Higher order time stepping for second order hyperbolic problems and optimal CFL conditions

J. Charles Gilbert and Patrick Joly

INRIA Rocquencourt, BP 105, 78153 Le Chesnay, France
Jean-Charles.Gilbert@inria.fr
Patrick.Joly@inria.fr

Summary. We investigate explicit higher order time discretizations of linear second order hyperbolic problems. We study the even order $(2m)$ schemes obtained by the modified equation method. We show that the corresponding CFL upper bound for the time step remains bounded when the order of the scheme increases. We propose variants of these schemes constructed to optimize the CFL condition. The corresponding optimization problem is analyzed in detail and the analysis results in a specific numerical algorithm. The corresponding results are quite promising and suggest various conjectures.

1 Introduction

We are concerned here with a very classical problem, namely the numerical approximation of second order hyperbolic problems, more precisely problems of the form

$$\frac{d^2u}{dt^2} + \mathcal{A}u = 0, \tag{1}$$

where $\mathcal{A}$ is a linear unbounded positive selfadjoint operator in some Hilbert space V. This appears to be the generic abstract form for a large class of partial differential equations in which u denotes a function $u(x,t)$ from $\Omega \subset \mathbb{R}^d \times \mathbb{R}^+$ in $\mathbb{R}^N$ and $\mathcal{A}$ is a second order differential operator in space, of elliptic nature. Such models are used for wave propagation in various domains of application, in particular in acoustics, electromagnetism, and elasticity [18].

During the past 4 decades, a considerable literature has been devoted to the construction of numerical methods for the approximation of (1). The most recent research deals with the construction of higher order in space and conservative methods for the space semi-discretization of (1) (see for instance [11] and the references therein). These methods lead us to consider a family

(indexed by $h > 0$, the approximation parameter which tends to 0 - typically the stepsize of the computational mesh) of problems of the form:

$$\frac{d^2 u_h}{dt^2} + \mathcal{A}_h u_h = 0, \tag{2}$$

where the unknown u_h is a function of time with value in some Hilbert space V_h (whose norm will be denoted $\|\cdot\|$, even if it does depend on h) and $\mathcal{A}_h$ denotes a bounded self-adjoint and positive operator in V_h (namely an approximation of the second order differential operator $\mathcal{A}$). Several approaches lead naturally to problems of the form (2), among which

- variational finite differences [10, 12, 1],
- finite element methods [8, 7],
- mixed finite element methods [9, 21],
- conservative discontinuous Galerkin methods [17, 15].

Of course, the norm of $\mathcal{A}_h$ blows up when h goes to 0, as

$$\|\mathcal{A}_h\| = O(h^{-2}).$$

It is well known that one has conservation of the discrete energy:

$$E_h(t) = \frac{1}{2} \left\| \frac{du_h}{dt} \right\|^2 + \frac{1}{2}\, a_h(u_h, u_h),$$

where $a_h(\cdot,\cdot)$ is the continuous symmetric bilinear form associated with $\mathcal{A}_h$. From the energy conservation result and the positivity of $\mathcal{A}_h$, one deduces a stability result: the norm of the solution $u_h(t)$ can be estimated in function of the norm of the Cauchy data:

$$u_{0,h} = u_h(0), \quad u_{1,h} = \frac{du_h}{dt}(0),$$

with constants independent of h. This is also a direct consequence of the formula:

$$u_h(t) = [\, \cos \mathcal{A}_h^{\frac{1}{2}} t \,]\; u_{0,h} + [\, \mathcal{A}_h^{-\frac{1}{2}} \sin \mathcal{A}_h^{\frac{1}{2}} t \,]\; u_{1,h},$$

which yields

$$\|u_h(t)\| \leq \|u_{0,h}\| + t\, \|u_{1,h}\|. \tag{3}$$

In what follows, we are interested in the time discretization of (2) by explicit finite difference schemes. More specifically, we are interested in the stability analysis of such schemes, i.e., in obtaining a priori estimates of the form (3) after time discretization. The conservative nature (i.e., the conservation of energy) of the continuous problem can be seen as a consequence of the time reversibility of this equation. That is why we shall favor centered finite difference schemes which preserve such a property at the discrete level.

The most well known scheme is the classical second order leap-frog scheme. Let us consider a time step $\Delta t > 0$ and denote by $u_h^n \in V_h$ an approximation of $u_h(t^n)$, $t^n = n\Delta t$. This scheme is

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h u_h^n = 0. \tag{4}$$

Of course, (4) must be completed by a start-up procedure using the initial conditions to compute u_h^0 and u_h^1. We omit this here for simplicity.

By construction, this scheme is second order accurate in time. Its stability analysis is well known and we have the (see for instance [18])

Theorem 1.1 *A necessary and sufficient condition for the stability of* (4) *is*

$$\frac{\Delta t^2}{4} \|\mathcal{A}_h\| \leq 1. \tag{5}$$

Remark 1.2 The condition (5) appears as an abstract CFL condition. In the applications to concrete wave equations, it is possible to get a bound for $\|\mathcal{A}_h\|$ of the form,

$$\|\mathcal{A}_h\| \leq \frac{4c_+^2}{h^2},$$

where c_+ is a positive constant. This one has the dimension of a propagation velocity and only depends on the continuous problem: it is typically related to the maximum wave velocity for the continuous problem. Therefore, a (weaker) sufficient stability condition takes the form,

$$\frac{c_+ \, \Delta t}{h} \leq 1.$$

In many situations, it is also possible to get a lower bound of the form (where $c_- \leq c_+$ also has the dimension of velocity),

$$\|\mathcal{A}_h\| \geq \frac{4c_-^2}{h^2},$$

so that a necessary stability condition is,

$$\frac{c_- \, \Delta t}{h} \leq 1.$$

□

Next we investigate one way to construct more accurate (in time) discretization schemes for (2). This is particularly relevant when the operator $\mathcal{A}_h$ represents a space approximation of the continuous operator $\mathcal{A}$ in $O(h^k)$ with $k > 2$: if one thinks about taking a time step proportional to the space step h (a usual choice which is in conformity with a CFL condition), one would

like to adapt the time accuracy to the space accuracy. In comparison to what has been done on the space discretization side, we found very few work in this direction, even though it is very likely that a lot of interesting solutions could probably be found in the literature on ordinary differential equations [16]. Most of the existing work is in the context of finite difference methods, compact schemes, etc: see for instance [12, 25, 10, 2] or [13, 26] in the context of first order hyperbolic problems.

The content of the rest of this paper is as follows. In section 2, we investigate a class of methods for the time discretization of (2), based on the so-called modified equation approach. These schemes can be seen as even higher order variations around the leap-frog scheme of which they preserve the main properties: explicit nature, time reversibility, energy conservation. It appears that the computational cost of one time step of the scheme of order $2m$ is m times larger than for one step of the second order scheme. This can be counterbalanced if one can use larger time steps than for the second order scheme. This is where the stability analysis plays a major role (section 2). This one shows that even though the maximum allowed time step increases with m (particularly for small even values of m), it remains uniformly bounded with m (theorem 2.4). In section 3, we investigate the question of constructing other schemes, conceived as modifications of the previous one, that should satisfy:

- the good properties of the schemes (explicitness, conservativity, etc) and the order of approximation are preserved,
- the maximal time step authorized by the CFL condition is larger.

We formulate this as a family of optimization problems that we analyze in detail. We are able to prove the existence and the uniqueness of the solution of these problems (corollary 3.8) and to give necessary and sufficient conditions of optimality (theorems 3.5 and 3.7) that we use to construct an algorithm for the effective computation of the solutions of these optimization problems. This algorithm, as well as the corresponding numerical results, are presented and discussed in section 4. Our first results are quite promising and show that the optimization procedure does allow us to improve significantly the CFL condition. However, the corresponding numerical schemes still have to be tested numerically. This will be the object of a forthcoming work.

2 Higher order schemes by the modified equation approach

2.1 The modified equation approach

It is possible to construct higher order schemes which remain explicit and centered. In particular, all the machinery of Runge-Kutta methods for ordinary differential equations [16] is available. Let us concentrate here on a classical

approach, the so-called modified equation approach [25, 5, 12]. For instance, to construct a fourth order scheme, we start by looking at the truncation error of (4),

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = \frac{d^2u_h}{dt^2}(t^n) + \frac{\Delta t^2}{12}\frac{d^4u_h}{dt^4}(t^n) + O(\Delta t^4).$$

Using the equation satisfied by u_h, we get the identity,

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -\mathcal{A}_h u_h(t^n) + \frac{\Delta t^2}{12}\mathcal{A}_h^2 u_h(t^n) + O(\Delta t^4),$$

which leads to the following fourth order scheme,

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h u_h^n - \frac{\Delta t^2}{12}\mathcal{A}_h^2 u_h^n = 0. \tag{6}$$

This one can be implemented in such a way that each time step involves only 2 applications of the operator $\mathcal{A}_h$, using Horner's rule,

$$u_h^{n+1} = 2u_h^n - u_h^{n-1} - \Delta t^2 \mathcal{A}_h \left(I - \frac{\Delta t^2}{12}\mathcal{A}_h\right) u_h^n.$$

