ON THE DERIVATION OF THE MODIFIED EQUATION FOR THE ANALYSIS OF LINEAR NUMERICAL METHODS

Romuald CARPENTIER[1], Armel de LA BOURDONNAYE[2]
and Bernard LARROUTUROU[3]

Abstract

The modified equation is a powerful tool for the error analysis of the numerical solution of partial differential equations. We present here a method which considerably simplifies the derivation of this equation in the linear case. Our method uses formal expansions, with no elimination step; it keeps the same simplicity when multistep Runge-Kutta schemes are used and in any space dimensions.

UNE METHODE DE CALCUL DES EQUATIONS EQUIVALENTES POUR L'ANALYSE DES METHODES NUMERIQUES LINEAIRES

Résumé

L'équation équivalente est un outil puissant d'analyse d'erreur pour la résolution numérique d'équations aux dérivées partielles. Nous présentons une méthode qui simplifie considérablement l'obtention de cette équation dans le cas linéaire. La méthode présentée utilise des séries formelles et ne nécessite aucune étape d'élimination; elle garde la même simplicité lorsque l'on utilise des schémas de Runge-Kutta et quelque soit la dimension spatiale.

[1] CERMICS, Sophia-Antipolis.

[2] CERMICS, Sophia-Antipolis

[3] Ecole Polytechnique, 91128 PALAISEAU Cedex and CERMICS, Sophia-Antipolis.

1 INTRODUCTION

The modified equation technique, which was introduced by Warming and Hyett [9], is a powerful tool for the analysis of the accuracy and stability of a numerical method aimed at solving a time-dependent problem governed by an evolution partial differential equation. For constant-coefficients linear partial differential equations, it allows a detailed analysis of the truncation error of the numerical methods. In particular, the effect, either dissipative or dispersive, of each error term can be interpreted using the modified equation, so that it allows detailed comparisons between different numerical methods; it may also sometimes be used as a tool for designing new numerical schemes (see e.g. [1, 9]). Lastly, the modified equation may also be used for the numerical analysis of some constant-coefficients nonlinear equations (see e.g. [5, 6, 7]), although the interpretation of the truncation error terms is less easy in the nonlinear case.

Let us briefly recall how the modified equation is derived, on a very simple example. Consider the explicit first-order upwind scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c \, \frac{u_j^n - u_{j-1}^n}{\Delta x} \ , \tag{1}$$

for the solution of the wave equation $w_t = -cw_x$, with $c > 0$; in (1), j and n are the spatial and temporal indices respectively, Δx and Δt are the mesh size and the time step, so that u_j^n is an approximation of $w(j\Delta x, n\Delta t)$.

The modified equation for the scheme (1) is a *formal* partial differential equation, which is derived from the difference equation:

$$\frac{u(x, t+\Delta t) - u(x,t)}{\Delta t} = -c \, \frac{u(x,t) - u(x-\Delta x, t)}{\Delta x} \ , \tag{2}$$

which mimics (1). Assuming that u is $\mathcal{C}^\infty$ in (2), one can deduce from (2) the following Taylor expansions at point (x,t):

$$u_t + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \cdots = -c \left(u_x - \frac{\Delta x}{2} u_{xx} + \frac{\Delta x^2}{6} u_{xxx} + \cdots \right) \ . \tag{3}$$

We will call this equation the *unresolved* modified equation. The goal is now to transform this equation (3), for Δt and Δx small, by replacing the time derivatives, except the first one, by spatial derivatives, using successive differentiations and substitutions. For instance, taking the partial derivative of (3) with respect to t and x, we obtain:

$$u_{tt} + \frac{\Delta t}{2} u_{ttt} + O(\Delta t^2) = -c \left(u_{xt} - \frac{\Delta x}{2} u_{xxt} + O(\Delta x^2) \right) \ , \tag{4}$$

$$u_{tx} + \frac{\Delta t}{2} u_{ttx} + O(\Delta t^2) = -c \left(u_{xx} - \frac{\Delta x}{2} u_{xxx} + O(\Delta x^2) \right) \ , \tag{5}$$

and we can eliminate the mixed derivative u_{tx} from these relations to get:

$$u_{tt} = c^2 u_{xx} + \frac{\Delta t}{2}\left(c u_{ttx} - u_{ttt}\right) + \frac{\Delta x}{2}\left(c u_{xxt} - c^2 u_{xxx}\right) + O(\Delta t, \Delta x)^2 \ , \tag{6}$$

from which, setting $\nu = \dfrac{c\Delta t}{\Delta x}$, we deduce a first form of the modified equation:

$$u_t = -c u_x + \frac{c\Delta x}{2}(1-\nu)u_{xx} + O(\Delta t, \Delta x)^2 \ . \tag{7}$$

