QUADRATIC STABILITY OF FLUX LIMITERS

Bruno Després

QUADRATIC STABILITY OF FLUX LIMITERS *

Bruno Després[1]

Abstract. We propose a novel approach to study the quadratic stability of 2D flux limiters for non expansive transport equations. The theory is developed for the constant coefficient case on a cartesian grid. The convergence of the fully discrete nonlinear scheme is established in 2D with a rate not less than $O(\Delta x^{\frac{1}{2}})$ in quadratic norm. It is a way to bypass the Goodman-Leveque obstruction Theorem. A new nonlinear scheme with corner correction is proposed. The scheme is formally second-order accurate away from characteristics points, satisfies the maximum principle and is proved to be convergent in quadratic norm. It is tested on simple numerical problems.

1991 Mathematics Subject Classification. 65M08, 65M12 .

The dates will be set by the publisher.

1. Introduction

A famous Godunov Theorem [18] states that a linear scheme for advection is only first-order accurate, except in trivial cases. Subsequently, Harten proposed in [22] a constructive notion for Total Variation Diminishing (TVD) numerical schemes which allows the construction of formally second-order accurate nonlinear numerical schemes for advection in space dimension $d = 1$ (1D). The theory of nonlinear TVD schemes is nowadays well established, see [17, 31] and references therein. However, in contrast with the success of the TVD theory in 1D, Goodman and Leveque proved in [19] the obstruction Theorem summarized as follows: *A TVD numerical scheme with compact stencil for the discretization of conservation laws in space dimension $d \geq 2$ is at most first-order accurate, except in trivial cases.* The proof [19] is by contradiction: one starts from a scheme which satisfies the TVD property and which has a compact stencil in space dimension $d = 2$ (2D); then one proves that this scheme is L^1 contracting for one dimensional profiles; it turns into the fact that the scheme is just first-order accurate by another important result [23]. It is a non constructive proof which does not propose any way to go beyond this limitation (nevertheless we immediately remark that L^2 stability of linear schemes for 1D advection can already be achieved at any order [10], contrary to linear stability in L^1 norm). To be precise, we follow the literature [19, 22] for which formal second-order and high-order accuracy means accuracy in the sense of local Taylor expansion for exact smooth solutions.

Since then, the development of numerical solvers has pursued its route [2, 17, 31] in space dimension $d \geq 2$, but without the compactness offered by TVD inequalities. This regrettable situation where there is a divorce between the development of high order numerical methods and the theory of numerical convergence is a direct corollary of the Goodman-Leveque Theorem. Quoting [19]: *While it is not logically necessary*

Keywords and phrases: Flux limiters, quadratic stability, quadratic convergence, Goodman-Leveque Theorem

* *The author thanks ANR-MUFFIN (ANR-19-CE46-0004) and CEA for support.*

[1] Laboratoire Jacques-Louis Lions, Sorbonne Université, 4 place Jussieu, 75005 Paris, France and Institut Universitaire de France, despres@ann.jussieu.fr

for a scheme to be TVD in order to be total variation stable, or to converge, we know of no method for computing weak solutions that is not TVD and yet can be shown to converge. Optimum positive schemes for transport by Roe and Sidilkover [28] are linear and first order so cannot constitute a solution. The Barth-Jespersen technique [1] on unstructured meshes guarantees the maximum principle, but the scheme is ultimately first-order accurate. One finds in Barth-Ohleberger [2] a very nice and comprehensive review on the topic of limiter techniques and on the consequences of the Goodman-Leveque obstruction theorem on the development of numerical methods, where the emphasis moves on the preservation of the maximum principle instead of TVD estimates. Other progresses have been made, in particular by systematically using bound-preserving or invariant-domain-preserving numerical methods. Bound-preserving numerical linear and nonlinear methods are extensively studied and developed in the works of Shu [6, 27, 29] and therein. Refer to [32] for the principles of bound-preserving Finite Volume methods, but even in the latest works, the convergence of fully discrete in time numerical schemes is not established. Domain invariant techniques and positivity-preserving techniques are also studied and developed in the contributions of Guermond-Popov et al [20, 21], still without proof of convergence for fully discrete schemes (see for example [21][Th. 4.8] for quite a technical statement on the stability of a family of bound preserving method for transport). In summary the Goodman-Leveque obstruction Theorem is still in the background, even in the the latest development.

The purpose of this work is to show that new quadratic stability estimates for nonlinear flux limiter techniques offer a strategy to go beyond the Goodman-Leveque obstruction Theorem, at least for transport equations on a cartesian grid. With quadratic stability, a control of nonlinear fully discrete numerical schemes is achieved in 2D for a large class of multidimensional flux limiters, allowing formally second-order accurate numerical schemes for which one can proves numerical convergence.

The problems embraced in our analysis must have some non expansive properties in quadratic norm, because the goal is to reproduce these properties at the discrete level. Such a general non expansive equation in general dimension writes

$$\partial_t u(\mathbf{x},t) + \nabla\cdot(\mathbf{a}(\mathbf{x})u(\mathbf{x},t)) = s(\mathbf{x}) \tag{1}$$

where the given velocity field is $\mathbf{a} \in W^{1,\infty}(\mathbb{R}^d)$ and the source is $s \in L^\infty(\mathbb{R}^d)$. For convection dominated flows in applied sciences, the calculation of accurate numerical solutions to this equation is of fundamental importance [16,17,26,31]. If the velocity field is divergent free, that is $\nabla\cdot\mathbf{a} = 0$, then the equation is non expansive in all $L^p(\mathbb{R}^d)$ norms and in particular in quadratic norm (this is $L^2(\mathbb{R}^d)$ norm). If the divergence is bounded $\nabla\cdot\mathbf{a} \in L^\infty(\mathbb{R}^d)$ the expansion is bounded by classical estimates. An equation similar to (1) can be obtained via the linearization of a truly nonlinear equation $\partial_t u_{\rm tot} + \nabla\cdot(\mathbf{b}(\mathbf{x})f(u_{\rm tot})) = 0$ where the final equation for the perturbation defined for exemple by $u_{\rm tot} = u_{\rm ref} + \varepsilon u + O(\varepsilon^2)$ can be written as $\partial_t u(\mathbf{x},t) + \nabla\cdot(\mathbf{a}(\mathbf{x},t)u(\mathbf{x},t)) = 0$: the velocity field $\mathbf{a}(\mathbf{x},t) = \mathbf{b}(\mathbf{x})f'(u_{\rm ref}(\mathbf{x},t))$ is a function of space and time; non expansive estimates can be obtained depending on the space-time regularity of the velocity field. Provided the regularity of the velocity field is sufficient, it is possible to approach weak solutions with smooth solutions (non smooth velocity fields low regularity such as in [5] are not considered in this work). For the simplicity of the theoretical developments, we will assume that the initial data has compact support with three bounded derivatives

$$u(0) = u_0 \in W_0^3(\mathbb{R}^d). \tag{2}$$

To concentrate on the main features of our approach and to minimize the amount of notations in 2D, the equation is simplified further. The model problem is 2D advection (that is transport at constant velocity) discretized on a cartesian grid

$$\partial_t u + p\partial_x u + q\partial_y u = 0, \qquad (x,y,t) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^+. \tag{3}$$

Additionally the advection velocity field is normalized

$$p \geq 0, \quad q \geq 0 \quad \text{and} \quad p+q = 1. \tag{4}$$

One dimensional configurations aligned with the discretization grid correspond to the limit cases $(p,q)=(1,0)$ or $(p,q)=(0,1)$. The equation (3) is already relevant for discretization of kinetic equations, for which $u=f(\mathbf{x},\mathbf{v},t)\geq 0$ represents the density of particles in function of the space variable, the time variable and the velocity $\mathbf{v}\in\mathbb{R}^d$ variable. The correspondance is $\mathbf{v}=\lambda(p,q)$ where $\lambda\geq 0$ is a rescaling factor to account for the normalization $p+q=1$. Considering the fierce activity nowadays in the field of numerical plasma physics [3, 7, 13, 15, 25], it is possible that the original methods for the numerical analysis of the advection equation (3) and perhaps the original scheme constructed at the end of this work for the purposes of numerical illustrations can already find immediate applications in this field. We think also of the applications of flux limiters with quadratic stability for the remap stage of Lagrange+remap schemes [8] applied to the numerical discretization of compressible Euler equations (a different technique is [24]). Radiation transport equations [21] could also benefit from the techniques of this work. Extension to non constant coefficients are evoked in the last Section.

The principle of the new quadratic estimates is exposed in 1D in Section 2 and the application to the convergence of the fully discrete schemes is explained in Section 3. The extension to the 2D case is performed in Section 4 where the quadratic stability and the convergence are proved in Theorem 11 which is the main theoretical contribution of this work. A new scheme is constructed in Section 5 with a notion of corner interaction which is natural in the context of this work. This scheme is formally second-order accurate except at characteristics points and satisfies the maximal principle. It converges in quadratic norm by virtue of the theoretical estimates, at a rate not less than $O(\Delta x^{\frac{1}{2}})$. Numerical results obtained with the new schemes are shown in Section 6, in particular it is clear that the theoretical estimate of convergence of Section 4 is largely suboptimal. It is also clear that the new non linear scheme is more performant than the linear Lax-Wendroff for truly weak solutions with BV regularity. Some natural extensions are evoked in last Section 7.

2. Quadratic estimates in 1D

Quadratic stability of flux limiters is easier to explain 1D than in 2D and the result of the analysis can be compared with the classical 1D theory of flux limiters [16, 17, 22, 31]. A generic scheme for the advection equation $\partial_t u+\partial_x u=0$ writes

$$\frac{\overline{U}_j-U_j}{\Delta t}+\frac{U_{j+\frac{1}{2}}-U_{j-\frac{1}{2}}}{\Delta x}=0,\qquad j\in\mathbb{Z},\tag{5}$$

with

$$U_{j+\frac{1}{2}}=U_j+\frac{1-\nu}{2}\varphi_{j+\frac{1}{2}}(U_{j+1}-U_j)\text{ for all }j\in\mathbb{Z}.\tag{6}$$

In these notations, U_j stand for U_j^n which is the numerical solution at time step $t_n=n\Delta t$ and $\overline{U}_j$ stands for U_j^{n+1} which is the numerical solution at time step $t_{n+1}=(n+1)\Delta t$. An explicit formulation is

$$\overline{U}_j=(1-\nu)U_j+\nu U_{j-1}-\frac{\nu(1-\nu)}{2}\left(\varphi_{j+\frac{1}{2}}(U_{j+1}-U_j)-\varphi_{j-\frac{1}{2}}(U_j-U_{j-1})\right).\tag{7}$$

The scalar $\varphi_{j+\frac{1}{2}}$ is the flux limiter between two consecutive cells (or mesh points). It is well known [26,31] that if $\varphi_{j+\frac{1}{2}}\equiv 0$ for all j then one obtains the upwind scheme, if $\varphi_{j+\frac{1}{2}}\equiv 1$ for all j then one obtains the Lax-Wendroff scheme. The theory of TVD flux limitation has been developed in [22,30] and is exposed in [16, 17, 26, 31]. For example the minmod limiter is defined by

$$\varphi_{j+\frac{1}{2}}=\text{minmod}\left(1,r_{j+\frac{1}{2}}\right)\text{ where }r_{j+\frac{1}{2}}=\frac{U_j-U_{j-1}}{U_{j+1}-U_j}.\tag{8}$$

The minmod function is defined by $\text{minmod}(a,b)=0$ for $ab\leq 0$ and $\text{minmod}(a,b)=\text{sign}(a)\min(|a|,|b|)$ for $ab\geq 0$.

Contrary to the classical references [16,17,26,31] where the analysis is l^1-based or TVD-based, our analysis is based on the quadratic discrete Lebesgue space

$$l^2 = \left\{ U = (U_j)_{j\in\mathbb{Z}} \text{ such that } \sum_{j\in\mathbb{Z}} |U_j|^2 < \infty \right\}.$$

In the next preliminary result, the l^2 norm of the numerical solution at the end of the time step is compared with the l^2 norm of the numerical solution at the beginning of the time step. We note $\Delta = \left(\Delta_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$ with $\Delta_{j+\frac{1}{2}} = U_{j+1} - U_j$.

Lemma 1. *Let $U \in l^2$. Then one has $\sum_{j\in\mathbb{Z}} |\overline{U}_j|^2 = \sum_{j\in\mathbb{Z}} |U_j|^2 - Q_\varphi(\Delta)$ where $Q_\varphi(\Delta)$ is the quadratic form*

$$\begin{array}{rcll} Q_\varphi(\Delta) & = & \nu(1-\nu)\sum_{j\in\mathbb{Z}} |\Delta_{j-\frac{1}{2}}|^2 & -\nu(1-\nu)\sum_{j\in\mathbb{Z}} \varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}\left((1-\nu)\Delta_{j+\frac{1}{2}} + \nu\Delta_{j-\frac{1}{2}}\right) \\ & & & -\frac{\nu^2(1-\nu)^2}{4}\sum_{j\in\mathbb{Z}} \left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2. \end{array} \tag{9}$$

The parameters of the quadratic form are noted $\varphi = \left(\varphi_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$.

