STABILITY THEORY FOR SOME SCALAR FINITE DIFFERENCE SCHEMES : VALIDITY OF THE MODIFIED EQUATIONS APPROACH [*]

Firas Dhaouadi[1], Emilie Duval[2], Sergey Tkachenko[3] and Jean-Paul Vila[4]

Abstract. In this paper, we discuss some limitations of the modified equations approach as a tool for stability analysis for a class of explicit linear schemes to scalar partial differential equations. We show that the infinite series obtained by Fourier transform of the modified equation is not always convergent and that in the case of divergence, it becomes unrelated to the scheme. Based on these results, we explain when the stability analysis of a given truncation of a modified equation may yield a reasonable estimation of a stability condition for the associated scheme. We illustrate our analysis by some examples of schemes namely for the heat equation and the transport equation.

Résumé. Dans ce papier, nous examinons quelques limites des équations modifiées, comme outil pour l'analyse de la stabilité d'une certaine classe de schémas numériques explicites linéaires pour des équations aux dérivées partielles scalaires. Nous montrons que la série obtenue par transformée de Fourier de l'équation modifiée n'est pas toujours convergente et que dans ce cas, son comportement n'est plus lié à celui du schéma. A partir de ces résultas, nous expliquons quand l'analyse de stabilité d'une troncature donnée peut donner des conditions de stabilité raisonnables. Cette analyse est illustrée par quelques exemples, notamment des schémas pour l'équation de la chaleur et pour l'équation de transport.

Introduction

Modified equations have been the subject of investigations and debates in numerical analysis. And yet, despite being relatively easy to establish, their use has been regarded with a lot of skepticism due to the lack of theoretical justification and the trickiness of their analysis. The first use of this technique for stability purposes is due to Hirt [1]. He provided the first practical examples for which modified equations give relevant information, in a heuristic manner. The pionneering work of Warming and Hyett [2] introduced how to obtain these equations in the general case and mostly clarified the link between the stability of the scheme and the modified equations. They provided a simple and efficient way to obtain the exact Von Neumann stability conditions from a finite number of the modified equation coefficients, for linear scalar schemes. An alternate way to obtain modified equations was introduced in [3]. The method permits to obtain the same modified equation through a series expansion of a an explicitly known function rather than the elimination technique of [2]. Other works that tried to tackle stability analysis through modified equations include [4, 5].

[1] Université Paul Sabatier, Institut de Mathématiques de Toulouse. e-mail : dhaouadi@insa-toulouse.fr
[2] Université Grenoble Alpes, Laboratoire Jean Kuntzmann. e-mail: emilie.duval@inria.fr
[3] Aix-Marseille Université, CNRS, IUSTI, UMR 7343. e-mail: sergey.tkachenko@univ-amu.fr
[4] Institut de Mathématiques de Toulouse, INSA Toulouse. e-mail: vila@insa-toulouse.fr

In this paper, we investigate the modified equations for some finite difference schemes, in an attempt to make it clear why this technique often fails to provide relevant information on stability. First, we present a reminder on how to obtain modified equations for a given linear scalar scheme. We explain the basics of the heuristic stability theory and its limitations through some examples. In the second part, we present the technical framework and tools needed for stability analysis. We then clarify the mathematical reasons behind the frequent failures of the heuristic stability theory by investigating the convergence of the Fourier transform of the modified equation. We show that the latter is conditionally convergent and only then does it give significant results. Then, we compare the stability conditions of truncated modified equations with the corresponding scheme in the region of convergence. Finally, we present some examples to justify our analysis.

1. On modified equations and heuristic stability

1.1. Obtaining the equations

Consider as an example, the linear scalar transport equation given by :

$$\frac{\partial \tilde{u}}{\partial t} + c\frac{\partial \tilde{u}}{\partial x} = 0 \tag{1}$$

with positive velocity c. In order to solve this equation with a finite difference scheme, we first introduce a uniform grid of points defined as usual by $(x_j = j\Delta x, t^n = n\Delta t)$ and we denote by $u_j^n = u(x_j, t^n)$ the value of the numerical solution in the corresponding grid point. Under these notations, take for instance the upwind Euler scheme for equation (1) :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \tag{2}$$

Generally, in order to get relevant information on the consistency of the scheme [6] or its convergence rate, we assume the existence of a smooth, infinitely differentiable numerical solution $u(x, t)$ that satisfies $u(x_j, t^n) = u_j^n$ in each grid point. Provided this solution, we expand each term of the scheme in Taylor series in the vicinity of (x_j, t^n) which leads to an equation containing an infinite number of partial derivatives :

$$\frac{\partial u}{\partial t} + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + \frac{\Delta t^2}{6}\frac{\partial^3 u}{\partial t^3} + \ldots = -c\frac{\partial u}{\partial x} + \frac{c\Delta x}{2}\frac{\partial^2 u}{\partial x^2} + \frac{c\Delta x^2}{6}\frac{\partial^3 u}{\partial t^3} + \ldots \tag{3}$$

This equation as is, is sufficient to prove consistency. In fact, we can clearly see that the in the limit $\Delta t \to 0$ and $\Delta x \to 0$ we recover the original transport equation. However, for further analysis of the numerical effects induced by the scheme, it would be more intuitive to consider an evolution equation with only space derivatives. To do that, we use the elimination procedure introduced by Warming and Hyett in [2], that is we repeatedly use linear combinations of the equation (3) and its derivatives in order to eliminate higher order time derivatives and finally obtain the equation :

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = c\frac{\Delta x}{2}\left(1 - c\frac{\Delta t}{\Delta x}\right)\frac{\partial^2 u}{\partial x^2} - \frac{\Delta x^2}{6}\left(1 - c\frac{\Delta t}{\Delta x}\right)\left(1 - 2c\frac{\Delta t}{\Delta x}\right)\frac{\partial^3 u}{\partial x^3}\ldots \tag{4}$$

This is called the modified equation associated with the upwind Euler scheme for the transport equation (1). It is worth noting that this is not a partial differential equation in the conventional sense as it does not have an order or a finite amount of partial derivatives. For a solution to exist, one also needs a proper definition of infinitely many boundary conditions. Therefore, we restrict our analysis to solutions that are periodic [2].

