Finite volume L^∞-stability for hyperbolic scalar problems

Stéphane Clain

Noname manuscript No.
(will be inserted by the editor)

Finite volume L^∞-stability for hyperbolic scalar problems

Stéphane Clain

the date of receipt and acceptance should be inserted later

Abstract A new formalism and tools are proposed to characterize high-order reconstructions in the finite volume method context. We introduce the notion of admissible reconstruction and investigate the maximum principle and positivity preserving properties for scalar hyperbolic problem using the new formalism. We show that the traditional limiting strategies cast in out formalism and provide new proves of the L^∞ stability.

Keywords finite volume scheme · L^∞ stability · maximum principle · high-order reconstruction · positivity preserving

1 Introduction

L^∞ stability is a fundamental property for scalar autonomous hyperbolic problem since the entropic solution has to respect the maximum principle [7,23] (MP property). Therefore, it seems desirable to design numerical schemes which also achieve such a statement and the maximum principle property has to be satisfied at the numerical level. The fundamental concept is the flux monotonicity property which provides the L^∞ stability. It is well established [16] that under an appropriated CFL condition depending on the numerical flux and the mesh characteristics, an explicit finite volume scheme (Euler forward time discretization for example) provides an approximation which respects the maximum principle. The main drawback of monotone numerical schemes is that we can only obtain first-order schemes characterized by a large amount of diffusion in the shock vicinity leading to an important (sometime severely) solution discrepancy. Higher-order

Stéphane Clain
Institut de Mathématiques de Toulouse, UMR CNRS 5219
118 route de Narbonne, 31062 Toulouse cedex, France
E-mail: stephane.clain@math.univ-toulouse.fr

reconstructions based on a local linear representation like the MUSCL method [4, 13] or polynomial reconstruction like ENO or WENO technique (see the review of Shu [20]) are very popular and employed in a wide panel of engineering problems. The major drawback is that we generally loose the maximum principle unless some restrictions (limiting procedures for instance) are employed to recover the MP property.
Important efforts have been realized during the four last decades upon the subject. Since the seventies, second-order reconstructions have been introduced in the 1D uniform mesh context and extensions have been developed in the eighties for 2D or 3D structured meshes. At last, in the early nineties adaptations to unstructured meshes have be realized. All the methods are mainly based on a spatial polynomial reconstruction coupling by a limiting procedure to enforce the maximum principle (MUSCL,ENO, WAF). A high-order TVD time scheme (Runge-Kuta method for instance) is then required to preserve the MP property [21]. More recent methods like ADER [8, 25] also use a time reconstruction strategy to provide a relevant high-order approximation.

Initiated with the series of paper of Van Leer [26], the limiting procedure for one-dimensional problems was based on the Total Variation Diminishing property leading to the second-order TVD schemes introduced by Harten [11, 12] (see also Sweby [24] and Boris and Book [5]). Unfortunately, Goodman and LeVeque [9] show that the TVD criterion is not adequate for higher dimension since a scheme which preserves the BV norm is reduced to a first-order one. A new concept for multi-dimensional hyperbolic scalar problem was then introduced by several authors: the positive coefficient schemes –or shortly positive schemes, also mentioned as monotone scheme in the Spekreijse paper [22]. The concept was firstly tackled by Jameson and Lax in [15] for one-dimensional uniform meshes but it is Spekreijse [22] who has really developed the idea of positive coefficient scheme for structured two-dimensional meshes. Extension for unstructured meshes (cell-centered version) was then proposed by Jameson [14] introducing the notion of Local Extrema Diminishing property. Basically, the updated value in cell C should be a convex combination of the former values situated in the vicinity of the cell. Such a property reveal to be easy to handle and investigations have been tackled to prove that specific reconstruction like the MUSCL one can be rewritten as a positive coefficient scheme (see for example the course of Barth [1] or Barth and Ohlberger [2]).

In the present paper, we follow a similar way since we prove the maximum principle property using the positive coefficient scheme approach. Nevertheless, we shall not deal with specific reconstruction but propose a generic framework introduced Clain and Clauzon [6]. Roughly speaking, we highlight two fundamental properties that a reconstruction operator should deserve to preserve the MP property without considering the manner the reconstruction is achieved (linear, quadratic representation). We show, as an example, that the classical MUSCL reconstructions cast in our general framework where we recover the stability results of Barth [1] and Park, Yoon, Kim [18]. An other point we shall deserve in the paper concerns the positivity preserving property for a class of hyperbolic system

such that the Euler isentropic problem. A surprising result proved by Perthame and Shu [19] or Linde and Roe [17] shows that a first-order positivity preserving numerical scheme turns to be automatically a second-order positivity preserving scheme as long as the reconstruction is obtained via a linear reconstruction –in fact the CFL condition has to be altered and time steps should sometime be very small in comparison with the space step. As a conclusion, the limiting procedure is not necessary to preserve the positivity with linear reconstruction. For more general reconstruction, up to the author knowledge, there is no positivity preserving results. We show here that the two fundamental properties upon which the MP property lies, lead to the positivity preserving property for the conservative variables.