Proof. The scheme (5-6) is written explicitly as

$$\overline{U}_j = U_j^{\text{up}} - \frac{\nu(1-\nu)}{2}\left(\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right)$$

where the upwind scheme is $U_j^{\text{up}} = (1-\nu)U_j + \nu U_{j-1}$ and the difference is $\Delta_{j+\frac{1}{2}} = U_{j+1} - U_j$ for all j. Denoting $R_j = |\overline{U}_j|^2 - |U_j^{\text{up}}|^2$, one can write

$$\begin{array}{rcl} |\overline{U}_j|^2 - |U_j|^2 & = & |U_j^{\text{up}}|^2 - |U_j|^2 + R_j \\ & = & -\nu(1-\nu)|U_j - U_{j-1}|^2 - \nu|U_j|^2 + \nu|U_{j-1}|^2 + R_j \\ & = & -\nu(1-\nu)|\Delta_{j-\frac{1}{2}}|^2 + \underbrace{\left(-\nu|U_j|^2 + \nu|U_{j-1}|^2\right)}_{\text{a difference}} + R_j \end{array} \tag{10}$$

where

$$\begin{array}{rcl} R_j & = & |\overline{U}_j|^2 - |U_j^{\text{up}}|^2 = 2(\overline{U}_j - U_j^{\text{up}})U_j^{\text{up}} + |\overline{U}_j - U_j^{\text{up}}|^2 \\ & = & -\nu(1-\nu)\left(\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right)U_j^{\text{up}} + \frac{\nu^2(1-\nu)^2}{4}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2 \\ & = & -\nu(1-\nu)\left(\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}U_{j+1}^{\text{up}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}U_j^{\text{up}}\right) \\ & & +\nu(1-\nu)\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}(U_{j+1}^{\text{up}} - U_j^{\text{up}}) + \frac{\nu^2(1-\nu)^2}{4}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2 \\ & = & \underbrace{-\nu(1-\nu)\left(\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}U_{j+1}^{\text{up}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}U_j^{\text{up}}\right)}_{\text{another difference}} \\ & & +\nu(1-\nu)\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}\left((1-\nu)\Delta_{j+\frac{1}{2}} + \nu\Delta_{j-\frac{1}{2}}\right) + \frac{\nu^2(1-\nu)^2}{4}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2. \end{array} \tag{11}$$

With these expressions, one can calculates $\sum_{j\in\mathbb{Z}}\left(|\overline{U}_j|^2 - |U_j|^2\right)$. The differences in (10-11) become telescopic by summation, so they vanish and the claim is obtained. □

The upwind scheme corresponds to $\varphi_{j+\frac{1}{2}} \equiv 0$: in this case the quadratic form $Q_{\text{up}} = Q_{\mathbf{0}}$ from (9) is non positive for all Δ provided $0 < \nu \leq 1$. One recovers without surprise the quadratic stability of the upwind scheme

under CFL condition (26). The Lax-Wendroff scheme corresponds to $\varphi_{j+\frac{1}{2}} \equiv 1$. One obtains he quadratic form $Q_{\rm lw} = Q_{\mathbf{1}}$ where

$$\begin{array}{rcl} Q_{\mathbf{1}}(\Delta) & = & \nu(1-\nu)\sum_{j\in\mathbb{Z}}\left(|\Delta_{j-\frac{1}{2}}|^2 - \Delta_{j+\frac{1}{2}}((1-\nu)\Delta_{j+\frac{1}{2}} + \nu\Delta_{j+\frac{1}{2}})\right) - \frac{\nu^2(1-\nu)^2}{4}\sum_{j\in\mathbb{Z}}\left|\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}\right|^2 \\ & = & \frac{\nu^2(1-\nu)}{2}\sum_{j\in\mathbb{Z}}\left|\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}\right|^2 - \frac{\nu^2(1-\nu)^2}{4}\sum_{j\in\mathbb{Z}}\left|\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}\right|^2 \\ & = & \frac{\nu^2(1-\nu^2)}{4}\sum_{j\in\mathbb{Z}}\left|\Delta_{j+\frac{1}{2}}\right|^2. \end{array}$$

One recovers, also without surprise, the quadratic stability of the Lax-Wendroff scheme under CFL condition (26). The natural question is to determine if there exists a general condition on the limiters $\varphi = \left(\varphi_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$ such that $Q_\varphi(\Delta) \leq 0$ for all $\Delta \in l^2$. A trivial answer is interpolation of $\varphi_{j+\frac{1}{2}}$ between 0 and 1. If the interpolation is uniform with respect to j, that is $\varphi_{j+\frac{1}{2}} = \psi \in [0,1]$ for all $j \in \mathbb{Z}$, then the result is evident. The key fact below is that the interpolation can be taken for $\varphi_{j+\frac{1}{2}}$ independently to $\varphi_{k+\frac{1}{2}}$ for all $j \neq k$.

The main technical difficulty concerns the quadratic form

$$H_\varphi(\Delta) = \sum_{j\in\mathbb{Z}}|\Delta_{j-\frac{1}{2}}|^2 - \sum_{j\in\mathbb{Z}}\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}\left((1-\nu)\Delta_{j+\frac{1}{2}} + \nu\Delta_{j-\frac{1}{2}}\right) - \frac{\nu}{4}\sum_{j\in\mathbb{Z}}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2. \quad (12)$$

This quadratic form is a part of $Q_\varphi(\Delta)$ since one has the formula

$$Q_\varphi(\Delta) = \nu(1-\nu)\left(H_\varphi(\Delta) + \frac{\nu^2}{4}\sum_{j\in\mathbb{Z}}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2\right). \quad (13)$$

Proposition 2. *Assume the CFL condition (26) and assume $0 \leq \varphi_{j+\frac{1}{2}} \leq 1$ for all $j \in \mathbb{Z}$. Then $H_\varphi(\Delta) \geq 0$ for all $\Delta \in l^2$.*

Proof. The condition $\Delta \in l^2$ guarantees the convergence of the sum in (9), so $H_\varphi(\Delta)$ is finite. The proof uses a rearrangement of (9), which is valid for $\Delta \in l^2$. Since $H_\varphi(\nu)$ is linear with respect to ν, we note $H_\varphi(\nu) = \alpha + \nu\beta$. One has $\alpha = \sum_{j\in\mathbb{Z}}(1-\varphi_{j+\frac{1}{2}})|\Delta_{j+\frac{1}{2}}|^2 \geq 0$ and

$$\begin{array}{rcl} \beta & = & \sum_{j\in\mathbb{Z}}\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}}\left(\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}\right) - \frac{1}{4}\sum_{j\in\mathbb{Z}}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - \varphi_{j-\frac{1}{2}}\Delta_{j-\frac{1}{2}}\right|^2 \\ & = & \sum_{j\in\mathbb{Z}}\left(\varphi_{j+\frac{1}{2}} - \frac{1}{2}\varphi_{j+\frac{1}{2}}^2\right)\Delta_{j+\frac{1}{2}}^2 + \sum_{j\in\mathbb{Z}}\left(-\varphi_{j+\frac{1}{2}} + \frac{1}{2}\varphi_{j+\frac{1}{2}}\varphi_{j-\frac{1}{2}}\right)\Delta_{j+\frac{1}{2}}\Delta_{j-\frac{1}{2}}. \end{array}$$

Consider

$$\gamma = \frac{1}{4}\sum_{j\in\mathbb{Z}}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - (2-\varphi_{j-\frac{1}{2}})\Delta_{j-\frac{1}{2}}\right|^2 - \sum_{j\in\mathbb{Z}}\left|(1-\varphi_{j+\frac{1}{2}})\Delta_{j+\frac{1}{2}}\right|^2.$$

An expansion shows that

$$\begin{array}{rcl} \gamma & = & \sum_{j\in\mathbb{Z}}\left(\frac{1}{4}\varphi_{j+\frac{1}{2}}^2 + \frac{1}{4}(2-\varphi_{j+\frac{1}{2}})^2 - (1-\varphi_{j+\frac{1}{2}})^2\right)\Delta_{j+\frac{1}{2}}^2 - \frac{1}{2}\sum_{j\in\mathbb{Z}}\varphi_{j+\frac{1}{2}}(2-\varphi_{j-\frac{1}{2}})\Delta_{j+\frac{1}{2}}\Delta_{j-\frac{1}{2}} \\ & = & \sum_{j\in\mathbb{Z}}\left(\varphi_{j+\frac{1}{2}} - \frac{1}{2}\varphi_{j+\frac{1}{2}}^2\right)\Delta_{j+\frac{1}{2}}^2 + \sum_{j\in\mathbb{Z}}\left(-\varphi_{j+\frac{1}{2}} + \frac{1}{2}\varphi_{j+\frac{1}{2}}\varphi_{j-\frac{1}{2}}\right)\Delta_{j+\frac{1}{2}}\Delta_{j-\frac{1}{2}}. \end{array}$$

So $\beta = \gamma$. Therefore

$$H_\varphi(\Delta) = \alpha + \nu\gamma = \sum_{j\in\mathbb{Z}}(1-\varphi_{j+\frac{1}{2}})\left(1 - \nu(1-\varphi_{j+\frac{1}{2}})\right)\left|\Delta_{j+\frac{1}{2}}\right|^2 + \frac{\nu}{4}\sum_{j\in\mathbb{Z}}\left|\varphi_{j+\frac{1}{2}}\Delta_{j+\frac{1}{2}} - (2-\varphi_{j-\frac{1}{2}})\Delta_{j-\frac{1}{2}}\right|^2. \quad (14)$$

The conditions of the claim insure $H_\varphi(\Delta) \geq 0$. □

Theorem 3. *Assume the conditions of Proposition 2. Then the scheme is stable in quadratic norm:* $\sum_{j\in\mathbb{Z}} |\overline{U_j}|^2 \leq \sum_{j\in\mathbb{Z}} |U_j|^2$.

Proof. The identity (13) yields the claim. □

Remark 1 (Control of discrete gradients)**.** *It is worthwhile for further developments to notice that (14-13) combined with a Cauchy-Schwarz inequality yield the inequality*

$$\sum_{j\in\mathbb{Z}} \left((1-\varphi_{j+\frac{1}{2}})|\Delta_{j+\frac{1}{2}}|^2 + |\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}|^2 \right) \leq C(\nu) Q_\varphi(\Delta) \tag{15}$$

for some constant $C(\nu) > 0$ *which depends of* $\nu \in (0,1)$*. This estimate shows a quadratic control of the second discrete gradient* $\Delta_{j+\frac{1}{2}} - \Delta_{j-\frac{1}{2}}$*. It also shows a quadratic control of the first gradient, multiplied by* $1-\varphi_{j+\frac{1}{2}}$*.*

3. Convergence in 1D

The convenient space for the numerical analysis is the l^2 space equipped with a quadratic norm weighted with the mesh size

$$V_{\Delta x} = \left\{ U = (U_j) \in l^2 \text{ with norm } \|U\|_{\Delta x}^2 = \Delta x \sum_{j\in\mathbb{Z}} |U_j|^2 < \infty \right\}.$$

The associated scalar product is denoted $(U,V)_{\Delta x} = \Delta x \sum_{j\in\mathbb{Z}} U_j V_j$ for $U, V \in V_{\Delta x}$.

The strategy of the proof of convergence relies on two ideas. The first idea corresponds to the semi-discrete limit regime which correspond to $O(\nu)$ terms in $Q_\varphi(\Delta)$: involved manipulations are similar to the ones in [6,27, 29,32]. The second idea is fully discrete and is necessary to treat the term which come from the discretization in time: it relies on sharp estimates which are new with respect to the literature.

3.1. The semi-discrete scheme

The semi-discrete scheme is continuous in time. It is obtained by letting $\Delta t \to 0$ or $\nu \to 0$ in (5-6). It writes

$$U_j'(t) + \frac{U_j(t) - U_{j-1}(t)}{\Delta x} + \frac{\Delta x}{2} \frac{\varphi_{j+\frac{1}{2}}(t)(U_{j+1}(t) - U_j(t)) - \varphi_{j-\frac{1}{2}}(t)(U_j(t) - U_{j-1}(t))}{\Delta x^2} = 0.$$

We remind the reader that, for simplicity, we consider a smooth initial data (2). Then the solution u is smooth as well. The interpolation of the smooth function $u \in W^3(\mathbb{R}\times\mathbb{R}^+)$ (with compact support in space) is denoted as $V_j(t) = u(j\Delta x, t)$. One defines the truncation error $r_j(t)$

$$r_j(t) = V_j'(t) + \frac{V_j(t) - V_{j-1}(t)}{\Delta x} + \frac{\Delta x}{2} \frac{\varphi_{j+\frac{1}{2}}(t)(V_{j+1}(t) - V_j(t)) - \varphi_{j-\frac{1}{2}}(t)(V_j(t) - V_{j-1}(t))}{\Delta x^2}.$$

It is easy to verify that

$$r_j(t) = O(\Delta x^2) + \frac{\Delta x}{2} \frac{(1-\varphi_{j+\frac{1}{2}}(t))(V_{j+1}(t) - V_j(t)) - (1-\varphi_{j-\frac{1}{2}}(t))(V_j(t) - V_{j-1}(t))}{\Delta x^2}. \tag{16}$$

The $O(\Delta x^2)$ is uniform with respect to the index j and the time t. Since u is compact in space, then only a finite number of $r_j(t)$ can be non zero at a given time t. The number of non zero terms is $O(\Delta x^{-1})$ uniformly with respect to the time t.