1.2. Heuristic stability and limitations

Besides proving consistency, the modified equation quantifies explicitly the additional numerical effects. For example, it tells that up to first order in Δx and Δt, the numerical solution rather satisfies a convection diffusion

equation with a numerical dissipation coefficient given by $c\frac{\Delta x}{2}\left(1 - c\frac{\Delta t}{\Delta x}\right)$. Note however that the sign of the latter can be negative in which case the equation becomes ill-posed in Sobolev spaces. It follows from this that a necessary and sufficient condition of stability for the solutions of this convection-diffusion equivalent is :

$$c\frac{\Delta t}{\Delta x} \leq 1 \tag{5}$$

which turns out to be exactly the CFL condition for the scheme (2). This gives a lot of potential for the modified equations to be a practical tool for stability analysis. Although the analysis is completely heuristic and has no rigorous foundation, its results make sense for a large class of schemes. However, there are also many examples for which this analysis fails to provide any practical stability condition. Consider for example the one dimensional heat equation :

$$\frac{\partial u}{\partial t} - \alpha\frac{\partial^2 u}{\partial x^2} = 0 \tag{6}$$

where $\alpha > 0$ is the diffusion coefficient. We discretize this equation using centered finite differences :

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} - \alpha\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = 0 \tag{7}$$

Using the same elimination procedure we obtain the following modified equation up to $4th$ order :

$$\frac{\partial u}{\partial t} - \alpha\frac{\partial^2 u}{\partial x^2} = \frac{\alpha\Delta x^2}{12}\left(1 - 6\frac{\Delta t}{\Delta x^2}\right)\frac{\partial^4 u}{\partial x^4} + \dots \tag{8}$$

If we choose to stop at the first non-zero truncation term as previously, the conclusions are already less straightforward as we have a non-vanishing second order term that comes from the heat equation itself and a $4th$ order term that comes from the numerical effects. Therefore in order to look into stability of this equation, let us shift to Fourier space. Let $v(k,t)$ be the Fourier transform of $u(x,t)$ then under these notations, equation (8) yields :

$$\frac{\partial v}{\partial t} = -\alpha k^2\left(1 - \frac{k^2\Delta x^2}{12}\left(1 - 6\frac{\Delta t}{\Delta x^2}\right)\right)v \tag{9}$$

It is reasonable to only consider the wavenumbers that are bounded by $|k\Delta x| \leq \pi$, since these are the only wavenumbers admissible by the discrete mesh. In this setting, it is easy to verify that :

$$-\alpha k^2\left(1 - \frac{k^2\Delta x^2}{12}\left(1 - 6\frac{\Delta t}{\Delta x^2}\right)\right) \leq 0 \quad \forall k \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \tag{10}$$

which implies that the considered truncation is unconditionally stable for all admissible Δt and Δx. This result is obviously erroneous as it is well known that the Von Neumann stability analysis proves that a necessary and sufficient stability condition for the scheme (7) is :

$$\alpha\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2} \tag{11}$$

Thus, the heuristic analysis of this truncation of the modified equation failed to provide any practical stability condition for the scheme and truncating the equation at higher orders does not seem to do any better. This is one of the examples that demonstrates limitations of the method. In what follows, we first introduce all the necessary notations and setting for stability analysis before attempting to explain why the stability of the truncation is sometimes incoherent with the stability of the scheme.

2. Theory of stability through modified equations

2.1. Notations and assumptions

We will consider for our analysis partial differential equations that are linear, first order in time and of arbitrary order in space :

$$\frac{\partial \tilde{u}}{\partial t} + \sum_{p=1}^{P} A_p \frac{\partial^p \tilde{u}}{\partial x^p} = 0 \tag{12}$$

where A_p are constants. We consider explicit in time schemes, that are consistent with equation (12) and can be written as :

$$u_j^{n+1} = u_j^n + \Delta t \sum_{p=-n_l}^{n_r} b_p\left(\Delta x\right) u_{j+p}^n \tag{13}$$

where n_l and n_r are the number of mesh points to the left and to the right of x_j respectively, used in every iteration. The coefficients b_p verify

$$\sum_{p=-n_l}^{n_r} b_p(\Delta x) = 0 \tag{14}$$

for consistency purposes, so that constant solutions, which are solution of the pde (12) remain as such for the scheme. Sometimes it is preferable to cast the scheme (13) into an equivalent form :

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{\Delta x^q} \sum_{p=-n_l}^{n_r} B_p\left(\Delta x\right) u_{j+p}^n \tag{15}$$

where q is the highest power of $1/\Delta x$ present in the summation. Very often, when using standard finite differences, the coefficients b_p are polynomials in $1/\Delta x$ and q is the highest degree among these polynomials which is frequently equal to the order P of the pde (12). This formulation can be practical for stability analysis since most stability conditions for explicit schemes are given by bounds on the quantity $\lambda_q = \Delta t/\Delta x^q$ in the limit $\Delta x \to 0$ and $\Delta t \to 0$. Therefore, setting λ_q as a non-vanishing parameter is a reasonable assumption that permits to reduce the number of free small parameters, that is we can take $\Delta t = g_q(\Delta x) = \lambda_q \Delta x^q$ and consider instead λ_q and Δx as free independent parameters. In what follows, in order to perform stability analysis in Fourier space, we consider a space continuous counterpart of the scheme (15) :

$$u^{n+1}(x) = u^n(x) + \lambda_q \sum_{p=-n_l}^{n_r} B_p\left(\Delta x\right) u^n(x + p\Delta x) \tag{16}$$