The paper is organized as follow. Section 2 is dedicated to high-order reconstruction where the two fundamental properties we shall employ in the sequel are defined. We then prove general maximum principal theorems independently of the manner the reconstruction is achieved. In the third section, we apply our general theorems to some popular reconstruction where we recover the classical stability results of Barth. At last, we consider the positivity preserving question in section 4 where we highlight the link with the MP property.

2 A general L^∞-stability result

We introduce in this section the new formalism we propose to analyse high-order reconstruction. The fundamental point is the notion of admissible reconstruction where we highlight the two properties that a reconstruction have to respect. We then prove that the L^∞-stability property stems from the two properties.

2.1 Mesh

We denote by $\mathcal{T}$ a conform mesh of $\mathbb{R}^2$ constituted of a collection of close non overlapping polygonal cells K_i, $i \in \mathcal{E}_{el}$ covering the whole space $\mathbb{R}^2$ and we denote by P_m, $m \in \mathcal{E}_{nd}$ the nodes (see figure 1). For any $K_i \in \mathcal{T}$, the set $\underline{\nu}(i) \subset \mathcal{E}_{el}$ contains the index j of elements $K_j \in \mathcal{T}$ which share a common side represented by $S_{ij} = K_i \cap K_j$. In the same way, the set $\overline{\nu}(i) \subset \mathcal{E}_{el}$ contains all the index j of elements $K_j \in \mathcal{T}$ such that $K_i \cap K_j \neq \emptyset$. In other word, $\bigcup_{j\in\overline{\nu}(i)} K_j$ is the corona formed by all the elements in contact with K_i. We also denote by $\mu(i)$ any intermediate index set such that

$$\underline{\nu}(i) \subset \mu(i) \subset \overline{\nu}(i) \text{ and } N_\mu = \max_{i\in\mathcal{E}_{el}} \#\mu(i).$$

At last, the subset $\lambda(i) \subset \mathcal{E}_{nd}$ represents the index set of the K_i the nodes. The quantities $|K_i|$, $|S_{ij}|$, $|PB|$ refer to the surface of K_i, the length of S_{ij} and segment $[PB]$ for any points B and P. Moreover, $\text{perim}(K_i) = \sum_{j\in\underline{\nu}(i)} |S_{ij}|$ is the perimeter of element K_i.

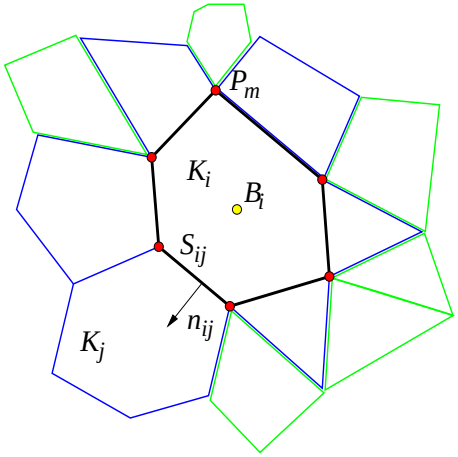

Fig. 1 Mesh notations. Index set $\underline{\nu}(i)$ corresponds to the blue elements while $\overline{\nu}(i)$ corresponds to the union of blue and green elements. Index set $\lambda(i)$ represents the red nodes.

In general, meshes can be very different from one to another hence we shall consider classes of meshes $\mathcal{M}(\alpha)$ characterized by a structural parameter α defined in the following (see definition 4 for example) whereas h is the mesh size parameter given by

$$h = \min_{\substack{K_i \in \mathcal{T} \\ j \in \underline{\nu}(i)}} \frac{|K_i|}{|S_{ij}|}. \tag{1}$$

The key point to distinguish the two parameters is that estimations involve coefficients which only depend on α and not on h. Moreover, we can easily exhibit sequences of meshes $\mathcal{T}_{h_k} \in \mathcal{M}(\alpha)$ such that $h_k \to 0$ with the same structural parameter.