Differentiating again (4) and (5) with respect to time and space makes it possible to further eliminate the mixed time-space derivatives. After several steps, we finally obtain:

$$\begin{aligned} u_t = -c u_x + & \ \frac{c\Delta x}{2}(1-\nu)u_{xx} - \frac{c\Delta x^2}{6}(2\nu^2 - 3\nu + 1)u_{xxx} \\ & + \frac{c\Delta x^3}{24}(6\nu^3 - 12\nu^2 + 7\nu - 1)u_{xxxx} + O(\Delta t, \Delta x)^4 \ . \end{aligned} \tag{8}$$

This is the modified equation, expanded up to order three in Δt and Δx. Formally, this is the partial differential equation which is actually solved by the numerical method (1). This equation shows the different terms of the truncation error of the numerical method and their interpretation (we see in (8) the dissipative first-order term, the dispersive second-order term and the dissipative third-order term); in particular, the modified equation (8) shows that the scheme (1) is first-order accurate, and it gives a necessary condition ($\nu \leq 1$) for the stability of the method.

If the final equation (8) is really of interest for the numerical analysis of the scheme (1), it appears however that its derivation is quite heavy and lengthy, even if the successive differentiations and eliminations can be handled using a symbolic computer algebra system as in [8]. In particular, the elimination process which leads from the *unresolved* equation (3) to the *resolved* modified equation (8) may well become much more intricate than in the above example when less simple schemes are considered, for instance in higher space dimensions or with multistep time integration methods. In such cases, even writing the difference equation (2) or the *unresolved* modified equation (3) may become a non trivial task: indeed, a spatially second order accurate scheme uses 5 points in one space dimension, but 9 points in two dimensions, and 33 points with a second-order Runge-Kutta scheme !

It is precisely the objective of this work to present a much simpler way of deriving the modified equation for a linear numerical method. Our method uses formal series expansions, without any elimination step; moreover, it has the advantage of keeping the same simplicity when multistep Runge-Kutta or predictor-corrector schemes are employed, and in any space dimensions.

2 THE MAIN RESULT

Our method for deriving the modified equation applies to any constant-coefficients linear numerical method. Let us consider a linear evolution partial differential equation of the following form, in one space dimension:

$$u_t = \sum_{K\geq 0} \gamma_K \frac{\partial^K u}{\partial x^K} \ , \tag{9}$$

where the right-hand-side summation is finite, and assume that the equation (9) is approximated on a uniform mesh using the explicit scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \sum_k A_k(\Delta_x) u_{j+k}^n \ , \tag{10}$$

(again with a finite right-hand-side summation). Then, introducing the function:

$$g_{\Delta x}(X) = \sum_k A_k(\Delta_x) e^{k\Delta x X} \ , \tag{11}$$

we state our main result:

PROPOSITION 1:

Assume that the scheme (10) *is consistent with equation* (9) *(in the classical finite-difference sense).*

Then the modified equation of the scheme (10) *writes:*

$$u_t = \sum_{k\geq 0} \alpha_k(\Delta t, \Delta x) \ \frac{\partial^k u}{\partial x^k} \ , \tag{12}$$

where $\sum_{k\geq 0} \alpha_k(\Delta t, \Delta x) X^k$ *is the formal series expansion of the function:*

$$\mathcal{F}(X) = \frac{\log\left(1 + \Delta t \ g_{\Delta x}(X)\right)}{\Delta t} \ . \bullet \tag{13}$$

The proof starts with the following Lemma:

LEMMA 1:

Assume that the scheme (10) *is consistent with equation* (9)*, and set:*

$$g_0(X) = \sum_{K\geq 0} \gamma_K X^K \ . \tag{14}$$

Then, the difference $g_{\Delta x}(X) - g_0(X)$ *formally tends to 0 as* Δx *tends to 0.* $\bullet$

PROOF: The consistency of the scheme (10) implies that, formally:

$$\lim_{\Delta x \to 0} \left(\sum_k A_k(\Delta x) v(x + k\Delta x) - \sum_{K \geq 0} \gamma_K \frac{d^K v}{dx^K} \right) = 0 \ , \tag{15}$$

for any $\mathcal{C}^\infty$ function $v(x)$. For $X \in I\!R$, we may take $v(x) = e^{Xx}$, so that (15) yields:

$$\lim_{\Delta x \to 0} \left(\sum_k A_k(\Delta x) e^{k\Delta x X} - \sum_{K \geq 0} \gamma_K X^K \right) = 0 \ , \tag{16}$$

which ends the proof. •

REMARK 1: It is also useful to see the above proof with a slightly different point of view, using formal series expansions. The consistency of the scheme (10) is usually expressed through Taylor expansions, i.e. one writes the Taylor expansion:

$$\sum_k \sum_{p \geq 0} A_k(\Delta x) \frac{(k\Delta x)^p}{p!} \ \frac{d^p v}{dx^p} \tag{17}$$

of the first term in (15), and one says that the scheme is consistent if:

$$\sum_k A_k(\Delta x) \frac{(k\Delta x)^K}{K!} = \gamma_K + O(\Delta x) \ \text{ for all } K \ . \tag{18}$$

