ESAIM: M2AN 48 (2014) 107–134
DOI: 10.1051/m2an/2013096

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

NON LINEAR SCHEMES FOR THE HEAT EQUATION IN 1D *

Bruno Després

Abstract. Inspired by the growing use of non linear discretization techniques for the linear diffusion equation in industrial codes, we construct and analyze various explicit non linear finite volume schemes for the heat equation in dimension one. These schemes are inspired by the Le Potier's trick [*C. R. Acad. Sci. Paris, Ser. I* **348** (2010) 691–695]. They preserve the maximum principle and admit a finite volume formulation. We provide a original functional setting for the analysis of convergence of such methods. In particular we show that the fourth discrete derivative is bounded in quadratic norm. Finally we construct, analyze and test a new explicit non linear maximum preserving scheme with third order convergence: it is optimal on numerical tests.

Mathematics Subject Classification. 65J05, 65M08, 65M12.

Received October 27, 2012. Revised June 9, 2013.
Published online December 18, 2013.

1. Introduction

Finite volume schemes are very convenient for complex applications [1, 2, 4, 11, 16, 26]. A very active field of research is nowadays the development of linear and non linear finite volume methods for the heat equation [2, 5, 12, 22, 26]. This topic has growing industrial importance in modern numerical techniques for complex applications. The fundamental problem of the non positivity of numerical discretization of the heat equation on distorted meshes can be traced back to Kershaw [18] for viscous lagrangian fluid dynamics. See [23] for a modern reference in the same direction. Monotonicity and control of the oscillations is also stringent for the computation of a physically sound numerical solution of linear or non linear diffusion equation of radionucleides in porous media [14]. An important reference is [16] for a benchmark comparison of the monotonicity properties of many numerical techniques for solving the diffusion equation in any dimension. See also [24]. Anisotropic diffusion yield the same kind of difficulty [3, 17]. These difficulties inspired Le Potier [9, 19] who designed ingenious non linear correction terms to guarantee the maximum principle in any dimension. But to our knowledge only partial convergence results are available in the literature [5]. For example a compactness technique is used

Keywords and phrases. Finite volume schemes, heat equation, non linear correction.

* *The author acknowledges the support of ANR under contract ANR-12-BS01-0006-01. Moreover this work was carried out within the framework of the European Fusion Development Agreement and the French Research Federation for Fusion Studies. It is supported by the European Communities under the contract of Association between Euratom and CEA. The views and opinions expressed herein do not necessarily reflect those of the European Commission.*

[1] LJLL, UPMC, Paris, France.

[2] Laboratoire Jacques-Louis Lions University Pierre et Marie Curie Boîte courrier 187 75252 Paris Cedex 05 France. despres@ann.jussieu.fr

Article published by EDP Sciences

in [9] to prove convergence, but of course without any order of convergence since it is a compactness technique. Last but not least the control of the maximum principle is fundamental in the analysis of elliptic linear and non linear partial differential equations [10, 13]. It is therefore highly desirable to extend this principle to discrete methods for the computation of a numerical solution to these equations.

Motivated by these applications and these new non linear numerical schemes for diffusion and heat equations, we investigate in this work a framework which provides quantitative orders of convergence for such non linear methods. At this stage it is worthwhile to make a comparison with the theory of non linear schemes for the advection equation [15, 21, 25, 28], for a which a beautiful and comprehensive theory based on TVD or TVB schemes is available in 1D for the advection equation. Such TVD schemes are based on a control of the discrete L^1 norm of the first derivative for TVD and TVB schemes [8, 15, 21]. It provides order of convergence in some cases [8]. We retain the idea that a control of discrete derivatives has been the key in the past to obtain a mathematical setting adapted to the analysis of non linear schemes for the advection equation.

Therefore a natural question is to control the norm of some discrete derivatives of non linear schemes for linear heat equation, more generally to establish a systematic method to obtain bounds on some discrete derivatives, and at the end to prove convergence. Quoting Droniou–Lepotier [9], *this theoretical study is not just a mathematical amusement since it leads us to an understanding of how to choose the parameters of the method in order to obtain good approximations of the solution.* Even if the ultimate goal is the analysis of 2D schemes which are used in practical applications, this is for the moment too complicated. This is why we concentrate in this work on the 1D non stationary situation[3].

Our model problem is the non stationary heat equation in dimension one

$$\begin{cases} \partial_t u - \partial_{xx} u = 0,\ t > 0,\ x \in \mathbb{R}, \\ u(0,x) = u_0(x),\ x \in \mathbb{R}. \end{cases} \tag{1.1}$$

We will show how to obtain control of discrete derivatives of non linear schemes using an approach proposed in [8] which is fundamentally Fourier based. In other words we show how to control some non linear terms with Fourier linear techniques: this apparent paradox is the cornerstone of this work. The structure of the main stability estimate is the following: if a certain function of the CFL number $\nu = \frac{\Delta t}{\Delta x^2}$ which is the product of three terms is less than one, that is if

$$Q \times \alpha \times E(\nu) \leq 1 - \varepsilon \tag{1.2}$$

where $Q > 0$ is a natural measure of the size of the non linear correction, $\alpha \geq 1$ is a constant characteristic of the Le Potier method (at the end of the analysis one always takes $\alpha = 1$) and $E(\nu)$ is a complicated function of the CFL number, then one obtains a simple control of the discrete derivatives of order k on which the non linear correction is based. The function $E(\nu)$ is a series and depends on the type of linear scheme that one considers: typically

$$E(\nu) = \nu \sum_{l \leq L} (1 - 4\nu)^l 2^k + \sum_{L+1 \leq l} \left(\frac{2l}{2l+k} \right)^l \left(\frac{k}{\nu(2l+k)} \right)^{\frac{k}{2}} \tag{1.3}$$

where $L = \left[\frac{k}{2} \left(\frac{1}{4\nu} - 1 \right) \right]$ is a threshold value justified in the core of the paper. This idea comes from [7, 8]. An important part of the paper will be devoted to obtain sharp estimates for $E(\nu)$ and related functions and to show that the product (1.2) is indeed smaller than 1.

The net result is this work is the design of a new scheme for the heat equation. This finite scheme scheme (3.20) is based on the Le Potier's trick and has enhanced approximation properties. It is explicit, non linear, preserves total mass, is maximum preserving under standard CFL condition: we are able to prove it converges at order 3 towards smooth solutions; the only restriction of our convergence theory is the CFL number which is for the moment slightly more restrictive than the usual one. Numerical results show this order of convergence is optimal. A second scheme with similar properties is designed in the appendix.

[3] With this respect the situation is similar to what is known for advection equation for which the 1D theory is much more developed than the multiD theory.

The organization is as follows. We first present the family of high order 1D schemes that we desire to analyze: we call theses schemes Le Potier or modified (Le Potier's) schemes. We explain how to use the Le Potier's trick to modify the schemes and to insure the maximum principle. After we will derive the fundamental and new *a priori* estimates (1.2)–(1.3) on which the convergence results of the final section are based. Numerical results are used to confirm the theoretical analysis. Some open problems are reviewed at the end.

2. Basic linear schemes

Let $\Delta x > 0$ is the mesh size of our finite difference or finite volume discretization method. We will consider square integrable numerical profiles $v = (v_j)_{j\in\mathbb{Z}}$ such that

$$\|v\| = \left(\Delta x \sum_{j\in\mathbb{Z}} v_j^2\right)^{\frac{1}{2}}. \tag{2.1}$$

It is convenient to define the space of square integrable numerical profiles

$$l_2 = \left\{v \in \mathbb{R}^{\mathbb{Z}};\ \|v\| < \infty\right\} \tag{2.2}$$

equipped with the natural scalar product

$$(u, v) = \Delta x \sum_{j\in\mathbb{Z}} u_j v_j. \tag{2.3}$$

We start from the finite volume form of a linear explicit discrete scheme

$$\Delta x \frac{\overline{u}_j - u_j}{\Delta t} - \left(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}\right) = 0, \qquad \forall j \in \mathbb{Z}, \tag{2.4}$$

where u_j is the current discrete solution in cell j and $f_{j+\frac{1}{2}}$ is the explicit numerical flux evaluated between cells j and $j+1$. Taking

$$f_{j+\frac{1}{2}} = \frac{u_{j+1} - u_j}{\Delta x},$$

one gets the classical three points linear scheme

$$\frac{\overline{u}_j - u_j}{\Delta t} - \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} = 0, \qquad j \in \mathbb{Z}.$$

This scheme is of order 1 in time and 2 in space. The Courant number is $\nu = \frac{\Delta t}{\Delta x^2}\cdot$ The three points scheme is stable in l_2 (actually it is stable in all discrete Lebesgue spaces) under CFL condition

$$2\nu \leq 1. \tag{2.5}$$

This monotone scheme is the fundamental brick for the heat equation with the lowest order of approximation in dimension one. The following examples display enhanced approximation properties for the heat equation. If one desires to establish a parallel with linear schemes for the advection equation, the fundamental brick is the upwind scheme while the Lax–Wendroff scheme which is second order with enhanced approximation properties.

2.1. Example 1

The second scheme that we consider is of order 1 in time and 4 in space. It is based on the observation that

$$\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{\Delta x^2} = \partial_{xx}u(x_j) + \frac{\Delta x^2}{12}\partial_{xxxx}u(x_j) + O\left(\Delta x^4\right)$$

for smooth functions. Therefore

$$\frac{u(x_{j+2})-2u(x_j)+u(x_{j-2})}{4\Delta x^2}=\partial_{xx}u(x_j)+\frac{\Delta x^2}{3}\partial_{xxxx}u(x_j)+O\left(\Delta x^4\right).$$

A linear combination yields

$$\frac{4}{3}\frac{u(x_{j+1})-2u(x_j)+u(x_{j-1})}{\Delta x^2}-\frac{1}{3}\frac{u(x_{j+2})-2u(x_j)+u(x_{j-2})}{4\Delta x^2}=\partial_{xx}u(x_j)+O\left(\Delta x^4\right).$$

A similar trick is used in [6]. That's why we will consider the scheme

$$\frac{\overline{u}_j-u_j}{\Delta t}-\frac{4}{3}\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2}+\frac{1}{3}\frac{u_{j+2}-2u_j+u_{j-2}}{4\Delta x^2}=0 \tag{2.6}$$

which is first order in time and fourth order in space. The corresponding flux is

$$f_{j+\frac{1}{2}}=\frac{4}{3}\frac{u_{j+1}-u_j}{\Delta x}-\frac{u_{j+2}+u_{j+1}-u_j-u_{j-1}}{12\Delta x}.$$

It is convenient for further developments to use a more compact variational formulation. We define the bilinear form

$$a(u,v)=\frac{4}{3}\Delta x\sum_j\frac{(u_{j+1}-u_j)(v_{j+1}-v_j)}{\Delta x^2}-\frac{1}{3}\Delta x\sum_j\frac{(u_{j+2}-u_j)(v_{j+2}-v_j)}{4\Delta x^2}. \tag{2.7}$$

With these notations the scheme (2.6) restricted to profiles in $\overline{u},u\in l_2$ can be rewritten under the variational form

$$\left(\frac{\overline{u}-u}{\Delta t},v\right)+a(u,v)=0,\quad\forall v\in l_2. \tag{2.8}$$

A classical result is the following.