Let us define the numerical error $E_j(t) = V_j(t) - U_j(t)$ which vanishes at initial time $E_j(0) = 0$. The error satisfies

$$E_j'(t) + \frac{E_j(t) - E_{j-1}(t)}{\Delta x} + \Delta x \frac{\varphi_{j+\frac{1}{2}}(t)(E_{j+1}(t) - E_j(t)) - \varphi_{j-\frac{1}{2}}(t)(E_j(t) - E_{j-1}(t))}{2\Delta x^2} = r_j(t). \tag{17}$$

Theorem 4. *Let $T > 0$. Assume $0 \leq \varphi_{j+\frac{1}{2}}(t) \leq 1$ for all $j \in \mathbb{Z}$ and $0 \leq t \leq T$. Then there exists a constant $C > 0$ such that $\|E(t)\|_{\Delta x} \leq C\Delta x^{\frac{1}{2}}$.*

Proof. Multiply (17) by $E_j(t)$ and sum over all j. One obtains

$$\frac{d}{dt}\|E(t)\|_{\Delta x}^2 + \frac{\Delta x^2}{2}\sum_{j\in\mathbb{Z}}(1 - \varphi_{j-\frac{1}{2}}(t))\left|\frac{E_j(t) - E_{j-1}(t)}{\Delta x}\right|^2 = (E(t), r(t))_{\Delta x}$$

where the residual can be estimated with (16). One gets

$$\begin{aligned}
&\tfrac{d}{dt}\|E(t)\|_{\Delta x}^2 + \tfrac{\Delta x^2}{2}\textstyle\sum_{j\in\mathbb{Z}}(1 - \varphi_{j-\frac{1}{2}}(t))\left|\frac{E_j(t) - E_{j-1}(t)}{\Delta x}\right|^2 \\
&\quad \leq C\Delta x^2\|E(t)\|_{\Delta x} - \tfrac{\Delta x^2}{2}\textstyle\sum_{j\in\mathbb{Z}}(1 - \varphi_{j-\frac{1}{2}}(t))\frac{E_j(t) - E_{j-1}(t)}{\Delta x}\frac{V_j(t) - V_{j-1}(t)}{\Delta x} \\
&\quad \leq C\Delta x^2\|E(t)\|_{\Delta x} + \tfrac{\Delta x^2}{4}\textstyle\sum_{j\in\mathbb{Z}}(1 - \varphi_{j-\frac{1}{2}}(t))\left|\frac{E_j(t) - E_{j-1}(t)}{\Delta x}\right|^2 \\
&\qquad\qquad + \tfrac{\Delta x^2}{4}\textstyle\sum_{j\in\mathbb{Z}}(1 - \varphi_{j-\frac{1}{2}}(t))\left|\frac{V_j(t) - V_{j-1}(t)}{\Delta x}\right|^2.
\end{aligned} \tag{18}$$

One has $\left|\frac{V_j(t) - V_{j-1}(t)}{\Delta x}\right| \leq C$ and only a $O(\Delta x^{-1})$ terms are non zero. One gets $\frac{d}{dt}\|E(t)\|_{\Delta x}^2 \leq C\Delta x^2\|E(t)\|_{\Delta x} + C\Delta x$. Then a Gronwall Lemma yields the claim. □

3.2. **The fully discrete scheme**

One combines the analysis of Section 3.1 and the inequality (15). The interpolation of the exact solution u on the pace-time grid is $V_j^n = u(j\Delta x, j)$. The truncation error is r_j^n

$$\frac{V_j^{n+1} - V_j^n}{\Delta t} + \frac{V_j^n - V_{j-1}^n}{\Delta x} + \frac{(1-\nu)}{2}\frac{\varphi_{j+\frac{1}{2}}^n(V_{j+1}^n - V_j^n) - \varphi_{j-\frac{1}{2}}^n(V_j^n - V_{j-1}^n)}{\Delta x} = r_j^n.$$

It is decomposed in two parts $r^n = s^n + t^n$. The first part is the truncation error of the Lax-Wendroff scheme

$$s_j^n = \frac{V_j^{n+1} - V_j^n}{\Delta t} + \frac{V_j^n - V_{j-1}^n}{\Delta x} + \frac{(1-\nu)}{2}\frac{(V_{j+1}^n - V_j^n) - (V_j^n - V_{j-1}^n)}{\Delta x}$$

so $s_j^n = O(\Delta x^2)$. The second part comes from the flux limitation

$$t_j^n = -\frac{(1-\nu)}{2}\frac{(1 - \varphi_{j+\frac{1}{2}}^n)(V_{j+1}^n - V_j^n) - (1 - \varphi_{j-\frac{1}{2}}^n)(V_j^n - V_{j-1}^n)}{\Delta x}. \tag{19}$$

The error $E_j^n = V_j^n - U_j^n$ vanishes at initial time (that is $E_j^0 = 0$ for all j) and satisfies $E_j^{n+1} = \overline{E}_j^n + \Delta t r_j^n$ where

$$\overline{E}_j^n = (1-\nu)E_j + \nu E_{j-1} - \frac{\nu(1-\nu)}{2}\left(\varphi_{j+\frac{1}{2}}(E_{j+1} - E_j) - \varphi_{j-\frac{1}{2}}(E_j - E_{j-1})\right).$$

One deduces that

$$\|E^{n+1}\|_{\Delta x}^2 = \|\overline{E}^n\|_{\Delta x}^2 + 2\Delta t\left(\overline{E}^n, r^n\right)_{\Delta x} + \Delta t^2\|r^n\|_{\Delta x}^2.$$

It is better to rewrite it as

$$\frac{\|E^{n+1}\|_{\Delta x}^2 - \|\overline{E}^n\|_{\Delta x}^2}{2\Delta t} = \left(\overline{E}^n, r^n\right)_{\Delta x} + \frac{\Delta t}{2}\|r^n\|_{\Delta x}^2$$

so that a comparison with (18) is possible. One obtains

$$\frac{\|E^{n+1}\|_{\Delta x}^2 - \|E^n\|_{\Delta x}^2}{2\Delta t} + \frac{1}{2\nu}Q_{\varphi^n}(\Delta^{E,n}) = \left(\overline{E}^n, r^n\right)_{\Delta x} + \frac{\Delta t}{2}\|r^n\|_{\Delta x}^2. \tag{20}$$

The quadratic form is evaluated with respect to $\Delta^{E,n} = \left(\Delta^{E,n}_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$ with $\Delta^{E,n}_{j+\frac{1}{2}} = E^n_{j+1} - E^n_j$. The last term in (20) does not pose any problem because by definition $\|r^n\|_{\Delta x} = O(1)$ and $\Delta t\|r^n\|_{\Delta x} = O(\Delta t)$. The other term is

$$\left(\overline{E}^n, r^n\right)_{\Delta x} = \left(\overline{E}^n, s^n\right)_{\Delta x} + \left(\overline{E}^n, t^n\right)_{\Delta x} \tag{21}$$

where $\left(\overline{E}^n, s^n\right)_{\Delta x} \leq C\Delta x^2\|E^n\|_{\Delta x}$ and

$$\left(\overline{E}^n, t^n\right)_{\Delta x} = -\frac{(1-\nu)}{2}\sum_{j\in\mathbb{Z}}\left((1-\nu)E^n_j + \nu E^n_{j-1} - \frac{\nu(1-\nu)}{2}\left(\varphi^n_{j+\frac{1}{2}}\Delta^{E,n}_{j+\frac{1}{2}} - \varphi^n_{j-\frac{1}{2}}\Delta^{E,n}_{j-\frac{1}{2}}\right)\right) \tag{22}$$

$$\times\left((1-\varphi^n_{j+\frac{1}{2}})D^n_{j+\frac{1}{2}} - (1-\varphi^n_{j-\frac{1}{2}})D^n_{j-\frac{1}{2}}\right)$$

with the notation $D^n_{j+\frac{1}{2}} = V^n_{j+1} - V^n_j$.

Lemma 5. *One has* $\left(\overline{E}^n, t^n\right)_{\Delta x} \leq C\left(Q_{\varphi^n}\left(\Delta^{E,n}\right)\right)^{\frac{1}{2}}\Delta x^{\frac{1}{2}}$.

Proof. From (22) one splits $\left(\overline{E}^n, t^n\right)_{\Delta x} = G^n_1 + G^n_2$ with

$$G^n_1 = -\frac{1-\nu}{2}\sum_{j\in\mathbb{Z}}\left((1-\nu)E^n_j + \nu E^n_{j-1}\right)\left((1-\varphi^n_{j+\frac{1}{2}})D^n_{j+\frac{1}{2}} - (1-\varphi^n_{j-\frac{1}{2}})D^n_{j-\frac{1}{2}}\right)$$

and

$$G^n_2 = -\frac{\nu(1-\nu)^2}{4}\sum_{j\in\mathbb{Z}}\left(\varphi^n_{j+\frac{1}{2}}\Delta^{E,n}_{j+\frac{1}{2}} - \varphi^n_{j-\frac{1}{2}}\Delta^{E,n}_{j-\frac{1}{2}}\right)\left((1-\varphi^n_{j+\frac{1}{2}})D^n_{j+\frac{1}{2}} - (1-\varphi^n_{j-\frac{1}{2}})D^n_{j-\frac{1}{2}}\right).$$

The first contribution is

$$G^n_1 = \frac{1-\nu}{2}\sum_{j\in\mathbb{Z}}\left((1-\nu)\Delta^{E,n}_{j+\frac{1}{2}} + \nu\Delta^{E,n}_{j-\frac{1}{2}}\right)(1-\varphi^n_{j+\frac{1}{2}})D^n_{j+\frac{1}{2}}$$

$$= -\frac{1-\nu}{2}\sum_{j\in\mathbb{Z}}(1-\varphi^n_{j+\frac{1}{2}})\Delta^{E,n}_{j+\frac{1}{2}}D^n_{j+\frac{1}{2}} - \nu\frac{1-\nu}{2}\sum_{j\in\mathbb{Z}}\left(\Delta^{E,n}_{j+\frac{1}{2}} - \Delta^{E,n}_{j-\frac{1}{2}}\right)(1-\varphi^n_{j+\frac{1}{2}})D^n_{j+\frac{1}{2}}.$$

Remark 1 and a Cauchy-Schwarz inequality yield

$$G^n_1 \leq C\left(Q_{\varphi^n}\left(\Delta^{E,n}\right)\right)^{\frac{1}{2}}\left(\sum_{j\in\mathbb{Z}}|D^n_{j+\frac{1}{2}}|^2\right)^{\frac{1}{2}} \leq C\left(Q_{\varphi^n}\left(\Delta^{E,n}\right)\right)^{\frac{1}{2}}\Delta x^{\frac{1}{2}}.$$

Similarly, Remark 1 and simple inequalities yield for the second contribution $G^n_2 \leq CQ_{\varphi^n}(\Delta^{E,n})^{\frac{1}{2}}\Delta x^{\frac{1}{2}}$ which ends the proof. □

Theorem 6. *Let $T > 0$. Assume the CFL condition (26) and assume $0 \leq \varphi^n_{j+\frac{1}{2}} \leq 1$ for all $j \in \mathbb{Z}$ and all $n \in \mathbb{N}$ such that $n\Delta t \leq T$. Then there exists a constant $C(\nu) > 0$ such that*

$$\max_{n\Delta t \leq T} \|E^n\|_{\Delta x} \leq C(\nu)\Delta x^{\frac{1}{2}}.$$

Proof. Consider the above material (20-21-22) and Lemma 5. Take $\varepsilon > 0$ a small parameter. A Cauchy-Schwarz inequality yields

$$\left(\overline{E}^n, t^n\right)_{\Delta x} \leq C\varepsilon Q_{\varphi^n}\left(\Delta^{E,n}\right) + \frac{c}{\varepsilon}\Delta x.$$

For ε small enough, the coefficient $C\varepsilon$ is less than the coefficient $\frac{1}{2\nu}$ in front of the quadratic form in the left hand side of (20). One obtains

$$\frac{\|E^{n+1}\|^2_{\Delta x} - \|E^n\|^2_{\Delta x}}{2\Delta t} \quad \leq \quad C\Delta x^2 \|E^n\|_{\Delta x} + C\Delta x + \frac{\Delta t}{2}\|r^n\|^2_{\Delta x} \tag{23}$$

where all constants depend on $\nu \in (0, 1)$. The end of the proof with a Gronwall inequality is standard. □

For example one obtains that the minmod scheme (5-7) with the minmod flux (8) is quadratically stable and convergent in quadratic norm. This is already a progress with respect to the proof with TVD estimates in [9] where the convergence of the minmod scheme is shown to be not less than $O(\Delta x^{\frac{1}{2}})$, but in the L^1 norm which is weaker than the quadratic norm.

4. Quadratic estimates in 2D

Now that quadratic stability of flux limiters has been established in 1D, we study a way to use this method for the design of flux limiters in higher dimension. At the end of the construction, it will provide a strategy to bypass the Goodman-Leveque obstruction by using quadratic stability instead of TVD stability. We will use the same approach as in 1D, that is we will modify a 2D Lax-Wendroff linear scheme with limiters and study the resulting scheme. However, in a preliminary stage, we need to define what we denote as the 2D Lax-Wendroff scheme. Due to its simplicity, it is probable the scheme is not new, even we do not know a reference in the literature.

One considers the same Finite Volume structure as in the Goodman-Leveque work [19]

$$\frac{\overline{U}_{i,j} - U_{i,j}}{\Delta t} + p\frac{U_{i+\frac{1}{2},j} - U_{i-\frac{1}{2},j}}{\Delta x} + q\frac{U_{i,j+\frac{1}{2}} - U_{i,j-\frac{1}{2}}}{\Delta x} = 0, \quad \text{for all } i, j \tag{24}$$

where the explicit numerical fluxes at time step n are denoted as $U_{i+\frac{1}{2},j}$ and $U_{i,j+\frac{1}{2}}$ for all possible i and j. The mesh size is the same in the horizontal and vertical directions, that is $\Delta x = \Delta y > 0$.

The basic first-order accurate numerical method is the upwind scheme for which the fluxes are

$$\text{Upwind fluxes: } U^{\text{up}}_{i+\frac{1}{2},j} = U^{\text{up}}_{i,j+\frac{1}{2}} = U_{i,j}. \tag{25}$$

The upwind scheme (24-25) is stable in quadratic norm (and in all Lebesgue norms) under the CFL condition

$$\nu = \frac{\Delta t}{\Delta x} \leq 1. \tag{26}$$

The upwind scheme is naturally consistant at first order $O(\Delta x + \Delta t) = O(\Delta x)$ with the advection equation. A way to derive the Lax-Wendroff scheme is to analyze with the modified equation technique the numerical dissipation of the upwind scheme, and then to subtract the first order numerical dissipation.