More generally, an explicit centered scheme of order $2m$ is given by,

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h^{(m)}(\Delta t)\, u_h^n = 0, \quad \mathcal{A}_h^{(m)}(\Delta t) = \mathcal{A}_h\, P_m(\Delta t^2 \mathcal{A}_h), \tag{7}$$

where the polynomial $P_m(x)$ is defined by,

$$P_m(x) = 1 + 2\sum_{l=1}^{m-1} (-1)^l \frac{x^l}{(2l+2)!}. \tag{8}$$

Indeed, a Taylor expansion gives:

$$u_h(t^{n\pm 1}) = u_h(t^n) + \sum_{k=1}^{2m+1} (\pm 1)^k \, \frac{\Delta t^k}{k!} \, \frac{d^k u_h}{dt^k}(t^n) + O(\Delta t^{2m+2})$$

so that

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = 2\sum_{k=1}^{m} \frac{\Delta t^{2k-2}}{2k!} \, \frac{d^{2k} u_h}{dt^{2k}}(t^n) + O(\Delta t^{2m}).$$

Since $\dfrac{d^{2k}u_h}{dt^{2k}}(t^n) = (-1)^k \mathcal{A}_h^k u_h(t^n)$, we also have:

$$\begin{aligned}&\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} \\ &\qquad = -\mathcal{A}_h u_h(t^n) + 2\sum_{k=2}^{m} (-1)^k \, \frac{\Delta t^{2k-2}}{2k!} \, \mathcal{A}_h^k u_h(t^n) + O(\Delta t^{2m}),\end{aligned}$$

or equivalently

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} + \mathcal{A}_h \left[u_h(t^n) + 2 \sum_{k=1}^{m-1} (-1)^k \, \frac{\Delta t^{2k}}{(2k+2)!} \, \mathcal{A}_h^k u_h(t^n) \right] = O(\Delta t^{2m}).$$

This identity leads to the scheme (7)-(8).

Using again Horner's rule for the representation of the polynomial P_m reduces the calculation of u_h^{n+1} to m successive applications of the operator $\mathcal{A}_h(\Delta t)$, according to the following algorithm

- Set $u_h^{n,0} = u_h^n$.
- Compute $u_h^{n,k} = u_h^{n,k-1} - 2 \, \dfrac{\Delta t^2 \, \mathcal{A}_h u_h^{n,k-1}}{(2k+1)(2k+2)}, \quad k = 1, \cdots, m,$
- Set $u_h^{n+1} = u_h^{n,m}$.

In other words, since the most expensive step of the algorithm is the application of the operator $\mathcal{A}_h$ (a matrix-vector multiplication in practice), the computational cost for one time step of the scheme of order $2m$ is only m times larger than the computational cost for one time step of the scheme of order 2.

2.2 Stability analysis

The stability analysis of the higher order scheme (7) is similar to the one of the second order scheme but it is complicated by the fact that one must verify that the operator $\mathcal{A}_h(\Delta t)$ is positive, which already imposes an upper bound on Δt.

Theorem 2.1 *A sufficient stability condition for scheme* (7) *is given by,*

$$\Delta t^2 \, \|\mathcal{A}_h\| \leq \alpha_m, \tag{9}$$

where we have defined

$$\alpha_m = \sup \{ \, \alpha \; / \; \forall \, x \in \; [0, \alpha], \; 0 \leq Q_m(x) \leq 4 \, \}, \tag{10}$$

with

$$Q_m(x) = x P_m(x) = x + 2 \sum_{l=1}^{m-1} (-1)^l \frac{x^{l+1}}{(2l+2)!}. \tag{11}$$

This condition is necessary as soon as the spectrum of $\mathcal{A}_h$ is the whole interval $[0, \|\mathcal{A}_h\|]$.

Proof. Using Von Neumann analysis [23] and spectral theory of self-adjoint operators (namely the spectral theorem [22]), it is sufficient to look at the (λ-parameterized) family of difference equations (u^n is now a sequence of complex numbers):

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} + \lambda\, P_m(\lambda \Delta t^2)\, u^n = 0, \quad \lambda \in \sigma(\mathcal{A}_h), \tag{12}$$

where $\sigma(\mathcal{A}_h)$ is the spectrum of $\mathcal{A}_h$. The characteristic equation of this recurrence is

$$r^2 - \left[2 - Q_m(\lambda \Delta t^2)\right] r + 1 = 0.$$

This is a second degree equation with real coefficients. The product of the roots being 1, the two solutions have modulus less than 1 - which is equivalent to the boundedness of u^n - if and only if the discriminant of this equation is nonpositive, in which case the roots belong to the unit circle. This leads to $Q_m(\lambda \Delta t^2)\ \left[4 - Q_m(\lambda \Delta t^2)\right] \geq 0$ or

$$0 \leq Q_m(\lambda \Delta t^2) \leq 4.$$

If (9) holds, since $\sigma(\mathcal{A}_h) \subset [\ 0, \|\mathcal{A}_h\|\ ]$, $\lambda \Delta t^2 \in [0, 4]$ which proves that (9) is a sufficient stability condition. The second part of the proof is left to the reader. □

Remark 2.2 The equality $\sigma(\mathcal{A}_h) = [\ 0, \|\mathcal{A}_h\|\ ]$ holds for instance when one uses a finite difference scheme of the wave equation with constant coefficients in the whole space. The Fourier analysis proves that the spectrum of $\mathcal{A}_h$ is, in this case, purely continuous. □

The finiteness of α_m for each m is quite obvious. However, its value is difficult to compute explicitly, except for the first values of m. One has in particular

$$\alpha_1 = 4, \quad \alpha_2 = 12, \quad \alpha_3 = 2(5 + 5^{\frac{1}{3}} - 5^{\frac{2}{3}}) \simeq 7.572, \quad \alpha_4 \simeq 21.4812, \quad \ldots \tag{13}$$

For the exact - but very complicated - expression of α_4, we refer to [8] or [18]; other values of α_m are given in the column "$k = 0$" of table 1 on page 22. It is particularly interesting to note that for the fourth order scheme, one is allowed to take a time step which is $\sqrt{\alpha_2/\alpha_1}$ ($\simeq 1.732$) times larger than for the second order scheme, which almost balances the fact that the cost of one time step is twice larger. In the same way, with the scheme of order 8, one can take a time step $\sqrt{\alpha_4/\alpha_1}$ ($\simeq 2.317$) times larger (while each time step costs four times more). Surprisingly, the scheme of order 6 seems less interesting: the stability condition is more constraining that for the fourth order scheme. From the theoretical point of view, it would be interesting to know the behaviour of α_m for large m. For this we first identify the limit behaviour of the polynomials $Q_m(x)$. One easily checks that

$$\lim_{m \to +\infty} Q_m(x) = Q_\infty(x) \equiv x + 2 \sum_{l=1}^{+\infty} (-1)^l \frac{x^{l+1}}{(2l+2)!} = 2\,(1 - \cos\sqrt{x}). \tag{14}$$

Remark 2.3 Setting $P_\infty(x) = 2\,\dfrac{1-\cos\sqrt{x}}{x}$ and taking (formally) the limit of (7) when $m \to +\infty$, we obtain the scheme:

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h\ P_\infty(\Delta t^2 \mathcal{A}_h) = 0. \tag{15}$$

This scheme is in fact an *exact scheme* for the differential equation (2). It suffices to remark that:

$$\left|\begin{array}{l} \sin\left(\mathcal{A}_h^{\frac12}\ t^{n+1}\right) - 2\sin\left(\mathcal{A}_h^{\frac12}\ t^{n}\right) + \sin\left(\mathcal{A}_h^{\frac12}\ t^{n-1}\right) \\ \qquad = -\left[2 - \cos\left(\mathcal{A}_h^{\frac12}\Delta t\right)\right]\ \sin\left(\mathcal{A}_h^{\frac12}\ t^{n}\right) \\ \cos\left(\mathcal{A}_h^{\frac12}\ t^{n+1}\right) - 2\cos\left(\mathcal{A}_h^{\frac12}\ t^{n}\right) + \cos\left(\mathcal{A}_h^{\frac12}\ t^{n-1}\right) \\ \qquad = -\left[2 - \cos\left(\mathcal{A}_h^{\frac12}\Delta t\right)\right]\ \cos\left(\mathcal{A}_h^{\frac12}\ t^{n}\right), \end{array}\right.$$

so that any solution of (2), of the form (for some a and b in V_h):

$$u_h(t) = \cos\left(\mathcal{A}_h^{\frac12} t\right)\ a + \sin\left(\mathcal{A}_h^{\frac12} t\right)\ b$$

satisfies:

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -\left(\mathcal{A}_h \Delta t^2\right)^{-1}\ \left[2 - \cos\left(\mathcal{A}_h^{\frac12}\Delta t\right)\right]\ \mathcal{A}_h u_h(t^n),$$

that is to say

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -\mathcal{A}_h\ P_\infty(\Delta t^2 \mathcal{A}_h).$$