Proposition 2.1. *The symmetric bilinear form (2.7) is non negative.*

Proof. One has $\Delta x\sum_j(u_{j+2}-u_j)^2\le 4\Delta x\sum_j(u_{j+1}-u_j)^2$ from which deduce that

$$a(u,u)\ge\frac{4\Delta x}{3}\sum_j\frac{(u_{j+1}-u_j)^2}{\Delta x^2}-\frac{\Delta x}{3}\sum_j\frac{(u_{j+1}-u_j)^2}{\Delta x^2}=\Delta x\sum_j\frac{(u_{j+1}-u_j)^2}{\Delta x^2}\ge 0. \tag{2.9}$$

An elementary upper bound is

$$a(u,u)\le\frac{16}{3\Delta x^2}\|u\|^2. \tag{2.10}$$

□

Lemma 2.2. *This scheme is stable in l_2 under the CFL condition*

$$\frac{16}{3}\nu\le 1. \tag{2.11}$$

Proof. Take the test function $v=\overline{u}$ in (2.8). One gets

$$\frac{1}{2}\|\overline{u}\|^2-\frac{1}{2}\|u\|^2+\frac{1}{2}\|\overline{u}-u\|^2=-\Delta t a(u,\overline{u})=-\frac{1}{2}\Delta t\ a(u,u)-\frac{1}{2}\Delta t\ a(\overline{u},\overline{u})+\frac{1}{2}\Delta t\ a(\overline{u}-u,\overline{u}-u).$$

Since the bilinear form a is non negative one has that

$$\|\overline{u}\|^2-\|u\|^2\le\Delta t\ a(\overline{u}-u,\overline{u}-u)-\|\overline{u}-u\|^2\le\left(\frac{16\nu}{3}-1\right)\|\overline{u}-u\|^2. \tag{2.12}$$

The CFL condition guarantees the non positivity of the right hand side. It ends the proof. □

2.2. Example 2

It is of course tempting to use the modified equation to design a scheme with enhanced consistency in time. Consider a smooth solution of the heat equation. One has

$$\frac{u(t_{n+1})-u(t_n)}{\Delta t}=\partial_t u(t_n)+\frac{\Delta t}{2}\partial_{tt}u(t_n)+O(\Delta t^2)=\partial_t u(t_n)+\frac{\Delta t}{2}\partial_x^{(4)}u(t_n)+O(\Delta t^2)$$

from which we deduce that the scheme

$$\frac{\overline{u}_j-u_j}{\Delta t}-\frac{4}{3}\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2}+\frac{u_{j+2}-2u_j+u_{j-2}}{12\Delta x^2}-\frac{\Delta t}{2}\frac{u_{j+2}-4u_{j+1}+6u_j-4u_{j-1}+u_{j-2}}{\Delta x^4}=0$$

is of order 2 in time and 4 in space. We refer to [27] page 43 where the same scheme is introduced mainly for theoretical purposes. The scheme can also be rewritten under the form

$$\frac{\overline{u}_j-u_j}{\Delta t}-\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2}+(1-6\nu)\frac{u_{j+2}-4u_{j+1}+6u_j-4u_{j-1}+u_{j-2}}{12\Delta x^2}=0.$$

It admits a variational reformulation: compute $u\in l_2$ such that

$$\left(\frac{\overline{u}-u}{\Delta t},v\right)+b(u,v)=0,\quad \forall v\in l_2 \tag{2.13}$$

where the bilinear form b is a correction of a (defined in (2.7))

$$b(u,v)=a(u,v)-\frac{\nu}{2}\widetilde{a}(u,v)$$

with

$$\widetilde{a}(u,v)=\Delta x\sum_j\frac{(u_{j+2}-2u_{j+1}+u_j)\,(v_{j+2}-2v_{j+1}+v_j)}{\Delta x^2}.$$

We notice that

$$0\le\widetilde{a}(u,u)\le 4\Delta x\sum_j\frac{(u_{j+1}-u_j)^2}{\Delta x^2}.$$

Therefore one gets using (2.9)

$$b(u,u)\ge(1-2\nu)\,\Delta x\sum_j\frac{(u_{j+1}-u_j)^2}{\Delta x^2}\ge 0.$$

Lemma 2.3. *The scheme (2.13) is stable in l_2 under the CFL condition (2.11).*

Proof. Since $b(u,u)\le a(u,u)$, the same proof as the one of lemma 2.2 holds. The final stability condition is the more restrictive one between (2.5) and (2.11). The proof is ended. □

3. Maximum principle and Le Potier's trick

The schemes (2.6)-(2.8) and (2.13) are high order. Unfortunately they do not preserve the maximum principle for all ν. For example the explicit formulation of (2.6) is

$$\overline{u}_j=\left(1-\frac{5\nu}{2}\right)u_j+\frac{4\nu}{3}\,(u_{j+1}+u_{j-1})-\frac{\nu}{12}\,(u_{j+2}+u_{j-2})\,. \tag{3.1}$$

Since the coefficients of the extreme parts are negative, the scheme cannot be maximum preserving: this is independent of the CFL condition.

C. Le Potier has proposed in [9,19] a way to introduce a non linear modification so as to recover the maximum principle in any dimension. It starts from the scheme rewritten under the form

$$\overline{u}_j - u_j = \nu \mathrm{Div}_j, \qquad \mathrm{Div}_j = \Delta x \left(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} \right).$$

The neighbors of a cell correspond by definition to non zero coefficient in the explicit stencil: for our example (3.1) it corresponds to

$$\mathcal{V}(j) = \{j-2, j-1, j+1, j+2\} \tag{3.2}$$

Let us define

$$\Sigma_j = \sum_{l \in \mathcal{V}(j)} |u_l - u_j| . \tag{3.3}$$

We note that $\Sigma_j \geq 0$ by definition. With this notation, the application of the Le Potier's trick to the discrete heat equation is to consider the non linear scheme

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j + \alpha\nu \sum_{l \in \mathcal{V}(j)} a_{jl}(u_l - u_j) \tag{3.4}$$

where $\alpha \geq 0$ is a coefficient to be prescribed and the non linear part of the scheme is given by

$$a_{jl} = a_{lj} = \frac{|\mathrm{Div}_j|}{\Sigma_j} + \frac{|\mathrm{Div}_l|}{\Sigma_l}$$

which is a non linear coefficient.

Remark 3.1. Since the stencil $\mathcal{V}(j)$ is large enough, there exists a local constant $C > 0$ such that $|\mathrm{Div}_j| \leq C\Sigma_j$. So it is always possible to remove the apparent singularity in the previous definition by taking for example

$$a_{jl} = a_{lj} = \lim_{\varepsilon \to 0^+} \left(\frac{|\mathrm{Div}_j|}{\Sigma_j + \varepsilon} + \frac{|\mathrm{Div}_l|}{\Sigma_l + \varepsilon} \right).$$

Anyway what really matters is the continuity of the product $a_{jl}(u_l - u_j)$ which is straightforward. Such continuity will be essential in (3.22). In summary we will systematically consider in the rest of this work that a_{jl} is a bounded quantity even if $\Sigma_j = 0$.

If one compares with TVD schemes for the advection equation, this term is very close to some extended slope indicator [8, 15, 21, 25, 28]. For example the minmod limiter writes $\varphi_{j+\frac{1}{2}} = \mathrm{minmod}\left(1, \frac{u_j - u_{j-1}}{u_{j+1} - u_j}\right)$ with corresponding limited Lax–Wendroff flux $f_{j+\frac{1}{2}} = \frac{1}{2}(1-\nu)(u_{j+1} - u_j)\varphi_{j+\frac{1}{2}}$. Under this form it is clear that a_{jl} acts as a limitation on the numerical value of the diffusion flux like $\varphi_{j+\frac{1}{2}}$ does on the advection flux.

By convention a_{jl} can be extended by zero

$$a_{jl} = 0 \text{ for } l \notin \mathcal{V}(j) \Longleftrightarrow j \notin \mathcal{V}(l),$$

so that the non linear scheme (3.4) can be rewritten under the more compact form

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j + \alpha\nu \sum_{l \in \mathbb{Z}} a_{jl}(u_l - u_j). \tag{3.5}$$

3.1. Stability

The non linear scheme (3.4) inherits the properties of the stationary Le Potier's schemes [9, 19].

Proposition 3.2. *The scheme (3.4) admits the finite volume formulation*

$$\Delta x \frac{\overline{u}_j - u_j}{\Delta t} - \left(f^{\rm tot}_{j+\frac{1}{2}} - f^{\rm tot}_{j-\frac{1}{2}} \right) = 0, \qquad \forall j \in \mathbb{Z},$$

where the total flux is the sum of the linear flux plus a non linear correction, that is $f^{\rm tot}_{j+\frac{1}{2}} = f_{j+\frac{1}{2}} + f^{\rm cor}_{j+\frac{1}{2}}$ *with*

$$f^{\rm cor}_{j+\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{l \le j+\frac{1}{2} \le m} a_{ml}(u_m - u_l), \qquad j \in \mathbb{Z}. \tag{3.6}$$

Remark 3.3. If l and m are two indices sufficiently far one to the other, then $a_{ml} = 0$. So only a finite number of terms enter in (3.6).

Proof. The linear part of the flux $f_{j+\frac{1}{2}}$ does not yield any difficulty. Concerning the non linear part one has by construction

$$f^{\rm cor}_{j+\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{j+1 \le m} a_{mj}(u_m - u_j) + \frac{\alpha}{\Delta x} \sum_{l \le j-1 \text{ and } j+1 \le m} a_{ml}(u_m - u_l).$$

The first sum is for $l = j$, the second sum is all other terms. One also has

$$f^{\rm cor}_{j-\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{l \le j-\frac{1}{2} \le m} a_{ml}(u_m - u_l) = \frac{\alpha}{\Delta x} \sum_{l \le j-1} a_{jl}(u_j - u_l) + \frac{\alpha}{\Delta x} \sum_{l \le j-1 \text{ and } j+1 \le m} a_{ml}(u_m - u_l)$$

where the first sum if for $m = j$ and the second sum is the rest. The difference is therefore

$$f^{\rm cor}_{j+\frac{1}{2}} - f^{\rm cor}_{j-\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{j+1 \le m} a_{mj}(u_m - u_j) - \frac{\alpha}{\Delta x} \sum_{l \le j-1} a_{jl}(u_j - u_l) = \frac{\alpha}{\Delta x} \sum_{l \in \mathbb{Z}} a_{jl}(u_l - u_j).$$

It ends the proof. □

Lemma 3.4. *The scheme (3.4) is stable in* l_2 *under CFL condition*

$$\left(\frac{16}{3} + 4\alpha \max_j \sum_l a_{jl} \right) \nu \le 1.$$

Proof. We define the bilinear form

$$a^{\rm tot}(u, v) = a(u, v) + \frac{\alpha}{\Delta x} a^{\rm cor}(u, v) \tag{3.7}$$

which is the sum of the classical bilinear form $a(u, v)$ that corresponds to the linear part of the scheme and which is non negative, and of the additional bilinear form

$$a^{\rm cor}(u, v) = -\frac{\alpha}{\Delta x} \sum_j \left(\sum_l a_{jl}(u_l - u_j) \right) v_j = \alpha \Delta x \sum_l \sum_j a_{jl} \frac{u_l - u_j}{\Delta x} \frac{v_l - v_j}{\Delta x}$$

which corresponds to the non linear coefficients a_{jl}. The equivalent of the stability inequality (2.12) is

$$\|\overline{u}\|^2 - \|u\|^2 \le \Delta t \ \left(a + \frac{\alpha}{\Delta x} a^{\rm cor} \right) (\overline{u} - u, \overline{u} - u) - \|\overline{u} - u\|^2 \le \left(\frac{16\nu}{3} + 4\alpha\nu \max_j \sum_l a_{jl} - 1 \right) \|\overline{u} - u\|^2$$

which proves the claim. □

Lemma 3.5. *Assume $\alpha \geq 1$. Assume the CFL condition*

$$\left(\sum_{l \in \mathcal{V}(j)} \left((\alpha + \varepsilon_{jl}) \frac{|\mathrm{Div}_j|}{\Sigma_j} + \alpha \frac{|\mathrm{Div}_l|}{\Sigma_l} \right) \right) \nu \leq 1 \tag{3.8}$$

where $\varepsilon_{jl} = \pm 1$ is defined below. Then the scheme (3.4) satisfies the maximum principle.

Remark 3.6. This result provides therefore a setting to obtain, at the discrete level, the maximal principle which is known to be fundamental at the continuous level for partial differential equations [10, 13].