Let $v(k, t)$ be the Fourier transform in space of $u(x, t)$. If we take $\theta = k\Delta x$, then the Fourier transform of equation (16) yields :

$$v^{n+1}(k) = \left(1 + \lambda_q \sum_{p=-n_l}^{n_r} B_p(\Delta x) \mathrm{e}^{\mathrm{i}p\theta}\right) v^n(k) = S(\theta, \lambda_q, \Delta x) v^n(k) \tag{17}$$

Since we will be operating most of the time in Fourier space, it seems necessary to recall consistency of the scheme in the same setting. Therefore if we assume that the exact solution to the pde (12) satisfies in Fourier space :

$$\tilde{v}(k, t + \Delta t) = \exp\left(\Delta t \sum_{p=1}^{P} (\mathrm{i}k)^p A_p\right) \tilde{v}(k, t) = \tilde{G}(\theta, \Delta t)\tilde{v}(k, t) \tag{18}$$

then the scheme (13) is consistent with the pde (12) to order s if and only if [7] :

$$\left| \frac{\tilde{G}(\theta, \Delta t) - S(\theta, \lambda_q, \Delta x)}{\Delta t} \right| = \mathcal{O}(\Delta t^s) \tag{19}$$

Lastly, we denote the modified equation associated to the scheme (13) by :

$$u_t = \sum_{p=1}^{\infty} \mu_p(\lambda_q, \Delta x) \frac{\partial^p u}{\partial x^p} \tag{20}$$

where μ_p are constants depending on Δx and λ_q.

2.2. Fourier stability analysis

In this part, we focus on the link between the modified equation and the scheme [2]. The amplification factor of the scheme is none other than the modulus of $S(\theta, \lambda_q, \Delta x)$. Now, in order to recover an equivalent expression in the continuous time for the modified equation, we apply the Fourier transform to (20). This implies that $v(k, t)$ satisfies the differential equation :

$$\frac{dv}{dt} = \left(\sum_{p=1}^{\infty} (ik)^p \mu_p(\lambda_q, \Delta x) \right) v(k, t) = \left(\sum_{p=1}^{\infty} \alpha_p(\lambda_q, \Delta x) \theta^p \right) v(k, t) = G(\theta, \lambda_q, \Delta x) v(k, t) \tag{21}$$

where $\alpha_p(\lambda_q, \Delta x) = \mu_p(\lambda_q, \Delta x) i^p / \Delta x^p$. For convenience, we will say that the modified equation is stable if, for an initial condition $v(k, t = 0) = v_0(k)$, the solution to the Cauchy problem :

$$\begin{cases} \dfrac{dv}{dt} = G(\theta, \lambda_q, \Delta x) v(k, t) \\ v(k, 0) = v_0(k) \end{cases} \tag{22}$$

that is given by :

$$v(k, t) = e^{tG(\theta, \lambda_q, \Delta x)} v_0(k) \tag{23}$$

remains bounded $\forall t < T$, where $T > 0$. Given this solution, one can obviously write :

$$v(k, t + \Delta t) = e^{\Delta t G(\theta, \lambda_q, \Delta x)} v(k, t) \tag{24}$$

Now, since the solution to the scheme (13) with the same initial condition is also an exact solution to the modified equation (20) [2], uniqueness of this solution gives :

$$e^{\Delta t G(\theta, \lambda_q, \Delta x)} = e^{\lambda_q \Delta x^q G(\theta, \lambda_q, \Delta x)} = S(\theta, \lambda_q, \Delta x) \tag{25}$$

It follows from that the proposition :

Proposition 2.1. *For any scheme that writes as (13) and that is consistent with (12), there exists a positive number θ_m that depends only on λ_q and Δx such that, $\forall \theta \in]-\theta_m, \theta_m[$ the expansion*

$$G(\theta, \lambda_q, \Delta x) = -\sum_{p=1}^{\infty} \frac{(1 - S(\theta, \lambda_q, \Delta x))^p}{p \lambda_q \Delta x^q}$$

holds and the series $\sum_{p=1}^{\infty} \alpha_p(\lambda_q, \Delta x) \theta^p$ converges to $\frac{1}{\lambda_q \Delta x^q} \ln(S(\theta, \lambda_q, \Delta x))$.

Proof. The scheme (13) is consistent with the pde (12) and so we have $|S(0, \lambda_q, \Delta x) - 1| = 0 < 1$. Since S is a continuous function with respect to θ, $\exists \theta_m(\lambda_q, \Delta x) > 0$ such that $|S(\theta, \lambda_q, \Delta x) - 1| < 1 \ \forall \theta \in] - \theta_m, \theta_m[$, and consequently, the principal logarithm $\ln(S(\theta, \lambda_q, \Delta x))$ defined by the series expansion :

$$\ln(S(\theta, \lambda_q, \Delta x)) = \ln(1 - (1 - S(\theta, \lambda_q, \Delta x))) = \sum_{p=1}^{\infty} -\frac{(1 - S(\theta, \lambda_q, \Delta x))^p}{p} \tag{26}$$

is convergent and $e^{\ln(S(\theta, \lambda_q, \Delta x))} = S(\theta, \lambda_q, \Delta x)$. This proves that the function :

$$G(\theta, \lambda_q, \Delta x) = \frac{1}{\lambda_q \Delta x^q} \ln(S(\theta, \lambda_q, \Delta x)) \tag{27}$$

is a solution of equation (25) for $\theta \in] - \theta_m, \theta_m[$. The exponential function is locally invertible in the vicinity of 0 and so this expansion is unique in this vicinity [3] and therefore $\forall \theta \in] - \theta_m, \theta_m[$. Moreover, since $S(\theta, \lambda_q, \Delta x)$ is entire (as a finite sum of exponential functions) then further expanding S into power series of θ for $\theta \in] - \theta_m, \theta_m[$ finally yields :

$$G(\theta, \lambda_q, \Delta x) = \sum_{p=1}^{\infty} \alpha_p \theta^p = \frac{1}{\lambda_q \Delta x^q} \ln(S(\theta, \lambda_q, \Delta x)) \tag{28}$$

and hence, $\sum_{p=1}^{\infty} \alpha_p(\lambda_q, \Delta x)\theta^p$ is the series expansion of $\frac{1}{\lambda_q \Delta x^q} \ln(S(\theta, \lambda_q, \Delta x))$. This concludes our proof. $\square$