But obviously, (18) allows us to write:

$$\sum_k \sum_{p \geq 0} A_k(\Delta x) \frac{(k\Delta x X)^p}{p!} = \sum_{K \geq 0} \gamma_K X^K + O(\Delta x) \ , \tag{19}$$

for any X, which yields (16). •

We can now achieve the proof of Proposition 1; it relies on applying the Fourier transform to the difference equation which mimics the numerical scheme;

PROOF of Proposition 1: Assume that $u(x,t)$ is bounded and satisfies the difference equation:

$$\frac{u(x, t + \Delta t) - u(x,t)}{\Delta t} = \sum_k A_k(\Delta_x) u(x + k\Delta x, t) \ . \tag{20}$$

We may introduce the Fourier transform $\hat{u}(\xi, \tau)$ of $u(x,t)$, and we get from (11) and (20):

$$\left(\frac{e^{i\tau\Delta t} - 1}{\Delta t} \right) \hat{u}(\xi, \tau) = g_{\Delta x}(i\xi) \hat{u}(\xi, \tau) \ . \tag{21}$$

We can observe here that expanding both sides of (21) in formal series gives:

$$\sum_{p\geq 1}(i\tau)^p\frac{\Delta t^{p-1}}{p!}\hat{u}(\xi.\tau)=\sum_k\sum_{p\geq 0}A_k(\Delta x)\frac{(k\Delta x)^p}{p!}(i\xi)^p\hat{u}(\xi.\tau)\ , \tag{22}$$

whose inverse Fourier transform yields the *unresolved* modified equation:

$$\sum_{p\geq 1}\frac{\Delta t^{p-1}}{p!}\frac{\partial^p u}{\partial t^p}=\sum_k\sum_{p\geq 0}A_k(\Delta x)\frac{(k\Delta x)^p}{p!}\frac{\partial^p u}{\partial x^p}\ . \tag{23}$$

However, the *resolution* is now elementary (and writing the expansions (22) and (23) is not even useful): (21) tells us that the distribution $\hat{u}(\xi,\tau)$ vanishes except on the manifold $\mathcal{V}$ defined by the relation $\dfrac{e^{i\tau\Delta t}-1}{\Delta t}=g_{\Delta x}(i\xi)$. Since we show below that the term $\Delta t g_{\Delta x}(i\xi)$ is small when Δt and Δx are small, and since the exponential function is bijective in the neighbourhood of 0, the manifold $\mathcal{V}$ is also defined for Δt and Δx small by the relation $i\tau=\dfrac{\log\left(1+\Delta t\ g_{\Delta x}(i\xi)\right)}{\Delta t}=\mathcal{F}(i\xi)$, where log denotes here the local inverse of the exponential function, which implies that the classical expansion $\log(1+\delta)=\sum_{p\geq 1}(-1)^{p-1}\dfrac{\delta^p}{p}$ holds true. We then get:

$$i\tau\hat{u}(\xi,\tau)-\mathcal{F}(i\xi)\hat{u}(\xi,\tau)=i\tau\hat{u}(\xi,\tau)-\sum_k\alpha_k(\Delta t,\Delta x)(i\xi)^k\hat{u}(\xi,\tau)=0\ , \tag{24}$$

and the inverse Fourier transform immediately gives the *resolved* modified equation (12).