Proof. Using the evident identity

$$\mathrm{Div}_j = \sum_{l \in \mathcal{V}(j)} \frac{\mathrm{Div}_j \, |u_l - u_j|}{\Sigma_j}$$

one rewrites the explicit scheme (3.5) under the form

$$\overline{u}_j = u_j + \nu \sum_{l \in \mathcal{V}(j)} \frac{\mathrm{Div}_j \, |u_l - u_j| + \alpha \, |\mathrm{Div}_j| \, (u_l - u_j)}{\Sigma_j} + \alpha \nu \sum_{l \in \mathcal{V}(j)} \frac{|\mathrm{Div}_l|}{\Sigma_l} (u_l - u_j).$$

The key observation is the following

- If $\mathrm{Div}_j(u_l - u_j) \geq 0$ then we set $\varepsilon_{jl} = 1$ so that

$$\mathrm{Div}_j \, |u_l - u_j| + \alpha \, |\mathrm{Div}_j| \, (u_l - u_j) = (\alpha + \varepsilon_{jl}) \, |\mathrm{Div}_j| \, (u_l - u_j).$$

- On the other hand if $\mathrm{Div}_j(u_l - u_j) < 0$ then we set $\varepsilon_{jl} = -1$ so that

$$\mathrm{Div}_j \, |u_l - u_j| + \alpha \, |\mathrm{Div}_j| \, (u_l - u_j) = (\alpha + \varepsilon_{jl}) \, |\mathrm{Div}_j| \, (u_l - u_j).$$

In both cases $\alpha + \varepsilon_{jl} \geq 0$ since $\alpha \geq 1$ by hypothesis. Therefore

$$\begin{aligned} \overline{u}_j &= u_j + \nu \sum_{l \in \mathcal{V}(j)} \left((\alpha + \varepsilon_{jl}) \frac{|\mathrm{Div}_j|}{\Sigma_j} + \alpha \frac{|\mathrm{Div}_l|}{\Sigma_l} \right) (u_l - u_j) \\ &= \left(1 - \nu \sum_{l \in \mathcal{V}(j)} \left((\alpha + \varepsilon_{jl}) \frac{|\mathrm{Div}_j|}{\Sigma_j} + \alpha \frac{|\mathrm{Div}_l|}{\Sigma_l} \right) \right) u_j + \nu \sum_{l \in \mathcal{V}(j)} \left((\alpha + \varepsilon_{jl}) \frac{|\mathrm{Div}_j|}{\Sigma_j} + \alpha \frac{|\mathrm{Div}_l|}{\Sigma_l} \right) u_l. \end{aligned}$$

It is a convex combination under CFL. It ends the proof. □

Let us consider a simple example in order to figure out the practical impact of (3.8) on the stability criterion. The worst guess reduces to

$$\mathrm{Card} \, \mathcal{V}(j) \times \sup_j \frac{|\mathrm{Div}_j|}{\Sigma_j} \times \sup_{jl} (|\alpha + \varepsilon_{jl}| + \alpha) \leq 1.$$

We consider the example described in (2.6) which corresponds to

$$\mathrm{Div}_j = -\frac{1}{12}(u_{j+2} + u_{j-2}) + \frac{4}{3}(u_{j+1} + u_{j-1}) - \frac{5}{2} u_j = -\frac{1}{12}(u_{j+2} - u_j) + \frac{4}{3}(u_{j+1} - u_j) + \frac{4}{3}(u_{j-1} - u_j) - \frac{1}{12}(u_{j-2} - u_j).$$

One has the bound

$$\frac{|\mathrm{Div}_j|}{\Sigma_j} \leq \frac{1}{12} + \frac{4}{3} + \frac{4}{3} + \frac{1}{12} = \frac{34}{12}.$$

So we obtain

$$\left(4 \times \frac{34}{12} \times (2\alpha + 1)\right) \nu \leq 1.$$

It is natural to choose the smallest value of α, that is $\alpha = 1$. One obtains the (sufficient) CFL condition

$$34\nu \leq 1.$$

It is quite a stringent condition in terms of time constraint.

Remark 3.7. In order to optimize the time step restriction, that is to be able to take Δt as large as possible, we will systematically use $\alpha = 1$ either in the theory or in the numerics. However we keep it in most of the analysis to be closer to Le Potier's notations.

A first natural question is to try to diminish the numerical value of the constant in this CFL condition: it will be done with the help of modified schemes developed hereafter.

3.2. Modified schemes

In many cases the initial linear scheme can be decomposed in two parts. A first linear part for which the maximum principle holds and a second linear part which does not naturally respect the maximum principle. We propose to call "modified schemes" such schemes where only the second part is modified by the seminal Le Potier's trick. This procedure which is very natural and has been developed also in [20] is a way to obtain a less severe CFL constraint, still guaranteeing the maximum principle.

3.3. A first modified scheme

For example the second example (2.6) can be rewritten as

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^0 + \nu \mathrm{Div}_j^1 \tag{3.9}$$

where the first part is defined by

$$\mathrm{Div}_j^0 = \frac{4}{3}\left(u_{j+1} - 2u_j + u_{j-1}\right)$$

and the second part is defined by

$$\mathrm{Div}_j^1 = -\frac{1}{3}\left(u_{j+2} - 2u_j + u_{j-2}\right).$$

The stencil associated to the second part is

$$\mathcal{V}_1(j) = \{j-2, j+2\}.$$

Since the first part Div_j^0 naturally corresponds to a scheme which satisfies the maximum principle, we need to modified only the second part. We obtain the scheme

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^0 + \nu \mathrm{Div}_j^1 + \alpha\nu \sum_{l \in \mathcal{V}_1(j)} a_{jl}^1 (u_l - u_j) \tag{3.10}$$

where $\alpha \geq 0$ is a coefficient to be prescribed and $a_{jl}^1 = a_{lj}^1 = \frac{\left|\mathrm{Div}_j^1\right|}{\Sigma_j^1} + \frac{\left|\mathrm{Div}_l^1\right|}{\Sigma_l^1}$. The quantity Σ_j^1 is defined in accordance by

$$\Sigma_j^1 = \sum_{l \in \mathcal{V}_1(j)} |u_l - u_j|. \tag{3.11}$$

3.4. A second modified scheme

However the decomposition (3.9) is not the only one. One can consider

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu \mathrm{Div}_j^3 \tag{3.12}$$

where the first part is now

$$\mathrm{Div}_j^2 = u_{j+1} - 2u_j + u_{j-1}$$

and the second part is

$$\mathrm{Div}_j^3 = -\frac{1}{12}\left(u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2}\right). \tag{3.13}$$

The stencil associated to this new second part is the total one

$$\mathcal{V}_3(j) = \mathcal{V}(j) = \{j-2, j-1, j+1, j+2\}.$$

Since the first part Div_j^2 satisfies the maximum principle, we need to modified only the second part. We obtain the scheme

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu \mathrm{Div}_j^3 + \alpha\nu \sum_{l \in \mathcal{V}_3(j)} a_{jl}^3 (u_l - u_j) \tag{3.14}$$

where $\alpha \geq 0$ is a coefficient to be prescribed and

$$a_{jl}^3 = a_{lj}^3 = \frac{\left|\mathrm{Div}_j^3\right|}{\Sigma_j^3} + \frac{\left|\mathrm{Div}_l^3\right|}{\Sigma_l^3}.$$

The quantity Σ_j^3 is defined by

$$\Sigma_j^3 = \sum_{l \in \mathcal{V}^3(j)} |u_l - u_j|. \tag{3.15}$$

3.5. A third modified scheme

Here we perform a more important modification. We start from the decomposition

$$u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2} = (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}).$$

We define

$$\Sigma_j^4 = |u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}| + |u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}|$$

and $a_{jl}^4 = a_{lj}^4 = \frac{\left|\mathrm{Div}_j^3\right|}{\Sigma_j^4} + \frac{\left|\mathrm{Div}_l^3\right|}{\Sigma_l^4}$ for $l \in \{j+1, j-1\} = \mathcal{V}_4(j)$ (otherwise $a_{jl} = 0$). Let us consider the scheme

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu \mathrm{Div}_j^3 + \alpha\nu a_{j,j+1}^4 (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - \alpha\nu a_{j,j-1}^4 (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}). \tag{3.16}$$

Proposition 3.8. *Set $\alpha = 1$. The scheme (3.16) is of Finite Volume type and satisfies the maximum principle under the standard CFL condition (2.5).*

Proof. We focus the second statement of the claim since the first statement is evident from (3.16). Performing the same kind of algebra as before we rewrite the second part of the linear flux as follows

$$\mathrm{Div}_j^3 = \frac{\mathrm{Div}_j^3 \, |u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}|}{\Sigma_j^4} + \frac{\mathrm{Div}_j^3 \, |u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}|}{\Sigma_j^4}.$$

Plugging in (3.16) one obtains

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu w (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - \nu z (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) \tag{3.17}$$

where

$$w = \left((\alpha + \varepsilon_{j,j+1}) \frac{\left|\mathrm{Div}_j^3\right|}{\Sigma_j^4} + \alpha \frac{\left|\mathrm{Div}_{j+1}^3\right|}{\Sigma_{j+1}^4} \right)$$

and

$$z = \left((\alpha - \varepsilon_{j,j-1}) \frac{\left|\mathrm{Div}_j^3\right|}{\Sigma_j^4} + \alpha \frac{\left|\mathrm{Div}_{j-1}^3\right|}{\Sigma_{j-1}^4} \right).$$

Here

$$\varepsilon_{j,j+1} = \begin{cases} +1 \text{ if } \mathrm{Div}_j^3(u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) > 0, \\ -1 \text{ if } \mathrm{Div}_j^3(u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) \le 0, \end{cases}$$

and

$$\varepsilon_{j,j-1} = \begin{cases} +1 \text{ if } \mathrm{Div}_j^3(u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) > 0, \\ -1 \text{ if } \mathrm{Div}_j^3(u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) \le 0. \end{cases}$$

By construction $\left|\mathrm{Div}_j^3\right| \le \frac{1}{12}\Sigma_j^4$ for all j, so $w, z \le \frac{1}{4}$. The expansion of formula (3.17) is

$$\overline{u}_j = (1 - 2\nu + 3\nu w + 3\nu z)u_j + \nu\,(1 - 3w - z)\,u_{j+1} + \nu\,(1 - w - 3z)\,u_{j-1} + \nu w u_{j+2} + \nu z u_{j-2} \tag{3.18}$$

from which the result is deduced since all weights are non negative under CFL $2\nu \le 1$: that is $1-2\nu+3\nu w+3\nu z \ge 0$, $1 - 3w - z \ge 0$, $1 - w - 3z$, $\nu w \ge 0$ and $\nu z \ge 0$. □

3.6. Second order in time

The second order in time scheme (2.13) admits the explicit form

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + (1 - 6\nu)\nu \mathrm{Div}_j^3 \tag{3.19}$$

which is very close to (3.12). So it is easy to modify (3.16) which becomes

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + (1 - 6\nu)\nu \mathrm{Div}_j^3 \tag{3.20}$$

$$+\alpha|1 - 6\nu|\nu a_{j,j+1}^4 (u_{j+2} - 3u_{j+2} + 3u_j - u_{j-1}) - \alpha|1 - 6\nu|\nu a_{j,j-1}^4 (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}).$$

Proposition 3.9. *Assume $6\nu \le 1$. The scheme (3.20) satisfies the maximum principle under the standard CFL condition (2.5).*

Proof. Change w (resp. z) in $|1 - 6\nu|w$ (resp. $|1 - 6\nu|z$) in (3.18). □

3.7. General formulation

Motivated by the previous examples, we will study the following family of modified schemes

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j + \alpha\nu g_j \tag{3.21}$$

where Div_j is a given linear stencil and

$$g_j = \sum_l a_{jl}(u_l - u_j)$$

is the non linear correction such that a_{jl} vanishes for $|j - l|$ large enough. We introduce natural notations.