Remark 2.2. Equality (28) also implies that the coefficients $\alpha_p(\lambda_q, \Delta x)$ are given by :

$$\alpha_p(\lambda_q, \Delta x) = \frac{\partial^p}{\partial \theta^p} \left(\frac{\ln(S(\theta, \lambda_q, \Delta x))}{p! \lambda_q \Delta x^q} \right) \Bigg|_{\theta=0} \tag{29}$$

Remark 2.3. θ_m is not the radius of convergence of the series $G(\theta, \lambda_q, \Delta x)$. If we denote by R the radius of convergence, then we have $R \geq \theta_m$. The exact radius is not trivial to find in practice in the general case, since $G(\theta, \lambda_q, \Delta x)$ is a series expansion of a composite function. However it is sometimes possible to give estimates or exact values of the radius for some examples as will be shown later.

Remark 2.4. In practice, when looking for convergence, we are mainly searching for constraints in λ_q and Δx, for which the series $G(\theta, \lambda_q, \Delta x)$ is convergent $\forall \theta \in [-\pi, \pi]$, that is we want $R > \pi$. A sufficient condition is $\theta_m > \pi$.

So far, we have shown that for fixed parameters Δx and λ_q, the series $G(\theta, \lambda_q, \Delta x)$ converges if $\theta \in] - \theta_m, \theta_m[$. This being a sufficient condition of convergence, we do not know what happens for $|\theta| \geq \theta_m$. On the other hand it is worth noting that, for $|\theta| > R$, the series is divergent and equality (25) does not hold anymore. Therefore, the modified equation stability is not linked to that of the scheme for $|\theta| > R$.

2.3. Scheme stability domain and series convergence domain

Generally, in the Von Neumann setting, stability conditions of the scheme are given by constraints linking λ_q and Δx so that the inequality

$$|S(\theta, \lambda_q, \Delta x)| \leq 1 \quad \forall \theta \in [-\pi, \pi] \tag{30}$$

is verified. These constraints define a region of stability $\mathcal{R}_s$ in the $(\lambda_q, \Delta x)$ plane, that is :

$$\mathcal{R}_s = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : \forall \theta \in [-\pi, \pi] : |S(\theta, \lambda_q, \Delta x)| \leq 1 \right\} \tag{31}$$

In the same manner, we can define a region $\mathcal{R}_c$ of the same plane, in which the series $G(\theta, \lambda_q, \Delta x)$ converges :

$$\mathcal{R}_c = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : \forall \theta \in [-\pi, \pi] : |G(\theta, \lambda_q, \Delta x)| < \infty \right\} \tag{32}$$

which is also equivalent to :

$$\mathcal{R}_c = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : R(\lambda_q, \Delta x) \geq \pi \right\} \tag{33}$$

In practice, it is not always possible to exactly determine $\mathcal{R}_c$. It is however easier to explicitly find a subset of this region, that is :

$$\Omega_c = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : \forall \theta \in [-\pi, \pi] : |1 - S(\theta, \lambda_q, \Delta x)| < 1 \right\} \subset \mathcal{R}_c \tag{34}$$

It is worth mentioning that Ω_c is always a non-empty set. Indeed, for $\lambda_q = 0$, we have $S(\theta, 0, \Delta x) = 1$ and consequently $\forall \Delta x \in \mathbb{R}_+$ there exists in $\mathbb{R}^2$ a neighborhood of $(\Delta x, \lambda_q)$ in which we have $|1 - S(\theta, \lambda_q, \Delta x)| < 1$. This means that we always have convergence for sufficiently small values of λ_q. Lastly, we denote by $\mathcal{R}_m$ the region of stability of the modified equation :

$$\mathcal{R}_m = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : \left| e^{\lambda_q \Delta x^q G(\theta, \lambda_q, \Delta x)} \right| \leq 1 \right\} = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}_+^2 : \mathrm{Re}(G(\theta, \lambda_q, \Delta x) \leq 0 \right\} \tag{35}$$

Provided these definitions, we can distinguish two cases :

(1) if $\mathcal{R}_s \subset \mathcal{R}_c$ then the stability of the modified equation provides reliable and complete information regarding the scheme stability.

(2) If any subset of $\mathcal{R}_s$ lies outside of the convergence domain $\mathcal{R}_c$, this means that there is information on the stability limit of the scheme that is missed by the modified equations since its Fourier transform is non existing outside of $\mathcal{R}_c$.

The following proposition is a direct consequence :

Proposition 2.5. *For any scheme that writes as (13) and that is consistent with (12) we have* $(\mathcal{R}_m \cap \mathcal{R}_c) \subset \mathcal{R}_s$, *that is if the modified equation is stable and its Fourier series is convergent then the scheme is also stable.*

This proves that inside the convergence domain $\mathcal{R}_c$, the stability of the modified equation is a sufficient stability condition for the scheme. Furthermore, we shall add that if $\mathcal{R}_s \subset \mathcal{R}_c$ then this condition is also necessary. This result, literally, is not of practical interest as the stability of the full series $S(\theta, \lambda_q, \Delta x)$ is either nontrivial or impossible to obtain. But we will show that the above classification permits to justify whether a truncated version of the modified equation yields significant information on stability.