It only remains to explain why the expansion in formal series is valid in (24), i.e. why the term $\Delta t\ g_{\Delta x}(i\xi)$ is formally small when Δt and Δx are small (notice indeed that the scheme coefficients $A_k(\Delta x)$ involve negative powers of the mesh spacing Δx, as in (1), so that difficulties may arise for small Δx). It follows indeed from (19) that $g_{\Delta x}(X)$ can be expanded in formal series under the form $g_{\Delta x}(X)=g_0(X)+\sum_{p\geq 1}\Delta x^p g_p(X)$, where the $g_p(X)$ are polynomials in X. The fraction $\mathcal{F}(X)$ then takes the form:

$$\mathcal{F}(X)=\frac{\log\left(1+\Delta t g_0(X)+\Delta t\sum_{p\geq 1}\Delta x^p g_p(X)\right)}{\Delta t}\ , \tag{25}$$

and it is perfectly valid to expand it when Δt and Δx are small under the form:

$$\mathcal{F}(X)=g_0(X)+\sum_{\substack{p+q\geq 1\\ r\geq 0}}\beta_{p,q,r}\Delta x^p\Delta t^q X^r\overset{def}{=}\sum_{k\geq 0}\alpha_k(\Delta t,\Delta x)X^k\ .\ \bullet \tag{26}$$

REMARK 2: The proof of Proposition 1 is both more rigorous and more constructive than the elimination method described in Section 1. In particular, it clearly gives the truncation error of the numerical scheme (10), since the modified equation (12) finally takes the form:

$$u_t = \sum_{K \geq 0} \gamma_K \frac{\partial^K u}{\partial x^K} + \sum_{\substack{p+q \geq 1 \\ r \geq 0}} \beta_{p,q,r} \Delta x^p \Delta t^q \frac{\partial^r u}{\partial x^r} \ . \bullet \tag{27}$$

3 SOME EXTENSIONS

The result of Proposition 1, which deals with explicit schemes in one-space dimension, using first-order accurate time integration, can be easily extended in several directions, which we now describe.

3.1 Multi-dimensional schemes

Proposition 1 can be extended with no difficulty to linear constant-coefficients numerical methods in two or three space dimensions. For instance, let us consider the following partial differential equation, in two space dimensions:

$$u_t = \sum_{K,M \geq 0} \gamma_{K,M} \frac{\partial^{K+M} u}{\partial x^K \partial y^M} \ , \tag{28}$$

approximated with the explicit scheme:

$$\frac{u_{j,l}^{n+1} - u_{j,l}^n}{\Delta t} = \sum_{k,m} A_{k,m}(\Delta_x, \Delta y) u_{j+k,l+m}^n \ . \tag{29}$$

We can then state (omitting the proof):

PROPOSITION 2:

Assume that the scheme (29) is consistent with equation (28). Then, its modified equation writes:

$$u_t = \sum_{k,m \geq 0} \alpha_{k,m}(\Delta t, \Delta x, \Delta y) \frac{\partial^{k+m} u}{\partial x^k \partial y^m} \ , \tag{30}$$

where $\sum_{k,m \geq 0} \alpha_{k,m}(\Delta t, \Delta x, \Delta y) X^k Y^m$ *is the formal series expansion of the function:*

$$\mathcal{F}(X) = \frac{\log\left(1 + \Delta t \ g_{\Delta x, \Delta y}(X, Y)\right)}{\Delta t} \ , \tag{31}$$

where $g_{\Delta x,\Delta y}(X,Y)$ is the following functions of two variables:

$$g_{\Delta x,\Delta y}(X,Y) = \sum_{k,m} A_{k,m}(\Delta_x, \Delta y) e^{k\Delta x X} e^{m \Delta y Y} \ . \bullet \tag{32}$$

REMARK 3: The method equally well extends to finite-element schemes, with a non diagonal mass matrix. For instance, the modified equation of the scheme:

$$\sum_{k,m} B_{k,m} \frac{u^{n+1}_{j+k,l+m} - u^n_{j+k,l+m}}{\Delta t} = \sum_{k,m} A_{k,m}(\Delta_x, \Delta y) u^n_{j+k,l+m} \tag{33}$$

(with, say, $\sum\limits_{k,m} B_{k,m} = 1$) takes the form (30), where $\sum\limits_{k,m \geq 0} \alpha_{k,m}(\Delta t, \Delta x, \Delta y) X^k Y^m$ is the formal series expansion of the function:

$$\mathcal{F}(X) = \frac{\log\left(1 + \Delta t \ \dfrac{g_{\Delta x,\Delta y}(X,Y)}{h_{\Delta x,\Delta y}(X,Y)}\right)}{\Delta t} \ , \tag{34}$$

with (32) and:

$$h_{\Delta x,\Delta y}(X,Y) = \sum_{k,m} B_{k,m} e^{k\Delta x X} e^{m \Delta y Y} \ . \bullet \tag{35}$$

3.2 Implicit schemes

The extension of Proposition 1 to implicit schemes is also straightforward. We can state:

PROPOSITION 3:

Assume that the scheme (10) *is consistent with equation* (9).