- T is the translation operator, that is

$$(Tu)_j = u_{j+1}, \quad u \in l_2.$$

- D is the difference operator, that is
$$D = T - I.$$
- We note for convenience the operator $A \in \mathcal{L}(l_2)$ such that
$$(Au)_j = -\mathrm{Div}_j, \quad \forall u \in l_2.$$

It has already been stressed in Remark 3.1 that the non linear correction is defined in a continuous manner. The continuity constant defined below plays an important role in the analysis developed in this work.

Definition 3.10. Let $k \in \mathbb{N}^\star$ be an integer naturally defined by the scheme. Throughout this paper, the continuity constant of g with respect to $D^k u$ will be denoted by $Q > 0$, that is
$$\|g\| \leq Q \|D^k u\|. \tag{3.22}$$

For example let us consider the scheme (3.16) for which the source $g = (g_j) \in l_2$ is defined by
$$g_j = a^4_{j,j+1}(u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - a^4_{j,j-1}(u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) \tag{3.23}$$

Proposition 3.11. *The source of the scheme (3.16) is such that* $\|g\| \leq \frac{1+\sqrt{2}}{12}\|D^4 u\|$, *that is* $Q = \frac{1+\sqrt{2}}{12}$ *and* $k = 4$.

Proof. We first notice that $\|A_3 u\| \leq \frac{1}{12}\|D^4 u\|$ where we have used natural notations compatible with (3.16). It remains to compute the continuity constant of g with respect to $A_3 u$. One has the decomposition $g_j = h_j + k_j + l_j$ with
$$h_j = \frac{(u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2})}{\Sigma^4_j} \left|\mathrm{Div}^3_j\right|,$$
$$k_j = \frac{u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}}{\Sigma^4_{j+1}} \left|\mathrm{Div}^3_{j+1}\right|$$
and
$$l_j = \frac{u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}}{\Sigma^4_{j-1}} \left|\mathrm{Div}^3_{j-1}\right|.$$
Since $|h_j| \leq |\mathrm{Div}^3_j|$, then $\|h\| \leq \|A_3 u\|$. One the other hand one has that $|k_j| \leq \alpha_{j+1}|\mathrm{Div}^3_{j+1}|$ and $|l_j| \leq \beta_{j-1}|\mathrm{Div}^3_{j-1}|$ where the coefficients
$$\alpha_{j+1} = \frac{|u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}|}{\Sigma^4_{j+1}} \text{ and } \beta_{j-1} = \frac{|u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}|}{\Sigma^4_{j-1}}$$
are the ones of a convex combination in the sense
$$0 \leq \alpha_j, \beta_j \text{ and } \alpha_j + \beta_j = 1 \qquad \forall j.$$
So
$$\|k + l\|^2 \leq \Delta x \sum_j \left(\alpha_{j+1}|\mathrm{Div}^3_{j+1}| + \beta_{j-1}|\mathrm{Div}^3_{j-1}|\right)^2$$
$$\leq \Delta x \sum_j 2\alpha^2_{j+1}|\mathrm{Div}^3_{j+1}|^2 + \Delta x \sum_j 2\beta^2_{j-1}|\mathrm{Div}^3_{j-1}|^2$$
$$\leq 2\Delta x \sum_j \left(\alpha^2_j + \beta^2_j\right)|\mathrm{Div}^3_j|^2 \leq 2\Delta x \sum_j |\mathrm{Div}^3_j|^2.$$
Therefore $\|k + l\| \leq \sqrt{2}\|A_3 u\|$ from which the result is deduced. □

Proposition 3.12. *The source of the scheme (3.20) is such that*

$$\|g\| \leq \frac{1+\sqrt{2}}{12}|1-6\nu|\|D^4u\|,$$

that is $Q = \frac{1+\sqrt{2}}{12}|1-6\nu|$ *and* $k=4$.

Proof. Evident from previous proposition and definition (3.20). The $|1-6\nu|$ is because we use the second order in time scheme. □

Due to the fact that non linear corrections may be rewritten under a finite volume form as stated in Proposition 3.2, there exists s such that

$$g = Ds. \tag{3.24}$$

For all schemes finite volume considered in this work, the continuity estimate (3.22) is also true for s. There exists $\widetilde{Q} > 0$ such that

$$\|s\| \leq \widetilde{Q}\|D^k u\|. \tag{3.25}$$

The continuity s with respect to $D^k u$ will also be of interest for the final convergence result. Nevertheless an important difference between (3.22) and and (3.25) is that the exact value of $\widetilde{Q}$ has no influence on the following convergence theory. This is why the verification of (3.25) can be a little simplified with respect to the one of (3.22).

Let us first detail the principle which is behind (3.25) and is merely a corollary of formula (3.6). By comparison of (3.21) and (3.22) one has that

$$\alpha\nu g_j = \frac{\Delta t}{\Delta x}\left(f^{\mathrm{cor}}_{j+\frac{1}{2}} - f^{\mathrm{cor}}_{j-\frac{1}{2}}\right)$$

which means using (3.6) that $g = Ds$ with $s_j = \sum_{l\leq j-\frac{1}{2}\leq m} a_{ml}(u_m - u_l)$ where the sum is over a finite number of terms. One gets

$$|s_j| \leq \sum_{l\leq j-\frac{1}{2}\leq m} a_{ml}|u_m - u_l|.$$

Since the continuity of g with respect to $D^k u$ is necessarily obtained through bounds for these terms $a_{ml}|u_m - u_l|$, the inequality (3.22) can be adapted to obtain (3.25). This general principle can be extended to the fourth order schemes (3.16) and (3.20).

Proposition 3.13. *The schemes (3.16) and (3.20) are such that* $g = Ds$ *with*

$$\|s\| \leq \widetilde{Q}\|D^4u\| \tag{3.26}$$

for some continuity constant $\widetilde{Q} \in \mathbb{R}^*$.

Proof. Concerning the first scheme, the formula (3.16) which is difference between two terms evidently implies

$$s_j = a^4_{j,j-1}(u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2})$$

$$= \frac{u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}}{\Sigma^4_j}\left|\mathrm{Div}^3_j\right| + \frac{u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}}{\Sigma^4_{j-1}}\left|\mathrm{Div}^3_{j-1}\right|.$$

Therefore $|s_j| \leq \left|\mathrm{Div}^3_j\right| + \left|\mathrm{Div}^3_{j-1}\right|$ which yields $\|s\| \leq 2\|\mathrm{Div}^3_j\| \leq \frac{1}{6}\|D^4u\|$ where we have used (3.13): that is $\widetilde{Q} = \frac{1}{6}$.

The scheme (3.20) is obtained from (3.16) after multiplication of the correction terms by the factor $1-6\nu$. In this case the inequality becomes $\|s\| \leq \frac{|1-6\nu|}{6}\|D^4u\|$. The proof is ended. □

Remark 3.14. Before going further it is fundamental to make a remark about the integer k. All non linear corrections considered in this work are also continuously controlled by $D^{k_0}u$ with $k_0 = 1$. See for example (3.14) where the non linear term is $a^3_{jl}(u_l - u_j)$. Since $\left|a^3_{jl}\right|$ is bounded by construction, the non linear term is bounded by the norm of the first discrete derivative: that is $\|g\| \leq C\|Du\|$ for some $C > 0$. However this case is useless in the context of this work because we will use contractivity arguments which are true only for $k > 2$, see Proposition 4.3 and Remark 4.4. In some sense it rules out the possibility to treat such non linear schemes in the TVD context, as pursued in [5].

4. Estimates

With the above notations and definition, any of the previous schemes can be rewritten as

$$\overline{u} = (I - \nu A)u + \alpha\nu g.$$

The Duhamel's formula can be used to express the solution at time step n in function of the initial solution and the correction

$$u^n = (I - \nu A)^n u^0 + \alpha\nu \sum_{p=0}^{n-1} (I - \nu A)^{n-1-p} g^p. \tag{4.1}$$

It shows that the solution at time step $n\Delta t$ is the sum of the standard discrete solution plus a contribution due the corrections. Our goal is to show that the global correction is small in some norm. Before we need some *a priori* estimates on the g^ps. These estimates will be obtained using a control of the non linear part of the Duhamel's formula by the linear part.

By application of the operator D^k to the Duhamel formula (4.1) and use of the commutativity $DA = AD$, one gets the identity

$$D^k u^n = (I - \nu A)^n D^k u^0 + \alpha\nu \sum_{p=0}^{n-1} (I - \nu A)^{n-1-p} D^k g^p. \tag{4.2}$$

Using the continuity of g^p with respect to $D^k u^p$, it yields the estimate

$$\left\|D^k u^n\right\| \leq \left\|(I - \nu A)^n D^k u^0\right\| + Q\alpha\nu \sum_{p=0}^{n-1} \left\|(I - \nu A)^{n-1-p} D^k\right\| \left\|D^k u^p\right\|.$$

A fundamental property is the following.

Proposition 4.1. *Assume l_2 contractivity of the linear part of the scheme. Assume there exists $\varepsilon > 0$ such that*

$$Q\alpha\nu \sum_{l=0}^{\infty} \left\|(I - \nu A)^l D^k\right\| \leq 1 - \varepsilon. \tag{4.3}$$

Then one has the stability estimate

$$\sup_{n\in\mathbb{N}} \left\|D^k u^n\right\| \leq \frac{1}{\varepsilon} \left\|D^k u^0\right\| \qquad \forall u^0 \in l_2.$$

Remark 4.2. The dissipativity of the scheme insures that $\lim_{l\to\infty} \|(I - \nu A)^l D^k\| = 0$ this property has been used in [7,8] to study non linear schemes for the transport equation. In this study, the dissipativity of the semi-group $(I - \nu A)^l$ is much stronger since it is a fundamental property of the heat equation. This dissipativity is essential to obtain a control of the non linear part of the Duhamel's formula.

Proof. The l_2 contractivity is in our case equivalent to the stability in l_2 with a stability constant equal to 1, that is $\|(I-\nu A)^l\| \le 1$ for all $l \in \mathbb{N}$. Let us define $Z_N = \sup_{n\le N} \|D^k u^n\|$ which satisfies the estimate

$$Z_N \le \|D^k u^0\| + \left(Q\alpha\nu \sum_{p=0}^{N-1} \| \left(I - \Delta t A\right)^{n-1-p} D^k \| \right) Z_{N-1}.$$

It turns into $Z_N \le \|D^k u^0\| + (1-\varepsilon) Z_N$ which shows that $Z_N \le \frac{1}{\varepsilon}\|D^k u^0\|$. Since it is true for all N, it shows the claim. □

For convenience we define the generic function

$$F(\nu) = Q\nu \sum_{l=0}^{\infty} \| \left(I - \nu A\right)^l D^k \|. \tag{4.4}$$

Since $\alpha = 1$ is used to guarantee the maximum property of the schemes, the criterion that we study is ultimately

$$F(\nu) < 1.$$

We will now study the function F for the different operators A and different ks. The estimates developed below are used to prove this estimate. The interested reader can jump first to Figure 4 where some numerical evaluations of these functions are displayed and then go back to the theory of the next section. The main message of Figure 4 is that the condition $F(\nu) < 1$ is naturally satisfied.

