerical Analysis*, 28:598–618, 2008.

[BJR05] E Becache, P Joly, and J Rodríguez. Space-time mesh refinement for elastodynamics. Numerical results. *Computer Methods in Applied Mechanics and Engineering*, 194(2-5):355–366, 2005.

[BM78] Ted Belytschko and Robert Mullen. Stability of explicit-implicit mesh partitions in time integration. *International Journal for Numerical Methods in Engineering*, 12(10):1575–1586, 1978.

[CFJ03a] F Collino, T Fouquet, and P Joly. A conservative space-time mesh refinement method for the 1-D wave equation. II. Analysis. *Numerische Mathematik*, 95(2):223–251, 2003.

[CFJ03b] F Collino, T Fouquet, and P Joly. A conservative space-time mesh refinement method for the 1-d wave equation. Part I: Construction. *Numerische Mathematik*, 95(2):197–221, 2003.

[Coh01] G Cohen. *Higher-order numerical methods for transient wave equations*. Springer-Verlag, 2001.

[DG09] J Diaz and M Grote. Energy conserving explicit local time-stepping for second-order wave equations. *SIAM Journal on Scientific Computing*, 31(3):1985–2014, 2009.

[Fre90] R Freund. On conjugate gradient type methods and polynomial preconditioners for a class of complex non-Hermitian matrices. *Numerische Mathematik*, 57(1):285–312, 1990.

[GJ08] JC Gilbert and P Joly. Higher order time stepping for second order hyperbolic problems and optimal CFL conditions. *Partial Differential Equations*, 16:67–93, 2008.

[JR10] P Joly and J Rodríguez. Optimized higher order time discretization of second order hyperbolic problems: Construction and numerical study. *Journal of Computational and Applied Mathematics*, 234(6):1953–1961, July 2010.

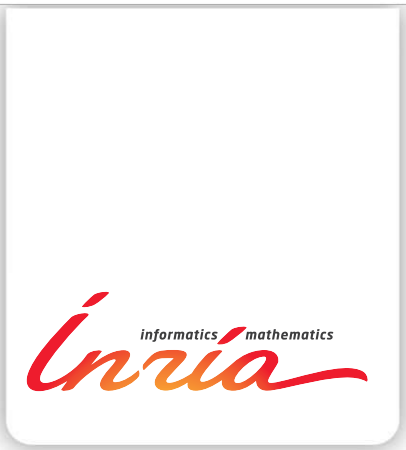
[Kar11] Samir Karaa. Finite element theta-schemes for the acoustic wave equation. *Advances in Applied Mathematics and Mechanics*, 3(2):181–203, 2011.

[KDE08] N A Kampanis, V A Dougalis, and J A Ekaterinaris. *Effective computational methods for wave propagation*. Chapman and Hall/CRC, 2008.

[LLL10] Hui Liang, M. Z. Liu, and Wanjin Lv. Stability of theta-schemes in the numerical solution of a partial differential equation with piecewise continuous arguments. *Applied Mathematics Letters. An International Journal of Rapid Publication*, 23(2):198–206, 2010.

[Ryl02] T Rylander. Stability of Explicit–Implicit Hybrid Time-Stepping Schemes for Maxwell's Equations. *Journal of Computational Physics*, 179(2):426–438, July 2002.

[SB87] Gregory R. Shubin and John 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*, 8(2):135–151, 1987.



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 Avenue de la Vieille Tour,
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399