□

Since $0 \le Q_\infty(x) \le 4$, if we define α_∞ by (19) for $m = +\infty$ we have $\alpha_\infty = +\infty$. Unfortunately, this does not mean, as we are going to see, that $\alpha_m \to +\infty$ when $m \to +\infty$. In fact, to describe the behaviour of α_m, we have to distinguish between the even and odd sequences α_{2m} and α_{2m+1}. Our first observation is that the convergence of the sequences $Q_{2m}(x)$ and $Q_{2m+1}(x)$ is monotone. Indeed, for $m \ge 1$:

$$Q_{2m-1}(x) - Q_{2m+1}(x) = 2\,\frac{x^{2m}}{4m!}\left[1 - \frac{x}{(4m+1)(4m+2)}\right]$$

which shows that $Q_{2m+1}(x)$ is a strictly decreasing sequence for large m:

$$Q_{2m+1}(x) < Q_{2m-1}(x) \quad \text{as soon as} \quad (4m+1)(4m+2) > x.$$

In particular, since $(4m+1)(4m+2) > \pi^2$ for $m \ge 1$:

$$Q_\infty(\pi^2) = \lim_{m\to+\infty} Q_{2m+1}(\pi^2) = 4 \quad \Longrightarrow \quad Q_{2m+1}(\pi^2) > 4,$$

which shows, using definition (10), that

$$\alpha_{2m+1} \leq \pi^2, \quad \text{for } m \geq 1.$$

Moreover, by definition de α_m, we know that $Q_m(\alpha_m) = 0$ or 4. On the other hand, since the sequence $Q_{2m+1}(x)$ is decreasing, for any $x \in [\,0, \pi^2\,]$, we have

$$Q_{2m+1}(x) > Q_\infty(x) = 2(1 - \cos\sqrt{x}) \text{ in } [0, \pi^2]$$

This makes impossible $Q_{2m+1}(\alpha_{2m+1}) = 0$, which implies that

$$Q_{2m+1}(\alpha_{2m+1}) = 4.$$

Finally the inequality:

$$Q_{2m+1}(x) < Q_1(x) = x$$

implies

$$Q_{2m+1}(x) < 4, \quad \forall\, x \in [0, 4],$$

which implies in particular

$$\alpha_{2m+1} > 4.$$

Let $\alpha_{\text{odd}} \in\ [4, 4\pi^2]$ be any accumulation point of α_{2m+1}, since the convergence of Q_m to Q_∞ is uniform in any compact set, we get:

$$Q_\infty(\alpha_{\text{odd}}) \quad \Longrightarrow \quad (\text{ since } \alpha_{\text{odd}} \in\ [4, \pi^2]\,) \quad \alpha_{\text{odd}} = \pi^2.$$

In the same way:

$$Q_{2m+2}(x) - Q_{2m}(x) = 2\,\frac{x^{2m+1}}{(4m+2)!}\left[1 - \frac{x}{(4m+3)(4m+4)}\right]$$

shows that the sequence $Q_{2m}(x)$ is strictly increasing for large m:

$$Q_{2m+2}(x) > Q_{2m}(x) \quad \text{as soon as} \quad (4m+3)(4m+4) > x.$$

In particular, as soon as $m \geq 1$,

$$Q_\infty(4\pi^2) = \lim_{m \to +\infty} Q_{2m}(4\pi^2) = 0 \quad \Longrightarrow \quad Q_{2m}(4\pi^2) < 0.$$

which shows that

$$\alpha_{2m} \leq 4\pi^2, \quad m \geq 1,$$

while the inequality $Q_{2m}(x) < 2(1 - \cos\sqrt{x}) \leq 4 \quad \text{in} \quad [0, \pi^2]$ for $m \geq 1$ implies that

$$Q_{2m}(\alpha_{2m}) = 0.$$

Finally the inequality, for $m > 1$

$$Q_{2m}(x) > Q_2(x) = x(1 - x/12) \quad \text{for} \quad x < 132$$

shows that $Q_{2m}(x) > 0$ for $x < 12$ which implies that

$$\alpha_{2m} \geq 12.$$

Let $\alpha_{\text{even}} \in [12, 4\pi^2]$ be any accumulation point of α_{2m}, we thus get:

$$Q_\infty(\alpha_{\text{even}}) = 0 \quad \Longrightarrow \quad (\text{since } \alpha_{\text{even}} \in [12, 4\pi^2]) \quad \alpha_{\text{even}} = 4\pi^2.$$

We have shown the following result:

Theorem 2.4 *Let α_m be defined by* (10)*, then:*

$$\lim_{m \to +\infty} \alpha_{2m} = 4\pi^2, \quad \lim_{m \to +\infty} \alpha_{2m+1} = \pi^2. \tag{16}$$

3 Modified higher order schemes: an optimization approach

For an integer k, we denote by $\mathbf{P}_k$ the set of polynomials of degree less or equal to k and define $\mathbf{P} \equiv \cup_{k \geq 0} \mathbf{P}_k$.

A general explicit scheme of order $2m$ is given by:

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \left[P_m(\Delta t^2 \mathcal{A}_h) + \Delta t^{2m} \mathcal{A}_h^m R_k(\Delta t^2 \mathcal{A}_h) \right] \mathcal{A}_h u_h^n = 0, \tag{17}$$

where $R_k \in \mathbf{P}_{k-1}$. The cost of this new scheme is a priori $(m+k)/m$ times larger than the cost of the scheme corresponding to $R_k = 0$. As in theorem 2.1, the stability condition of this new scheme is:

$$\frac{\Delta t^2}{4} \|\mathcal{A}_h\| \leq \alpha_m(R_k), \tag{18}$$

where we have defined,

$$\alpha_m(R) = \sup \{ \alpha \ / \ \forall x \in [0, \alpha],\ 0 \leq x [P_m(x) + x^m R(x)] \leq 4 \}. \tag{19}$$

The natural idea, in some sense, to get an optimal scheme would be to solve the optimization problem:

$$\text{Find } R_{m,k} \in \mathbf{P}_{k-1} \ / \ \alpha_m(R_{m,k}) = \sup_{R \in \mathbf{P}_{k-1}} \alpha_m(R). \tag{20}$$

Then, assuming that this problem has a solution $R_{m,k}$, one gets the optimal CFL constant for the schemes in the class, namely

$$\alpha_{m,k} = \alpha_m(R_{m,k}). \tag{21}$$

Clearly, since $\mathbf{P}_{k-1} \subset \mathbf{P}_k$, $\alpha_{m,k}$ increases with k. We have also $\alpha_{m,k} > 0$, since $P_m(0) = 1$ $(m \geq 1)$.

For what follows, it is useful to introduce the following affine map

$$\left| \begin{array}{l} \psi_m \ : \mathbf{P} \to \mathbf{P} \\ \qquad R \to \psi_m(R) = Q_m + x^{m+1}\ R, \end{array} \right. \tag{22}$$

where we recall that $Q_m(x) = x\ P_m(x)$. Note that ψ_m maps $\mathbf{P}_{k-1}$ into $\mathbf{P}_{m+k}$.

Lemma 3.1 *The function $R \in \mathbf{P}_{k-1} \mapsto \alpha_m(R) \in \mathbb{R}_+^*$ has the following properties.*

- *It goes to 0 at infinity:*

$$\lim_{\|R\| \to +\infty} \alpha_m(R) = 0.$$

- *It is upper semi-continuous:*

$$R_n \to R \text{ in } \mathbf{P}_{k-1} \quad \Longrightarrow \quad \alpha_m(R) \geq \limsup\ \alpha_m(R_n).$$

Proof. Let $r_j(R)$ denote the coefficient of x^j in $R \in \mathbf{P}_{k-1}$ and consider $R_n \in \mathbf{P}_{k-1}$ such that:

$$\|R_n\|_\infty \equiv \sup_{0 \leq j \leq k-1} |r_j(R_n)| \quad \to \quad +\infty.$$

Referring to the fact that $\mathbf{P}_{k-1}$ is finite dimensional, one can find a subsequence (still denoted R_n for simplification) and a fixed non zero polynomial $\varphi \in \mathbf{P}_{k-1}$ such that, as soon as $\varphi(x) \neq 0$,

$$R_n(x) \sim \|R_n\|_\infty\ \varphi(x) \quad (n \to +\infty).$$

For such positive values of x, $[\psi_m(R_n)](x) \notin [0,4]$ for sufficiently large n which means that $\alpha_m(R_n) < x \Longrightarrow \limsup \alpha_m(R_n) < x$. Since φ is a non zero polynomial, one can find arbitrarily small values of such x so that $\limsup \alpha_m(R_n) \leq 0$. As $\alpha_m(R_n)$ is a sequence of positive real numbers, this means that $\alpha_m(R_n)$ tends to 0.

On the other hand, let $R_n \in \mathbf{P}_{k-1}$ be a sequence converging to R. Let ε be any arbitrarily small positive number. By the uniform convergence of R_n to R in the interval $I_R(\varepsilon) = [0, \alpha(R) + \varepsilon]$ we have:

$$\lim_{n \to +\infty} \|\psi_m(R_n) - 2\|_{L^\infty\left(I_R(\varepsilon)\right)} = \|\psi_m(R) - 2\|_{L^\infty\left(I_R(\varepsilon)\right)} > 2.$$

Thus, there exists an integer N_ε such that:

$$n \geq N_\varepsilon \quad \Longrightarrow \quad \|\psi_m(R_n) - 2\|_{L^\infty\left(I_R(\varepsilon)\right)} > 2 \quad \Longrightarrow \quad \alpha_m(R_n) < \alpha_m(R) + \varepsilon.$$

Therefore

$$\limsup\ \alpha_m(R_n) \leq \alpha_m(R) + \varepsilon,$$

which yields (ε being arbitrarily small) $\limsup \alpha_m(R_n) \leq \alpha_m(R)$. □

The classical existence theory in analysis [24, theorem 2.7.11] leads an existence result.