4.1. Basic operator

Some of the justifications of the bounds used hereafter for general operators A are greatly simplified if one can first prove (4.3) with

$$A_1 = -T + 2I - T^{-1}. \tag{4.5}$$

For a given k, the Fourier symbol of $\left(I - \nu A_1\right)^l D^k$ is

$$\lambda_l(\theta) = \left(1 + \nu(\mathrm{e}^{i\theta} - 2 + \mathrm{e}^{-i\theta})\right)^l \left(\mathrm{e}^{i\theta} - 1\right)^k, \qquad \theta \in \mathbb{R}.$$

One has that

$$|\lambda_l(\theta)| = \left|1 - 4\nu \sin^2 \frac{\theta}{2}\right|^l \left|2 \sin \frac{\theta}{2}\right|^k.$$

We assume

$$4\nu \le 1 \tag{4.6}$$

which is twice more restrictive than (2.5). It will simplify a lot the analysis and is not a real restriction. We perform the change of variable $y = \left|2 \sin \frac{\theta}{2}\right|$. Let us define the function

$$f_l^\nu(y) = (1 - \nu y^2)^l y^k$$

so that

$$\|(I - \nu A_1)^l D^k\| = \max_{0\le y\le 2} f_l^\nu(y). \tag{4.7}$$

Let us set

$$\mu_l(\nu) = \max_{0\le y\le 2} f_l^\nu(y) \qquad l \in \mathbb{N}.$$

We finally define

$$E_1(\nu) = \nu \sum_{l=0}^{\infty} \mu_l(\nu).$$

Proposition 4.3. *Assume $k < 2$. Then $E_1(\nu) = +\infty$.*

Proof. Note that μ_l, E_1, ... can be studied for real positive k as well, not only integer values. By definition $\mu_l(\nu) \geq f_l^\nu(z)$ for any $z \in [0,2]$. So

$$E_1(\nu) \geq \nu \sum_{l=0}^{\infty} (1-\nu z^2)^l z^k = z^{k-2} \quad \forall z \in [0,2].$$

If $k < 2$ the right hand side is singular for $z = 0$. It ends the proof. □

Remark 4.4. It will be showed hereafter that $k = 2$ is also singular. So it must be understood that $k > 2$ in the rest of the paper.

Elementary properties of f_l^ν and $\mu_l(\nu)$ are the following: one has

$$f_{l+1}^\nu(y) \leq f_l^\nu(y), \qquad 0 \leq y \leq 2$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\nu} f_l^\nu(y) \leq 0, \qquad 0 \leq y \leq 2.$$

Therefore

$$\mu_{l+1}(\nu) \leq \mu_l(\nu) \qquad \forall l \in \mathbb{N}.$$

Next we study the function f_l^ν. One has

$$\frac{\mathrm{d}}{\mathrm{d}y} f_l^\nu(y) = -2l\nu y(1-\nu y^2)^{l-1} y^k + k(1-\nu y^2)^l y^{k-1} = (1-\nu y^2)^{l-1} y^{k-1} \left(-2l\nu y^2 + k - k\nu y^2\right).$$

We define

$$y_l^2 = \frac{k}{\nu(2l+k)}$$

so that f_l^ν increases from $y = 0$ to y_l, then decreases from y_l to $\left(\frac{1}{\nu}\right)^{\frac{1}{2}}$. We note that

$$0 \leq \cdots \leq y_{l+1}^2 \leq y_l^2 \leq \cdots \leq y_0^2 = \frac{1}{\nu}$$

and that $4 \leq \frac{1}{\nu}$ due to the CFL condition (4.6). So

- Either $y_l < 2$ and
$$\mu_l(\nu) = f_l^\nu(y_l) = \left(\frac{2l}{2l+k}\right)^l \left(\frac{k}{\nu(2l+k)}\right)^{\frac{k}{2}}.$$
- Or $y_l \geq 2$ and
$$\mu_l(\nu) = f_l^\nu(2) = (1-4\nu)^l 2^k.$$

The transition is for $L(\nu)$ such that

$$\frac{k}{\nu(2(L(\nu)+1)+k)} < 4 \leq \frac{k}{\nu(2L(\nu)+k)}$$

that is

$$\frac{k}{2}\left(\frac{1}{4\nu} - 1\right) - 1 < L(\nu) \leq \frac{k}{2}\left(\frac{1}{4\nu} - 1\right) \tag{4.8}$$

which means that

$$L(\nu) = \left[\frac{k}{2}\left(\frac{1}{4\nu} - 1\right)\right].$$

Proposition 4.5. *Assuming $k > 2$, one has the bound $E_1(\nu) \leq h(\nu)$ with*

$$h(\nu) = 2^{k-2}\left(1 - (1-4\nu)^{\frac{k}{8\nu}}\right) + \lambda_k(\nu)\left(\frac{2^{k-2}k}{k-2} + \nu 2^k\right). \tag{4.9}$$

where

$$\lambda_k(\nu) = \left(\frac{2\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+2}{2\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+2+k}\right)^{\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1}.$$

The proof which is only technical is postponed in the appendix.

Remark 4.6. It can be checked this estimate is sharp for small ν. In particular one has that the limit value in 0^+ is given by $E_1(\nu) = h(0)$ and

$$h(0) = \frac{2^k}{4} + \mathrm{e}^{-\frac{k}{2}}\frac{2^{k-1}}{k-2}. \tag{4.10}$$

Indeed the first term admits the limit value $(1-4\nu)^{\frac{k}{8\nu}}$ tends to $\mathrm{e}^{-\frac{k}{2}}$ when ν vanishes. Concerning the second term one noticed that $\lambda_k(\nu) = \left(\frac{2(L+1)}{2(L+1)+k}\right)^{L+1}$ with $L \to \infty$ in the regime $\nu \to 0^+$. Therefore $\lambda_k(0^+) = \mathrm{e}^{-\frac{k}{2}}$ by continuity: it yields (4.10). This numerical value can be checked in the numerical Figure 3.

Remark 4.7. The same proof shows that $E_1(\nu) = +\infty$ for $k = 2$, because Z_2 (see below) diverges.

4.2. Fourth order operator

Next we consider the operator

$$A_2 = -T + 2I - T^{-1} + \frac{1}{12}\left(T^2 - 4T + 6I - 4T^{-1} + T^{-2}\right) \tag{4.11}$$

which corresponds to the scheme (2.8) or (3.1). For a given k, the Fourier symbol of $(I - \nu A_2)^l D^k$ is

$$\lambda_l(\theta) = \left(1 + \nu(\mathrm{e}^{i\theta} - 2 + \mathrm{e}^{-i\theta}) - \frac{1}{12}\nu\left(\mathrm{e}^{2i\theta} - 4\mathrm{e}^{i\theta} + 6 - 4\mathrm{e}^{-i\theta} + \mathrm{e}^{-2i\theta}\right)\right)^l \times \left(\mathrm{e}^{i\theta} - 1\right)^k.$$

One has that

$$|\lambda_l(\theta)| = \left|1 - 4\nu\sin^2\frac{\theta}{2} - \frac{4}{3}\nu\sin^4\frac{\theta}{2}\right|^l \left|2\sin\frac{\theta}{2}\right|^k.$$

The method of analysis is very similar to the previous one. We assume for simplicity of the analysis that

$$\frac{16}{3}\nu \leq 1. \tag{4.12}$$

Let us define the function

$$g_l^\nu(y) = (1 - \nu y^2 - \frac{1}{12}\nu y^4)^l y^k, \qquad y = 2|\sin\frac{\theta}{2}|,$$

so that

$$\|(I - \nu A_2)^l D^k\| = \max_{0\leq y\leq 2} g_l^\nu(y). \tag{4.13}$$

Let us set

$$\sigma_l(\nu) = \max_{0\leq y\leq 2} g_l^\nu(y)$$

together with

$$E_2(\nu) = \nu\sum_{l=0}^{\infty}\sigma_l(\nu).$$

Proposition 4.8. *Assume the CFL condition (4.12). Then*

- $\sigma_l(\nu) \leq \mu_l(\nu)$ *for all* l*: it yields* $E_2 \leq E_1$ *and*

$$E_2 \in L^\infty \left[0, \frac{3}{16}\right].$$

- $\sigma_{l+1}(\nu) \leq \sigma_l(\nu)$ *for all* l.

Proof. Evident since $g_l^\nu(y) \leq f_l^\nu(y)$ for $\frac{16}{3}\nu \leq 1$. □

The extremal point of g_l is z_l such that $g_l'(z_l) = 0$. That is

$$(2l+k)z_l^2 + \left(\frac{k}{12} + \frac{l}{3}\right) z_l^4 = \frac{k}{\nu}.$$

One sees of course that $z_l \leq y_l$. The solution is

$$z_l^2 = \frac{-(2l+k) + \sqrt{(2l+k)^2 + 4\left(\frac{k}{12} + \frac{l}{3}\right)\frac{k}{\nu}}}{2\left(\frac{k}{12} + \frac{l}{3}\right)},$$

or also in the conjugate form

$$z_l^2 = \frac{2\frac{k}{\nu}}{(2l+k)z + \sqrt{(2l+k)^2 + 4\left(\frac{k}{12} + \frac{l}{3}\right)\frac{k}{\nu}}}.$$

The value of $\sigma_l(\nu)$ is as follows

- Either $z_l < 2$ so

$$\sigma_l(\nu) = \left(1 - \nu z_l^2 - \frac{4}{3}\nu z_l^4\right)^l z_l^k.$$

- Or $n_l \geq 2$ and

$$\sigma_l(\nu) = g_l^\nu(2) = \left(1 - \frac{16}{3}\nu\right)^l 2^k.$$

The transition is the largest M such that

$$(2l+k)z_M^2 + \left(\frac{k}{12} + \frac{l}{3}\right) z_M^4 < \frac{k}{\nu}.$$

Therefore

$$E_2(\nu) = \sum_{l \leq M} \left(1 - \frac{16}{3}\nu\right)^l 2^k + \sum_{M+1 \leq l} \left(1 - \nu z_l^2 - \frac{4}{3}\nu z_l^4\right)^l z_l^k. \tag{4.14}$$

4.3. Full fourth order operator

Finally we consider the Fourier symbol of $(I - \nu A_3)^l D^k$ of the operator

$$A_3 = -T + 2I - T^{-1} + \frac{1-6\nu}{12}\left(T^2 - 4T + 6I - 4T^{-1} + T^{-2}\right) \tag{4.15}$$

which corresponds to the second order in time and fourth order in space linear scheme. The scheme is full fourth order in the sense that second order in time corresponds to fourth order in space under CFL condition $6\nu \leq 1$. We define the function

$$E_3(\nu) = \nu|1 - 6\nu| \sum_{l=0}^{\infty} \|(I - \nu A_3)^l D^k\|.$$

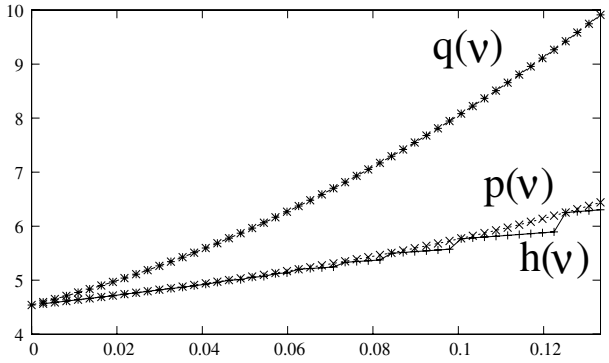

FIGURE 1. Numerical illustration of the bounds $h \leq p \leq q$ on the interval $\nu \in [0 : \frac{1}{6}]$.

Using the same method as before, it can be checked that

$$E_3(\nu) = \nu|1-6\nu| \sum_{l \leq N} \left(1 - \left(4 + \frac{4}{3}(1-6\nu)\right)\nu\right)^l 2^k + \nu|1-6\nu| \sum_{N+1 \leq l} \left(1 - \nu w_l^2 - \frac{4}{3}\nu(1-6\nu)w_l^4\right)^l w_l^k$$

where the extremal point of the function $y \mapsto h_l^\nu(y) = \left(1 - \nu y^2 - \frac{(1-6\nu)}{12}\nu y^4\right)^l y^k$ is

$$w_l^2 = \frac{2\frac{k}{\nu}}{(2l+k) + \sqrt{(2l+k)^2 + 4\left(\frac{k}{12} + \frac{l}{3}\right)\frac{k}{\nu}(1-6\nu)}}.$$

The transition is the largest N such that

$$(2l+k)w_N^2 + \left(\frac{k}{12} + \frac{l}{3}\right)(1-6\nu)w_N^4 < \frac{k}{\nu}.$$

Since $h_l^\nu(y) \leq f_l^\nu(y)$ one has that

$$E_3(\nu) \leq |1-6\nu| E_1(\nu). \tag{4.16}$$

4.4. Application to the scheme (3.16)

In the following we apply the various inequalities to the case $k = 4$ and $Q = \frac{1+\sqrt{2}}{12}|1-6\nu|$ which allowed a complete analysis of the scheme (3.20) (see also proposition 3.12). Our first task is to show that $F_3(\nu) = QE_3(\nu)$ is such that $F_3 < 1$ so that the stability bound $\|D^4 u^p\| \leq C\|D^4 u^0\|$ $\forall p$ holds from proposition (4.1). This will be performed with a series of elementary bounds for three functions h, p and q. These bounds $h \leq p \leq q$ are illustrated on Figure 1, on which one sees that the three functions have the same continuous limit $4 + 4e^{-2}$ at $\nu = 0$.

