HAL open science

Non linear finite volume schemes for the heat equation in 1D.

Bruno Després

▶ To cite this version:

Bruno Després. Non linear finite volume schemes for the heat equation in 1D.. 2012. hal-00714781

HAL Id: hal-00714781
https://hal.science/hal-00714781

Submitted on 5 Jul 2012

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

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

NON LINEAR FINITE VOLUME SCHEMES FOR THE HEAT EQUATION IN 1D

BRUNO DESPRÉS

Abstract. We construct various explicit non linear finite volume schemes for the heat equation in dimension one. These schemes are inspired by the Le Potier's trick [CRAS Paris, I 348, 2010]. They preserve the maximum principle and admit a finite volume formulation. We provide a functional setting for the analysis of convergence of such methods. Finally we construct, analyze and test a new explicit non linear maximum preserving scheme: we prove third order convergence: it is optimal on numerical tests.

1. Introduction

Finite volume schemes are very convenient for complex applications [9, 18, 1, 3, 12, 2]. A very active field of research is nowadays the development of linear and non linear finite volume methods for the heat equation [2, 4, 10, 16, 18]. A major inspiration is the work of Le Potier [13, 8] who has 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 [4]. In our mind this is fundamentally related to the lack of an a priori estimate which controls the norm of some discrete derivatives of such schemes. In this paper we will show how to derive such an a priori estimate for some non linear schemes for the discretization of the model non stationary heat equation in dimension one

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

The long term ambition of this work is to establish, for the heat equation, the same kind of setting that was fruitful in the recent past for non linear schemes for the advection equation [20, 17, 11, 15]. Indeed 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 [11, 7, 15]. In this sense a natural question is to generalize this methodology to 1D non linear schemes for the heat equation, trying to establish a systematic method to obtain bounds on some discrete derivatives, and after that to prove convergence. We will show how to obtain these estimates using an approach proposed in [7] which is fundamentally Fourier based. In order 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

Date: ?
2000 *Mathematics Subject Classification.* 65M06, 65M12.
Key words and phrases. Finite volume schemes, heat equation, non linear correction.
The author was supported in part by UPMC Grant Onuplafu and by FR-FCM.

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, 6]. 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 (3.19) 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.

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. Some numerical results are used to confirm the theoretical analysis.

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 $\nu = \frac{\Delta t}{\Delta x^2}$. 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

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

for smooth functions. Therefore

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

A linear combination yields

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

A similar trick is used in [5]. 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 $\Delta x \sum_j (u_{j+2} - u_j)^2 \leq 4\Delta x \sum_j (u_{j+1} - u_j)^2$ from which deduce that

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

An elementary upper bound is

$$a(u,u) \leq \frac{16}{3\Delta x^2} \|u\|^2. \tag{2.9}$$

□

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

$$\frac{16}{3}\nu \leq 1. \tag{2.10}$$

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

$$\|\overline{u}\|^2 - \|u\|^2 \leq \Delta t\ a(\overline{u} - u, \overline{u} - u) - \|\overline{u} - u\|^2 \leq \left(\frac{16\nu}{3} - 1\right) \|\overline{u} - u\|^2. \tag{2.11}$$

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

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

from which we deduce that the scheme

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

is of order 2 in time and 4 in space. We refer to [19] page 43 where the same scheme is introduced mainly for theoretical purposes. The scheme can also be rewritten under the form

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

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

$$\left(\frac{\overline{u}-u}{\Delta t}, v\right) + b(u,v) = 0, \quad \forall v \in l_2 \tag{2.12}$$

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

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

We notice that

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

Therefore

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

under condition (2.5).

Lemma 2.3. *The scheme (2.12) is stable in l_2 under the CFL condition (2.10).*

Proof. Since $b(u,u) \leq a(u,u)$, the same proof as the one of lemma 2.2 holds. The final stability condition is the more restrictive one between (2.5) and (2.10). The proof is ended. □

3. Maximum principle and Le Potier's trick

The schemes (2.6)-(2.8) and (2.12) 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 [13, 8] 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. One sees that a_{jl} is a bounded quantity by construction.

If one compares with TVD schemes, this term is very close to some extended slope indicator. 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 [13, 8].