2.4. **Link between the stability of the scheme and the stability of a truncation**

Instead of the full series expansion, let us consider a truncated modified equation to an arbitrary order $N > P$:

$$u_t = \sum_{p=1}^{N} \mu_p(\lambda_q, \Delta x) \frac{\partial^p u}{\partial x^p}, \quad \mu_N \neq 0 \tag{36}$$

We recall that our main purpose is to know when does the stability of this truncation provide relevant information on the stability of the corresponding scheme. This differs from the approach of Warming and Hyett [2] in the sense that they showed how to reconstruct the exact Von Neumann amplification factor using a finite amount of coefficients μ_p without actually analyzing the stability of the truncated version. Under the Fourier transform, the previous equation writes :

$$\frac{dv}{dt} = \left(\sum_{p=1}^{N} \alpha_p(\lambda_q, \Delta x) \theta^p \right) v(k, t) = P_N(\theta, \lambda_q, \Delta x) v(k, t) \tag{37}$$

Here, $P_N(\theta, \lambda_q, \Delta x)$ is a polynomial of θ of degree N which is trivially a truncation of the series $G(\theta, \lambda_q, \Delta x)$. In the same manner as previously, the ordinary differential equation (37) yields :

$$v(k, t + \Delta t) = e^{\Delta t P_N(\theta, \lambda_q, \Delta x)} v(k, t) = S_N(\theta, \lambda_q, \Delta x) v(k, t) \tag{38}$$

In contrast to the full series, the stability conditions of the truncation are obtainable in most cases through the analysis of the polynomial function $P_N(\theta, \lambda_q, \Delta x)$. Let $R_N(\theta, \lambda_q, \Delta x)$ be the rest of the series defined by :

$$R_N(\theta, \lambda_q, \Delta_x) = G(\theta, \lambda_q, \Delta_x) - P_N(\theta, \lambda_q, \Delta_x) = \sum_{p=N+1}^{\infty} \alpha_p(\lambda_q, \Delta x)\theta^p \tag{39}$$

In the convergence domain $\mathcal{R}_c$, the rest $R_N(\theta, \lambda_q, \Delta x)$ is bounded and we have :

$$\lim_{N \to +\infty} R_N(\theta, \lambda_q, \Delta x) \equiv 0 \quad \text{and} \quad \lim_{N \to +\infty} S_N(\theta, \lambda_q, \Delta x) \equiv S(\theta, \lambda_q, \Delta x) \tag{40}$$

In this setting we can state the following result :

Proposition 2.6. *Assume an initial condition satisfying* $supp(v_0) \in [-M, M]$ *and an arbitrary truncation order* $N > P$, *then for any* $(\Delta x, \lambda_q) \in \mathcal{R}_c$, *if the truncated modified equation is stable in the sense that there exists* $C > 0$ *such that :*

$$|S_N(\theta, \lambda_q, \Delta x)| \le 1 + C\Delta t \tag{41}$$

then the scheme is also stable in the same sense.

Proof. For $(\Delta x, \lambda_q) \in \mathcal{R}_c$ we have :

$$S(\theta, \lambda_q, \Delta x) = e^{\Delta t G(\theta, \lambda_q, \Delta x)} = S_N(\theta, \lambda_q, \Delta x)e^{\Delta t R_N(\theta, \lambda_q, \Delta x)}$$

Thus, since $v^n = S^n v_0$, we can write :

$$\begin{aligned}
|v^n| = |S^n v_0| \le |S_N|^n \left| e^{n \sum_{N+1}^{\infty} \alpha_p \theta^p} v_0 \right| \\
\le (1 + C\Delta t)^n \left| e^{nA\theta^{N+1}} v_0 \right| \\
\le (1 + C\Delta t)^n \left| e^{nA(k\Delta x)^{N+1}} v_0 \right| \\
\le (1 + C\Delta t)^n \left| e^{nA\Delta x^{N+1-q}\Delta t k^{N+1}/\lambda_q} v_0 \right| \\
\le e^{CT} \left| e^{AT\Delta x^{N+1-q}M^{N+1}/\lambda_q} v_0 \right| \\
\le e^{CT} e^{AT\Delta x^{N+1-q}M^{N+1}/\lambda_q} |v_0|
\end{aligned}$$

That is, v^n is L^2-stable for initial conditions that are of compact support in the frequency domain, that is $k \in [-M, M]$.   □

3. Examples

3.1. Heat equation - centered finite differences

Consider the centered finite differences scheme for the heat equation. It can be cast into the form :

$$u_i^{n+1} = u_i^n + \alpha\lambda_2 \left(u_{i-1}^n - 2u_i^n + u_{i-1}^n \right) \tag{42}$$

We take for simplicity $\alpha = 1$. The modified equations associated to this scheme up to $8th$ order for example is given by :

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\Delta x^2}{12}(1 - 6\lambda_2)\frac{\partial^4 u}{\partial x^4} + \frac{\Delta x^4}{360}\left(1 - 30\lambda_2\left(1 - 4\lambda_2\right)\right)\frac{\partial^6 u}{\partial x^6} + \frac{\Delta x^6}{20160}\left(1 - 42\lambda_2\left(3 - 40\lambda_2\left(1 - 3\lambda_2\right)\right)\right)\frac{\partial^8 u}{\partial x^8} + \dots$$