Then the modified equation of the implicit scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \sum_k A_k(\Delta_x) u^{n+1}_{j+k} \tag{36}$$

writes:

$$u_t = \sum_{k \geq 0} \alpha_k(\Delta t, \Delta x) \frac{\partial^k u}{\partial x^k} \ , \tag{37}$$

where $\sum\limits_{k \geq 0} \alpha_k(\Delta t, \Delta x) X^k$ is the formal series expansion of the function:

$$\mathcal{F}(X) = \frac{\log\left(1 - \Delta t \ g_{\Delta x}(X)\right)}{(-\Delta t)} \ . \bullet \tag{38}$$

PROOF: From the difference equation:

$$\frac{u(x,t)-u(x,t-\Delta t)}{\Delta t}=\sum_k A_k(\Delta_x)u(x+k\Delta x,t) \ , \tag{39}$$

we get:

$$\hat{u}(\xi,\tau)\left(\frac{1-e^{-i\tau\Delta t}}{\Delta t}-g_{\Delta x}(i\xi)\right)=0 \ , \tag{40}$$

so that $\hat{u}(\xi,\tau)$ vanishes except on the manifold $\mathcal{V}$ defined by $i\tau=\dfrac{\log\left(1-\Delta t\ g_{\Delta x}(i\xi)\right)}{(-\Delta t)}$. •

REMARK 4: Extending the statement of Proposition 3 to semi-implicit schemes, it is easy to see that the modified equation of the following semi-implicit method:

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\sum_k A_k(\Delta_x)u_{j+k}^n+\sum_k B_k(\Delta_x)u_{j+k}^{n+1} \ , \tag{41}$$

takes the form (37), where $\sum_{k\geq 0}\alpha_k(\Delta t,\Delta x)X^k$ is the formal series expansion of the function:

$$\mathcal{F}(X)=\frac{1}{\Delta t}\ \log\left(\frac{1+\Delta t\ g_{\Delta x}(X)}{1-\Delta t\ h_{\Delta x}(X)}\right) \ , \tag{42}$$

with $g_{\Delta x}$ and $h_{\Delta x}$ defined as:

$$g_{\Delta x}(X)=\sum_k A_k(\Delta_x)e^{k\Delta xX} \ , \quad h_{\Delta x}(X)=\sum_k B_k(\Delta_x)e^{k\Delta xX} \ . \tag{43}$$

The proof of this fact uses the existence of two polynomials $g_0(X)$ and $h_0(X)$ such that both differences $g_{\Delta x}(X)-g_0(X)$ and $h_{\Delta x}(X)-h_0(X)$ formally tend to 0 as Δx tends to 0; the existence of these polynomials can be shown to follow from the consistency of the scheme (41). •

3.3 Multistep schemes

Let us lastly show that the method can be extended while keeping its simplicity to Runge-Kutta schemes. We state here:

PROPOSITION 4:

Assume that the scheme (10) *is consistent with equation* (9). *When the* N^{th}*-order Runge-Kutta method is applied to the scheme* (10)*, the modified equation writes:*

$$u_t=\sum_{k\geq 0}\alpha_k(\Delta t,\Delta x)\frac{\partial^k u}{\partial x^k} \ , \tag{44}$$

where $\sum_{k\geq 0} \alpha_k(\Delta t, \Delta x) X^k$ *is the formal series expansion of the function:*

$$\mathcal{F}(X) = \frac{\log\left(1 + \sum_{K=1}^{N} \frac{[\Delta t \; g_{\Delta x}(X)]^K}{K!}\right)}{\Delta t} \; . \bullet \qquad (45)$$

PROOF: Let us write the scheme (10) in condensed form as $\frac{u_j^{n+1} - u_j^n}{\Delta t} = (G(u^n))_j$. For the sake of simplicity, we will only consider the second-order Runge-Kutta scheme, which writes:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = (G(u^n))_j + \frac{\Delta t}{2} (G \circ G(u^n))_j \; . \qquad (46)$$

Proposition 4 is then a consequence of the next Lemma. •

LEMMA 2:

Let G^1 *and* G^2 *be two linear schemes, defined by* $\left(G^1(u)\right)_j^n = \sum_k A_k^1(\Delta_x) u_{j+k}^n$ *and* $\left(G^2(u)\right)_j^n = \sum_k A_k^2(\Delta_x) u_{j+k}^n$*, and let* $g_{\Delta x}^1(X)$ *and* $g_{\Delta x}^2(X)$ *be the corresponding functions associated with* G^1 *and* G^2 *respectively using* (11).