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

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

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

$$f^{\mathrm{cor}}_{j+\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{l \leq j+\frac{1}{2} \leq m} a_{ml}(u_m - u_l). \tag{3.6}$$

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

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

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

$$f^{\mathrm{cor}}_{j-\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{l \leq j-\frac{1}{2} \leq m} a_{ml}(u_m - u_l)$$

$$= \frac{\alpha}{\Delta x} \sum_{l \leq j-1} a_{jl}(u_j - u_l) + \frac{\alpha}{\Delta x} \sum_{l \leq j-1 \text{ and } j+1 \leq m} a_{ml}(u_m - u_l)$$

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

$$f^{\mathrm{cor}}_{j+\frac{1}{2}} - f^{\mathrm{cor}}_{j-\frac{1}{2}} = \frac{\alpha}{\Delta x} \sum_{j+1 \leq m} a_{mj}(u_m - u_j)$$

$$-\frac{\alpha}{\Delta x}\sum_{l\leq j-1} a_{jl}(u_j-u_l)=\frac{\alpha}{\Delta x}\sum_{l\in\mathbb{Z}} a_{jl}(u_l-u_j).$$

It ends the proof. □

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

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

Proof. We define the bilinear form

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

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

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

which corresponds to the non linear coefficients a_{jl}. The equivalent of the stability inequality (2.11) is

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

which proves the claim. □

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

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

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

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

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

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

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

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

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

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

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

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 3.6. 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 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 [14] 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

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

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

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu \mathrm{Div}_j^3 + \alpha\nu \sum_{l \in \mathcal{V}_3(j)} a_{jl}^3 (u_l - u_j) \tag{3.13}$$

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

$$\Sigma_j^3 = \sum_{l \in \mathcal{V}^3(j)} \left|u_l - u_j\right|. \tag{3.14}$$

3.5. **A third modified scheme.** Here we perform a more important modification. We start from the decomposition

$$\begin{aligned} & u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2} \\ & = (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) \end{aligned}$$

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

$$\begin{aligned} \overline{u}_j &= u_j + \nu \mathrm{Div}_j^2 + \nu \mathrm{Div}_j^3 \\ &+ \alpha\nu a_{j,j+1}^4 (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - \alpha\nu a_{j,j-1}^4 (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}). \end{aligned} \tag{3.15}$$

Proposition 3.7. *Set $\alpha = 1$. The scheme (3.15) 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 (3.15). Performing the same kind of algebra as before we rewrite the second part of the linear flux as follows

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

Plugging in (3.15) one obtains

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + \nu w (u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}) - \nu z (u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}) \tag{3.16}$$

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

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

and

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

By construction $\left|\mathrm{Div}_j^3\right| \leq \frac{1}{12}\Sigma_j^4$ for all j, so $w, z \leq \frac{1}{4}$. The expansion of formula (3.16) is

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

$$+\nu\left(1 - 3w - z\right)u_{j+1} + \nu\left(1 - w - 3z\right)u_{j-1} + \nu w u_{j+2} + \nu z u_{j-2}$$

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

3.6. **Second order in time.** The second order in time scheme (2.12) admits the explicit form

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j^2 + (1 - 6\nu)\nu \mathrm{Div}_j^3 \tag{3.18}$$

which is very close to (3.12). So it is easy to modify (3.15) which becomes

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

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

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

Proposition 3.8. *Assume $6\nu \leq 1$. The scheme (3.19) 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 (3.17). □

3.7. **General formulation.** Motivated by the previous examples, we will study the following family of modified schemes

$$\overline{u}_j = u_j + \nu \mathrm{Div}_j + \alpha\nu g_j \tag{3.20}$$

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

$$(Au)_j = -\mathrm{Div}_j, \quad \forall u \in l_2.$$

A quantity that plays an important role in the analysis developed in this work is the following.

Definition 3.9. Let $q \in \mathbb{N}^\star$ be an integer naturally defined by the scheme. Throughout this paper, the continuity constant of g with respect to $D^q u$ will be denoted by $Q > 0$, that is

$$\|g\| \leq Q \|D^q u\|.$$

For example let us consider the scheme (3.15) for which the source $g = (g_j) \in l_2$ is defined by

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

Proposition 3.10. *The source of the scheme (3.15) is such that* $\|g\| \leq \frac{1+\sqrt{2}}{12}\|D^4 u\|$, *that is* $Q = \frac{1+\sqrt{2}}{12}$ *and* $q = 4$.