Straightforward computations yield :

$$S(\theta, \lambda_2, \Delta x) = 1 + 4\lambda_2 \sin^2(\theta/2) \tag{43}$$

$$\mathcal{R}_s = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}^2 : \lambda_2 \leq \frac{1}{2} \right\} \quad ; \quad \Omega_c = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}^2 : \lambda_2 \leq \frac{1}{4} \right\} \tag{44}$$

$$\Delta t P_8(\theta, \lambda_2, \Delta x) = -\lambda_2 \left(\theta^2 - \frac{1 - 6\lambda_2}{12}\theta^4 + \frac{1 - 30\lambda_2(1 - 4\lambda_2)}{360}\theta^6 - \frac{1 - 42\lambda_2\left(3 - 40\lambda_2\left(1 - 3\lambda_2\right)\right)}{20160}\theta^8 \right) \tag{45}$$

The domains of stability and convergence only depend on the parameter λ_2 independently of Δx. This permits to take an arbitrary value of $\Delta x = 1$ and carry on the analysis, based on only λ_2. Furthermore since $|S(\theta, \lambda_2, \Delta x)|$ is an even function with respect to θ it suffices to look at $\theta \in [0, \pi]$. Figure 1 shows a comparison between $|S(\theta, \lambda_2, \Delta x)|$ and $|S_N(\theta, \lambda_2, \Delta x)|$ for $N = 2$ and $N = 8$ in two cases. In the limit of stability $\lambda_2 = 1/2$, the scheme is stable but the series $G(\theta, \lambda_2, \Delta x)$ is not convergent for $\theta > R = \pi/2$ (See Appendix A). As displayed in the left-hand side of figure 1, the curves begin aligned in the low frequencies, and then the truncation curves begin to diverge completely from the function $|S(\theta, \lambda_2, \Delta x)|$ once θ surpasses the threshold R. This is not the case for $\lambda_2 = 1/4$. For this value we can calculate the radius of convergence $R = \pi$ (See Appendix A). In fact, we can see on the right-hand side of the figure that for λ_2, the curves remain very close $\forall \theta \in [0, \pi]$. For $N = 8$, the two curves almost overlap.
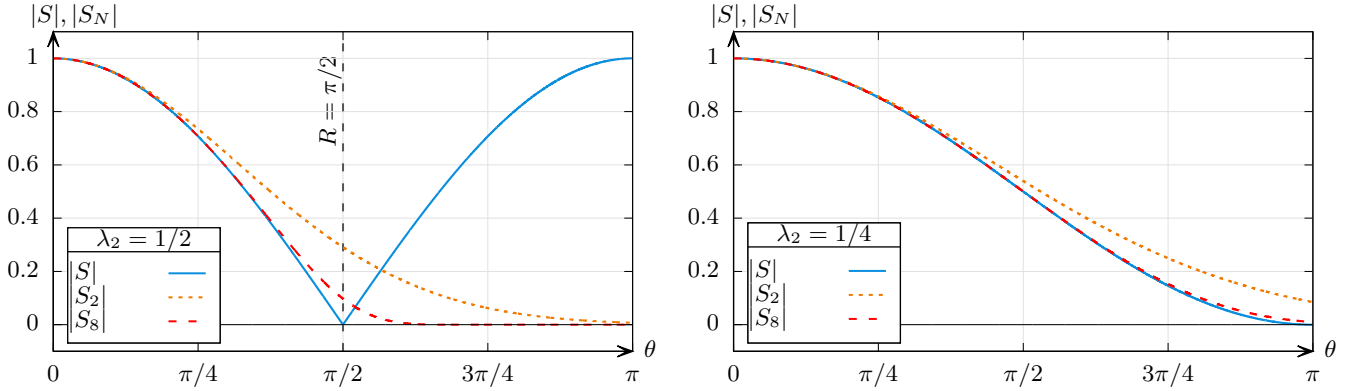

FIGURE 1. Plot of the function $|S(\theta, \lambda_2, \Delta x)|$ along $|S_2(\theta, \lambda_2, \Delta x)|$ and $|S_8(\theta, \lambda_2, \Delta x)|$ for the values of $\lambda_2 = 1/2$ (left) and $\lambda_2 = 1/4$ (right). We can see that for $\lambda_2 = 1/2$, which lies outside of the convergence domain, the truncation curves stray away from the curve of S starting from $\theta = R$. For $\lambda_2 = 1/4$, the truncations match well with S.

3.2. Transport equation - Upwind Euler

The scheme writes :

$$u_i^{n+1} = u_i^n - c\lambda_1(u_i^n - u_{i-1}^n) \tag{46}$$

We take $c = 1$. In this case, the modified equation up to $4th$ order for example is given by :

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = (1 - \lambda_1)\left(\frac{\Delta x}{2}\frac{\partial^2 u}{\partial x^2} - \frac{\Delta x^2}{6}(1 - 2\lambda_1)\frac{\partial^3 u}{\partial x^3} + \frac{\Delta x^3}{24}(1 - 6\lambda_1(1 - \lambda_1))\frac{\partial^4 u}{\partial x^4} \right) + \dots \tag{47}$$

Straightforward computations yield :

$$S(\theta, \lambda_1, \Delta x) = 1 - \lambda_1 \left(1 - e^{-i\theta}\right) \tag{48}$$

$$\mathcal{R}_s = \left\{(\lambda_q, \Delta x) \in \mathbb{R}^2 : \lambda_1 \leq 1\right\} \quad ; \quad \Omega_c = \left\{(\lambda_q, \Delta x) \in \mathbb{R}^2 : \lambda_1 \leq \frac{1}{2}\right\} \tag{49}$$

$$\Delta t P_4(\theta, \lambda_1, \Delta x) = -i\lambda_1\theta + \lambda_1(1 - \lambda_1)\left(-\frac{1}{2}\theta^2 + \frac{1}{6}(1 - 2\lambda_1)i\theta^3 + \frac{1}{24}(1 - 6\lambda_1(1 - \lambda_1))\theta^4\right) + \ldots \tag{50}$$