Remark 1 Non conform meshes could be also considered but more complex notations should be employed to generalize the stability results. Therefore, we prefer to restrict the study to the conform mesh case for the sake of simplicity. □

2.2 Generic first-order monotone scheme

We consider a general scalar hyperbolic problem cast in the conservative form

$$\partial_t u + \partial_{x_1} f_1(u) + \partial_{x_2} f_2(u) = 0, \tag{2}$$

where f_1 and f_2 are C^1 real value functions defined on $\mathbb{R}$ which can be reduced to the admissible domain of solution u if necessary. Let $u^0 \in L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$,

$u \in L^\infty(\mathbb{R}^2 \times]0,+\infty[) \cap C^0([0,\infty[, L^1(\mathbb{R}^2))$ is a solution if u satisfies equation (2) in a weak sense with the initial condition $u(.,t=0) = u^0$ as in [7].

We now detail the numerical approximation. For a given time t^n and a cell $K_i \in \mathcal{T}$, we denote by

$$u_i^n \approx \frac{1}{|K_i|} \int_{K_i} u(.,t^n)\, dx$$

an approximation of the mean value of u on cell $K_i \in \mathcal{T}$ at time t^n while the initial condition is given by

$$u_i^0 = \frac{1}{|K_i|} \int_{K_i} u^0\, dx. \tag{3}$$

The generic first-order explicit finite volume scheme provides an approximation at time $t^{n+1} = t^n + \Delta t$ by

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} g(u_i^n, u_j^n, n_{ij}). \tag{4}$$

where $g(u_i, u_j, n_{ij})$ is the numerical flux across S_{ij} following the outward normal vector direction n_{ij}. We assume that the numerical flux satisfies the following properties:

(a) *regularity:* function g is continuous, differentiable with respect to the first and the second argument and $\partial_1 g$, $\partial_2 g$ are continuous functions;
(b) *consistency:* the numerical flux is consistent with the physical flux (f_1, f_2):

$$g(u_i, u_i, n_{ij}) = f_1(u_i) n_{ij,1} + f_2(u_i) n_{ij,2}; \tag{5}$$

(c) *monotony:* the numerical flux is monotone:

$$\partial_1 g(u_i, u_j, n_{ij}) \geq 0, \qquad \partial_2 g(u_i, u_j, n_{ij}) \leq 0. \tag{6}$$

Note that the consistency condition implies the conservation property

$$\sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} g(u_i, u_i, n_{ij}) = 0. \tag{7}$$

Remark 2 The flux conservation across the interface is usually satisfied by the numerical flux:

$$g(u_i, u_j, n_{ij}) = -g(u_j, u_i, n_{ji}). \tag{8}$$

which implies equivalence

$$\partial_1 g(u_i, u_j, n_{ij})) \geq 0 \Leftrightarrow \partial_2 g(u_i, u_j, n_{ij}) \leq 0.$$

Nevertheless, the flux conservation property is not necessary to provide the stability of the scheme and, as we shall see, only properties (5), (6) and (7) are required. □

Remark 3 We can restrict the regularity assumption on g requiring that $\partial_1 g$ and $\partial_2 g$ are only bounded functions. □

Remark 4 Since the domain is the whole plane $\mathbb{R}^2$, we do not introduce any boundary condition in order to simplify our analysis. Nevertheless, one can consider bounded close domain Ω with reflexion condition for example using the ghost cells technique as in [6] where a virtual mesh is employed. □

2.3 Reconstructions

Let $u \in L^\infty(\mathbb{R}^2)$, we denote by u_i an approximation of the mean value of u on element $K_i \in \mathcal{T}$ and u_h corresponds to the constant piecewise representation given by

$$u_h = \sum_{i \in \mathcal{E}_{el}} u_i \mathbb{1}_{K_i},$$

where $\mathbb{1}_{K_i} = 1$ on K_i and zero-value elsewhere.
The reconstruction operator provides new real values on sides S_{ij} using the values u_i on cells K_i. Formally, for a given mesh $\mathcal{T}$ we define the one-point reconstruction operator $\mathcal{R}(\mathcal{T})$ by

$$(u_i)_{K_i \in \mathcal{T}} \overset{\mathcal{R}(\mathcal{T})}{\rightarrow} (u_{ij})_{K_i \in \mathcal{T},\ j \in \underline{\nu}(i)}, \tag{9}$$

where u_{ij} and u_{ji} corresponds to approximations of u on both sides of S_{ij} since the reconstruction is discontinuous across the interface.