Then, the function associated by (11) *with the composed scheme* $G^1 \circ G^2$ *is simply the product* $g_{\Delta x}^1(X) g_{\Delta x}^2(X)$. •

PROOF: It suffices to realize that $\left(G^1 \circ G^2(u^n)\right)_j = \sum_k \sum_m A_k^1 A_m^2 u_{j+k+m}^n$, so that the associated function is $g_{\Delta x}(X) = \sum_k \sum_m A_k^1 A_m^2 e^{(k+m)\Delta x X}$, and Lemma 2 readily follows. •

REMARK 5: We can also write (45) under the form:

$$\mathcal{F}(X) = \frac{\log\left(\exp\left(\Delta t \; g_{\Delta x}(X)\right) - \sum_{K=N+1}^{\infty} \frac{[\Delta t \; g_{\Delta x}(X)]^K}{K!}\right)}{\Delta t} \; . \qquad (47)$$

Therefore, the expansion up to order N writes:

$$\mathcal{F}(X) = g_0(X) + \sum_{p=1}^{N} \Delta x^p g_p(X) + O(\Delta t, \Delta x)^N \; . \qquad (48)$$

This shows not only that the time error is of order N when using the N^{th} order Runge-Kutta scheme, but also that the expansion of the truncation error up to order p reduces when $p \leq N$ to the expansion at the same order of $g_{\Delta x}(X) - g_0(X)$. •

REMARK 6: Notice that Lemma 2 allows us to consider not only Runke-Kutta schemes, but also predictor-corrector two-steps methods, where different spatial schemes are used in each of the two steps. •

4 EXAMPLES

4.1 Runge-Kutta schemes

To illustrate the above results, let us examine how the modified equation is influenced when Runge-Kutta schemes are used, for a given spatial scheme.

We are going to write the expansion of $\mathcal{F}(X) = \sum_{k\geq 0} \alpha_k X^k$ which gives the modified equation, for the best-used Runge-Kutta (i.e., for $1 \leq N \leq 4$). Once a spatial scheme is given, we expand the function $g_{\Delta x}(X)$ under the form $g_{\Delta x}(X) = g_0(X) + \sum_{p\geq 1} \Delta x^p g_p(X)$ (with for instance $g_1 = 0$ if the scheme is spatially second-order accurate). In two space dimensions, a similar expansion can also be written: assuming that the aspect ratio $\dfrac{\Delta y}{\Delta x}$ is constant, one may indeed write $g_{\Delta x, \Delta y}(X,Y) = g_0(X,Y) + \sum_{p\geq 1} \Delta x^p g_p(X,Y)$, where the g_p are now polynomials of two variables. Then, the expansion giving the modified equation takes the following forms, up to third order in Δt and Δx, for the Runge-Kutta schemes up to fourth-order:

$$\begin{aligned}\mathcal{F}_{(N=1)} = g_0 \quad & +\Delta x g_1 - \frac{\Delta t}{2} g_0^2 \\ & +\Delta x^2 g_2 - \Delta t \Delta x g_0 g_1 + \frac{\Delta t^2}{3} g_0^3 \\ & +\Delta x^3 g_3 - \frac{\Delta t \Delta x^2}{2} g_1^2 - \Delta t \Delta x^2 g_0 g_2 + \Delta t^2 \Delta x g_0^2 g_1 - \frac{\Delta t^3}{4} g_0^4 \\ & +O(\Delta t, \Delta x)^4 \ .\end{aligned} \tag{49}$$

$$\begin{aligned}\mathcal{F}_{(N=2)} = g_0 \quad & +\Delta x g_1 + \Delta x^2 g_2 - \frac{\Delta t^2}{6} g_0^3 \\ & +\Delta x^3 g_3 - \frac{\Delta t^2 \Delta x}{2} g_0^2 g_1 + \frac{\Delta t^3}{8} g_0^4 \\ & +O(\Delta t, \Delta x)^4 \ .\end{aligned} \tag{50}$$

$$\begin{aligned}\mathcal{F}_{(N=3)} = g_0 \quad & +\Delta x g_1 + \Delta x^2 g_2 + \Delta x^3 g_3 - \frac{\Delta t^3}{24} g_0^4 \\ & +O(\Delta t, \Delta x)^4 \ .\end{aligned} \tag{51}$$

$$\begin{aligned}\mathcal{F}_{(N=4)} = g_0 \quad & +\Delta x g_1 + \Delta x^2 g_2 + \Delta x^3 g_3 \\ & +O(\Delta t, \Delta x)^4 \ .\end{aligned} \tag{52}$$

Also, we know from Proposition 3 that the expansion giving the modified equation for the backward Euler implicit scheme is simply obtained by substituting $-\Delta t$ instead of Δt in (49).