In contrast to the previous example for the heat equation, we could not compute explicitly the radius of convergence for values of interest of λ_1. Nevertheless, it is possible to extend Ω_c to cover all the values of λ_1 that are in $\mathcal{R}_s$ (See appendix B). Hence, as shown in figure 2, for values of $\lambda_1 \leq 1$, the amplification factor of the truncation to $6th$ order $|S_6(\theta, \lambda_1, \Delta x)|$ seems to be a very good approximation of the scheme amplification factor $|S(\theta, \lambda_1, \Delta x)|$, even for values of λ_1 that are in the vicinity of the stability threshold.
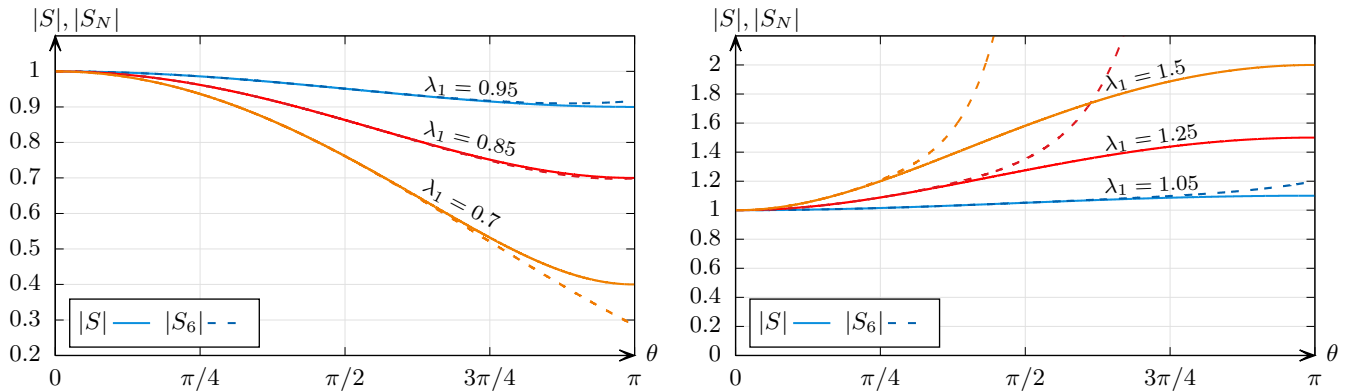

FIGURE 2. Plot of the function $|S(\theta, \lambda_1, \Delta x)|$ along $|S_6(\theta, \lambda_1, \Delta x)|$ for different values of λ_1 in the stable region (left) and in the unstable region (right). The continuous lines stand for the amplification factor $|S(\theta, \lambda_1, \Delta x)|$ and the dashed lines represent the truncated modified equation amplification factor $|S_6(\theta, \lambda_1, \Delta x)|$. We can see that for $\lambda_1 \leq 1$ the truncation curves match well with the exact amplification factor even in the boundaries of stability. Inversely, it seems according to the right-hand graphic that $\lambda_1 = 1$ marks the threshold of convergence.

CONCLUSION AND PERSPECTIVES

We could explain throughout this work that one of the main reasons behind the failures of the modified equations technique is non other than series divergence. Although the analysis only provides sufficient conditions in general, it lifts some well-known ambiguities and provides some assumptions required by the technique to be of a more justified practical use. We show that in these settings, the stability of a truncation gives already reasonable approximations to stability conditions of the scheme. An extension of the obtained results to the case of systems is underway. It would be also interesting to extend this approach to cover a larger class of explicit and also implicit schemes.

A. Computing some radii of convergence

In each of the following cases, we would like to check the convergence radius of the series :

$$\ln(S(\theta, \lambda_q, \Delta x) = \sum_{p=0}^{\infty} \frac{\alpha_p}{\lambda_q \Delta x^q} \theta^p = \sum_{p=0}^{\infty} a_p \theta^p \tag{51}$$

For that, we use the root convergence test, that is the radius of convergence R is given by :

$$\lim_{p \to +\infty} |a_p|^{1/p} = \frac{1}{R} \tag{52}$$

Centered scheme for the heat equation : $\lambda_2 = \frac{1}{2}$

We have :

$$S(\theta, \lambda_2, \Delta x) = 1 - 2\sin(\theta/2)^2 \tag{53}$$

In this case, it is possible to obtain an explicit expression of the pth term of the sequence a_p which is given by :

$$\begin{cases} a_0 = 0 & \\ a_{2p} = -\dfrac{(-4)^p E_{2p-1}(0)}{2(2p)!} & \forall p \geq 1 \\ a_{2p+1} = 0 & \forall p \geq 0 \end{cases} \tag{54}$$

where $E_p(x)$ denotes the pth Euler polynomial function defined by :

$$\frac{2e^{xt}}{e^t + 1} = \sum_{n=0}^{\infty} E_n(x) \frac{t^n}{n!} \tag{55}$$

Since the series has only even order coefficients, it is equivalent and more convenient to check the convergence of the series :

$$\sum_{p=0}^{\infty} b_p \phi^p \tag{56}$$

where $\phi = \theta^2$ and $b_p = a_{2p}$. We proceed as follows in order to check the radius of convergence. We use Bernoulli's number [8] :

$$B_{2p} = \frac{-p E_{2p-1}(0)}{(2^{2p} - 1)} \tag{57}$$

and plug this ansatz into b_p to obtain :

$$|b_p| = \left| \frac{(4)^p E_{2p-1}(0)}{2(2p)!} \right| = \left| \frac{(4)^p (2^{2p} - 1)}{2p(2p)!} B_{2p} \right| \underset{p \to \infty}{\sim} \left| \frac{2^{2p}}{p(\pi)^{2p}} \right|$$

Hence we can write :

$$|b_p| \underset{p \to \infty}{\sim} \frac{1}{p} \left(\frac{2}{\pi} \right)^{2p} \tag{58}$$

and so :

$$\lim_{p \to +\infty} |b_p|^{1/p} = \left(\frac{2}{\pi} \right)^2 \tag{59}$$

This means that the series (56) has a radius $R_\phi = \pi^2/4$ which implies that the radius of convergence of the series $G(\theta, 1/2, \Delta x)$ is $R = \pi/2$.