Corollary 3.2 (existence of a solution) *The optimization problem* (20) *has* (*at least*) *one solution.*

Clearly, the function $R \to \alpha_m(R)$ is not continuous. Let us consider for instance the case when $m = 1$ and $k = 1$. Then, the function $\alpha_1(R)$ can be identified to the function of the real variable r defined by

$$\alpha_1(r) = \sup \{ \alpha \ / \ \forall\, x \in [0, \alpha],\ 0 \leq x - rx^2 \leq 4 \}. \tag{23}$$

It is straightforward to compute that:

$$\alpha_1(r) = \frac{1 - \sqrt{1 - 16r}}{2r} \quad \text{if} \quad r < \frac{1}{16}, \quad \text{and} \quad \alpha_1(r) = \frac{1}{r} \quad \text{if} \quad r \geq \frac{1}{16}.$$

It is clear that α_1 is discontinuous at $r = 1/16$ since (see also figure 1)

$$\alpha_1(1/16) = 16 \quad \text{and} \quad \lim_{r \uparrow 1/16} \alpha_1(r) = 8.$$

Note that for $r = 1/16$ the graph of the polynomial $x - rx^2$ is tangent to the

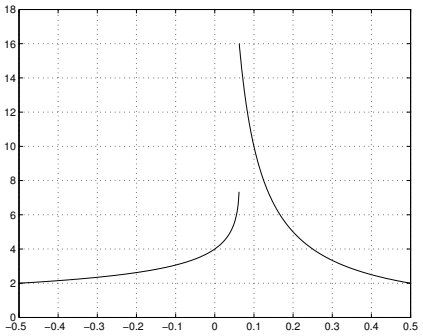

Fig. 1. Graph of the function $\alpha_1(r)$

line $y = 4$ at $x = 8 < \alpha_1(1/16) = 16$. This is an illustration of a more general property.

Lemma 3.3 *Let D_k be the set of polynomials $R \in \mathbf{P}_{k-1}$ such that*

$$\exists\, x_* \in]0, \alpha_m(R)[\quad / \quad [\psi_m(R)](x_*) = 0 \text{ or } 4. \tag{24}$$

The function $R \to \alpha_m(R)$ is discontinuous at every point of D_k and continuous everywhere else.

Proof. Let $R \in D_k$ be such that $[\psi_m(R)](x_*) = 4$ for some $x_* \in]0, \alpha_m(R)[$ (a similar argument works if $[\psi_m(R)](x_*) = 0$). For any $\varepsilon > 0$, $\psi_m(R + \varepsilon) = \psi_m(R) + \varepsilon x^{m+1} > 4$ in a small neighborhood of x_*. This implies that $\alpha_m(R + \varepsilon) < x_* < \alpha_m(R)$, hence the discontinuity of α_m at R.

On the other hand, let $R \in \mathbf{P}_{k-1} \backslash D_k$ and consider a sequence of polynomials $R_n \in \mathbf{P}_{k-1}$ converging to R. Since

$$\left| \frac{[\psi_m(R_n)](x) - [\psi_m(R)](x)}{x^{m+1}} \right| = |R_n(x) - R(x)| \to 0,$$

uniformly in $x \in [0, \alpha_m(R)]$, there exists an integer N_1 such that $[\psi_m(R)](x) - x^{m+1} \leq [\psi_m(R_n)](x) \leq [\psi_m(R)](x) + x^{m+1}$, for $n \geq N_1$ and $x \in [0, \alpha_m(R)]$. These inequalities and the fact that $[\psi_m(R)](0) = 0$ and $[\psi_m(R)]'(0) = 1$ imply that there is an $\varepsilon_1 > 0$ such that $[\psi_m(R_n)](x) \in [0, 4]$, for $n \geq N_1$ and $x \in [0, \varepsilon_1]$. In other words,

$$\text{for } n \geq N_1, \quad \alpha_m(R_n) \geq \varepsilon_1.$$

For any $\varepsilon \in]0, \varepsilon_1]$, small enough, and $J_R(\varepsilon) = [\varepsilon, \alpha_m(R) - \varepsilon]$, there holds

$$\|\psi_m(R) - 2\|_{L^\infty\left(J_R(\varepsilon)\right)} < 2.$$

Then there exists an integer $N_\varepsilon \geq N_1$ such that, for $n \geq N_\varepsilon$:

$$\|\psi_m(R_n) - 2\|_{L^\infty\left(J_R(\varepsilon)\right)} < 2 \quad \text{or} \quad \alpha_m(R_n) > \alpha_m(R) - \varepsilon.$$

Now $\varepsilon > 0$ is arbitrary small, so that $\liminf \alpha_m(R_n) \geq \alpha_m(R)$. The continuity of α_m at R follows, since α_m is upper semi-continuous by lemma 3.1. □

Lemma 3.4 *The set of solutions of the optimization problem* (20) *is a convex subset of* D_k.

Proof. Let us first prove that any local maximum of α_m belongs to D_k. Indeed, it is easy to see that, if $R \notin D_k$, the function

$$t \in \mathbb{R} \mapsto \alpha_m(R + t)$$

is continuous and strictly monotone in the neighborhood of the origin. This shows that R cannot be a local maximum of α_m.

Let R_1 and R_2 be two solutions of (20):

$$\alpha_m(R_1) = \alpha_m(R_2) = \alpha_{m.k} \equiv \sup_{R \in \mathbf{P}_{k-1}} \alpha_m(R).$$

By definition of α_m

$$\forall\, x \leq \alpha_{m,k}, \quad 0 \leq [\psi_m(R_1)](x) \leq 4 \quad \text{and} \quad 0 \leq [\psi_m(R_2)](x) \leq 4.$$

Therefore, since ψ_m is an affine function, for any $t \in [\,0,1]$, there holds

$$\forall\, x \le \alpha_{m,k}, \quad 0 \le \big[\psi_m\big(tR_1 + (1{-}t)R_2\big)\big]\,(x) \le 4.$$

Hence

$$\alpha_m\big(tR_1 + (1{-}t)R_2\big) = \alpha_{m,k}.$$

In other words, any point of the segment $[R_1, R_2]$ is a solution of (20), i.e., the set of solutions of (20) is convex. □

As a consequence of lemma 3.3 and lemma 3.4, we know that any solution R of (20) is such that

$$\mathcal{T}_R \equiv \{\tau \in \,]0, \alpha_{m,k}[\quad / \quad [\psi_m(R)]\,(\tau) = 0 \text{ or } 4\}$$

is nonempty. Let us call *tangent point* an element of $\mathcal{T}_R$. Theorem 3.5 below is more precise, since it claims that there is at least $M \ge k$ tangent points τ_j at which $\psi_m(R)$ takes *alternatively* the values 0 and 4. For any R, it is convenient to construct and enumerate these tangent points in decreasing order:

$$\tau_{M+1} = 0 < \tau_M < \cdots < \tau_1 < \tau_0 = \alpha_{m,k}.$$

The selected subset $\{\tau_1, \tau_2, \ldots, \tau_M\} \subset \mathcal{T}_R$ is built as follows. Let us start by setting

$$\tau_0 = \alpha_{m,k} \quad \text{and} \quad s_0 = \begin{cases} -1 \text{ if } [\psi_m(R)]\,(\tau_0) = 4 \\ +1 \text{ if } [\psi_m(R)]\,(\tau_0) = 0. \end{cases} \tag{25}$$

The points $\tau_j \in \mathcal{T}_R$, $j = 1, \ldots, M$ and their number M are determined by the following recurrence. For $j \ge 1$, set

1) set $s_j = -s_{j-1}$, (26)

2) if this is possible, take τ_j as the largest $\tau \in \,]\,0, \tau_{j-1}[$ such that (27)

$$[\psi_m(R)]\,(\tau_j) = \begin{cases} 4 & \text{if } s_j = -1 \\ 0 & \text{if } s_j = +1. \end{cases}$$

The procedure stops when there is no relevant τ_j in step 2 above (it must stop because of the polynomial nature of $\psi_m(R)$). In the proof of theorem 3.5 below, s_j is actually the sign at τ_j of a certain function φ that is added to a potential solution R.

A priori, because of the chosen selection procedure, it may occur that $M = 0$, even though the number of tangent points is nonzero. The next theorem shows that this is not the case for a local maximum.

Theorem 3.5 (a necessary optimality condition) *Let R be a local maximum of* (20)*. Then the number M of alternate tangent points selected by the procedure* (25)–(27) *satisfies $M \ge k$.*

Proof. We proceed by contradiction, assuming that $M \leq k-1$. For $j = 0, \ldots, M-1$, one can find a point

$$\tau_{j+\frac{1}{2}} \in]\tau_{j+1}, \tau_j[\text{ such that } [\psi_m(R)]\left(]\tau_{j+1}, \tau_{j+\frac{1}{2}}]\right) \subset]0,4[. \tag{28}$$

Consider the polynomial φ defined at $x \in \mathbb{R}$ by

$$\varphi(x) = s_0 \prod_{j=0}^{M-1} \left(x - \tau_{j+\frac{1}{2}}\right).$$

Hence $\varphi \equiv s_0$ if $M = 0$. This polynomial is of degree $M \leq k-1$, so that it is a possible increment to R in $\mathbf{P}_{k-1}$. For $t > 0$, consider the polynomial $p_t = \psi_m(R + t\varphi)$, which verifies for all $x \in \mathbb{R}$:

$$p_t(x) = [\psi_m(R)](x) + tx^{m+1}\varphi(x).$$

We shall get a contradiction and conclude the proof if we show that, for any small $t > 0$, $p_t(x) \in]0,4[$ for $x \in]0, \alpha_{m,k}]$ (since then $\alpha_m(R+t\varphi) > \alpha_{m,k}$ and R would not be a local maximum).