Proposition 4.9. *Assume $k = 4$ and $\nu \leq \frac{1}{6}$. Then the function h defined in (4.9) is such that $h \leq p$ where*

$$p(\nu) = 4\left(1 - (1-4\nu)^{\frac{1}{2\nu}}\right) + (1-4\nu)^{\frac{1}{2\nu}-2}(8+16\nu).$$

Proof. The function h is the sum of two terms. Concerning the first term we notice that our hypotheses imply that $\frac{k}{8\nu} = \frac{1}{2\nu}$. So

$$2^{k-2}\left(1 - (1-4\nu)^{\frac{k}{8\nu}}\right) = 4\left(1 - (1-4\nu)^{\frac{1}{2\nu}}\right).$$

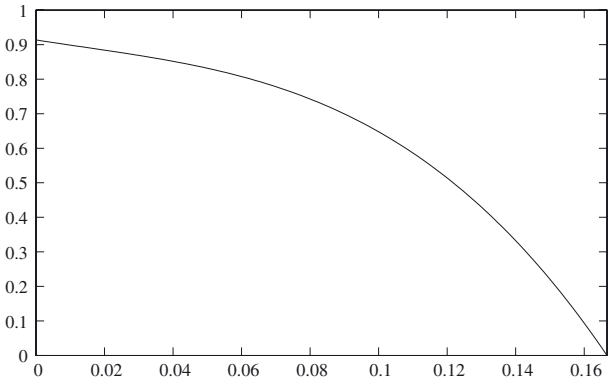

FIGURE 2. Plot of $\frac{1+\sqrt{2}}{12}(1-6\nu)q(\nu)$ in the range $\nu \in [0, \frac{1}{6}]$. The function is monotone decreasing.

Concerning the second contribution, we notice that the value of the function $\lambda_k(\nu)$ can be expressed as $\lambda_k(\nu) = \left(\frac{x}{x+u}\right)^x \equiv v(x)$ with $x = \left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1$ and $u = \frac{k}{2}$. It is immediate to check that v is a decreasing function for positive x. Since $x \geq y \equiv \frac{k}{2}\left(\frac{1}{4\nu}-1\right)$, it yields

$$\lambda_k(\nu) = v(x) \leq v(y) = \left(\frac{k\left(\frac{1}{4\nu}-1\right)}{k\left(\frac{1}{4\nu}-1\right)+k}\right)^{\frac{k}{2}\left(\frac{1}{4\nu}-1\right)}.$$

With $k = 4$ it turns into $\lambda_4(\nu) \leq (1-4\nu)^{\frac{1}{2\nu}-2}$ from the result is deduced after summation of the two contributions. □

Proposition 4.10. *One has $p \leq q$ with*

$$q(\nu) = 4 + 4\mathrm{e}^{-2}(1+24\nu)(1+12\nu), \qquad \nu \leq \frac{1}{6}.$$

Proof. One first has that

$$(1-4\nu)^{\frac{1}{2\nu}-2} = \mathrm{e}^{\frac{2}{y}(1-y)\ln(1-y)}, \qquad y = 4\nu.$$

Since $y = 4\nu \leq 1$ one has immediately $(1-y)\ln(1-y) \leq (1-y)(-y) = -y+y^2$: so

$$(1-4\nu)^{\frac{1}{2\nu}-2} \leq \mathrm{e}^{-2+2y} = \mathrm{e}^{-2+8\nu}.$$

Therefore

$$p(\nu) = 4 + (1-4\nu)^{\frac{1}{2\nu}-2}\left(8+16\nu-4(1-4\nu)^2\right)$$

$$= 4 + (1-4\nu)^{\frac{1}{2\nu}-2}\left(4+48\nu-64\nu^2\right) \leq 4 + (1-4\nu)^{\frac{1}{2\nu}-2}(4+48\nu) \leq 4 + 4\mathrm{e}^{-2}\mathrm{e}^{8\nu}(1+12\nu).$$

On the interval $[0, \frac{1}{6}]$, one has that $\mathrm{e}^{8\nu} \leq 1+24\nu$ which is easy to show by convexity. □

Proposition 4.11. *On the interval $[0, \frac{1}{6}]$, the function $(1-6\nu)q(\nu)$ decays from $4+4\mathrm{e}^{-2}$ until 0.*

Proof. For such a simple function, we consider that a plot (with gnuplot) is enough. We refer to Figure 2.

□

Lemma 4.12. *Assume $6\nu \leq 1$. The function F_3 associated to the scheme (3.16) is such that*

$$F_3(\nu) \leq \frac{1+\sqrt{2}}{12}\left(4+4\mathrm{e}^{-2}\right) \approx 0.91\ldots$$

As a consequence the fourth order discrete derivative of the numerical solution is bounded uniformly with respect to the iteration number n and to the CFL number ν: there exists $C>0$ such that

$$\|D^4 u^n\| \leq C\|D^4 u^0\| \qquad n \in \mathbf{N}.$$

Remark 4.13. The constant is $C \approx \frac{1}{1-0.91\ldots} \leq \frac{1}{0.08} = 12.5$. See the fundamental proposition 4.1.

Proof. By construction

$$F_3(\nu) = QE_3(\nu) \leq Q(1-6\nu)E_1(\nu) \leq Q(1-6\nu)q(\nu) \leq Qq(0) = \frac{1+\sqrt{2}}{12}\left(4+4\mathrm{e}^{-2}\right).$$

Numerical application show that

$$\frac{1+\sqrt{2}}{12}\left(4+4\mathrm{e}^{-2}\right) \approx 0.913647\cdots < 1.$$

Therefore $F_3 < 1$ over the range $\nu \in \left[0, \frac{1}{6}\right]$. See also the plot in Figure 4. □

5. Convergence

First we consider any scheme (4.1) such that an estimate on the norm of g^p can be obtained. And after we particularize the estimates for the scheme (3.20) and obtain the main convergence result of this work.

Let us go back to the Duhamel formula (4.1) and consider the reminder

$$R = \nu \sum_{p=0}^{n-1} (I-\nu A)^{n-1-p} g^p, \qquad A = A_1, A_2 \text{ or } A_3.$$

As explained previously in proposition 3.13, the non linear corrections can be rewritten under the finite volume form $g^p = Ds^p$ with the continuity estimate $\|s^p\| \leq \widetilde{Q}\|D^k u\|$. So one can write

$$R = \nu \sum_{p=0}^{n-1} \left((I-\nu A)^{n-1-p} D\right) s^p.$$

Using moreover that $g_l^\nu \leq f_l^\nu$ (resp. $h_l^\nu \leq f_l^\nu$) for all k, one can upper bound using A_1 whatever A and gets

$$\|R\| \leq \widetilde{Q}\left(\nu \sum_{p=0}^{n-1} \|(I-\nu A)^{n-1-p} D\|\right)\|D^k u_0\| \leq \widetilde{Q}\left(\underbrace{\nu \sum_{p=0}^{n-1} \|(I-\nu A_1)^{n-1-p} D\|}_{=Q_n}\right)\|D^k u_0\|.$$

Proposition 5.1. *Let $T>0$. Using the previous assumptions, there exists $C_3>0$ such that*

$$\|Q_n\| \leq \frac{C_3}{\Delta x}, \qquad n\Delta t \leq T.$$

Proof. The term between parenthesis is very similar to the series analyzed previously, but here $k = 1$ so that the infinite series cannot converge, see proposition 4.3. In order to bound this term we consider the truncated series

$$Q_n \leq \nu \sum_{l=0}^{\frac{T}{\Delta t}} \mu_l(\nu).$$

Using (A.1) one gets

$$Q_n \leq \nu \underbrace{\sum_{l \leq L} (1-4\nu)^l 2}_{=Z_1} + \nu \underbrace{\sum_{L+1 \leq n}^{\frac{T}{\Delta t}} \left(\frac{2l}{2l+1}\right)^l \left(\frac{1}{\nu(2l+1)}\right)^{\frac{1}{2}}}_{=Z_2}, \quad L = \left[\frac{1}{2}\left(\frac{1}{4\nu} - 1\right)\right].$$

As before the first term is bounded since an elementary summation yields $\nu Z_1 \leq \nu \frac{2}{4\nu} = \frac{1}{2}$. The second term is bounded by exactly the same method as in the proof of (4.9). The only but essential difference is the sum H which becomes now

$$H' = \nu \sum_{\nu(L+2) \leq x_l \leq \nu \frac{T}{\Delta t}} \left(\frac{1}{2x_l + \nu}\right)^{\frac{1}{2}}.$$

We obtain $H' \leq \int_{\frac{1}{8}-\frac{\nu}{2}}^{\nu \frac{T}{\Delta t}} \frac{\mathrm{d}x}{(2x+\nu)^{\frac{1}{2}}} \quad (= I)$. For small Δt the integral diverges like

$$I \approx C\sqrt{\nu \frac{T}{\Delta t}} \leq \frac{C'(T)}{\Delta x} \tag{5.1}$$

thanks to the CFL condition. It ends the proof. □

Remark 5.2. Inequality (5.1) can be obtained directly from the fundamental inequality derived in [7,8].

Let $\Pi_{\Delta x}$ be the point-wise projector of a smooth function onto the mesh grid. Take the initial data as $u^0 = \Pi_{\Delta x} u(0)$.

Theorem 5.3. *Consider the scheme (3.20). Assume $u_0 \in H^4(\mathbb{R})$. Assume the CFL condition $6\nu \leq 1$. Let $T > 0$. Then there exists a constant $C > 0$ such that*

$$\|u^n - \Pi_{\Delta x} u(n\Delta t)\| \leq C\Delta x^3 |u_0|_{H^4(\mathbb{R})}, \qquad \forall n,\ n\Delta t \leq T. \tag{5.2}$$

Proof. In this case $k = 4$. It is an easy matter to show that the regularity assumption implies

$$\|D^4 u^0\| \leq c'' |u_0|_{H^4(\mathbb{R})} \Delta x^4.$$

Therefore $\|R\| \leq K|u_0|_{H^4(\mathbb{R})}\Delta x^{4-1} = K|u_0|_{H^4(\mathbb{R})}\Delta x^3$. On the other hand the linear part of the scheme is fourth order in space and second order in time. Therefore

$$\|(I - \nu A)^n u^0 - \Pi_{\Delta x} u(n\Delta t)\| \leq C\Delta x^4 |u_0|_{H^4(\mathbb{R})}.$$

Since $u^n = (I - \nu A)^n u^0 + R$, the triangular inequality shows the result. □

This theorem can be adapted to take into account the others Le Potier or modified schemes considered in this work. In particular the curve F_2 in Figure 4 is under threshold 1. So the schemes (3.16) also converges at order 3.