Lemma 7. *The modified equation of the upwind scheme (24-25) is*

$$\partial_t u + p\partial_x u + q\partial_y u - \frac{p\Delta x}{2}\partial_{xx} u - \frac{q\Delta x}{2}\partial_{yy} u + \frac{\Delta t}{2}\left(p\partial_x + q\partial_y\right)^2 u = 0, \tag{27}$$

or equivalently

$$\partial_t u + p\partial_x u + q\partial_y u - (1-\nu)\frac{\Delta x}{2}\left(p\partial_x + q\partial_y\right)^2 u - pq\frac{\Delta x}{2}\left(\partial_x - \partial_y\right)^2 u = 0. \tag{28}$$

Proof. With the notations $V_j^n = u(x_j, t_n)$ for all j and n, one has the expansions

$$\left\{\begin{array}{l} p\frac{V_{i,j}^n - V_{i-1,j}^n}{\Delta x} = p\partial_x u_{i,j}^n - \frac{p\Delta x}{2}\partial_{xx} u_{i,j}^n + O(\Delta x^2), \\ q\frac{V_{i,j}^n - V_{i,j-}^n}{\Delta x} = q\partial_y u_{i,j}^n - \frac{q\Delta x}{2}\partial_{yy} u_{i,j}^n + O(\Delta x^2), \\ \frac{V_{i,j}^{n+1} - V_{i,j}^n}{\Delta t} = \partial_t u_{i,j}^n - \frac{\Delta t}{2}\partial_{tt} u_{i,j}^n + O(\Delta t^2). \end{array}\right. \tag{29}$$

For u a solution to (24), one has $\partial_t u = -(p\partial_x + q\partial_y)u$ and $\partial_{tt} u = (p\partial_x + q\partial_y)^2 u$. Plugging (29) in (24-25) , one gets

$$\partial_t u + p\partial_x u + q\partial_y u - \frac{p\Delta x}{2}\partial_{xx} u - \frac{q\Delta x}{2}\partial_{yy} u + \frac{\Delta t}{2}\left(p\partial_x + q\partial_y\right)^2 u = O(\Delta x^2 + \Delta t^2)$$

which is the modified equation at second order. The tensorial nature of the numerical diffusion is better revealed after rearrangement of the second terms in the modified equation

$$\begin{aligned} &-p\partial_{xx} u - q\partial_{yy} u + \nu(p\partial_x + q\partial_y)^2 u \\ &= -(p\partial_x + q\partial_y)^2 u + (p^2 - p)\partial_{xx} u + 2pq\partial_{xy} u + (q^2 - q)\partial_{yy} u + \nu(p\partial_x + q\partial_y)^2 u \\ &= -(1-\nu)(p\partial_x + q\partial_y)^2 u - pq\left(\partial_{xx} - 2\partial_{xy} + \partial_{yy}\right) u \\ &= -(1-\nu)(p\partial_x + q\partial_y)^2 u - pq\left(\partial_x - \partial_y\right)^2 u. \end{aligned}$$

It gives (28). □

The modified equation (28) couples an advection equation with an anisotropic diffusion equation. The theory of discrete anisotropic equations made recent progress on the basis of Selling's decomposition of symmetric tensors [4]. In this work, we use a more direct approach to analyze the tensorial nature of the numerical diffusion. The first contribution $(1-\nu)\frac{\Delta x}{2}\left(p\partial_x + q\partial_y\right)^2 u$ is the diffusion in the direction of the flow, it vanishes for $\nu = 1$. The second contribution $-pq\frac{\Delta x}{2}\left(\partial_x - \partial_y\right)^2 u$ is the diffusion in the direction at angle $\frac{3}{4}\pi$ with the horizontal direction. It is independent of the time step and is purely a two-dimensional grid effect. This term has no counterpart in dimension one.

We define the 2D Lax-Wendroff type scheme for which the fluxes are

$$\text{Lax-Wendroff fluxes: } \left\{\begin{array}{l} U_{i+\frac{1}{2},j}^{\text{lw}} = \frac{U_{i,j} + U_{i+1,j}}{2} - \nu\frac{U_{i+1,j} - pU_{i,j} - qU_{i+1,j-1}}{2}, \\ U_{i,j+\frac{1}{2}}^{\text{lw}} = \frac{U_{i,j} + U_{i,j+1}}{2} - \nu\frac{U_{i,j+1} - pU_{i-1,j+1} - qU_{i,j}}{2}, \end{array}\right. \tag{30}$$

It can be rewritten in another equivalent form which is more adapted to our purposes. We note the difference in the direction of the flow

$$\Delta_{i,j} = U_{i,j} - pU_{i-1,j} - qU_{i,j-1}$$

and we define modified fluxes

$$\text{Modified fluxes: } \left\{\begin{array}{l} U_{i+\frac{1}{2},j}^{\text{mod}} = U_{i,j} + \frac{1-\nu}{2}\Delta_{i+1,j}, \\ U_{i,j+\frac{1}{2}}^{\text{mod}} = U_{i,j} + \frac{1-\nu}{2}\Delta U_{i,j+1}, \end{array}\right. \tag{31}$$

We consider also the corner difference

$$\Delta_{i+\frac{1}{2},j+\frac{1}{2}} = U_{i,j+1} - U_{i+1,j}. \tag{32}$$

It allows to defined the scheme

$$\frac{\overline{U}_{i,j} - U_{i,j}}{\Delta t} + p\frac{U^{\rm mod}_{i+\frac{1}{2},j} - U^{\rm mod}_{i-\frac{1}{2},j}}{\Delta x} + q\frac{U^{\rm mod}_{i,j+\frac{1}{2}} - U^{\rm mod}_{i,j-\frac{1}{2}}}{\Delta x} + \frac{pq}{2}\frac{\Delta_{i-\frac{1}{2},j+\frac{1}{2}} - \Delta_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta x} = 0 \tag{33}$$

Lemma 8. *The scheme obtained with the fluxes (30) in the standard Finite Volume formulation (24) is equal to the scheme (33) with the Finite Volume fluxes (31) and the corner correction (32). It is second order accurate in space and time and does not respect the maximum principle.*

Proof. The total incoming flux of (30)-(24) is

$$\begin{array}{rcl} pU^{\rm lw}_{i-\frac{1}{2},j} + qU^{\rm lw}_{i,j-\frac{1}{2}} & = & \frac{pU_{i-1,j}+qU_{i,j-1}}{2} + \frac{U_{i,j}}{2} - \frac{\nu}{2}\Delta U_{i,j} \\ & = & pU_{i-1,j} + qU_{i,j-1} + \frac{1-\nu}{2}\Delta_{i,j} = pU^{\rm mod}_{i-\frac{1}{2},j} + qU^{\rm mod}_{i,j-\frac{1}{2}}, \end{array}$$

so it is equal to the total incoming flux of (31-33).
The total outgoing flux of (30)-(24) is

$$\begin{array}{rcl} pU^{\rm lw}_{i+\frac{1}{2},j} + qU^{\rm lw}_{i,j+\frac{1}{2}} & = & \frac{U_{i,j}}{2} + \frac{pU_{i+1,j}+qU_{i,j+1}}{2} - \frac{\nu}{2}p\Delta_{i+1,j} - \frac{\nu}{2}q\Delta U_{i,j+1} \\ & = & U_{i,j} + \frac{1-\nu}{2}p\Delta U_{i+1,j} + \frac{1-\nu}{2}q\Delta_{i,j+1} \\ & & -\frac{q}{2}\left(U_{i,j+1} - pU_{i-1,j+1} - qU_{i,j}\right) \\ & & -\frac{U_{i,j}}{2} + \frac{pU_{i+1,j}+qU_{i,j+1}}{2} - \frac{p}{2}\left(U_{i+1,j} - pU_{i,j} - qU_{i+1,j-1}\right) \\ & = & pU^{\rm mod}_{i-\frac{1}{2},j} + qU^{\rm mod}_{i,j-\frac{1}{2}} - pqU_{i,j} + \frac{pq}{2}U_{i+1,j-1} + \frac{pq}{2}U_{i-1,j+1} \\ & = & pU^{\rm mod}_{i-\frac{1}{2},j} + qU^{\rm mod}_{i,j-\frac{1}{2}} + \frac{pq}{2}\left(\Delta_{i-\frac{1}{2},j+\frac{1}{2}} - \Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right) \end{array}$$

which is the sum of the total outgoing fluxes plus the corner correction in (33). So both schemes are the same. The Lax-Wendroff fluxes (30) are easily justified by the modified equation, so the scheme is second-order accurate (a confirmation is in the numerical Section 6). For $p = 1$ and $q = 0$, the scheme is identical to the classical Lax-Wendroff scheme which does not respect the maximum principle. □

The 2D spaces for the numerical analysis are extension of the 1D spaces

$$l^2 = \left\{ U = (U_{i,j})_{i,j\in\mathbb{Z}} \text{ such that } \sum_{ij} |U_{i,j}|^2 < \infty \right\}$$

and

$$V_{\Delta x} = \left\{ U \in l^2 \text{ with the norm } \|U\|^2_{\Delta x} = \Delta x \sum_{ij} |U_{i,j}|^2 \right\}.$$

4.1. **2D limiters**

The Lax-Wendroff scheme is written (33) as a Finite Volume scheme with additional corner interactions. This formulation is adapted to the tensorial nature of the numerical diffusion visible in the modified equation (28). The introduction of flux limiters $\varphi_{i+\frac{1}{2},j}$, $\varphi_{i,j+\frac{1}{2}}$ and $\varphi_{i+\frac{1}{2},j+\frac{1}{2}}$ is done my mimicking the 1D procedure

(7-8). It yields the scheme

$$\frac{\overline{U}_{i,j}-U_{i,j}}{\Delta t}+p\frac{U^{\lim}_{i+\frac{1}{2},j}-U^{\lim}_{i-\frac{1}{2},j}}{\Delta x}+q\frac{U^{\lim}_{i,j+\frac{1}{2}}-U^{\lim}_{i,j-\frac{1}{2}}}{\Delta x}+\frac{pq}{2}\frac{C^{\lim}_{i-\frac{1}{2},j+\frac{1}{2}}-C^{\lim}_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta x}=0 \tag{34}$$

where

$$\begin{cases} U^{\lim}_{i+\frac{1}{2},j}=U_{i,j}+\frac{1-\nu}{2}\varphi_{i+1,j}\Delta_{i+1,j},\\ U^{\lim}_{i,j+\frac{1}{2}}=U_{i,j}+\frac{1-\nu}{2}\varphi_{i,j+1}\Delta_{i,j+1},\\ C^{\lim}_{i+\frac{1}{2},j+\frac{1}{2}}=\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}}. \end{cases} \tag{35}$$

The cell-based limiter $\varphi_{i,j}$ is the equivalent of $\varphi_{j+\frac{1}{2}}$ is 1D (while $\Delta_{i,j}$ is the equivalent of $\Delta_{j-\frac{1}{2}}$). The corner-based limiter limiter $\varphi_{i-\frac{1}{2},j+\frac{1}{2}}$ is a new term related to the tensorial nature of numerical diffusion in 2D. Inspired by the quadratic properties in 1D, we study the properties of the scheme when the flux limiters satisfy the bounds

$$0\leq\varphi_{i,j},\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\leq 1 \qquad \text{for all } i,j. \tag{36}$$

We will assume without restriction that $U\in V_{\Delta x}$.

Lemma 9. *One has* $\sum_{ij}|\overline{U}_{i,j}|^2=\sum_{ij}|U_{i,j}|^2-Q_\varphi(\Delta)$ *where the quadratic form is*

$$\begin{aligned} Q_\varphi(\Delta)&=\nu(1-\nu)\sum_{ij}\left|\Delta_{i,j}\right|^2+\nu pq\sum_{ij}\left|\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\right|^2\\ &-\nu(1-\nu)\sum_{ij}\varphi_{i,j}\Delta_{i,j}\left((1-\nu)\Delta_{i,j}+\nu p\Delta_{i-1,j}+\nu q\Delta_{i,j-1}\right)\\ &-\nu pq\sum_{ij}\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\left((1-\nu)\Delta_{i-\frac{1}{2},j+\frac{1}{2}}+\nu p\Delta_{i-\frac{3}{2},j+\frac{1}{2}}+\nu q\Delta_{i-\frac{1}{2},j-\frac{1}{2}}\right)\\ &-\sum_{ij}\left|\frac{\nu(1-\nu)}{2}\left(p\varphi_{i+1,j}\Delta_{i+1,j}+q\varphi_{i-1,j}\Delta_{i-1,j}-\varphi_{i,j}\Delta_{i,j}\right)\right.\\ &\left.+\frac{\nu pq}{2}\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}-\varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right)\right|^2 \end{aligned}$$

Proof. The structure of the calculation is the same as in the proof of Lemma 1. One decompose

$$Q_\varphi(\Delta)=\underbrace{\sum_{ij}|U_{i,j}|^2-\sum_{ij}|U^{\rm up}_{i,j}|^2}_{=Q^1}+\underbrace{\sum_{ij}|U^{\rm up}_{i,j}|^2-\sum_{ij}|\overline{U}_{i,j}|^2}_{=Q^2}.$$

• Calculation of Q^1.
Since $U^{\rm up}_{i,j}=(1-\nu)U_{i,j}+\nu pU_{i-1,j}+\nu qU_{i,j-1}$, one has

$$\begin{aligned} &\textstyle\sum_{ij}|U^{\rm up}_{i,j}|^2=\sum_{ij}\left((1-\nu)^2+\nu^2p^2+\nu^2q^2\right)|U_{i,j}|^2\\ &\quad+2\textstyle\sum_{ij}\left((1-\nu)\nu pU_{i,j}U_{i-1,j}+(1-\nu)\nu qU_{i,j}U_{i,j-1}+\nu^2pqU_{i-1,j}U_{i,j-1}\right), \end{aligned}$$

and one also has

$$\begin{aligned} \nu(1-\nu)\sum_{ij}|U_{i,j}-pU_{i-1,j}-qU_{i,j-1}|^2&=\nu(1-\nu)\sum_{ij}\left(1+p^2+q^2\right)|U_{i,j}|^2\\ &-2\nu(1-\nu)pU_{i,j}U_{i-1,j}-2\nu(1-\nu)qU_{i,j}U_{i,j-1}+2\nu(1-\nu)pqU_{i-1,j}U_{i,j-1}. \end{aligned}$$