We shall only consider the case when $[\psi_m(R)](\alpha_{m,k}) = 4$, since the reasoning is similar when $[\psi_m(R)](\alpha_{m,k}) = 0$. Then $s_0 = -1$ by (25).

- On the interval $]\tau_{1/2}, \alpha_{m,k}]$, $\psi_m(R)$ is greater than a positive constant (since it is positive on $]\tau_1, \alpha_{m,k}]$ by definition of τ_1 and $\tau_1 < \tau_{1/2} < \alpha_{m,k}$). On the other hand, on this interval, φ is negative (since $s_0 < 0$) and bounded. Therefore, for $t > 0$ small enough, $p_t \in]0,4[$ on this interval.
- Since $\varphi(\tau_{1/2}) = 0$, $p_t(\tau_{1/2}) = [\psi_m(R)](\tau_{1/2})$, which is in $]0,4[$ by definition of $\tau_{1/2}$ in (28).
- On the interval $]\tau_{3/2}, \tau_{1/2}[$, $\psi_m(R)$ is less than a constant < 4 (since it is < 4 on $]\tau_2, \tau_{1/2}]$ by definition of τ_2 and $\tau_{1/2}$, see (27) and (28), and $\tau_2 < \tau_{3/2} < \tau_1 < \tau_{1/2}$). On the other hand, φ is positive and bounded on this interval. Therefore, for $t > 0$ small enough, $p_t \in]0,4[$ on this interval.

We proceed similarly for the other points $\tau_{j+1/2}$ $(j = 1, \ldots, M-1)$ and intervals $]\tau_{j+3/2}, \tau_{j+1/2}[$ $(j = 1, \ldots, M-2)$. Let us now consider the interval $]0, \tau_{M-1/2}[$, which contains tangent points that are all at $y = 0$ or all at $y = 4$.

- If $s_M > 0$ then, on the considered interval, the tangent points are all at $y = 0$, $\psi_m(R)$ is less than a constant < 4, and φ is positive. It results that, for $t > 0$ small enough, $p_t(\cdot) \in]0,4[$ on the interval.
- If $s_M < 0$ then, on the considered interval, the tangent points are all at $y = 4$, $\psi_m(R)$ is positive, and φ is negative. Since the map $x \mapsto [\psi_m(R)](x)/x = 1 + c_1 x + \ldots$ is greater than a positive constant on the considered interval, the map $x \mapsto [\psi_m(R)](x)/x + tx^m\varphi(x) = p_t(x)/x$ is also positive on the interval for $t > 0$ sufficiently small. It results that, for $t > 0$ small enough, $p_t(\cdot) \in]0,4[$ on the considered interval. □

Our next result shows that the necessary optimality conditions of theorem 3.5 are also sufficient. We shall need the following lemma on polynomials.

Lemma 3.6 *If $P \in \mathbf{P}_{k-1}$ takes alternatively nonnegative and nonpositive values at $k+1$ successive distinct points, then $P = 0$.*

Proof. Without loss of generality, we can assume that, for points $x_0 < x_1 < \cdots < x_k$, there hold

$$(-1)^j P(x_j) \geq 0, \quad \text{for } j = 0, 1, \ldots, k. \tag{29}$$

Let us introduce the set of indices

$$\mathbf{I}(P) = \{j \in \{0, 1, \ldots, k\} \ / \ P(x_j) = 0\}.$$

When $\mathbf{I}(P) = \{0, 1, \ldots, k\}$ (resp. $\mathbf{I}(P) = \emptyset$), the conclusion is straightforward since then P has $k+1$ (resp. k) roots.

Suppose now that $\mathbf{I}(P) \neq \emptyset$ and $\mathbf{I}(P) \neq \{0, 1, \ldots, k\}$. Let us introduce the Lagrange interpolation polynomials associated with the x_j's:

$$P_l(x) = \prod_{\substack{j \in \mathbf{I}(P) \\ j \neq l}} \frac{(x - x_j)}{(x_l - x_j)}.$$

Note that all the P_l's belong to $\mathbf{P}_{k-1}$ since $\mathbf{I}(P)$ contains at most k points. For $\varepsilon > 0$, we introduce

$$P_\varepsilon = P + \varepsilon \sum_{l \in \mathbf{I}(P)} (-1)^l \, P_l$$

and note that

$$\forall \ j \ \in \mathbf{I}(P), \quad (-1)^j P_\varepsilon(x_j) = \varepsilon > 0.$$

On the other hand, since $P_\varepsilon \to P$ uniformly on $[x_0, x_k]$, there exists an $\varepsilon_0 > 0$ such that:

$$\forall \varepsilon < \varepsilon_0, \quad \forall \ j \ \notin \mathbf{I}(P), \quad (-1)^j P_\varepsilon(x_j) > 0.$$

Therefore, for $\varepsilon < \varepsilon_0$, P_ε satisfies (29) with moreover $\mathbf{I}(P_\varepsilon) = \emptyset$. This implies that $P_\varepsilon = 0$. By taking the limit when ε tends to 0, we get $P = 0$ (actually this contradicts the fact that $\mathbf{I}(P)$ can be nonempty and different from $\{0, \ldots, k\}$). □

Theorem 3.7 (a sufficient condition of optimality) *Suppose that $P = \psi_m(R)$, for some $R \in \mathbf{P}_{k-1}$, have k tangent points $\{\tau_j\}_{j=1}^k$ such that $0 < \tau_k < \cdots < \tau_1 < \tau_0 = \alpha_m(R)$ and $P(\tau_j) + P(\tau_{j+1}) = 4$ for $j = 0, \ldots, k-1$. Then R is optimal for problem* (20).

Proof. Let $P_{m,k} = \psi_m(R_{m,k})$ be an optimal polynomial (corollary 3.2). The difference $D = R - R_{m,k} \in \mathbf{P}_{k-1}$ takes at $x > 0$ the value

$$D(x) = \frac{P(x) - P_{m,k}(x)}{x^{m+1}}.$$

Since $R_{m,k}$ is optimal, $P_{m,k}(\tau_j) \in [0, 4]$ for $j = 0, \ldots, k$. Then $D(\tau_j) \geq 0$ (resp. $D(\tau_j) \leq 0$) when $P(\tau_j) = 4$ (resp. $P(\tau_j) = 0$). Since $P(\tau_j)$, $j = 0, \ldots, k$, alternates in $\{0, 4\}$, we have shown that

$$(-1)^j \left(P(\tau_0) - 2\right) D(\tau_j) \geq 0, \quad \text{for } j = 0, \ldots, k.$$

These inequalities tell us that $D \in \mathbf{P}_{k-1}$ satisfies the conditions of lemma 3.6. Therefore $D = 0$, proving the R is optimal. □

The necessary and sufficient optimality conditions of theorems 3.5 and 3.7 will be used to determine the optimal polynomials in section 4. We conclude this section with two corollaries of these optimality conditions. The first one deals with the uniqueness of the solution. The second one provides a full description of the optimal polynomials when $m = 1$, relating them to the Chebyshev polynomials of the first kind [6, 20, 27].

Corollary 3.8 (uniqueness of the solution) *The maximization problem* (20) *has one and only one solution. It has no other local maximum.*

Proof. Existence has been quoted in corollary 3.2. Uniqueness is is actually a by-product of the proof of theorem 3.7, where it is shown that if a polynomial $P = \psi_m(R)$, for some $R \in \mathbf{P}_{k-1}$, satisfies the optimality conditions (this is the case for any local maximum, by theorem 3.5), then R is equal to an arbitrarily fixed solution. Hence there cannot be more than one solution or local maximum. □

Corollary 3.9 (optimal polynomials when $m = 1$) *For $k \geq 0$,*

$$\alpha_{1,k} = 4(k+1)^2 \tag{30}$$

and the optimal polynomial $\psi_1(R_{1,k})$ takes at $x \in [0, \alpha_{1,k}]$ the value

$$[\psi_1(R_{1,k})](x) = 2\left[1 - T_{k+1}\left(1 - \frac{2x}{\alpha_{1,k}}\right)\right], \tag{31}$$

where T_k denotes the Chebyshev polynomial of the first kind and degree k, which verifies $T_k(x) = \cos(k \arccos x)$, for $x \in [-1, 1]$.

Proof. Let $\alpha_{1,k}$ be defined by (30) and let φ be the function defined at $x \in [0, \alpha_{1,k}]$ by the right hand side of (31). The fact that $\varphi \equiv \psi_1(R_{1,k})$ will result from the following observations.

- $\varphi \in \psi_1(\mathbf{P}_{k-1})$. Indeed, $\varphi \in \mathbf{P}_{k+1}$. On the other hand, the above formula of T_k shows that $T_k'(1) = k^2$, so that $\varphi'(0) = 4T_{k+1}'(1)/\alpha_{1,k} = 1$, which indicates that the coefficient of x in φ is the one of Q_1.
- The formula of T_k clearly shows that $\varphi(x) \in [0,4]$ for $x \in [0, \alpha_{1,k}]$. On the other hand, $\varphi(\alpha_{1,k}) = 2[1 + (-1)^k]$ and $\varphi'(\alpha_{1,k}) = 4T_{k+1}'(-1)/\alpha_{1,k} = (-1)^k$, so that φ gets out of $[0,4]$ at $x = \alpha_{1,k}$.
- The formula of T_k shows that

$$\varphi(\tau) = 0 \text{ when } \tau = 2(k{+}1)^2 \left(1 - \cos\frac{2j\pi}{k{+}1}\right),\ 0 < 2j < k{+}1,$$
$$\varphi(\tau) = 4 \text{ when } \tau = 2(k{+}1)^2 \left(1 - \cos\frac{(2j{+}1)\pi}{k{+}1}\right),\ 0 < 2j{+}1 < k{+}1,$$

  in which $j \in \mathbb{N}$. Therefore, φ has k tangent points in $]0, \alpha_{1,k}[$, at which φ takes alternatively the value 4 and 0.