Proof. We first notice that $\|A^3 u\| \leq \frac{1}{12}\|D^4 u\|$ where we have used natural notations compatible with (3.15). 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

$$\alpha_{j+1} = \frac{|u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}|}{\Sigma^4_{j+1}} \text{ and } \beta_{j-1} = \frac{|u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}|}{\Sigma^4_{j-1}}$$

satisfy

$$0 \leq \alpha_j, \beta_j \text{ and } \alpha_j + \beta_j = 1 \qquad \forall j.$$

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 3.11. *The source of the scheme (3.19) is such that*

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

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

Proof. Evident from previous proposition and definition (3.19). The $|1 - 6\nu|$ is because we use the second order in time scheme. □

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^q to the Duhamel formula (4.1), one gets the identity

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

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

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

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*

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

Then one has the stability estimate

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

Remark 4.2. The dissipativity of the scheme insures that $\lim_{l\to\infty} \| (I - \nu A)^l D^q\| = 0$ this property is been used in [7, 6] 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 $\|(I - \nu A)^l\| \leq 1$ for all $l \in \mathbb{N}$. Let us define $Z_N = \sup_{n\leq N} \|D^q u^n\|$ which satisfies the estimate

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

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

For convenience we define the generic function

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

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 3 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 3 is that the condition $F(\nu)$ 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

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

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

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

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

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

Let us set

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

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 all the rest of the paper.

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

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

and

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

Therefore

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

Next we study the function f_l. One has

$$\begin{aligned}\frac{d}{dy} f_l^\nu(y) &= -2l\nu y(1-\nu y^2)^{l-1} y^k + k(1-\nu y^2)^l y^{k-1} \\ &= (1-\nu y^2)^{l-1} y^{k-1} \left(-2l\nu y^2 + k - k\nu y^2\right).\end{aligned}$$

We define

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

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

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

and that $4 \le \frac{1}{\nu}$ 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 $y_l \ge 2$ and

$$\mu_l(\nu) = f_l^\nu(2) = (1-4\nu)^l 2^k.$$

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

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

that is

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

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 $E_1(\nu) \le h(\nu)$ 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

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

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

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

Indeed the first term admits the limit value $(1-4\nu)^{\frac{k}{8\nu}}$ tends to $e^{-\frac{k}{2}}$ 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 $\lambda_k(0^+) = e^{-\frac{k}{2}}$ by continuity: it yields (4.10). This numerical value can be checked in the numerical figure 2.

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

Proof. We obtain the formula for the sum of norms (4.7)

$$\sum_{l=0}^{\infty} \mu_l(\nu) = \underbrace{\sum_{l \le L} (1-4\nu)^l 2^k}_{=Z_1} + \underbrace{\sum_{L+1 \le l} \left(\frac{2l}{2l+k}\right)^l \left(\frac{k}{\nu(2l+k)}\right)^{\frac{k}{2}}}_{=Z_2}. \tag{4.11}$$

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

$$\nu Z_1 \le 2^{k-2}\left(1 - (1-4\nu)^{\frac{k}{8\nu}}\right).$$

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

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

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

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

So $\nu Z_2 \le k\lambda_k(\nu) G$ with

$$G = \frac{\nu}{k} \sum_{x_l \le \frac{\nu}{k}(L+1)} \left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}}$$

in which we recognize the 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

$$G = \frac{\nu}{k}\left(\frac{1}{2x_{L+1}+\nu}\right)^{\frac{k}{2}} + \underbrace{\frac{\nu}{k} \sum_{\frac{\nu}{k}(L+2) \le x_l} \left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}}}_{=H}. \tag{4.12}$$

The first term can be bounded using

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

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 $\frac{\nu}{k}\left(\frac{1}{2x_l+\nu}\right)^{\frac{k}{2}} \leq \int_{x_l-\frac{\nu}{k}}^{x_l} \frac{dx}{(2x+\nu)^{\frac{k}{2}}}$. Therefore $H \leq \int_{\frac{\nu}{k}(L+1)}^{\infty} \frac{dx}{(2x+\nu)^{\frac{k}{2}}}$. 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

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

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. □

4.2. **Fourth order operator.** Next we consider the operator

$$A_2 = -T+2I-T^{-1}+\frac{1}{12}\left(T^2-4T+6I-4T^{-1}+T^{-2}\right) \tag{4.13}$$