One gets by addition

$$|U_{i,j}^{\rm up}|^2+\nu(1-\nu)\sum_{ij}|U_{i,j}-pU_{i-1,j}-qU_{i,j-1}|^2$$

$$=\sum_{ij}\left((1-\nu)^2+\nu^2p^2+\nu^2q^2+\nu(1-\nu)+\nu(1-\nu)p^2+\nu(1-\nu)q^2\right)|U_{i,j}|^2+2\nu pqU_{i-1,j}U_{i,j-1}$$

$$=\sum_{ij}\left((1-\nu)+\nu p^2+\nu q^2\right)|U_{i,j}|^2+2\nu pqU_{i-1,j}U_{i,j-1}$$

$$=\sum_{ij}|U_{i,j}|^2-\nu pq\sum_{ij}|U_{i-1,j}-U_{i,j-1}|^2.$$

So

$$Q^1=\nu(1-\nu)\sum_{ij}|U_{i,j}-pU_{i-1,j}-qU_{i,j-1}|^2+\nu pq\sum_{ij}|U_{i-1,j}-U_{i,j-1}|^2$$

$$=\nu(1-\nu)\sum_{ij}|\Delta_{i,j}|^2+\nu pq\sum_{ij}\left|\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\right|^2\geq 0.$$

• Calculation of Q^2.
One has

$$Q^2=\underbrace{-2\sum_{ij}(\overline{U}_{i,j}-U_{i,j}^{\rm up})U_{i,j}^{\rm up}}_{=Q^3}-\sum_{ij}|\overline{U}_{i,j}-U_{i,j}^{\rm up}|^2$$

where

$$\overline{U}_{i,j}=U_{i,j}^{\rm up}-\frac{\nu(1-\nu)}{2}\left(p\varphi_{i+1,j}\Delta_{i+1,j}+q\varphi_{i-1,j}\Delta_{i-1,}j-\varphi_{i,j}\Delta_{i,j}\right)-\frac{\nu pq}{2}\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}-\varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right).$$

The first term in Q^2 is

$$Q^3=-\frac{\nu(1-\nu)}{2}\sum_{ij}\left(p\varphi_{i+1,j}\Delta_{i+1,j}+q\varphi_{i-1,j}\Delta_{i-1,}j-\varphi_{i,j}\Delta_{i,j}\right)\left((1-\nu)U_{i,j}+\nu pU_{i-1,j}+\nu qU_{i,j-1}\right)$$

$$-\frac{\nu pq}{2}\sum_{ij}\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}-\varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right)\left((1-\nu)U_{i,j}+\nu pU_{i-1,j}+\nu qU_{i,j-1}\right).$$

Rearrangement yields

$$Q^3=\frac{\nu(1-\nu)}{2}\sum_{ij}\varphi_{i,j}\Delta_{i,j}\left((1-\nu)\Delta_{i,j}+\nu p\Delta_{i-1,j}+\nu q\Delta_{i,j-1}\right)$$

$$+\frac{\nu pq}{2}\sum_{ij}\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\left((1-\nu)\Delta_{i-\frac{1}{2},j+\frac{1}{2}}+\nu p\Delta_{i-\frac{3}{2},j+\frac{1}{2}}+\nu q\Delta_{i-\frac{1}{2},j-\frac{1}{2}}\right).$$

Other calculations are immediate. □

To analyze the sign of the quadratic form in Lemma 9, it is possible to rely on the one-dimensional result of Proposition 2 for the simpler quadratic form (12). Let us define four quantities $H^{1,2,3,4}$. The first one H^1 concerns the cell-based differences analyzed in the horizontal direction by means of the 1D quadratic form (12)

$$H^1=\sum_{ij}|\Delta_{i,j}|^2-\sum_{ij}\varphi_{i,j}\Delta_{i,j}\left((1-\nu)\Delta_{i,j}+\nu\Delta_{i-1,j}\right)-\frac{\nu}{4}\sum_{ij}|\varphi_{i,j}\Delta_{i,j}-\varphi_{i-1,j}\Delta_{i-1,j}|^2.$$

The second one H^2 concerns the cell-based differences analyzed in the vertical direction by means of (12)

$$H^2 = \sum_{ij} |\Delta_{i,j}|^2 - \sum_{ij} \varphi_{i,j}\Delta_{i,j}\left((1-\nu)\Delta_{i,j} + \nu\Delta_{i,j-1}\right) - \frac{\nu}{4}\sum_{ij} \left|\varphi_{i,j}\Delta_{i,j} - \varphi_{i,j-1}\Delta_{i,j-1}\right|^2.$$

The third one H^3 concerns the corner-based differences analyzed in the horizontal direction

$$H^3 = \sum_{ij} |\Delta_{i+\frac{1}{2},j+\frac{1}{2}}|^2 - \sum_{ij} \varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}}\left((1-\nu)\Delta_{i+\frac{1}{2},j+\frac{1}{2}} + \nu\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\right)$$

$$-\frac{\nu}{4}\sum_{ij} \left|\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}} - \varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\right|^2.$$

Finally the fourth one H^4 concerns the corner-based differences analyzed in the vertical direction

$$H^4 = \sum_{ij} |\Delta_{i+\frac{1}{2},j+\frac{1}{2}}|^2 - \sum_{ij} \varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}}\left((1-\nu)\Delta_{i+\frac{1}{2},j+\frac{1}{2}} + \nu\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right)$$

$$-\frac{\nu}{4}\sum_{ij} \left|\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}} - \varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right|^2.$$

One has the decomposition

$$Q_\varphi(\Delta) = pH^1 + qH^2 + pH^3 + qH^4 + H^5 \tag{37}$$

where the additional term is

$$\begin{array}{rcl} H^5 & = & \frac{\nu^2(1-\nu)p}{4}\sum_{ij}\left|\varphi_{i,j}\Delta_{i,j} - \varphi_{i-1,j}\Delta_{i-1,j}\right|^2 \\ & + & \frac{\nu^2(1-\nu)q}{4}\sum_{ij}\left|\varphi_{i,j}\Delta_{i,j} - \varphi_{i,j-1}\Delta_{i,j-1}\right|^2 \\ & + & \frac{\nu^2p^2q}{4}\sum_{ij}\left|\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}} - \varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}}\right|^2 \\ & + & \frac{\nu^2pq^2}{4}\sum_{ij}\left|\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}} - \varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right|^2 \\ & - & \sum_{ij}\left|\frac{\nu(1-\nu)}{2}\left(p\varphi_{i+1,j}\Delta_{i+1,j} + q\varphi_{i-1,j}\Delta_{i-1,j} - \varphi_{i,j}\Delta_{i,j}\right)\right. \\ & & \left. + \frac{\nu pq}{2}\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta_{i-\frac{1}{2},j+\frac{1}{2}} - \varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right)\right|^2 \end{array} \tag{38}$$

The term H^5 is by definition $O(\nu^2)$, while $H^{1,2,3,4}$ are affine with respect to the Courant number ν. Moroever H^5 is homogeneous of degree 2 with respect to all products $(\varphi_{i,j}\Delta_{i,j})_{ij}$ and $\left(\varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta_{i+\frac{1}{2},j-\frac{1}{2}}\right)_{ij}$. Under the conditions (36) and assuming the CFL condition, one already has $H^{1,2,3,4} \geq 0$ as a corollary of Proposition 2. It remains to study the sign of H^5.

Lemma 10. *Under the previous conditions, one has $H^5 \geq 0$.*

Proof. Let us begin with notations which are useful to synthesize the structure of H^5. For $a = (a_{i,j}) \in l^2$ defined at centers, one defines the operator L such that $(La)_{i,j} = a_{i,j} - pa_{i-1,j} - qa_{i,j-1}$ and the operator M such that $(Ma)_{i,j} = a_{i-1,j} - a_{i,j-1}$. For $b = (b_{i+\frac{1}{2},j+\frac{1}{2}}) \in l^2$ defined at vertices, one defines with similar notations $(Lb)_{i+\frac{1}{2},j+\frac{1}{2}} = b_{i+\frac{1}{2},j+\frac{1}{2}} - pb_{i-\frac{1}{2},j+\frac{1}{2}} - qb_{i+\frac{1}{2},j-\frac{1}{2}}$ and $(Mb)_{i+\frac{1}{2},j+\frac{1}{2}} = b_{i-\frac{1}{2},j+\frac{1}{2}} - b_{i+\frac{1}{2},j-\frac{1}{2}}$. Let T the translation operator such that $a = Tb$ means $a_{i,j} = b_{i+\frac{1}{2},j+\frac{1}{2}}$. The adjoints operators L^t and M^t are defined with respect to the natural quadratic scalar product in l^2. One has for example $(L^ta)_{i,j} = a_{i,j} - pa_{i+1,j} - qa_{i,j+1}$.

One uses $a = (a_{i,j})$ where $a_{i,j} = \varphi_{i,j}\Delta_{i,j}$ is defined at centers and $b = (b_{i+\frac{1}{2},j+\frac{1}{2}})$ where $b_{i+\frac{1}{2},j+\frac{1}{2}} = \varphi_{i+\frac{1}{2},j+\frac{1}{2}}\Delta_{i+\frac{1}{2},j+\frac{1}{2}}$ is defined at vertices. Then one writes

$$\begin{aligned} H^5 &= \tfrac{\nu^2(1-\nu)}{4}\sum_{ij} p|a_{i,j}-a_{i+1,j}|^2 + q|a_{i,j}-a_{i,j+1}|^2 \\ &\quad + \tfrac{\nu^2 pq}{4}\sum_{ij} p|b_{i,j}-b_{i-1,j}|^2 + q|b_{i,j}-b_{i,j-1}|^2 - \frac{\nu^2}{4}\left\|(1-\nu)L^t a - pqTMb\right\|_{l^2}^2. \end{aligned}$$

One uses the identity $p\alpha^2+q\beta^2 = (p\alpha+q\beta)^2 + pq(\alpha-\beta)^2$ to transform the first two lines H^5. One obtains

$$H^5 = \tfrac{\nu^2(1-\nu)}{4}\left(\|L^t a\|_{l^2}^2 + pq\|M^t a\|_{l^2}^2\right) + \tfrac{\nu^2 pq}{4}\left(\|Lb\|_{l^2}^2 + pq\|Mb\|_{l^2}^2\right) - \tfrac{\nu^2}{4}\|(1-\nu)L^t a - pqTMb\|_{l^2}^2.$$

An expansion of the last square yields

$$\begin{aligned} \|(1-\nu)L^t a - pqTMb\|_{l^2}^2 &= (1-\nu)^2\|L^t a\|_{l^2}^2 - 2(1-\nu)pq(L^t a, TMb)_{l^2} + p^2q^2\|Mb\|_{l^2}^2 \\ &= (1-\nu)^2\|L^t a\|_{l^2}^2 - 2(1-\nu)pq(M^t a, TLb)_{l^2} + p^2q^2\|Mb\|_{l^2}^2 \end{aligned}$$

because the double product is rearranged as $(L^t a, TMb)_{l^2} = (M^t a, TLb)_{l^2}$ using the commutation property $LTM = MTL$. So one has the formula

$$\begin{aligned} H^5 &= \frac{\nu^3(1-\nu)}{4}\|L^t a\|_{l^2}^2 + \frac{\nu^2(1-\nu)pq}{4}\|M^t a\|_{l^2}^2 + \frac{\nu^2 pq}{4}\|Lb\|_{l^2}^2 + \frac{\nu^2(1-\nu)pq}{2}(M^t a, TLb)_{l^2} \\ &= \frac{\nu^3(1-\nu)}{4}\|L^t a\|_{l^2}^2 + \frac{\nu^3(1-\nu)pq}{4}\|M^t a\|_{l^2}^2 + \frac{\nu^2 pq}{4}\|(1-\nu)M^t a + TLb\|_{l^2}^2. \end{aligned}$$

Since it is the sum of three non negative terms, the claim is obtained. □

Theorem 11. *Under the CFL condition (26), the scheme (35-36) is quadratically stable and is convergent with an order at least $O(\Delta x^{\frac{1}{2}})$ in quadratic norm.*

Proof. Stability is a consequence of (37) and Lemma 10. To show convergence, it is sufficient to adapt Section 3.2. One considers that the truncation error $r^n = s^n - t^n$ is a sum of the truncation error $s^n = O(\Delta x^2)$ of the Lax-Wendroff scheme plus the contribution (39) of the limiters: it denoted as $t^n = \left(t^n_{i,j}\right)$, like (19) in 1D. The latter is function of all $1-\varphi_{i,j}$ and all $1-\varphi_{i+\frac{1}{2},j+\frac{1}{2}}$. The identity (20) also holds in 2D and it is sufficient to estimate $(\overline{E}^n, t^n)_{\Delta x}$ as in (21). We give below the main steps of the proof.