Using the last observation and the fact that $\varphi(\alpha_{1,k}) = 2[1 + (-1)^k]$ (= 0 if k is odd and = 4 if k is even) show that φ satisfies the sufficient optimality conditions (theorem 3.7). Hence $\varphi = \psi_1(R_{1,k})$. □

Remark 3.10 A natural question is whether the number of tangent points of an optimal polynomial $\psi_m(R_{m,k})$ can be greater than k. The answer to this question depends actually on the coefficients of x^0, ..., x^m, which are fixed in the optimization process. We do not know the answer when the coefficients are those of the polynomial Q_m, but for other coefficients the number of tangent points can be greater than k. The argument is the following. Let $[\psi_{m-1}(R_{m-1,2})](x) = Q_{m-1}(x) + x^m(r_0 + r_1 x)$ be the optimal polynomial with $m-1$ fixed and 2 free coefficients. By the previous theorem, it has at least 2 tangent points. Now, consider the function $\tilde{\psi}_m$ obtained by replacing in ψ_m defined by (22), Q_m by the polynomial $x \mapsto Q_{m-1}(x) + r_0 x^m$. Clearly the optimal polynomial associated with $\tilde{\psi}_m$ on $\mathbf{P}_0$ is $\tilde{\psi}_m(\tilde{R}_{m,1})$ where $\tilde{R}_{m,1}$ is the constant r_1. Therefore $\tilde{\psi}_m(\tilde{R}_{m,1}) = \psi_{m-1}(R_{m-1,2})$ has 2 tangent points, although the minimization has been done on $\mathbf{P}_0$. □

Remark 3.11 When checking optimality by looking at the alternate character of $[\psi_m(R)](\tau_j)$ in $\{0,4\}$, one has to include the point $\tau_0 = \alpha_m(R)$. In particular, when $k = 1$, a polynomial with a single tangent point may not be optimal. An example with $m = 4$ and $k = 1$ is shown in figure 2. The optimal polynomial, given by

$$[\psi_4(R_{4,1})](x) = x - \frac{x^2}{12} + \frac{x^3}{360} - \frac{x^4}{20160} + rx^5 \quad \text{with } r \simeq 4.28\ 10^{-7},$$

is represented by the solid (blue) curve; the dashed (black) curve is Q_4. The optimal polynomial $[\psi_4(R_{4,1})]$ has only one tangent point $\tau_1 \simeq 33,39$, while $\tau_0 = \alpha_{4,1} \simeq 44.03$. As predicted by theorem 3.5, $[\psi_4(R_{4,1})](\tau_1) + [\psi_4(R_{4,1})](\tau_0) = 4$.

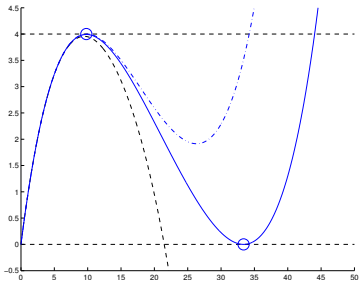

Fig. 2. Checking the sufficient condition of optimality for $m = 4$ and $k = 1$

Now, by increasing r to $r \simeq 5.13\ 10^{-7}$, one gets the dash-dotted (blue) curve, which has a tangent point at $\tau_1 \simeq 9.88$, but is not optimal since the value of the polynomial at this point does not satisfy $[\psi_4(R_{4,1})](\tau_1)+[\psi_4(R_{4,1})](\tau_0) = 4$ (for this polynomial $\tau_0 \simeq 34.22$). □

4 Computational issues

4.1 An algorithm based on the parametrization by the tangent points

In the numerical results discussed below, the optimal polynomial is searched by its k alternate tangent points $(\tau_j)_{1\leq j\leq k}$, with $\tau_1 > \tau_2 > \cdots > \tau_k$, whose existence is ensured by theorem 3.5. These points are determined in the following manner. For $\tau = (\tau_1, \ldots, \tau_k)$, let $R(\tau)$ be the polynomial in $\mathbf{P}_{k-1}$ satisfying

$$\psi_m(R(\tau)) = v \in \mathbb{R}^k,$$

in which the components of v take alternatively the values 0 and 4. Whether one has to impose $v_1 = 0$ or $v_1 = 4$ is further discussed below. The coefficients $r = (r_0\ \cdots\ r_{k-1})^\top$ of $R(\tau)$ are uniquely determined by the equation above, which can also be written

$$\begin{pmatrix} \tau_1^{m+1} & \cdots & \tau_1^{m+k} \\ \vdots & & \vdots \\ \tau_k^{m+1} & \cdots & \tau_k^{m+k} \end{pmatrix} r = v - \begin{pmatrix} [\psi_m(0)]\,(\tau_1) \\ \vdots \\ [\psi_m(0)]\,(\tau_k) \end{pmatrix}. \tag{32}$$

Next, let us introduce the function $F : \tau \in \mathbb{R}^k \mapsto F(\tau) \in \mathbb{R}^k$, where the components of $F(\tau)$ are the derivatives of the polynomial $\psi_m(R(\tau))$ at the τ_j's:

$$F(\tau) = \begin{pmatrix} [\psi_m(R(\tau))]'(\tau_1) \\ \vdots \\ [\psi_m(R(\tau))]'(\tau_k) \end{pmatrix}.$$

Obviously, there holds $F(\tau) = 0$ if τ is the vector of the alternate tangent points of the optimal polynomial. We propose to determine the root(s) τ of F by Newton's method (see [14, 4] for instance). The procedure could have been improved by using a version of Newton's method that exploits inequalities (see for example [19, 3] and the references thereof) to impose $\tau_1 > \tau_2 > \cdots > \tau_k$ as well as the curvature of the solution polynomial at the tangent points: $[\psi_m(R(\tau))]''(\tau_j)(2 - v_j) \geq 0$, for $1 \leq j \leq k$. We have not adopted this additional sophistication, however.

The Newton method requires the computation of $F'(\tau)$. If we denote by $r_l(\tau)$, $1 \leq l \leq k$, the coefficients of $R(\tau)$, by δ_{ij} the Kronecker symbol, and by $V_k(\tau)$ the Vandermonde matrix of order k, there holds

$$\begin{aligned}\frac{\partial F_i}{\partial \tau_j}(\tau) &= \delta_{ij}\,[\psi_m(R(\tau))]''(\tau_i) + \sum_{l=1}^{k} \frac{\partial r_l}{\partial \tau_j}(\tau)(m+l)\tau_i^{m+l-1} \\ &= \delta_{ij}\,[\psi_m(R(\tau))]''(\tau_i) + \\ &\qquad [\mathrm{Diag}(\tau_1^m, \ldots, \tau_k^m) V_k(\tau)\,\mathrm{Diag}((m+1), \ldots, (m+k)) r'(\tau)]_{ij}\,.\end{aligned}$$

To get an expression of $r'(\tau)$, let us differentiate with respect to τ_j the identity $[\psi_m(R(\tau))](\tau_i) = v_i$. It results

$$\delta_{ij}[\psi_m(R(\tau))]'(\tau_i) + \begin{pmatrix}\tau_i^{m+1} & \cdots & \tau_i^{m+k}\end{pmatrix} \frac{\partial r}{\partial \tau_j}(\tau) = 0.$$

Denoting by $M(\tau)$ the coefficient matrix of the linear system (32), we get

$$\begin{aligned} r'(\tau) &= -M(\tau)^{-1}\,\mathrm{Diag}\left([\psi_m(R(\tau))]'(\tau_1), \ldots, [\psi_m(R(\tau))]'(\tau_k)\right) \\ &= -M(\tau)^{-1}\,\mathrm{Diag}(F(\tau)). \end{aligned}$$

Therefore

$$\begin{aligned} F'(\tau) = \mathrm{Diag}\left([\psi_m(R(\tau))]''(\tau_1), ..., [\psi_m(R(\tau))]''(\tau_k)\right) - \\ \mathrm{Diag}(\tau_1^m, ..., \tau_k^m) V_k(\tau)\,\mathrm{Diag}((m+1), ..., (m+k)) M(\tau)^{-1}\,\mathrm{Diag}(F(\tau)). \end{aligned}$$

Observe that at a solution τ^* the second term above vanishes, so that $F'(\tau^*)$ is diagonal. It is also nonsingular if the second derivatives $[\psi_m(R(\tau^*))]''(\tau_j^*)$ are nonzero. Around such a solution, Newton's method is therefore well defined.

In the numerical results presented below, we have used the solver of nonlinear equations `fsolve` of Matlab (version 7.2), which does not take into account the inequality constraints. The vector v has been determined by adopting the following heuristics. We have *assumed* that the optimal polynomial is negative for all $x < 0$ (it has unit slope at $x = 0$), which implies that r_k, the coefficient of x^{m+k} of the optimal polynomial, has the sign $(-1)^{m+k+1}$; if the assumption

is correct, the optimal polynomial should get out of the interval at $y = 0$ if $m + k$ is even and at $y = 4$ if $m + k$ is odd; according to theorem 3.5, one should therefore take $v_1 = 4 - \varepsilon_v$ if $m + k$ is even and $v_1 = \varepsilon_v$ if $m + k$ is odd. The value of ε_v is taken nonnegative and as close as possible to 0. A positive value of ε_v is usually necessary for counterbalancing rounding errors. The other values of v_i alternate in $\{\varepsilon_v, 4 - \varepsilon_v\}$. The initial point τ is chosen by trials and errors, or according to suggestions made in the discussion below.