Remark 5.4. A case of interest is the scheme (3.14) for which one can prove that $Q = \frac{1}{4}$ and $q = 4$. A proof is given in the appendix. Considering the curves F_2 and F_3 of Figure 4 that are normalized for $Q = \frac{1+\sqrt{2}}{12}$, one can apply a correction factor $\lambda = \frac{4}{1+\sqrt{2}} \approx 1.24\ldots$ One gets that $\widetilde{F}_3 = \lambda F_3 < 1$ for all ν, and that $\widetilde{F}_2 = \lambda F_2 < 1$ for $\nu < .15$ approximatively. On this range it yields a control of the fourth order discrete derivative, and therefore

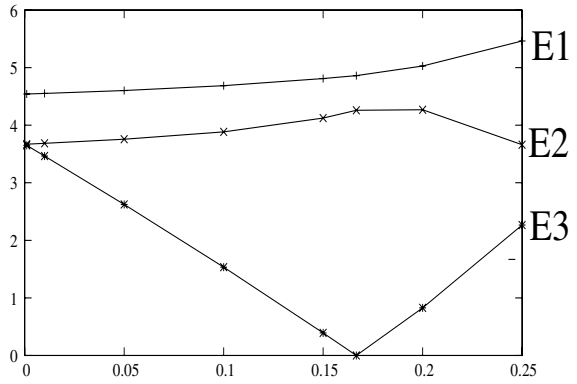

FIGURE 3. The constant $E_{1,2,3}$ for three different set of parameters.

a proof of convergence at order three if one considers the use of the non linear correction (3.14) with the $1-6\nu$ term.

Remark 5.5. The curve F_3 is smaller than the limit value in a range $[0, C[$ where the constant $C > \frac{1}{6}$ can be identified with numerical experiments. We infer that the scheme converges at order 3 in this larger range also.

Remark 5.6. The main open problem is the adaptation of such theorem of convergence in dimension two and greater. This is fully open problem.

6. Numerical tests

We perform simple numerical tests to assess the properties of the numerical schemes developed in this work.

6.1. Functions E and F

In Figure 3 we plot the numerical value of $E_1(\nu) = \nu \sum_{l=0}^{10^6} \mu_l(\nu)$, $E_2(\nu) = \nu \sum_{l=0}^{10^6} \nu_l(\nu)$ and $E_3(\nu) = \nu|1-6\nu| \sum_{l=0}^{10^6} \sigma_l(\nu)$. We observe that these quantities are pretty constant, less than 6 for $\nu \leq \frac{1}{4}$. The computed value $E_1(0)$ is very close to the exact value $\approx 4 + 4\mathrm{e}^{-2} \approx 4.5413\ldots$ Next in Figure 4 we plot

$$F_i = \frac{1+\sqrt{2}}{12} E_i, \qquad i = 1, 2, 3.$$

We observe that $F_2 < 1$ and $F_3 < 1$ for $\nu \leq \frac{1}{4}$. On the other hand $F_1 < 1$ is true only for approximatively $\nu \leq .18$. By inspection of the graphics, it is clear that the function F_3 is decreasing and bounded as stated in lemma 4.12.

6.2. Stability test

We consider the numerical solution of the heat equation on a 10 cells mesh. The initial data is a discrete Dirac mass. Such initial profiles are very convenient to illustrate the maximum principle. We observe in Figure 5 that the two fourth order in space linear schemes do not preserve the maximum principle. The three points scheme and the new third order non linear scheme preserve it.

6.3. Accuracy test

We solve the heat equation on the interval $[0, 1]$ with periodic boundary conditions. The initial data is $u_0(x) = \cos(2\pi x)$ so that the exact solution is $u(t) = \mathrm{e}^{-4\pi^2 t} u_0$. We measure in Table 1 and Figure 6 the relative error in l^2 norm at time $t = 0.1$ in function of the number of cells. As predicted by the theory, the modified scheme based on a the second order in time and fourth order in space scheme converges at order 3.

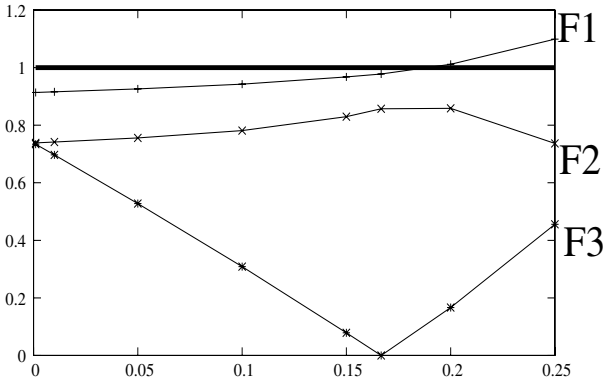

FIGURE 4. The functions $F_{1,2,3}$ for three different set of parameters. $F_1(\nu) = \frac{1+\sqrt{2}}{12} E_1(\nu)$ is for theoretical understanding. $F_2(\nu) = \frac{1+\sqrt{2}}{12} E_2(\nu)$ corresponds to the scheme (3.16). $F_3(\nu) = \frac{1+\sqrt{2}}{12} E_3(\nu)$ corresponds to the scheme (3.20). What is important is to be under the threshold 1 (in bold).

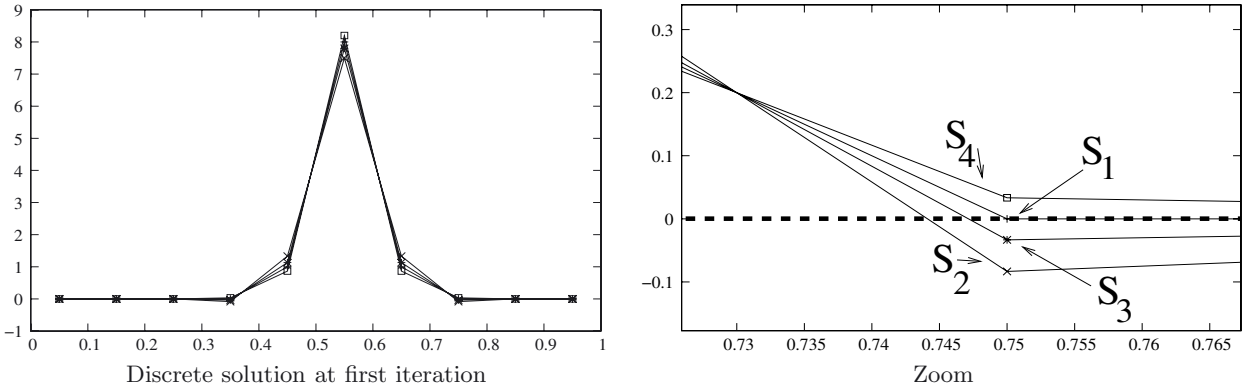

FIGURE 5. Numerical solution calculated by the four different schemes defined in Table 1: one iteration. One sees that the three points schemes and the modified scheme satisfy the maximum principle. The two fourth order in space schemes do not.

TABLE 1. Error in function of the mesh size for the four different schemes: S_1= second order in space and first order in time; S_2= fourth order in space and first order in time; S_3= fourth order in space and second order in time; S_4= S_3+non linear correction.

cells	S_1	S_2	S_3	S_4
10	0.051417	0.075126	0.0027395	0.02549113
20	0.012956	0.019319	0.0001710	0.00286890
40	0.003247	0.004863	0.00001068	0.00034257
80	0.000811	0.001217	0.0000006676	0.00004171
160	0.000202	0.000304	0.00000004172	0.00000515
Order	≈ 2	≈ 2	≈ 4	≈ 3

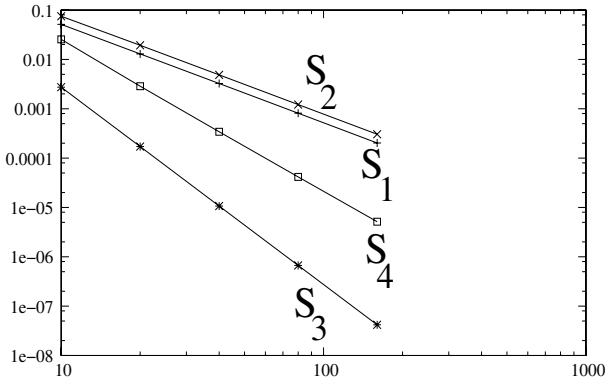

FIGURE 6. Error of the different schemes with respect to the number of cells (10, 20, 40, 80, 160). Log-scaled. Same as Table 1.

7. Open problems

We conclude with a short review of open problems, for which a positive answer would have great interest for practical computations.

Firstly it is natural to try to extend this work to stationary diffusion, at least because most non stationary diffusion problems are solved by a series of stationary implicit problems. What can be expected is to transform the condition (4.3) under the form of a contraction estimate for a non linear system of algebraic equations. Once such contraction argument is obtained, a fixed point argument proves the well posedness of the non linear stationary equation. It could be an improvement over more involved Brouwer type theorems.

The design of limited schemes at higher order, for example the initial linear schema at order 6 and the limited non linear scheme at order 5, is fully open.

Multidimensional estimates of convergence are of course much more involved since Fourier type techniques cannot work on general unstructured meshes. In the same vein, it will be necessary to develop new ideas or techniques to be able to get quantitative estimates of convergence for linear anisotropic diffusion problems discretized by means of non linear schemes. Nevertheless it can be postulated that the mechanism identified in this work, that is the non linear term may be control by the contractivity properties of the linear operator, could be an help in that direction.

Another set of interesting problems lies in the numerical analysis of non linear schemes for advection diffusion equations $\partial_t u + a\partial_x u = \nu\partial_{xx} u$. For the moment TVD theory adapted to the transport part of the equation is not compatible with the kind of arguments developed in this work for the diffusive part of the equation.

Appendix A. Proof of Proposition 4.5

We detail the formula (4.7) for the sum of norms

$$\sum_{l=0}^{\infty} \mu_l(\nu) = \underbrace{\sum_{l\le L} (1-4\nu)^l 2^k}_{=Z_1} + \underbrace{\sum_{L+1\le l} \left(\frac{2l}{2l+k}\right)^l \left(\frac{k}{\nu(2l+k)}\right)^{\frac{k}{2}}}_{=Z_2}. \tag{A.1}$$

One has the identity $\nu Z_1 = 2^{k-2}\left(1-(1-4\nu)^{L+1}\right)$. Since

$$L+1 = \left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1 \le \frac{k}{2}\left(\frac{1}{4\nu}-1\right)+1 \le \frac{k}{8\nu}-\frac{k-2}{2} \le \frac{k}{8\nu},$$

one has that $\nu Z_1 \leq 2^{k-2}\left(1-(1-4\nu)^{\frac{k}{8\nu}}\right)$. The second term can be analyzed as a staircase Riemann approximation of a convergent integral. Indeed let us define $x_l = \nu\frac{l}{k}$, so that νZ_2 can be rewritten as

$$\nu Z_2 = k\left(\frac{\nu}{k}\sum_{x_l \leq \frac{\nu}{k}(L+1)}\left(\frac{1}{1+\frac{\nu}{2x_l}}\right)^{\frac{kx_l}{\nu}}\left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}}\right). \tag{A.2}$$

Let $\gamma = \frac{k}{2} > 0$: the function $z \mapsto \left(\frac{1}{1+z}\right)^{\frac{\gamma}{z}}$ is increasing. Since $\frac{\nu}{2x_l} \leq \frac{\nu}{2\left(\frac{\nu}{k}(L+1)\right)}$, one obtains that

$$\left(\frac{1}{1+\frac{\nu}{2x_l}}\right)^{\frac{kx_l}{\nu}} \leq \left(\frac{1}{1+\frac{\nu}{2\frac{\nu}{k}(L+1)}}\right)^{\frac{k\frac{\nu}{k}(L+1)}{\nu}} = \lambda_k(\nu)$$

for all x_l. So all these terms can be upper estimated by $\lambda_k(\nu)$ and are now outside of the remaining sum $G = \frac{\nu}{k}\sum_{x_l \leq \frac{\nu}{k}(L+1)}\left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}}$. So the equality (A.2) becomes an inequality $\nu Z_2 \leq k\lambda_k(\nu)G$. The term G is a staircase approximation of the Riemann's integral of the decreasing function $x \mapsto \left(\frac{1}{2x+\nu}\right)^{\frac{k}{2}}$. It is convenient to isolate the first term, that is

$$G = \frac{\nu}{k}\left(\frac{1}{2x_{L+1}+\nu}\right)^{\frac{k}{2}} + \underbrace{\frac{\nu}{k}\sum_{\frac{\nu}{k}(L+2) \leq x_l}\left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}}}_{=H}. \tag{A.3}$$