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

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

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

$$\frac{16}{3}\nu \leq 1. \tag{4.14}$$

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

$$\|(I-\nu A_2)^l D^k\| = \max_{0\leq y\leq 2} g_l(y). \tag{4.15}$$

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 (4.14). 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

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

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

$$E_2(\nu) = \sum_{l \leq M} \left(1 - \frac{16}{3}\nu\right)^l 2^k + \sum_{M+1 \leq l} \left(1 - \nu z_l^2 - \frac{4}{3}\nu z_l^4\right)^l z_l^k. \tag{4.16}$$

4.3. **Full fourth order operator.** Finally we consider the Fourier symbol of $(I-\nu A_3)^l D^k$ of the operator

$$A_3 = -T + 2I - T^{-1} + \frac{1-6\nu}{2}\left(T^2 - 4T + 6I - 4T^{-1} + T^{-2}\right) \tag{4.17}$$

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

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

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

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

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

$$E_3(\nu) \leq |1-6\nu| E_1(\nu). \tag{4.18}$$

4.4. **Application to the scheme (3.15).** 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 (3.19) (see also proposition 3.11).

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). A first simple and convenient estimate that can be derived from (4.9) is the following.

Proposition 4.9. *Assume $k=4$ and $\nu \leq 16$. 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. Indeed our hypotheses imply that $\frac{k}{8\nu} = \frac{1}{2\nu}$ on the one hand. So

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

One the other hand one has $\left[\frac{k}{2}\left(\frac{1}{4\nu}-1\right)\right] \geq \frac{k}{2}\left(\frac{1}{4\nu}-1\right)-1$: since the function λ_k is monotone decreasing, it yields

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

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

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

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

Proof. One first has that

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

Since

$$(1-y)\ln(1-y) = (1-y)\left(-y - \frac{1}{2}y^2 - \frac{1}{3}y^3 + \dots\right)$$

$$= -y + \left(1-\frac{1}{2}\right)y^2 + \left(\frac{1}{2}-\frac{2}{3}\right)y^3 + \dots \leq -y + y^2, \qquad y = 4\nu \leq 1,$$

then

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

Therefore

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

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

$$\leq 4 + 4e^{-2}e^{8\nu}(1+12\nu)$$

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

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

Proof. For such a simple function, we consider that a plot (with gnuplot) is enough. We refer to figure 1.

□

Lemma 4.12. *Assume $6\nu \leq 1$. The function F_3 associated to the scheme (3.15) is such that*

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

As a consequence the fourth order discrete derivative is bounded uniformly with respect to n and v: there exists $C>0$ such that

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

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

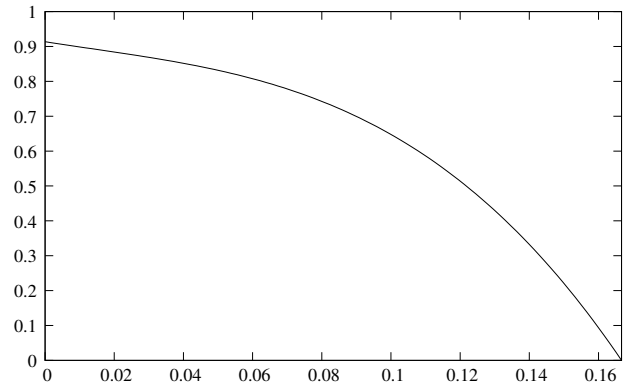


FIGURE 1. Plot of $\frac{1+\sqrt{2}}{12}(1-6\nu)q(\nu)$ in the range $\nu \in [0, \frac{1}{6}]$. The function is monotone decreasing.

Proof. By construction

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

Numerical application show that

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

Therefore $F_3 < 1$ over the range $\nu \in \left[0, \frac{1}{6}\right]$. See also the plot in figure 3. □

5. Convergence

Now we consider any scheme (4.1) such that an estimate on the norm of g^p can be obtained. Let us go back to the Duhamel formula (4.1) and consider the reminder

$$R = \nu \sum_{p=0}^{n-1} (I-\nu A)^{n-1-p} g^p, \qquad A = 1_1, A_2 \text{ or } A_3.$$

Due to the fact that corrections compatible with the Le Potier's technique may be rewritten under a finite volume form as stated in proposition 3.2, there exists $s^p \in l_2$ such that

$$g^p = Ds^p, \qquad \|s^p\| \leq C_2\|D^4u\|.$$

Therefore one has

$$\|R\| \leq C_2\left(\nu \sum_{p=0}^{n-1} \|(I-\nu A)^{n-1-p}D\|\right)\|D^4u_0\|$$
$$\leq C_2\left(\underbrace{\nu \sum_{p=0}^{n-1} \|(I-\nu A_1)^{n-1-p}D\|}_{=F_n}\right)\|D^4u_0\|.$$

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

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

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

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

Using (4.11) one gets

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

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 $H' \leq \int_{\frac{1}{8} - \frac{\nu}{2}}^{\nu \frac{T}{\Delta t}} \frac{dx}{(2x+\nu)^{\frac{1}{2}}} \quad (= I)$. 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, 6].