The proposed approach has advantages (+) and disadvantages (−).

+ The problem has few variables (just k).
+ The problem looks well conditioned, provided the second derivatives at the tangent points are reasonable, which seems to be the case.
− There is no guarantee that the solution found is the optimal one since a zero of F will not be a solution to the original problem if the polynomial gets out of $[0, 4]$ at a point τ_0 less than τ_1. An example of this situation is given in figure 3. However, if $\tau_0 > \tau_1$ and if $[\psi_m(R)](\tau_0)+[\psi_m(R)](\tau_1) = 4$,

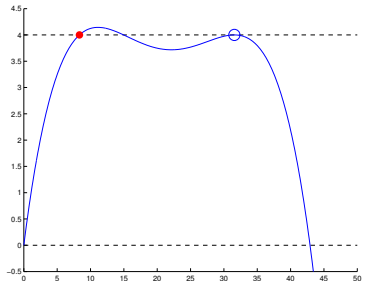

Fig. 3. A zero of F that is not an optimal polynomial ($m = 3$, $k = 1$)

the sufficient optimality conditions of theorem 3.7 guarantee that R is the solution.

− The solution polynomial may get out of the interval $[0, 4]$ near a tangent point due to the lack of precision of the solution, which has motivated the use of the small $\varepsilon_v > 0$.
− Obtaining the convergence to a zero of F (not only a stationary point τ^* of $\|F\|_2^2$, hence verifying $F'(\tau^*)^\top F(\tau^*) = 0$) depends on the initialization of the iterative process.

4.2 Numerical results

Computing $\alpha_{m,k}$

Table 1 shows the computed values of $\alpha_{m,k}$, for $1 \leq m \leq 8$ and $0 \leq k \leq 8$.

Table 1. Computed values of the first $\alpha_{m,k}$'s

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
$m=1$	4.00	16.00	36.00	64.00	100.00	144.00	196.00	256.00	324.00
$m=2$	12.00	32.43	60.56	96.61	140.64	192.66	252.67	320.68	396.69
$m=3$	7.57	23.40	45.72	75.06	111.58	155.38	206.51	265.04	331.00
$m=4$	21.48	44.03	73.45	110.01	153.83	204.98	263.51	329.49	402.92
$m=5$	9.53	31.61	58.23	90.77	129.90	175.84	228.71	288.59	355.23
$m=6$	30.72	57.23	89.78	128.89	174.84	227.71	287.61	354.59	428.71
$m=7$	9.85	37.37	68.93	108.35	151.08	199.56	255.61	317.90	357.95
$m=8$	37.08	70.89	107.67	150.35	199.32	254.89	317.22	386.35	462.27

The computed solutions were always satisfying the optimality conditions, so that we are pretty confident in the values of $\alpha_{m,k}$ in the table. In particular, the small $\varepsilon_v > 0$ hardly modifies these values.

The column $k = 0$ of table 1 corresponds to the polynomials Q_m defined by (11), for which the first values of the $\alpha_{m,0}$'s were already given in (13) (there denoted α_m). We observe that the convergence of $\alpha_{2m+1,0}$ (resp. $\alpha_{2m,0}$) to $\pi^2 \simeq 9.87$ (resp. $4\pi^2 \simeq 39.48$), predicted by theorem 2.4, is rather fast. On the other hand, we observe that the values $\alpha_{m,k}$ can be made spectacularly larger than $\alpha_{m,0}$, which was our objective.

We have verified that the optimal polynomials corresponding to $m = 1$ are indeed related to the Chebyshev polynomials through formula (31), as claimed by corollary 3.9. This fact can be observed in the first row of the table, whose values of $\alpha_{1,k}$ are indeed those given by (30).

Another observation is that the oscillating behaviour of α_m with m, highlighted in the analysis leading to theorem 2.4, is recovered in the sequences $\{\alpha_{m,k}\}_{m\geq 1}$. The reason is similar. The first positive stationary point of the optimal polynomial, which is close to the one of Q_∞, is (resp. is not) a tangent point when m is odd (resp. even). This observation leads to the following conjecture: if we denote by $\tau_{m,k,j}$ the jth tangent point of the optimal polynomial $\psi_m(R_{m,k})$ $(1 \leq j \leq k)$, then, when m goes to infinity, $\tau_{2m+1,k,j}$ (resp. $\tau_{2m,k,j}$) converges the jth (resp. $(j+1)$th) positive stationary point of Q_∞, the polynomial defined by (14). More specifically

$$\tau_{2m+1,k,j} \to j^2\pi^2 \quad \text{and} \quad \tau_{2m,k,j} \to (j+1)^2\pi^2, \quad \text{when } m \to \infty. \tag{33}$$

In practice, these values can be used to choose a good starting point for the algorithm when m is large.

The diagonal schemes $k = m$

We have found interesting to have a particular look at the case $k = m$. First it gives a computational effort per time step that is twice the one for the original $(2m)$th order scheme, which corresponds to $k = 0$. The second reason is more related to intuition: if one wants to get $\alpha_{m,k}$ roughly proportional to m^2, we have to control the first m maxima or minima of the optimal polynomial $\psi_m(R_{m,k})$, for which we think that we need m parameters, which corresponds to $k = m$. Below, we qualify such a scheme as *diagonal*.

Figure 4 shows the optimal polynomials $\psi_m(R_{m,m})$, for $m = 1, \ldots, 8$. The tangent points are quoted by (blue) circles on the graphs, while the $\alpha_{m,m}$'s are quoted by (red) dots.

Table 2. Asymptotic behaviour of the diagonal schemes

m	$\dfrac{\Delta t_{m,m}}{\Delta t_{m,0}}$	$\dfrac{C_{m,m}(T)}{C_{1,0}(T)}$	$\dfrac{2\alpha_{m,m}}{m^2\pi^2}$
1	2.00	1.00	3.24
2	3.89	1.03	3.07
3	4.33	1.39	1.69
4	6.20	1.29	1.95
5	6.63	1.51	1.43
6	8.48	1.42	1.62
7	8.91	1.57	1.31
8	10.75	1.49	1.46
∞		1.80	1.00

Table 2 investigates the asymptotic behaviour of the diagonal schemes.

- Its first column highlights the growth of the ratio between the maximum time step allowed by the stability analysis in a diagonal scheme $\Delta t_{m,m}$ and in the second order scheme $\Delta t_{1,0}$. According to section 2.2, there holds

$$\frac{\Delta t_{m,m}}{\Delta t_{1,0}} = \left(\frac{\alpha_{m,m}}{\alpha_{1,0}}\right)^{1/2} = \frac{\alpha_{m,m}^{1/2}}{2}. \tag{34}$$

- The computational cost $C_{m,m}(T)$ of the diagonal scheme of order $2m$ on an integration time T is proportional to the computational cost $C^1_{m,m}$ of one time step multiplied by the number of time steps. Hence, assuming that the largest time step allowed by the stability analysis is taken, one has