Due to $\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1 \geq \frac{k}{8\nu}-\frac{k}{2}$, the first term can be bounded using

$$2x_{L+1}+\nu = \frac{2\nu}{k}\left(\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1\right)+\nu \geq \frac{2\nu}{k}\left(\frac{k}{8\nu}-\frac{k}{2}\right)+\nu \geq \frac{1}{4}$$

which yields $\frac{\nu}{k}\left(\frac{1}{2x_{L+1}+\nu}\right)^{\frac{k}{2}} \leq \frac{\nu 2^k}{k}$. The rest can be bounded using $\frac{\nu}{k}\left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}} \leq \int_{x_l-\frac{\nu}{k}}^{x_l}\frac{\mathrm{d}x}{(2x+\nu)^{\frac{k}{2}}}$. Therefore $H \leq \int_{\frac{\nu}{k}(L+1)}^{\infty}\frac{\mathrm{d}x}{(2x+\nu)^{\frac{k}{2}}}$. Since

$$\frac{\nu}{k}(L+1) = \frac{\nu}{k}\left(\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right]+1\right) \geq \frac{\nu}{k}\frac{k}{2}\left(\frac{1}{4\nu}-1\right) = \frac{1}{8}-\frac{\nu}{2}$$

one gets that

$$H \leq \int_{\frac{1}{8}-\frac{\nu}{2}}^{\infty}\frac{\mathrm{d}x}{(2x+\nu)^{\frac{k}{2}}} = \frac{4^{\frac{k}{2}-1}}{k-2} = \frac{2^{k-2}}{k-2}.$$

That is $\nu Z_2 \leq k\lambda_k(\nu)\left(\nu\frac{2^k}{k}+\frac{2^{k-1}}{k-2}\right)$. It ends the proof.

Appendix B. Constant Q for the scheme (3.14)

The control of the fourth order discrete difference used in the scheme (3.14) can be analyzed as in propositions (3.11) and (3.12). This scheme is also a modified (Le Potier's) scheme, but closer to the initial Le Potier's scheme than (3.16) or (3.20).

For convenience we start from the fourth order in space and second order in time linear scheme, plus the non linear correction

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + (1-6\nu)\nu \mathrm{Div}_j^3 + \nu|1-6\nu| \sum_{l\in\mathcal{V}_3(j)} a_{jl}^3 (u_l - u_j) \tag{B.1}$$

where all terms are defined in (3.14): see also below. The main point is to determine the continuity constant of the non linear correction $g = (g_j)$ where

$$g_j = \sum_{l\in\mathcal{V}_3(j)} a_{jl}^3 (u_l - u_j),$$

and

$$\begin{cases} a_{jl}^3 = a_{lj}^3 = \frac{\left|\mathrm{Div}_j^3\right|}{\Sigma_j^3} + \frac{\left|\mathrm{Div}_l^3\right|}{\Sigma_l^3}, \\ \Sigma_j^3 = \sum_{l\in\mathcal{V}^3(j)} |u_l - u_j|, \\ \mathcal{V}^3(j) = \{j+2, j+1, j-1, j-2\}. \end{cases}$$

One has that $g_j = h_j + k_{j+2} + l_{j+1} + m_{j-1} + n_{j-2}$ where the first term $h = (h_j)$ is

$$h_j = \frac{\sum_{l\in\mathcal{V}^3(j)} (u_l - u_j)}{\Sigma_j^3} \left|\mathrm{Div}_j^3\right| \Longrightarrow |h_j| \le \left|\mathrm{Div}_j^3\right| \Longrightarrow \|h\| \le \left\|\mathrm{Div}^3\right\|$$

and the other terms are

$$\begin{cases} k_{j+2} &= a_{j+2} \left|\mathrm{Div}_{j+2}^3\right|, \ a_{j+2} = \frac{u_{j+2}-u_j}{\Sigma_{j+2}^3}, \\ l_{j+1} &= b_{j+1} \left|\mathrm{Div}_{j+1}^3\right|, \ b_{j+1} = \frac{u_{j+1}-u_j}{\Sigma_{j+1}^3}, \\ m_{j-1} &= c_{j-1} \left|\mathrm{Div}_{j-1}^3\right|, \ c_{j-1} = \frac{u_{j-1}-u_j}{\Sigma_{j-1}^3}, \\ n_{j-2} &= d_{j-1} \left|\mathrm{Div}_{j-2}^3\right|, \ d_{j-2} = \frac{u_{j-2}-u_j}{\Sigma_{j-2}^3}. \end{cases}$$

One has the relation

$$|a_j| + |b_j| + |c_j| + |d_j| = \frac{|u_j - u_{j-2}| + |u_j - u_{j-1}| + |u_j - u_{j+1}| + |u_j - u_{j+2}|}{\Sigma_j^3} = 1 \quad \forall j.$$

Since

$$\begin{aligned} &\|k+l+m+n\|^2 \\ &\le \Delta x \sum_j \left(|a_{j+2}| \left|\mathrm{Div}_{j+2}^3\right| + |b_{j+1}| \left|\mathrm{Div}_{j+1}^3\right| + |c_{j-1}| \left|\mathrm{Div}_{j-1}^3\right| + |d_{j-2}| \left|\mathrm{Div}_{j-2}^3\right| \right)^2 \\ &\le 4\Delta x \sum_j |a_{j+2}|^2 \left|\mathrm{Div}_{j+2}^3\right|^2 + 4\Delta x \sum_j |b_{j+1}|^2 \left|\mathrm{Div}_{j+1}^3\right|^2 \\ &+4\Delta x \sum_j |c_{j-1}|^2 \left|\mathrm{Div}_{j-1}^3\right|^2 + 4\Delta x \sum_j |d_{j-2}|^2 \left|\mathrm{Div}_{j-2}^3\right|^2 \\ &\le 4\Delta x \sum_j \left(|a_j|^2 + |b_j|^2 + |c_j|^2 + |d_j|^2 \right) \left|\mathrm{Div}_j^3\right|^2 \le 4\|\mathrm{Div}^3\|^2. \end{aligned}$$

Therefore

$$\|g\| \le \|h\| + \|k+l+m+n\| \le (1+\sqrt{4})\|\mathrm{Div}^3\| = 3\|\mathrm{Div}^3\| \le \frac{1}{4}\|D^4 u\|$$

by definition of Div^3, see (3.12).

We now refer Remark 5.4 where it is proved that this numerical value $Q = \frac{|1-6\nu|}{4}$ is sufficiently small to be sure that the control of the fourth order difference holds. It yields the convergence at order 3 of this scheme. Notice however that the time step to achieve the maximum principle is *a priori* more stringent than for (3.20).

Acknowledgements. The author deeply wishes to thank Christophe Le Potier and Clément Cancès for valuable comments.

References

[1] I. Aavatsmark, T. Barkve, O. Boe, T. Mannseth, Discretization on unstructured grids for inhomogeneous, anisotropic media. Part I: Derivation of the methods. *SIAM J. Sci. Comput.* **19** (1998) 1700–1716.

[2] F. Boyer, F. Hubert, Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Numer. Anal.* **46** (2008) 3032–3070.

[3] F. Brezzi, K. Lipnikov, M. Shashkov, V. Simoncini, A new discretization methodology for diffusion problems on generalized polyhedral meshes. *Comput. Meth. Appl. Mech. Eng.* **196** (2007) 3682–3692.

[4] C. Buet, B. Després and E. Franck, Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes Numerische Mathematik, Online First (2012).

[5] C. Cancès, M. Cathala, C. Le Potier, Monotone coercive cell-centered finite volume schemes for anisotropic diffusion equations, online *Numer. Math.* (2013).

[6] G. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*. Springer-Verlag (2001)

[7] B. Després, Convergence of non-linear finite volume schemes for linear transport. In *Notes from the XIth Jacques-Louis Lions Hispano-French School on Numerical Simulation in Physics and Engineering*. Grupo Anal. Teor. Numer. Modelos Cienc. Exp. Univ. Cadiz (2004) 219–239.

[8] B. Després, Lax theorem and Finite Volume schemes. *Math. Comput.* **73** (2004) 1203–1234.

[9] J. Droniou, C. Le Potier, Construction and convergence study of local-maximum-principle preserving schemes for elliptic equations. *SIAM J. Numer. Anal.* **49** (2011) 459–490.

[10] L.C. Evans, Partial Differential Equations. Rhode Island: American Mathematical Society, Providence (1988).

[11] R. Eymard, T. Gallouët and R. Herbin, Finite Volume Methods, vol. 7 of *Handbook of Numerical Analysis*. Edited by P.G. Ciarlet and J.L. Lions. North Holland (2000) 713–1020.

[12] R. Eymard, T. Gallouët and R. Herbin, Discretisation of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. SUSHI: a scheme using stabilization and hybrid interfaces, *IMA J Numer Anal.* **30** (2010) 1009–1043.

[13] D. Gilbarg, N. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Springer, New York (1983).

[14] A. Genty and C. Le Potier, Maximum and minimum principles for radionuclide transport calculations in geological radioactive waste repository: comparison between a mixed hybrid finite element method and finite volume element discretizations. *Transp. Porous Media* **88** (2011) 65–85.

[15] E. Godlewski and P.-A. Raviart, Numerical approximation of hyperbolic systems of conservation laws, vol. 118 of *Applied Mathematical Sciences*. Springer (1996).

[16] R. Herbin, F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids, in: *5th International Symposium on Finite Volumes for Complex Applications*, edited by V.R. Eymard and J.M. Herard. Wiley (2008) 659–692.

[17] Hermeline F., A finite volume method for approximating 3D diffusion operators on general meshes. *J. Comput. Phys.* **228** (2009) 5763–5786.

[18] Kershaw D., Differencing of the diffusion equation in Lagrangian hydrodynamic codes. *J. Comput. Phys.* **39** (1981) 375–395.

[19] C. Le Potier, Correction non linéaire et principe du maximum pour la discrétisation d'opérateurs de diffusion avec des schémas volumes finis centrés sur les mailles. *C. R. Acad. Sci. Paris, Ser. I* **348** (2010) 691–695.

[20] C. Lepotier, private communication (2012).

[21] R.J. Leveque, Numerical Methods for Conservation Laws, *Lectures in Mathematics*. ETH-Zurich Birkhauser-Verlag, Basel (1990).

[22] K. Lipnikov, M. Shashkov, I. Yotov, Local flux mimetic finite difference methods. *Numer. Math.* **112** (2009) 115–152.

[23] K. Lipnikov and M. Shashkov, A framework for developing a mimetic tensor artificial viscosity for Lagrangian hydrocodes on arbitrary polygonal meshes. *J. Comput. Phys.* **229** (2010) 7911–7941.

[24] K. Lipnikov, G. Manzini and D. Svyatskiy, Analysis of the monotonicity conditions in the mimetic finite difference method for elliptic problems. *J. Comput. Phys.* **230** (2011) 2620–2642.

[25] P.L. Roe, Characteristic-based schemes for the Euler equations. *Ann. Rev. Fluid Mech.* **18** (1986) 337–365.

[26] Z. Sheng, J. Yue, G. Yuan, Monotone Finite volume schemes of non-equilibrium radiation diffusion equations of distorted meshes, *SIAM J. Sci. Comput.* **31** (2009) 2915–2934.

[27] Yu.I. Shokin, The method of differential approximation, Springer-Verlag (1983).

[28] P. Sweby, High-resolution schemes using flux limiters for hyperbolic conservation-laws. *SIAM J. Numer. Anal.* **21** (1984) 995–1011.