Centered scheme for the heat equation : $\lambda_2 = \frac{1}{4}$

We have :

$$S(\theta, \lambda_2, \Delta x) = 1 - \sin(\theta/2)^2 \tag{60}$$

The expression of the *pth* term a_p is given by :

$$\begin{cases} a_0 = 0 \\ a_{2p} = -\dfrac{(-1)^p E_{2p-1}(0)}{(2p)!} & \forall p \geq 1 \\ a_{2p+1} = 0 & \forall p \geq 0 \end{cases} \tag{61}$$

Using the same previous notations as in (56), we have :

$$|b_p| = \left| \frac{(E_{2p-1}(0))}{(2p)!} \right| = \left| \frac{(2^{2p} - 1)}{p(2p)!} B_{2p} \right| \underset{p \to \infty}{\sim} \left| \frac{2}{p\pi^{2p}} \right| \tag{62}$$

which yields :

$$\lim_{p \to +\infty} |b_p|^{1/p} = \frac{1}{\pi^2}. \tag{63}$$

This gives the radius of convergence of the series $G(\theta, 1/4, \Delta x)$ is $R = \pi$.

B. PROOF OF CONVERGENCE FOR $\lambda_1 \leq 1$ OF UPWIND EULER FOR TRANSPORT EQUATION

Since $\Omega_c = \left\{ (\lambda_q, \Delta x) \in \mathbb{R}^2 : \lambda_1 \leq \frac{1}{2} \right\}$, then the series $G(\theta, \lambda_1, \Delta x)$ is convergent $\forall \lambda_1 \leq 1/2$ and we have :

$$e^{\Delta t G(\theta, \lambda_1, \Delta x)} = S(\theta, \lambda_1, \Delta x)$$

Since stability only depends on the modulus of $S(\theta, \lambda_1, \Delta x)$ it is sufficient for our analysis to consider the equality :

$$e^{\Delta x \lambda_1 \text{Re}(G(\theta, \lambda_1, \Delta x))} = |S(\theta, \lambda_1, \Delta x)| \tag{64}$$

Next, we show the following symmetry :

$$|S(\theta, 1/2 - \lambda_1, \Delta x)| = |S(\theta, 1/2 + \lambda_1, \Delta x)|$$

Indeed we have :

$$S(\theta, 1/2 - \lambda_1, \Delta x) = 1 - (1/2 - \lambda_1)\left(1 - e^{-i\theta}\right) = (1/2 + \lambda_1) + (1/2 - \lambda_1)e^{-i\theta}$$
$$S(\theta, 1/2 + \lambda_1, \Delta x) = 1 - (1/2 + \lambda_1)\left(1 - e^{-i\theta}\right) = (1/2 - \lambda_1) + (1/2 + \lambda_1)e^{-i\theta}$$
$$= \left((1/2 + \lambda_1) + (1/2 - \lambda_1)e^{i\theta}\right)e^{-i\theta}$$
$$= e^{-i\theta}\overline{S(\theta, 1/2 - \lambda_1, \Delta x)}$$

where the bar denotes the complex conjugate. Hence $|S(\theta, 1/2 - \lambda_1, \Delta x)| = |S(\theta, 1/2 + \lambda_1, \Delta x)|$. This implies through equality (64) that :

$$\Delta x(1/2 - \lambda_1)\text{Re}(G(\theta, 1/2 - \lambda_1, \Delta x)) = \Delta x(1/2 + \lambda_1)\text{Re}(G(\theta, 1/2 + \lambda_1, \Delta x)) \tag{65}$$

which also implies in terms of coefficients :

$$(1/2 - \lambda_1)\alpha_{2p}(1/2 - \lambda_1, \Delta x) = (1/2 + \lambda_1)\alpha_{2p}(1 + \lambda_1, \Delta x) \quad \forall p \geq 1 \tag{66}$$

Therefore, since the series is convergent for $0 \leq \lambda_1 \leq 1/2$, it is also convergent for $1/2 \leq \lambda_1 \leq 1$.

References

[1] Cyril W Hirt. Heuristic stability theory for finite-difference equations. *Journal of Computational Physics*, 2(4):339–355, 1968.

[2] R. F. Warming and B. J. Hyett. The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods. *Journal of Computational Physics*, 179:159–179, 1974.

[3] Romuald Carpentier, Armel de La Bourdonnaye, and Bernard Larrouturou. On the derivation of the modified equation for the analysis of linear numerical methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 31(4):459–470, 1997.

[4] Jiequan Li and Zhicheng Yang. The von neumann analysis and modified equation approach for finite difference schemes. *Applied Mathematics and Computation*, 225:610–621, 2013.

[5] LD Tyler. Heuristic analysis of convective finite difference techniques. In *Proceedings of the Second International Conference on Numerical Methods in Fluid Dynamics*, pages 314–319. Springer, 1971.

[6] Robert D Richtmyer and Keith W Morton. Difference methods for initial-value problems. *Malabar, Fla.: Krieger Publishing Co.,| c1994, 2nd ed.*, 1994.

[7] Robert Dautray and Jacques-Louis Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 6 Evolution Problems II*. Springer Science & Business Media, 2012.

[8] Eric W. Weisstein. "euler polynomial" from mathworld–a wolfram web resource.