4.2 Analysing the truncation error

Let us consider again the example of Section 1. For the upwind scheme (1), we have written in (7) and (8) the modified equation expanded up to first order and up to third order respectively. Comparing these two relations, we notice that the first error term, which involves the second derivative u_{xx}, keeps the same coefficient in both expansions. In the same way, we could observe that the coefficients of the third and fourth derivatives in (8) would not be affected by expanding further the modified equation.

This is a particularly nice situation, in particular for the usual substitution-elimination method, since the coefficients of the first derivatives of u are obtained as the first terms in the expansion in Δt and Δx of the modified equation. This property is rather general: writing again the modified equation as in Remark 2:

$$u_t = \sum_{K \geq 0} \gamma_K \frac{\partial^K u}{\partial x^K} + \sum_{\substack{p+q \geq 1 \\ r \geq 0}} \beta_{p,q,r} \Delta x^p \Delta t^q \frac{\partial^r u}{\partial x^r} \ , \tag{53}$$

it is easy to see from (25) that, if $g_0(0) = g_{\Delta x}(0) = 0$[1], then $r \geq q$ for all non zero terms in (53), so that the complete coefficient of the error term involving the derivative $\dfrac{\partial^q u}{\partial x^q}$ is obtained by expanding the modified equation only up to order q in Δt. The situation is even better for the scheme (1), where $g_{\Delta x}(X)$ has the form $g_{\Delta x}(X) = \sum\limits_{p \geq 1} a_p \Delta x^p X^{p+1}$: it can indeed be shown from (25) that $r = p + q + 1$ for all non zero terms in (53), as one could guess from (8).

On the opposite, when $g_0(0) \neq 0$, obtaining the coefficients of the low-order derivatives in the modified equation is not as simple. Consider for instance the equation $w_t = aw - cw_x$, approximated using the explicit scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = a u_j^n - c \ \frac{u_j^n - u_{j-1}^n}{\Delta x} \ . \tag{54}$$

The substitution-elimination process of Section 1 is particularly tedious and lengthy in this case (see e.g. [3, 4]). Up to first order in Δx and Δt, one obtains the following expansion of the modified equation:

$$u_t = au - cu_x - \frac{a^2 \Delta t}{2} u + ac\Delta t u_x + \left(\frac{c\Delta x}{2} - \frac{c^2 \Delta t}{2} \right) u_{xx} + O(\Delta x, \Delta t)^2 \ , \tag{55}$$

[1] Notice that, in contrats with what is usually claimed (see e.g. [2, 9]), the relation $g_{\Delta x}(0) = 0$ does not necessary follow from the consistency of the numerical scheme as soon as $g_0(0) = 0$. For instance, when $g_0(X) = -cX$, i.e. for the wave equation, the (somewhat strange) scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c \ \frac{(1 + \Delta x^2) u_j^n - u_{j-1}^n}{\Delta x}$$

is consistent in the usual sense (see Remark 1 or the proof of Lemma 1), but it satisfies $g_{\Delta x}(0) = -c\Delta x$.

whereas the expansion up to second order in Δx and Δt yields:

$$\begin{aligned} u_t = au - cu_x \quad & - \left(\frac{a^2 \Delta t}{2} - \frac{a^3 \Delta t^2}{3} \right) u + \left(ac\Delta t - a^2 c \Delta t^2 \right) u_x \\ & + \left(\frac{c\Delta x}{2} - \frac{c^2 \Delta t}{2} + ac^2 \Delta t^2 - \frac{ac\Delta x \Delta t}{2} \right) u_{xx} \\ & + \left(\frac{c^2 \Delta t \Delta x}{2} - \frac{c^3 \Delta t^2}{3} \right) u_{xxx} + O(\Delta x, \Delta t)^3 \; . \end{aligned} \tag{56}$$

We therefore see that the coefficients of each derivative (including the "zero-th order derivative") is influenced by a further expansion of the modified equation (and this is confirmed by the expression (25), which for instance shows in this case that $\beta_{0,q,0} \neq 0$ for all q in the expansion (53)). Thus, the elimination process is unable to give the *full* coefficients of the first derivatives in the modified equation of the scheme (54), whereas the expression of these coefficients readily follows from Proposition 1. Indeed, we have here:

$$\mathcal{F}(X) = \frac{1}{\Delta t} \log \left[1 + \Delta t \left(a - c \left(\frac{1 - e^{-\Delta x X}}{\Delta x} \right) \right) \right] \; , \tag{57}$$

and the coefficients α_0, α_1 and α_2 in (12) are simply given as:

$$\alpha_0 = \mathcal{F}(0) \; , \quad \alpha_1 = \mathcal{F}'(0) \; , \quad \alpha_2 = \frac{\mathcal{F}''(0)}{2} \; , \tag{58}$$