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 (3.19). 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. 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 3 is under threshold 1. So the schemes (3.15) also converges at order 3.

Remark 5.4. A case of interest is the scheme (3.13) 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 3 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 a proof of convergence at order three if one considers the use of the non linear correction (3.13) 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.

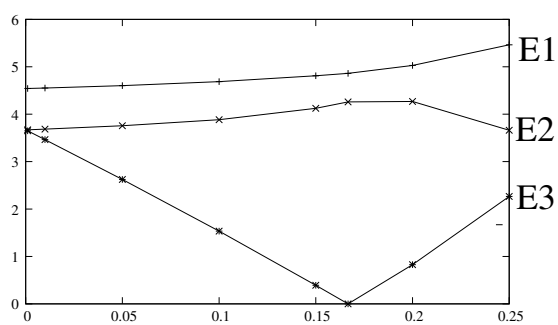


Figure 2. The constant $E_{1,2,3}$ for three different set of parameters.

6.1. **Functions E and F.** In figure 2 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 $\approx 4 + 4e^{-2} \approx 4.5413\ldots$. Next in figure 3 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.

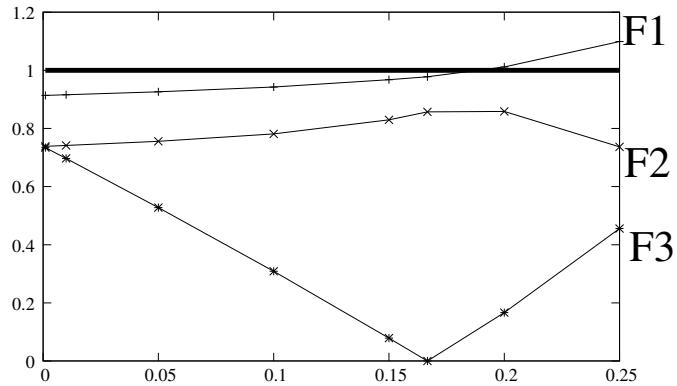
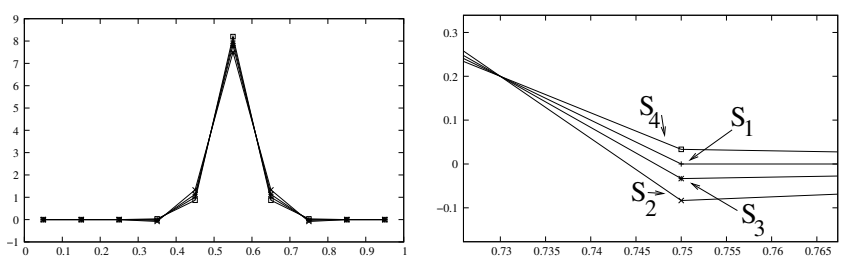


FIGURE 3. The functions $F_{1,2,3}$ for three different set of parameters. $F_1(\nu) = \frac{1+\sqrt{2}}{12} E_1(\nu)$ is for theoretical understanding. $F_2(\nu) = \frac{1+\sqrt{2}}{12} E_2(\nu)$ corresponds to the scheme (3.15). $F_3(\nu) = \frac{1+\sqrt{2}}{12} E_3(\nu)$ corresponds to the scheme (3.19). What is important is to be under the threshold 1 (in bold).

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 4 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.



Discrete solution at first iteration Zoom

FIGURE 4. Numerical solution calculated by the four different schemes: 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.

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

$$u(t) = e^{-4\pi^2 t} u_0.$$

We measure in table 1 and figure 5 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.

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

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.