• **First step.** One writes $E^{\mathrm{up},n}_{i,j} = (1-\nu)E^n_{i,j} + \nu p E^n_{i-1,j} + \nu q E^n_{i,j-1}$, $\overline{E}^n_{i,j} = E^{\mathrm{up,n}}_{i,j} + F_{i,j}$ and

$$F^n_{i,j} = -\frac{\nu(1-\nu)}{2}\left(p\varphi_{i+1,j}\Delta^{E,n}_{i+1,j} + q\varphi_{i-1,j}\Delta^{E,n}_{i-1,j} - \varphi_{i,j}\Delta^{E,n}_{i,j}\right) - \frac{\nu pq}{2}\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}\Delta^{E,n}_{i-\frac{1}{2},j+\frac{1}{2}} - \varphi_{i+\frac{1}{2},j-\frac{1}{2}}\Delta^{E,n}_{i+\frac{1}{2},j-\frac{1}{2}}\right).$$

The total residual $r^n = s^n + t^n$ is split in two terms. The first is the residual of the linear Lax-Wendroff scheme so it is $O(\Delta x^2)$. The second one is the departure to the Lax-Wendroff residual (compare with (19))

$$\begin{aligned} t^n_{i,j} &= -\frac{\nu(1-\nu)}{2}\frac{p(1-\varphi_{i+1,j})D_{i+1,j} + q(1-\varphi_{i-1,j})D_{i-1,j} - (1-\varphi_{i,j})D_{i,j}}{\Delta x} \\ &\quad - \frac{\nu pq}{2}\frac{(1-\varphi_{i-\frac{1}{2},j+\frac{1}{2}})D_{i-\frac{1}{2},j+\frac{1}{2}} - (1-\varphi_{i+\frac{1}{2},j-\frac{1}{2}})D_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta x}. \end{aligned} \tag{39}$$

One has

$$\left(\overline{E}^n, t^n\right)_{\Delta x} = (E^{\mathrm{up},n}, t^n)_{\Delta x} + (F^n, t^n)_{\Delta x}.$$

• **Second step.** One has by rearrangement

$$(E^{\mathrm{up},n},t^n)_{\Delta x}=\frac{\nu(1-\nu)}{2}\sum_{ij}\left((1-\nu)\Delta_{i,j}^{E,n}+\nu p\Delta_{i-1,j}^{E,n}+\nu q\Delta_{i,j-1}^{E,n}\right)(1-\varphi_{i,j})D_{i,j}$$

$$+\frac{\nu pq}{2}\sum_{ij}\left((1-\nu)\Delta_{i+\frac12,j-\frac12}^{E,n}+\nu p\Delta_{i-\frac12,j-\frac12}^{E,n}+\nu q\Delta_{i+\frac12,j-\frac32}^{E,n}\right)(1-\varphi_{i+\frac12,j-\frac12})D_{i+\frac12,j-\frac12}$$

• **Third step.** From the decomposition (37) and Remark 1, one obtains the bound

$$\begin{array}{l}\sum_{ij}\Big((1-\varphi_{i,j}^n)|\Delta_{i,j}^{E,n}|^2+(1-\varphi_{i+\frac12,j+\frac12}^n)|\Delta_{i+\frac12,j+\frac12}^{E,n}|^2\\ +p|\Delta_{i,j}^{E,n}-\Delta_{i-1,j}^{E,n}|^2+q|\Delta_{i,j}^{E,n}-\Delta_{i,j-1}^{E,n}|^2\\ +p|\Delta_{i+\frac12,j+\frac12}^{E,n}-\Delta_{i-\frac12,j+\frac12}^{E,n}|^2+q|\Delta_{i+\frac12,j+\frac12}^{E,n}-\Delta_{i+\frac12,j-\frac12}^{E,n}|^2\Big)\quad\le C(\nu)Q_{\varphi^n}(\Delta^{E,n})\end{array}$$

• **Fourth step.** One has the fact that $|D_{i,j}|+\left|D_{i+\frac12,j-\frac12}\right|\le C\Delta x$ with only $O(\Delta x^{-1})$ being non zero. So one obtains by a Cauchy-Schwarz inequality that

$$(E^{\mathrm{up},n},t^n)_{\Delta x}\le C(\nu)Q_{\varphi^n}(\Delta^{E,n})^{\frac12}\Delta x^{\frac12}.$$

• **Fifth step.** One has more directly $(F^n,t^n)_{\Delta x}\le C(\nu)Q_{\varphi^n}(\Delta^n)^{\frac12}\Delta x^{\frac12}$. One obtains the 2D equivalent of Lemma 5. It yields the proof after an estimate (23) which is the same as in Theorem 6. □

5. Construction of a 2D minmod limiter

The objective in this section is to show that the above considerations can be used to construct a practical procedure for the definition of the limiters (35-36). The proposed method relies in a first step on the LBV method [11,12] which is on extension of the TVD criterion on general grid. This method is well suited to design the limiters $(\varphi_{i,j})$ centered in the cells. The second step is devoted to the definition of the limiters $\left(\varphi_{i+\frac12,j+\frac12}\right)$ for the corner interactions.

5.1. Construction of cell-based limiters $\varphi_{i,j}$

The LBV semi-norm is defined in [12] for general meshes. We review the *LBV* framework then we use it to define the cell-based limiters $\varphi_{i,j}$s. Another idea will be used to define the corner-based limiters.

5.1.1. *The LBV semi-norm on a cartesian mesh*

A simple definition of the LBV semi-norm on a cartesian mesh is the following.

Definition 12. *On a cartesian mesh, the LBV semi-norm is*

$$|U|_{\mathrm{LBV}}=\Delta x^2\sum_{i,j}\left|\frac{U_{i,j}-pU_{i-1,j}-qU_{i,j-1}}{\Delta x}\right|=\Delta x\sum_{i,j}\left|U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right|. \tag{40}$$

The interest of the LBV semi-norm is that it is endowed with an appropriate multidimensional Harten calculus [22]. Let us consider the local minimal value and the local maximal value

$$m_{i,j}=\min\left(U_{i,j},pU_{i-1,j}+qU_{i,j-1}\right)\ \text{and}\ M_{i,j}=\max\left(U_{i,j},pU_{i-1,j}+qU_{i,j-1}\right).$$

Let us assume that the numerical solution at the end of time step verifies for all i, j a double inequality

$$m_{i,j} \leq \overline{U}_{i,j} \leq M_{i,j}, \tag{41}$$

which can be rewritten as

$$\overline{U}_{i,j} = (1 - d_{i,j})U_{i,j} + d_{i,j}\left(pU_{i-1,j} + qU_{i,j-1}\right) \tag{42}$$

where

$$0 \leq d_{i,j} \leq 1. \tag{43}$$

Lemma 13. *Assume (41) for all i, j. Then $|\overline{U}|_{\mathrm{LBV}} \leq |U|_{\mathrm{LBV}}$.*

Proof. From (42), one has

$$\begin{array}{rcl}
\overline{U}_{i,j} - \left(p\overline{U}_{i-1,j} + q\overline{U}_{i,j-1}\right) & = & U_{i,j} - \left(pU_{i-1,j} + qU_{i,j-1}\right) \\
& - & d_{i,j}\left(U_{i,j} - \left(pU_{i-1,j} + qU_{i,j-1}\right)\right) \\
& + & d_{i-1,j}p\left(U_{i-1,j} - \left(pU_{i-2,j} + qU_{i-1,j-1}\right)\right) \\
& + & d_{i,j-1}q\left(U_{i,j-1} - \left(pU_{i-1,j-1} + qU_{i,j-2}\right)\right) \\
& = & \left(1 - d_{i,j}\right)\left(U_{i,j} - \left(pU_{i-1,j} + qU_{i,j-1}\right)\right) \\
& + & pd_{i-1,j}\left(U_{i-1,j} - \left(pU_{i-2,j} + qU_{i-1,j-1}\right)\right) \\
& + & qd_{i,j-1}\left(U_{i,j-1} - \left(pU_{i-1,j-1} + qU_{i,j-2}\right)\right).
\end{array}$$

One obtains the inequality

$$\begin{array}{rcl}
\left|\overline{U}_{i,j} - \left(p\overline{U}_{i-1,j} + q\overline{U}_{i,j-1}\right)\right| & \leq & \left(1 - d_{i,j}\right)\left|U_{i,j} - \left(pU_{i-1,j} + qU_{i,j-1}\right)\right| \\
& + & pd_{i-1,j}\left|U_{i-1,j} - \left(pU_{i-2,j} + qU_{i-1,j-1}\right)\right| \\
& + & qd_{i,j-1}\left|U_{i,j-1} - \left(pU_{i-1,j-1} + qU_{i,j-2}\right)\right|.
\end{array}$$

Summation over all indices yields

$$\sum_{i,j}\left|\overline{U}_{i,j} - \left(p\overline{U}_{i-1,j} + q\overline{U}_{i,j-1}\right)\right| \leq \sum_{i,j}\left(\left(1 - d_{i,j}\right) + pd_{i,j} + qd_{i,j}\right)\left|U_{i,j} - \left(pU_{i-1,j} + qU_{i,j-1}\right)\right|$$

which yields the claim since $p + q = 1$. □

The upwind scheme is LVD, indeed it admits the explicit formulation $\overline{U}_{i,j} = (1-\nu)U_{i,j} + \nu\left(pU_{i-1,j} + qU_{i,j-1}\right)$ which satisfies the criterion of Lemma 13. The two-dimensional Lax-Wendroff fluxes (30) generate a scheme which is not LVD for $0 < \nu < 1$: this is evident, since the Lax-Wendroff scheme is not TVD in dimension one which corresponds to $p = 1$ and $q = 0$. The interesting question is to find a constructive method for the design of the classical Finite Volume numerical fluxes on the edges $U^n_{i+\frac{1}{2},j}$ and $U^n_{i,j+\frac{1}{2}}$ such that the LVD criterion is satisfies. Below we use the method proposed by Lagoutière.

Lemma 14. *Assume the usual CFL condition $0 < \nu \leq 1$. Then the upwind fluxes and Lax-Wendroff fluxes (30) satisfy for all i, j*

$$m_{i,j} \leq pU_{i-\frac{1}{2},j} + qU_{i,j-\frac{1}{2}} \leq M_{i,j}. \tag{44}$$

Proof. Evident. □

Lemma 15. *Consider the explicit version of the scheme (24)*

$$\overline{U}_{i,j} = U_{i,j} - \nu\left(pU_{i+\frac{1}{2},j} + qU_{i,j+\frac{1}{2}}\right) + \nu\left(pU_{i-\frac{1}{2},j} + qU_{i,j-\frac{1}{2}}\right). \tag{45}$$

Assume the numerical fluxes satisfies two family of inequalities which are, firstly the incoming conditions (44) for all i, j, and secondly the following outgoing conditions for all i, j

$$\begin{cases} U_{i,j} + \frac{1-\nu}{\nu}\left(U_{i,j} - M_{i,j}\right) \le pU_{i+\frac{1}{2},j} + qU_{i,j+\frac{1}{2}}, \\ U_{i,j} + \frac{1-\nu}{\nu}\left(U_{i,j} - m_{i,j}\right) \ge pU_{i+\frac{1}{2},j} + qU_{i,j+\frac{1}{2}}. \end{cases} \tag{46}$$

Then $|\overline{U}|_{\mathrm{LBV}} \le |U|_{\mathrm{LBV}}$.

Inequalities (44) are called incoming because they concern incoming fluxes $U_{i-\frac{1}{2},j}$ and $U_{i,j-\frac{1}{2}}$ in a given cell with indices i and j. Inequalities (46) are called outgoing because they concern outgoing fluxes $U_{i+\frac{1}{2},j}$ and $U_{i,j+\frac{1}{2}}$. The first inequalities can also be interpreted as a kind of consistency requirement of the incoming flux.

Proof of Lemma 15. From (45), (44) and (46), one obtains the upper bound

$$\overline{U}_{i,j} \le U_{i,j} - \nu\left(U_{i,j} + \frac{1-\nu}{\nu}\left(U_{i,j} - M_{i,j}\right)\right) + \nu M_{i,j}$$

that is $\overline{U}_{i,j} \le M_{i,j}$ after simplifications. Similarly one has

$$\overline{U}_{i,j} \ge U_{i,j} - \nu\left(U_{i,j} + \frac{1-\nu}{\nu}\left(U_{i,j} - m_{i,j}\right)\right) + \nu m_{i,j}.$$

that is $\overline{U}_{i,j} \ge m_{i,j}$ after simplifications. □

Lemma 16. *Assume $0 < \nu \le 1$. Then the upwind fluxes satisfy all inequalities (44) and (46).*

Proof. The bounds (44) are already verified in Lemma 14. The remaining inequalities (46) are evident because the upwind fluxes yield $pU_{i+\frac{1}{2},j} + qU_{i,j+\frac{1}{2}} = U_{i,j}$. □

5.1.2. *Explicit formula for $\varphi_{i,j}$*

Next we consider the scheme (34) defined by the minmod type limiters

$$\varphi_{i,j} = \mathrm{minmod}\left(1, r_{i-\frac{1}{2},j}, r_{i,j-\frac{1}{2}}\right) \tag{47}$$

where

$$\begin{cases} r_{i-\frac{1}{2},j} = \dfrac{U_{i,j} - pU_{i-1,j} - qU_{i,j-1}}{U_{i-1,j} - pU_{i-2,j} - qU_{i-1,j-1}}, \\ r_{i,j-\frac{1}{2}} = \dfrac{U_{i,j} - pU_{i-1,j} - qU_{i,j-1}}{U_{i,j-1} - pU_{i-1,j-1} - qU_{i,j-2}} \end{cases} \tag{48}$$

and by

$$\varphi_{i+\frac{1}{2},j+\frac{1}{2}} = 0. \tag{49}$$

The latter means that the correction terms at corners are not taken into account. They will be reintroduced later.

Lemma 17. *The scheme (34-35) with (47-49) satisfies the special form (41) of the maximum principle, is LVD and is stable in quadratic norm. It is formally only first order accurate.*

Proof. By definition, the requirements (36) are satisfied so quadratic stability holds. The scheme is only first-order accurate because of (49), that is the corner contributions miss with respect to the formulation (31-33) of the second-order accurate Lax-Wendroff scheme.

To show that it is LVD and satisfies the maximum principle under the form (41), one relies on the verification of of (44-46). One has

$$pU_{i-\frac{1}{2},j}+qU_{i,j-\frac{1}{2}}=pU_{i-1,j}+qU_{i,j-1}+\varphi_{i,j}\left(U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right)$$

$$=\left(1-\varphi_{i,j}\right)\left(U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right)+\varphi_{i,j}U_{i,j}\in[m_{i,j},M_{i,j}]$$

so the double inequality (44) holds for the total incoming flux.