$$C_{m,m}(T) \simeq \frac{C^1_{m,m} T}{\Delta t_{m,m}}.$$

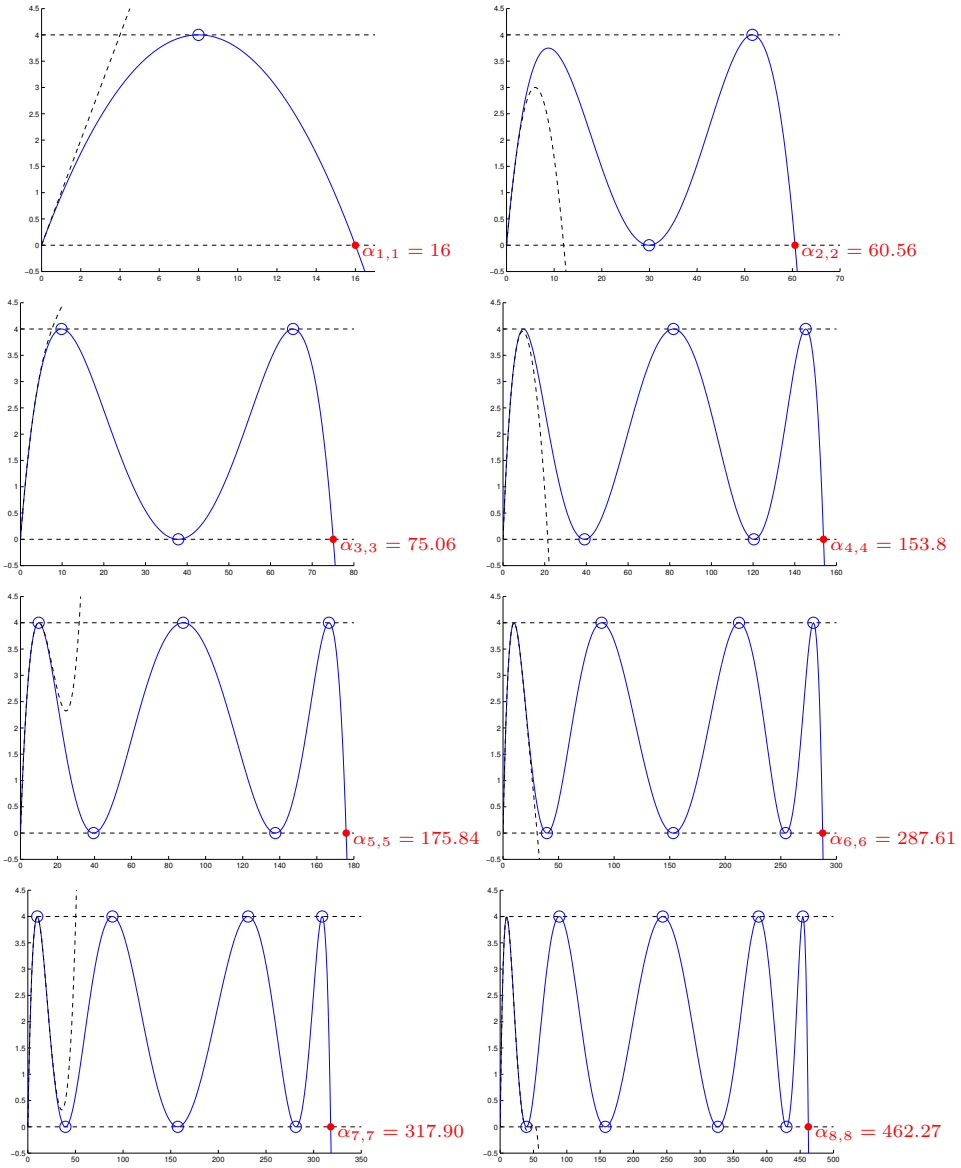

Fig. 4. The polynomials $Q_m = \psi_m(0)$ (dashed curves) and the optimal polynomials $\psi_m(R_{m,m})$ for $m = 1, \ldots, 8$ (solid curves)

A similar expression holds for the computational cost $C_{1,0}(T)$ of the second order scheme, with $C^1_{m,m}$ and $\Delta t_{m,m}$ replaced by $C^1_{1,0}$ and $\Delta t_{1,0}$, respectively. The second column of table 2 gives the ratio of these two costs. Using (34) and the fact that $C^1_{m,m} \simeq 2mC^1_{1,0}$ (each time step of the diagonal scheme requires $2m$ times more operator multiplications than each time step of the second order scheme), the ratio can be estimated by

$$\frac{C_{m,m}(T)}{C_{1,0}(T)} \simeq \frac{4m}{\alpha^{1/2}_{m,m}}.$$

The numbers in the second column of table 2 suggest that this ratio is bounded. If the conjecture (35) below is correct, it should converge to $4\sqrt{2}/\pi \simeq 1.80$, when m goes to infinity.

- Taking $k = m$ and $j = \lceil m/2 \rceil$ in (33), and assuming that $\alpha_{m,m} \sim 2\tau_{m,m,\lceil m/2 \rceil}$ (suggested by the approximate symmetry of the optimal polynomials) lead us to the following conjecture:

$$\frac{\alpha_{m,m}}{m^2} \to \frac{\pi^2}{2}, \quad \text{when } m \to \infty. \tag{35}$$

This conjecture is explored numerically in the third column of table 2. Note that it does not distinguish between even and odd values of m, at least asymptotically. However, looking at the $\alpha_{m,m}$'s on the diagonal of table 1, it appears that the even values of $k = m$ look more interesting than the odd ones.

5 Conclusion

In this paper, we have analyzed the stability of higher order time discretization schemes for second order hyperbolic problems based on the modified equation approach. We have in particular proven that the upper bound for the time step (the CFL limit) remains uniformly bounded for large m ($2m$ is the order of the scheme). On the basis of this information, we have proposed the construction of new schemes that are seen as modifications of the previous ones and are designed in order to optimize the CFL condition: this is formulated as an optimization problem in a space of polynomials of given degree. Despite some unpleasant properties (the objective function is nonconvex and even discontinuous at the solution!), this problem can be fully analyzed. In particular, we prove the existence and uniqueness of the solution and give necessary and sufficient conditions of optimality. These conditions are exploited to design an algorithm for the effective numerical solution of the optimization problem. The obtained results are more than satisfactory with respect to our original objective. They suggest some conjectures that would mean that we would be able to produce schemes of arbitrary high order in time and whose computational cost would be almost independent of the order.

Of course, this is a preliminary work and much has still to be done, including the following items.

- The effective efficiency of the new schemes should be tested on realistic wave propagation problems.
- The impact of the modification of the initial schemes (the ones which are based on the modified equation technique) on the effective accuracy (we are only guaranteed that the order of approximation is preserved) should be analyzed thorough numerical dispersion studies.
- Our various theoretical conjectures should be addressed in a rigorous way.

These will be the subjects of forthcoming works.

References

1. R. M. Alford, K. R. Kelly, Boore D. M. (1974). Accuracy of finite difference modeling of the acoustic wave equation. *Geophysics*, 39, 834–842.
2. Laurent Anné, Patrick Joly, Quang Huy Tran (2000). Construction and analysis of higher order finite difference schemes for the 1D wave equation. *Comput. Geosci.*, 4(3), 207–249.
3. S. Bellavia, B. Morini (2005). An interior global method for nonlinear systems with simple bounds. *Optimization Methods and Software*, 20, 1–22.
4. J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, C. Sagastizábal (2006). *Numerical Optimization – Theoretical and Practical Aspects* (second edition). Universitext. Springer Verlag, Berlin.
5. Romuald Carpentier, Armel de La Bourdonnaye, Bernard Larrouturou (1997). On the derivation of the modified equation for the analysis of linear numerical methods. *RAIRO Modél. Math. Anal. Numér.*, 31(4), 459–470.
6. E.W. Cheney (1966). *Introduction to Approximation Theory*. McGraw-Hill Book Company.
7. M. J. S. Chin-Joe-Kong, W. A. Mulder, M. Van Veldhuizen (1999). Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation. *J. Engrg. Math.*, 35(4), 405–426.
8. G. Cohen, P. Joly, J. E. Roberts, N. Tordjman (2001). Higher order triangular finite elements with mass lumping for the wave equation. *SIAM J. Numer. Anal.*, 38(6), 2047–2078 (electronic).
9. Gary Cohen, Sandrine Fauqueux (2005). Mixed spectral finite elements for the linear elasticity system in unbounded domains. *SIAM J. Sci. Comput.*, 26(3), 864–884 (electronic).
10. Gary Cohen, Patrick Joly (1996). Construction analysis of fourth-order finite difference schemes for the acoustic wave equation in nonhomogeneous media. *SIAM J. Numer. Anal.*, 33(4), 1266–1302.
11. Gary C. Cohen (2002). *Higher-order numerical methods for transient wave equations*. Scientific Computation. Springer-Verlag, Berlin.
12. M. A. Dablain (1986). The application of high order differencing for the scalar wave equation. *Geophysics*, 51, 54–56.
13. Stéphane Del Pino, Hervé Jourdren (2006). Arbitrary high-order schemes for the linear advection and wave equations: application to hydrodynamics and aeroacoustics. *C. R. Math. Acad. Sci. Paris*, 342(6), 441–446.

14. P. Deuflhard (2004). *Newton Methods for Nonlinear Problems – Affine Invariance and Adaptative Algorithms*. Computational Mathematics 35. Springer, Berlin.
15. Loula Fezoui, Stéphane Lanteri, Stéphanie Lohrengel, Serge Piperno (2005). Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes. *M2AN Math. Model. Numer. Anal.*, 39(6), 1149–1176.
16. E. Hairer, G. Wanner (1996). *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition. Stiff and differential-algebraic problems.
17. J. S. Hesthaven, T. Warburton (2002). Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, 181(1), 186–221.
18. Patrick Joly (2003). Variational methods for time-dependent wave propagation problems. In *Topics in computational wave propagation*, volume 31 of *Lect. Notes Comput. Sci. Eng.*, pages 201–264. Springer, Berlin.
19. Ch. Kanzow (2001). An active set-type Newton method for constrained nonlinear systems. In M.C. Ferris, O.L. Mangasarian, J.S. Pang (editors), *Complementarity: applications, algorithms and extensions*, pages 179–200. Kluwer Acad. Publ., Dordrecht.
20. P. Lascaux, R. Théodor (1986). *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur*. Masson, Paris.
21. Sebastien Pernet, Xavier Ferrieres, Gary Cohen (2005). High spatial order finite element method to solve Maxwell's equations in time domain. *IEEE Trans. Antennas and Propagation*, 53(9), 2889–2899.
22. Michael Reed, Barry Simon (1978). *Methods of modern mathematical physics. IV. Analysis of operators*. Academic Press [Harcourt Brace Jovanovich Publishers], New York.
23. Robert D. Richtmyer, K. W. Morton (1967). *Difference methods for initial-value problems*. Second edition. Interscience Tracts in Pure and Applied Mathematics, No. 4. Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney.
24. L. Schwartz (1991). *Analyse I – Théorie des Ensembles et Topologie*. Hermann, Paris.
25. Gregory R. Shubin, John B. Bell (1987). A modified equation approach to constructing fourth-order methods for acoustic wave propagation. *SIAM J. Sci. Statist. Comput.*, 8(2), 135–151.
26. E. F. Toro, V. A. Titarev (2005). ADER schemes for scalar non-linear hyperbolic conservation laws with source terms in three-space dimensions. *J. Comput. Phys.*, 202(1), 196–215.
27. E.W. Weisstein (2006). Chebyshev polynomial of the first kind. *MathWorld*. `http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirstKind.html`.