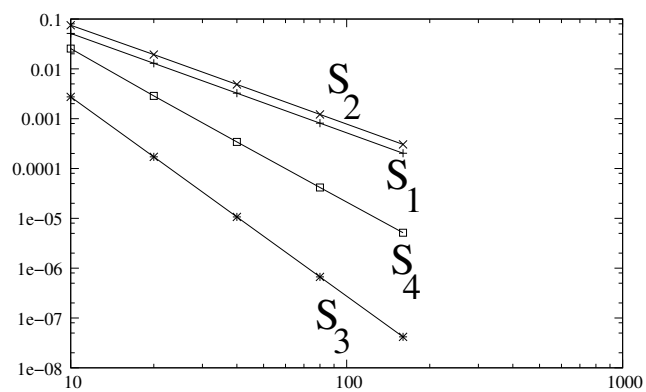


FIGURE 5. Error of the different schemes with respect to the number of cells (10, 20, 40, 80, 160). Log-scaled. Same as table 1.

APPENDIX A. CONSTANT Q FOR THE SCHEME (3.13)

The control of the fourth order discrete difference used in the scheme (3.13) can be analyzed as in propositions 3.10 and 3.11. This scheme is also a modified (Le Potier's) scheme, but closer to the initial Le Potier's scheme than (3.15) or (3.19).

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

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

where all terms are defined in (3.13): see also below. The main point is to determine the continuity constant of the non linear correction $g = (g_j)$ where

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

and

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

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

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

and the other terms are

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

One has the relation

$$\begin{array}{c} |a_j| + |b_j| + |c_j| + |d_j| \\ = \dfrac{|u_j - u_{j-2}| + |u_j - u_{j-1}| + |u_j - u_{j+1}| + |u_j - u_{j+2}|}{\Sigma^3_j} = 1 \quad \forall j. \end{array}$$

Since

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

Therefore

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

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 (3.19).

Acknowledgements The author deeply wishes to thank Christophe Le Potier and Clément Cances 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 (05) (September 1998) 17001716.

[2] F. Boyer, F. Hubert : Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities, SIAM Journal on Numerical Analysis, Vol. 46, N 6, pp. 3032–3070, (2008)

[3] Christophe Buet, Bruno Després and Emmanuel Franck, Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes Numerische Mathematik, Online First, 16 March 2012.

[4] C. Cances, M. Cathala, C. Le Potier, Monotone coercive cell-centered finite volume schemes for anisotropic diffusion equations, in preparation. Private communication.

[5] G. Cohen, Higher-Order Numerical Methods for Transient Wave Equations, Springer-Verlag, 2001

[6] 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, pages 219239. Grupo Anal. Teor. Numer. Modelos Cienc. Exp. Univ. Cadiz, 2004.

[7] B. Després, Lax theorem and Finite Volume schemes, Math. Comp. 73 (2004), 1203-1234.

[8] J. Droniou, C. Le Potier, Construction and convergence study of local-maximum-principle preserving schemes for elliptic equations, SIAM J. Numer. Anal. 49 (2) (2011) 459490

[9] R. Eymard, T. Gallouët and R. Herbin, Finite Volume MEthods, Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, North Holland 7, 713-1020, 2000.

[10] 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, Advance Access published on June 16, 2009

[11] Godlewski E. and Raviart P.-A., Numerical approximation of hyperbolic systems of conservation laws, Applied Mathematical Sciences, Volume 118, Springer, 1996.

[12] 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, June 0813, 2008.

[13] 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 (1112) (2010) 691695.

[14] C. Lepotier, private communication, 2012.

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

[16] K. Lipnikov, M. Shashkov, I. Yotov, Local flux mimetic finite difference methods, Numer. Math. 112 (2009) 115152.

[17] P.L. Roe, (1986), Characteristic-based schemes for the Euler equations, Ann. Rev. Fluid Mech. 18: 337365

[18] Z. Sheng, J. Yue, G. Yuan, Monotone Finite volume schemes of non-equilibrium radiation diffusion equations of distorted meshes, SIAM J. Sci. Comput. Vol. 31,No. 4, 2915-2934, 2009.

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

[20] P. Sweby, High-resolution schemes using flux limiters for hyperbolic conservation-laws, Siam J. Numer. Analysis, 21, 995-1011, 1984.

LJLL, UPMC, Paris, France

Current address: Laboratoire Jacques-Louis Lions Université Pierre et Marie Curie Boîte courrier 187 75252 Paris Cedex 05 France

E-mail address: `despres@ann.jussieu.fr`