Moreover one has by construction (47-48) that

$$\varphi_{i+1,j}\left(U_{i+1,j}-pU_{i,j}-qU_{i+1,j-1}\right)=\alpha_{i,j}\left(U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right)\quad 0\leq\alpha_{i,j}\leq 1$$

and

$$\varphi_{i+1,j}\left(U_{i+1,j}-pU_{i,j}-qU_{i+1,j-1}\right)=\beta_{i,j}\left(U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right)\quad 0\leq\beta_{i,j}\leq 1.$$

So one can write

$$pU_{i+\frac{1}{2},j}+qU_{i,j+\frac{1}{2}}=U_{i,j}+\frac{1-\nu}{2}\left(p\alpha_{i,j}+q\beta_{i,j}\right)\left(U_{i,j}-pU_{i-1,j}-qU_{i,j-1}\right).$$

Then one has the lower bound

$$pU_{i+\frac{1}{2},j}+qU_{i,j+\frac{1}{2}}\geq U_{i,j}+\frac{1-\nu}{2}\left(p\alpha_{i,j}+q\beta_{i,j}\right)\left(U_{i,j}-M_{i,j}\right)\geq U_{i,j}+\frac{1-\nu}{2}\left(U_{i,j}-M_{i,j}\right)$$

from which the first inequality of (46) is deduced. Similar manipulations yield the second inequality of (46). Therefore the scheme is LVD. □

Next we desire to show the asymptotic value $\varphi_{i,j}\approx 1$ for smooth solutions. This is the first step is showing that the scheme is formally second order, which is one element of the refutation of the Goodman-Leveque obstruction Theorem. The second step will to construct the corner limiter $\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\approx 1$, contrary to (49) which is analyzed in this section. We consider the smooth solution u issued from an initial data (2) and the discrete data obtained by interpolation

$$U_{i,j}=u(i\Delta x,j\Delta x)\text{ for all }(i,j).\tag{50}$$

Lemma 18. *Consider the scheme (34-35) with the limiters (47-49) and with the initialization (2-50). Assume the solution is locally non characteristic*

$$\left(a\partial_x u+b\partial_y u\right)(i\Delta x,j\Delta x)\neq 0.$$

Then $\varphi_{i,j}=1+O(\Delta x)$.

Proof. We note $x_i=i\Delta x$ and $y_j=j\Delta x$. From (48), a Taylor expansion shows

$$U_{i,j}-pU_{i-1,j}-qU_{i,j-1}=\Delta x\left(p\partial_x u+q\partial_y u\right)(x_i,y_j)+O(\Delta x^2).$$

Of course one also has $U_{i-1,j}-pU_{i-2,j}-qU_{i-1,j-1}=\Delta x\left(p\partial_x u+q\partial_y u\right)(x_i,y_j)+O(\Delta x^2)$. The non characteristic condition yields that $r_{i-\frac{1}{2},j}=1+O(\Delta x)$. Similarly $r_{i,j-\frac{1}{2}}=1+O(\Delta x)$. The claim is obtained by (47). □

5.2. **Construction of corner-based limiters** $\varphi_{i+\frac{1}{2},j+\frac{1}{2}}$

The LVD requirement is restricted to the definition $\varphi_{i,j}$ and is a poor interest for the design of $\varphi_{i+\frac{1}{2},j+\frac{1}{2}}$. This is why a different method is used to construct explicitly the corner-based limiters in a way such that $\varphi_{i+\frac{1}{2},j+\frac{1}{2}}\approx 1$ for smooth solutions.

Since it is important for many applications, we concentrate hereafter on the fulfillment of the maximum principle for the numerical solution at the end of the time step (as in many references cited in the introduction). That is we ask that

$$m_{i,j}^{\rm abs} = \min\left(U_{i,j}, U_{i-1,j}, U_{i,j-1}\right) \leq \overline{U}_{i,j} \leq M_{i,j}^{\rm abs} = \max\left(U_{i,j}, U_{i-1,j}, U_{i,j-1}\right) \tag{51}$$

where $\overline{U}_{i,j}$ at the end of the time step if obtained from the numerical solution defined at (47-48) which is denoted as $\widehat{U}_{i,j}$. That is $\widehat{U}_{i,j}$ takes of the standard fluxes through the edges while $\overline{U}_{i,j}$ incorporates the corner corrections. Note that $m_{i,j}^{\rm abs} \leq m_{i,j} \leq U_{i,j} \leq M_{i,j} \leq M_{i,j}^{\rm abs}$.

The formula for $\widehat{U}_{i,j}$ is

$$\frac{\widehat{U}_{i,j} - U_{i,j}}{\Delta t} + p\frac{U_{i+\frac{1}{2},j}^{\rm lim} - U_{i-\frac{1}{2},j}^{\rm lim}}{\Delta x} + q\frac{U_{i,j+\frac{1}{2}}^{\rm lim} - U_{i,j-\frac{1}{2}}^{\rm lim}}{\Delta x} = 0. \tag{52}$$

It is followed by corner corrections

$$\frac{\overline{U}_{i,j} - \widehat{U}_{i,j}}{\Delta t} + \frac{pq}{2}\frac{C_{i-\frac{1}{2},j+\frac{1}{2}}^{\rm lim} - C_{i+\frac{1}{2},j-\frac{1}{2}}^{\rm lim}}{\Delta x} = 0. \tag{53}$$

One gets from (52-53)

$$C_{i-\frac{1}{2},j+\frac{1}{2}}^{\rm lim} - C_{i+\frac{1}{2},j-\frac{1}{2}}^{\rm lim} = \frac{2}{pq\nu}\left(\widehat{U}_{i,j} - \overline{U}_{i,j}\right).$$

To fulfill (51) we require

$$\left\{\begin{array}{l} \frac{1}{pq\nu}\left(\widehat{U}_{i,j} - M_{i,j}^{\rm abs}\right) \leq C_{i-\frac{1}{2},j+\frac{1}{2}}^{\rm lim} \leq \frac{1}{pq\nu}\left(\widehat{U}_{i,j} - m_{i,j}^{\rm abs}\right), \\ \frac{1}{pq\nu}\left(\widehat{U}_{i,j} - M_{i,j}^{\rm abs}\right) \leq -C_{i+\frac{1}{2},j-\frac{1}{2}}^{\rm lim} \leq \frac{1}{pq\nu}\left(\widehat{U}_{i,j} - m_{i,j}^{\rm abs}\right). \end{array}\right. \tag{54}$$

It leads to a definition of the corner-based limiters

$$\varphi_{i-\frac{1}{2},j+\frac{1}{2}} = \min\left(\varphi_{i-\frac{1}{2},j+\frac{1}{2}}^{+}, \varphi_{i-\frac{1}{2},j+\frac{1}{2}}^{-}\right) \tag{55}$$

where

$$\left\{\begin{array}{l} \varphi_{i-\frac{1}{2},j+\frac{1}{2}}^{+} = {\rm minmod}\left(1, \frac{\frac{1}{pq\nu}\left(\widehat{U}_{i,j} - m_{i,j}^{\rm abs}\right)}{\Delta_{i-\frac{1}{2},j+\frac{1}{2}}}\right) + {\rm minmod}\left(1, \frac{\frac{1}{pq\nu}\left(\widehat{U}_{i,j} - M_{i,j}^{\rm abs}\right)}{\Delta_{i-\frac{1}{2},j+\frac{1}{2}}}\right), \\ \varphi_{i+\frac{1}{2},j-\frac{1}{2}}^{-} = {\rm minmod}\left(1, \frac{\frac{1}{pq\nu}\left(\widehat{U}_{i,j} - m_{i,j}^{\rm abs}\right)}{-\Delta_{i+\frac{1}{2},j-\frac{1}{2}}}\right) + {\rm minmod}\left(1, \frac{\frac{1}{pq\nu}\left(\widehat{U}_{i,j} - M_{i,j}^{\rm abs}\right)}{-\Delta_{i+\frac{1}{2},j-\frac{1}{2}}}\right). \end{array}\right. \tag{56}$$

Note that $\widehat{U}_{i,j} - m_{i,j}^{\rm abs} \geq 0$ and $\widehat{U}_{i,j} - M_{i,j}^{\rm abs} \leq 0$. Therefore one term vanishes systematically in both lines of (56).

Lemma 19. *The new scheme with corner limitation (56) satisfies the maximum principle. Assume the solution is locally not an extrema*

$$|\partial_x u(x_i, y_j)| + |\partial_y u(x_i, y_j)| > 0.$$

Then it is formally second-order.

Proof. For the simplicity of the analysis, we consider that the previous stage for the construction of $\widehat{U}_{i,j}$ is performed with optimal limitation, that is $\varphi_{i,j} \equiv 1$. Then the previous optimally removes the diffusion in the direction of the flow visible in the modified equation (28). So one can write

$$\widehat{U}_{i,j} = u(x_i, y_j) - \Delta x\nu(p\partial_x + q\partial_y)u(x_i, y_j) + pq\frac{\Delta x^2\nu}{2}\left(\partial_x - \partial_y\right)^2 u(x_i, y_j) + O(\Delta x^3).$$

Let us analyze the first ratio $R = \frac{\frac{1}{pq\nu}\left(\widehat{U}_{i,j} - m_{i,j}^{\text{abs}}\right)}{\Delta_{i-\frac{1}{2},j+\frac{1}{2}}}$. For the simplicity of the analysis we assume that $\partial_y u(x_i, y_j) > \partial_x u(x_i, y_j)$ which is not a restriction. One can write $m_{i,j}^{\text{abs}} = u(x_i, y_j) + \Delta x \min\left(0, -\partial_x u(x_i, y_j), -\partial_y u(x_i, y_j)\right) + O(\Delta x^2)$ that is

$$m_{i,j}^{\text{abs}} = u(x_i, y_j) - \Delta x \max\left(0, \partial_y u(x_i, y_j)\right) + O(\Delta x^2).$$

So

$$\widehat{U}_{i,j} - m_{i,j}^{\text{abs}} = \Delta x \left(\max\left(0, \partial_y u(x_i, y_j)\right) - \nu(p\partial_x + q\partial_y)u(x_i, y_j)\right) + O(\Delta x^2).$$

One also has

$$\Delta_{i-\frac{1}{2},j+\frac{1}{2}} = -\Delta x \left(\partial_x - \partial_y\right) u(x_i, y_j) + O(\Delta x^2)$$

One gets the approximation

$$R = \frac{\max\left(0, \partial_y u(x_i, y_j)\right) - \nu(p\partial_x + q\partial_y)u(x_i, y_j)}{pq\nu\left(\partial_y u(x_i, y_j) - \partial_x u(x_i, y_j)\right)} + O(\Delta x).$$

It can be reorganized as

$$R = \frac{1}{q} + \frac{\max\left(0, \partial_y u(x_i, y_j)\right) - \nu\partial_y u(x_i, y_j)}{pq\nu\left(\partial_y u(x_i, y_j) - \partial_x u(x_i, y_j)\right)} + O(\Delta x)$$

Asymptotically for $\Delta x \to 0$ and $q < 1$, one has $R \geq \frac{1}{q} > 1$. Considering (56) one obtains $\varphi^+_{i-\frac{1}{2},j+\frac{1}{2}} = 1$ for Δx small enough. Other cases are similar so one gets that $\varphi^+_{i\pm\frac{1}{2},j\mp\frac{1}{2}} = 1$ for small Δx. One recovers the Lax-Wendroff scheme at the limit which is second order in space and time. □

At the end of this construction, the scheme (34-35) with the edge limiters (47-49) and the corner limiters (56) is convergent in quadratic norm with an order not less than $O(\Delta x^{\frac{1}{2}})$, is formally second order away from characteristic points and satisfies the maximum principle.

6. Numerical illustrations

We provide some numerical illustrations of the capabilities of the numerical method that has been developed in the above theoretical Sections. The results show without ambiguity that the $O(\Delta x^{\frac{1}{2}})$ is pessimistic. We compare the results calculated with the Lax-Wendroff scheme (30-33) and with the new nonlinear scheme (34-35)+(47-49)+(56). The error is reported in L^1, L^2 and L^∞ norms with a procedure explained in Section 6.1.

The domain of calculation is the academic square with periodic boundary conditions, i.e. the torus $\mathbb{T} = [0,1]^2_{\text{per}}$. In what follows we report error measurements in function of the mesh size in different norms, and deduce the order of convergence. One observes that the order of convergence of the nonlinear scheme is between 1 and the order of convergence of the Lax-Wendroff scheme for the first three test problems with smooth exact solutions. It seems to be the price to pay to obtain the maximum principle. However for the last test for which the initial data is the indicatrix function of a square, the initial data has not the smoothness required by (2). On the contrary it is a BV profile and the solution to the equation must be interpreted in the weak sense. In this case the new nonlinear scheme is more accurate that the Lax-Wendroff scheme, both in L^1 and L^2 norms.

6.1. **Simple test**

The initial data is $u_0(x, y) = \cos 2\pi(x + 2y)$. The velocity vector is defined by $p = q = 1/2$. The final time of observation is $T = 2$, so the final solution is equal to the initial solution, that is $u(T) = u_0$. The CFL is $\nu = 0.25$. The asymptotic order of accuracy is approximated with the formula

$$\text{order} \approx \frac{\log(E(\Delta x) - \log(E(\Delta x/2)}{\log 2}.$$

where $E(\Delta x)$ and $\log(E(\Delta x/2)$ are evaluated with the two finest meshes. The results are in Table 1.

One observes order 2 with the Lax-Wendroff scheme and on order between 1 and 1.5 for the nonlinear scheme. This was expected because it is a known behavior of the 1D minmod scheme. Indeed the accuracy is formally shown to be equal to 2, but away from characteristic points. Around characteristic points, the accuracy is much less because the limitation procedure is activated to satisfy the maximum principle.

	LW L^1	LW L^2	LW L^∞	NL L^1	NL L^2	NL L^∞
Δx =1/20	0.825	0.815	0.809	0.706	0.710	0.730
Δx =1/40	0.219	0.219	0.219	0.258	0.289	0.345
Δx =1/80	0.0553	0.0553	0.0552	0.0994	0.1078	0.147
Δx =1/160	0.0138	0.0138	0.0138	0.0344	0.0395	0.0629
order $\approx$	2	2	2	1.46	1.45	1.22

TABLE 1. Error measurements for $u_0(x, y) = \cos 2\pi(x + 2y)$.