so that a straightforward calculation leads to the following form of the modified equation:

$$\begin{aligned} u_t = \quad & \frac{\log(1 + a\Delta t)}{\Delta t} u - \frac{c}{1 + a\Delta t} u_x + \frac{1}{2} \left(\frac{c\Delta x}{1 + a\Delta t} - \frac{c^2 \Delta t}{(1 + a\Delta t)^2} \right) u_{xx} \\ & + \left(\frac{c^2 \Delta t \Delta x}{2} - \frac{c^3 \Delta t^2}{3} \right) u_{xxx} + O(\Delta x, \Delta t)^3 \; , \end{aligned} \tag{59}$$

where the rest $O(\Delta x, \Delta t)^3$ now involves only the third and higher order derivatives of u. This shows therefore another advantage of the method presented in this paper: (59) gives the complete form of the first derivative terms in the truncation error.

5 CONCLUSIONS

We have presented a very simple method for the derivation of the modified equation of any linear numerical method solving an evolution constant-coefficients linear partial differential equation. The method is much simpler than the usual technique, which derives the modified equation through Taylor expansions by a lengthy substitution and elimination process. The modified equation can be explicitly derived using our method for any linear scheme involving two time levels, in any space dimensions and for various time integration methods, either by hand or using a computer system for symbolic algebra.

As a conclusion, it is useful to summarize our method by exhibiting its relation with the Von Neumann stability analysis. The above method is indeed as simple as, and very close to the method for evaluating the amplification factor in the stability analysis: inserting $u_j^n = G(i\xi)^n e^{ij\xi\Delta x}$ in the scheme (10), one obtains with our notations $G(i\xi) = 1 + \Delta t \; g_{\Delta x}(i\xi)$, that is:

$$G(i\xi) = \exp\left(\Delta t \mathcal{F}(i\xi)\right) \quad \text{or} \quad \mathcal{F}(i\xi) = \frac{\log[G(i\xi)]}{\Delta t} \; . \tag{60}$$

These relations, which hold more generally for all linear numerical methods examined in the previous sections, summarize our "recipe" for deriving the modified equation. Note also that they imply the known fact that $|G(i\xi)| \leq 1$ for all ξ (i.e. the scheme (10) is stable) if and only if the coefficients of the modified equation satisfy:

$$\sum_{k\geq 0} (-1)^k \alpha_{2k} \xi^{2k} \leq 0 \;\; \text{for all } \xi \text{ in } I\!R \; . \tag{61}$$

REMARK 7: After completion of this work, we were made aware that Chang [2] already noticed that the modified equation can be obtained from a function $\mathcal{F}$ satisfying $e^{\Delta t \mathcal{F}(i\xi)} = 1 + \Delta t \; g_{\Delta x}(i\xi)$. Chang rigorously proved the existence of such a function $\mathcal{F}$ under the restrictive assumptions that $g_0(0) = g_{\Delta x}(0) = 0$. But he used the existence of $\mathcal{F}$ for analysis purposes only, and *not* as a practical tool for deriving the modified equation, which he still constructed using the elimination method. •

References

[1] D. A. ANDERSON, J. C. TANNEHILL & R. H. PLETCHER, "Computational fluid mechanics and heat transfer", Hemisphere, Mc Graw-Hill, (1984).

[2] S. C. CHANG, "A critical analysis of the modified equation technique of Warming and Hyett", J. Comp. Phys., **86**, pp. 107-126, (1990).

[3] D. CHARGY, "Etude numérique d'écoulements réactifs transsoniques", Thesis, ENPC, Paris, (1991).

[4] N. GLINSKY, "Simulation numérique d'écoulements hypersoniques réactifs hors-équilibre chimique, Thesis, Université de Nice, (1990).

[5] J. GOODMAN & A. MAJDA, "The validity of the modified equation for nonlinear shock waves", J. Comp. Phys., **58**, pp. 336-348, (1985).

[6] S. LANTERI, “Simulation d’écoulements aérodynamiques instationnaires sur une architecture SIMD massivement parallèle”, Thesis, Université de Nice, (1991).

[7] R. PEYRET, “Résolution numérique des systèmes hyperboliques – Application à la dynamique des gaz”, ONERA Report 1977-5, (1977).

[8] M. SPIRIDONOVA & J. A. DESIDERI, “Symbolic computations for the analysis of finite-difference schemes by the modified equation approach”, *unpublished.*

[9] R. F. WARMING & F. HYETT, “The modified equation approach to the stability and accuracy analysis of finite-difference methods”, J. Comp. Phys., **14**, (2), p. 159, (1974).