6.2. **Stationary solution**

The velocity vector is still $p = q = 1/2$. Here $u_0(x, y) = \cos 2\pi(x - y)$ which is a stationary solution $u(t) = u_0$ for all $t \geq 0$ since $(p\partial_x + q\partial_y)u_0 \equiv 0$. Such non trivial stationary solution cannot exist in 1D. The final time of observation is $T = 2$. The CFL is $\nu = 0.25$. The results are provided in Table 2. The order of convergence is around 3 for the 2 schemes in all norms, except for the nonlinear scheme for which the order is $\approx \frac{5}{2}$ in maximum norm.

	LW L^1	LW L^2	LW L^∞	NL L^1	NL L^2	NL L^∞
Δx =1/20	0.00597	0.00597	0.00597	0.0143	0.0150	0.0203
Δx =1/40	0.00075	0.00075	0.00075	0.00189	0.00232	0.00421
Δx =1/80	0.000095	0.000095	0.000095	0.00022	0.00030	0.00070
order $\approx$	2.98	2.98	2.98	3.1	2.95	2.58

TABLE 2. Error measurements for $u_0(x, y) = \cos 2\pi(x - y)$.

6.3. **A Gaussian**

This problem is more challenging for the Lax-Wendroff scheme which develops natural oscillations because this scheme is not able to respect the maximum principle for steep profiles. The velocity vector is now given by $p = 1/3$ and $q = 2/3$. The CFL is $\nu = 0.5$. The initial data is $u_0(x, y) = \exp(-100r^2)$ with $r^2 = (x - 0.5)^2 + (y - 0.5)^2$. A plot of the solutions is displayed in Figure 1 on the coarse mesh at $T = 3$. It shows the oscillation of the Lax-Wendroff scheme. On the contrary the new nonlinear scheme respects perfectly the maximum principle, in accordance with its theoretical properties. The error are displayed in Table 3 with similar conclusions as for the previous tests.

	LW L^1	LW L^2	LW L^∞	NL L^1	NL L^2	NL L^∞
Δx =1/40	0.823	0.499	0.438	0.461	0.386	0.454
Δx =1/80	0.221	0.182	0.182	0.195	0.161	0.186
Δx =1/160	0.0557	0.0492	0.0535	0.0594	0.0517	0.0731
Δx =1/320	0.0139	0.0124	0.0136	0.0170	0.0164	0.0312
order $\approx$	2	1.98	1.97	1.80	1.65	1.22

TABLE 3. Error measurements for the Gaussian.

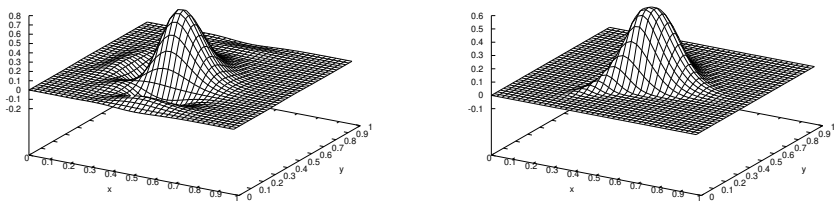

FIGURE 1. Advection of Gaussian. On the left calculated with the linear Lax-Wendroff scheme. On the right calculated with limitation of the fluxes. Th maximum principle is not respected with the linear scheme. It is satisfied with the limitation procedure.

6.4. **Indicatrix function**

The interest of an unconditional respect of the maximum principle is better exemplified with an initial data which is an indicatrix function. Indeed, high-order linear schemes cannot satisfy the maximum principle and in practice, important oscillations are generated. Here we choose $u_0(x,y) = 1$ if $\max(|x-0.5|, |y-0.5|) < 0.2$ and $u_0(x,y) = 0$ otherwise. It is the indicatrix function of a square. The other datas are as in the two first tests. The solutions are shown in Figure 2. The errors are reported in Table 4. No convergence is observed in L^∞ because the solution is a weak solution with BV regularity only. The nonlinear scheme is more accurate than the Lax-Wendroff scheme in L^1 and L^2 norms. It seems to convergence at order $\frac{2}{3}$ in L^1 and $\frac{1}{3}$ in L^2. This fact is easy to interpret by interpolation of L^2 between L^1 and L^∞.

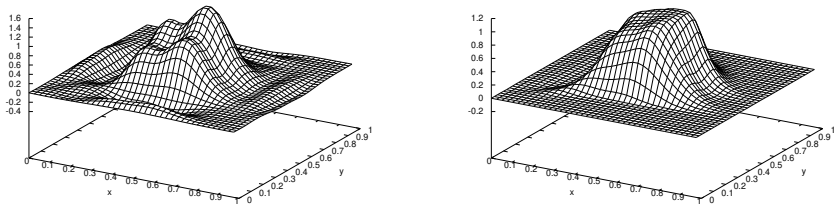

FIGURE 2. Advection of square indicatrix function. On the left calculated with the linear Lax-Wendroff scheme. On the right calculated with limitation of the fluxes. Th maximum principle is not respected with the linear scheme. It is satisfied with the limitation procedure.

7. POSSIBLE EXTENSIONS

Corner corrections were developed in this work only for the purposes of having compact notations amenable for the refutation of the Goodman-Leveque obstruction theorem by means of quadratically stable limiters. We do not known if corner corrections are logically necessary for the obtention of quadratic stability of flux limitation techniques. Hereafter we evoke two possibilities to extend the modified equation to more interesting problems.

	LW L^1	LW L^2	LW L^∞	NL L^1	NL L^2	NL L^∞
Δx =1/20	0.856	0.487	0.773	0.718	0.479	0.767
Δx =1/40	0.713	0.470	0.857	0.454	0.381	0.791
Δx =1/80	0.508	0.381	0.860	0.280	0.297	0.801
Δx =1/160	0.346	0.305	0.869	0.174	0.233	0.808
Δx =1/320	0.234	0.245	0.876	0.109	0.184	0.812
Δx =1/640	0.157	0.197	0.882	0.0683	0.145	0.814
order $\approx$	0.57	0.31	0	0.67	0.34	0

TABLE 4. Error measurements for the indicatrix function.

Let us firstly consider the 2D equation with smooth variable coefficients

$$\partial_t u + \partial_x(a(y)u) + \partial_y(b(x)u) = \partial_t u + (a(y)\partial_x + b(x)\partial_y b(x))u = 0.$$

This equation is relevant in plasma physics [3, 13, 15, 25]. It is easy to check that the equivalent equation of the upwind scheme on a cartesian grid is

$$\partial_t u + (a(y)\partial_x + b(x)\partial_y)u - \frac{\Delta x}{2}\partial_x(|a(y)|\partial_x u) - \frac{\Delta x}{2}\partial_y(|b(x)|\partial_y u) + \frac{\Delta t}{2}(\partial_x a(y) + \partial_y b(x))^2 = 0.$$

With this formulation, it is easy to construct a conservative (or divergent) Lax-Wendroff like scheme with variable coefficients by modifying (34). One can for exemple discretize p and q at half index positions. More complicated coefficients $a(x, y)$ and $b(x, y)$ are possible also.

The other situation concerns a 3D equation with constant coefficients on a cartesian grid $\partial_t u + p\partial_x u + q\partial_y u + r\partial_z u = 0$ where $p, q, r \geq 0$ and $p + q + r = 1$. The modified equation of the upwind scheme is

$$\partial_t u + (p\partial_x + q\partial_y + r\partial_z)u - \frac{\Delta x}{2}\left(p\partial_{xx} + q\partial_{yy} + r\partial_{zz}\right)u + \frac{\Delta t}{2}(p\partial_x + q\partial_y + r\partial_z)^2 u = 0.$$

It can be rewritten as

$$\partial_t u + (p\partial_x + q\partial_y + r\partial_z)u - \frac{\Delta x}{2}(1-\nu)(p\partial_x + q\partial_y + r\partial_z)^2 u$$

$$-\frac{\Delta x}{2}pq(\partial_x - \partial_y)^2 u - \frac{\Delta x}{2}pr(\partial_x - \partial_z)^2 u - \frac{\Delta x}{2}qr(\partial_y - \partial_z)^2 u = 0.$$

The development of corner correction techniques in 3D should be possible from this formulation. We also think of using the recent theory [4] in order to get more insights into the tensorial nature of the anisotropic diffusion operator.

The methods and proofs were developed on a cartesian grid, but for the sole purposes of the simplicity of the mathematical developments. It is reasonable to foresee that transport on unstructured grids can also be addressed with quadratically stable limiters. It will nevertheless require a convenient mathematical apparatus to generalize the quadratic forms to such more challenging configurations.

REFERENCES

[1] T. Barth and D.C. Jespersen, The design and application of upwind schemes on unstructured meshes, American Institute for Aeronautics and Astronautics, Report 89-0366:1-12, 1989.

[2] T. Barth and M. Ohlberger, Finite Volume methods: foundation and analysis, Encyclopedia of Computational Mechanics, John Wiley & Sons, 2004.

[3] J. Bernier, F. Casas and N. Crouseilles, Nicolas, Splitting methods for rotations: application to Vlasov equations. SIAM J. Sci. Comput. 42-2, A666-A697, 2020.

[4] F. Bonnans, G. Bonnet and J.-M. Mirebeau, Second order monotone finite differences discretization of linear anisotropic differential operators, Hal preprint hal-03084046v2, 2021.
[5] F. Bouchut, F. James and S. Mancini, Uniqueness and weak stability for multi-dimensional transport equations with one-sided Lipschitz coefficient. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 4 -1, 1-25, 2005.
[6] W. Cao, C.-W. Shu, Y. Yang and Z. Zang, Superconvergence of discontinuous Galerkin method for linear hyperbolic equations in one space dimension, SIAM Journal on Numerical Analysis 56(2):732-765, 2018.
[7] N. Crouseilles, M. Mehrenberger, E. Sonnendrucker, Conservative semi-Lagrangian schemes for Vlasov equations, Journal of Computational Physics, Volume 229, Issue 6, Pages 1927-1953, 2010.
[8] B. Després, Numerical Methods for Eulerian and Lagrangian Conservation Laws, Springer Verlag, 2017.
[9] B. Després, Lax Theorem and Finite Volume schemes, Math. of Comp., Volume 73, Number 247, 1203-1234, 2003.
[10] B. Despres, Uniform asymptotic stability of Strang's explicit compact schemes for linear advection, SIAM J. Numer. Anal. 47, no. 5, 3956-3976, 2009.
[11] B. Després and F. Lagoutière, Generalized Harten formalism and longitudinal variation diminishing schemes for linear advection on arbitrary grids. M2AN Math. Model. Numer. Anal. 35, 6, 1159-1183, 2001.
[12] B. Després and F. Lagoutière, A longitudinal variation diminishing estimate for linear advection on arbitrary grids. C. R. Acad. Sci. Paris Sér. I Math. 332 (2001), no. 3, 259–263.
[13] L. Einkemmer and A. Ostermann., Convergence analysis of a discontinuous Galerkin/Strang splitting approximation for the Vlasov-Poisson equations. SIAM J. Numer. Anal. 52-22, 757-778, 2014.
[14] R. Eymard, T. Gallouet R. Herbin, Finite volume methods. Handbook of Numerical Analysis, North Holland: Amsterdam, 7:713-1020, 2000.
[15] F. Filbet and E. Sonnendrücker, Comparison of Eulerian Vlasov solvers. Comput. Phys. Comm. 150-3, 247-266, 2003.
[16] E. Godlewski and P.-A. Raviart, Hyperbolic systems of conservation laws, Mathematics and Applications 3/4, Ellipses, Paris, 1991.
[17] E. Godlewski and P.-A. Raviart, Numerical Approximation of Hyperbolic Systems of Conservation Laws, Applied Mathematical Sciences, 118, Springer, 1996.
[18] S. Godunov, A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations, Mat. Sbornik, 47, 271-306, translated US Joint Publ. Res. Service, JPRS 7226, 1969.
[19] J. Goodman and R.J. LeVeque, On the accuracy of stable schemes for 2D scalar conservation laws. Math. Comp. 45, 171, 15-21, 1985.
[20] J.-L. Guermond, M. Maier, B. Popov and I. Tomas, Second-order invariant domain preserving approximation of the compressible Navier-Stokes equations, Comput. Methods Applied Mech. Engin., 375, 2021.
[21] J.-L. Guermond, B. Popov and J. Ragusa, Positive and Asymptotic Preserving Approximation of the Radiation Transport Equation, SIAM J. Numer. Anal., 58 1, 519-540, 2020.
[22] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput. Phys., 49 (2): 357-393, 1983.
[23] A. Harten, J. M. Hyman and P. D. Lax, On finite-difference approximations and entropy conditions for shocks, Comm. Pure Appl. Math., 29, 297-322, 1976.
[24] A. Kurganov, Y. Liu and V. Zeitlin, Numerical dissipation switch for two-dimensional central-upwind schemes, M2AN Volume 55, Number 3, 713-734, 2021.
[25] G. Latu, M. Mehrenberger, Y. Guclu, M. Ottaviani and E. Sonnendrucker, Field-aligned interpolation for semi-Lagrangian gyrokinetic simulations. J. Sci. Comput. 74 (2018), no. 3, 1601–1650.
[26] R. J. Leveque, Finite Volume Methods for Hyperbolic Problems, Cambridge University Press, 2012.
[27] J. Lu, Y. Liu and C.-W. Shu, An oscillation-free discontinuous Galerkin method for scalar hyperbolic conservation laws, SIAM Journal on Numerical Analysis 59(3), 1299-1324. ,2021.
[28] P.L. Roe and D. Sidilkover, Optimum positive linear schemes for advection in two and three dimensions SIAM journal on numerical analysis 29 (6), 1542-1568, 1992.
[29] C.-W. Shu, Bound-preserving high order finite volume schemes for conservation laws and convection-diffusion equations, in Finite volumes for complex applications VIII, vol. 199 of Springer, 2017, pp. 3-14.
[30] P.K. Sweby, High resolution schemes using flux-limiters for hyperbolic conservation laws, SIAM J. Numer. Anal., 21 (5): 995-1011, 1984.
[31] E. F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics- A Practical Introduction, Springer, 2009.
[32] X. Zhang and C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments, Proc. R. Soc., 467, 2752-2776, 2001.