Remark 5 We can also considered a R-points reconstruction

$$(u_i)_{K_i \in \mathcal{T}} \overset{\mathcal{R}(\mathcal{T})}{\rightarrow} (u_{ij,r})_{K_i \in \mathcal{T},\ j \in \underline{\nu}(i),\ r=1,\dots,R}$$

where the values $u_{ij,r}$ and $u_{ji,r}$ correspond to approximations of u at several collocation points $X_{ij,k}$ on the side S_{ij} (the Gauss points for instance). For the sake of simplicity, we only present the stability results for the one-point reconstruction case but extension will be mentioned for the R-points reconstruction. □

Remark 6 Note that we do not require the reconstruction to satisfy some consistent property with the gradient *i.e.* to be exact with linear functions for instance. Of course, such a property is desirable if one would like to construct a second-order scheme but it is not necessary in the stability context. □

We now define the new finite volume scheme with the one-point reconstruction setting

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} g(u_{ij}^n, u_{ji}^n, n_{ij}). \tag{10}$$

For any constant piecewise function u_h^n on $\mathcal{T}$, we define the Euler forward scheme operator

$$u_h^n \rightarrow u_h^{n+1} = \mathcal{H}(u_h^n) = \mathcal{H}(u_h^n; \mathcal{T}, \mathcal{R})$$

where u_i^{n+1} is given by relation (10) on each element $K_i \in \mathcal{T}$.

Remark 7 Extension to R-points reconstruction can also be defined with

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \sum_{r=1}^{R} \zeta_r g(u_{ij,r}^n, u_{ji,r}^n, n_{ij}). \tag{11}$$

where ζ_r are non-negative convex weight coefficients for the numerical integration of $f(u).n_{ij}$ on side S_{ij} with $\sum_{r=1}^{R} \zeta_r = 1$. Consequently, relation (11) can be written as a convex combination of R relations of type (10) where we substitute u_{ij} and u_{ji} with $u_{ij,r}$ and $u_{ji,r}$ respectively. □

2.4 A maximum principle theorem

L^∞-stability property is based on the positive coefficient scheme concept. We want to write u_i^{n+1} as a mean of the former values at time t^n in the form

$$u_i^{n+1} = u_i + \sum_{j \in \mu(i)} \alpha_{ij}(u_j - u_i),$$

with $\alpha_{ij} \geq 0$ and $\sum_{j \in \mu(i)} \alpha_{ij} \leq 1$ where $\underline{\nu}(i) \subset \mu(i) \subset \overline{\nu}(i)$. To this end, we have to impose some specific conditions on the reconstruction leading to the notion of admissible reconstruction operator.

Definition 1 (μ-local descrete extrema) Let $K_i \in \mathcal{T}$, and $\mu(i)$ an index set such that $\underline{\nu}(i) \subset \mu(i) \subset \overline{\nu}(i)$. We define the μ-local discrete minimum and maximum on the stencil $\{i\} \cup \mu(i)$ by

$$m_{i,\mu}^n = \min_{j \in \mu(i)} (u_i^n, u_j^n), \qquad M_{i,\mu}^n = \max_{j \in \mu(i)} (u_i^n, u_j^n),$$

and we associate two corresponding indexes $k_m, k_M \in \{i\} \cup \mu(i)$ such that

$$u_{k_m} = m_{i,\mu}, \quad u_{k_M} = M_{i,\mu}. \square$$

Definition 2 (μ-admissible reconstruction operator) Let $\mathcal{T} \in \mathcal{M}(\alpha)$ with $\alpha > 0$. The reconstruction operator $\mathcal{R} = \mathcal{R}(\mathcal{T})$ is μ-admissible with respect to the structural parameter α if the two following properties are satisfied.

a) There exists $C_\theta = C_\theta(\alpha) \geq 0$ and coefficients θ_{ijk}, $i \in \mathcal{E}_{el}$, $j \in \underline{\nu}(i)$, $k \in \mu(i)$ with

$$0 \leq \theta_{ijk} \leq C_\theta,$$

such that

$$u_{ij} - u_j = \sum_{k \in \mu(i)} \theta_{ijk}(u_k - u_j). \tag{12}$$

b) There exists $C_\omega = C_\omega(\alpha) \geq 0$ and coefficients ω_{ijk}, $i \in \mathcal{E}_{el}$, $j \in \underline{\nu}(i)$, $k \in \mu(i)$ with

$$0 \leq \omega_{ijk} \leq C_\omega,$$

such that

$$u_{ij} - u_i = -\sum_{k \in \mu(i)} \omega_{ijk}(u_k - u_i). \qquad \square \tag{13}$$

The following more restrictive definition is also considered.

Definition 3 (convex μ-admissible reconstruction operator) The reconstruction operator $\mathcal{R} = \mathcal{R}(\mathcal{T})$ is said convex if definition 2 is satisfied with

$$\sum_{k \in \mu(i)} \theta_{ijk} = 1. \tag{14}$$

Remark 8 Note that relation (14) implies

$$u_{ij} = \sum_{k \in \mu(i)} \theta_{ijk} u_k. \tag{15}$$

and $C_\theta \leq 1$. $\square$

Remark 9 We precise that $\mathcal{R}$ is a μ-admissible reconstruction operator since we can choose coefficients θ_{ijk} and ω_{ijk} with $k \in \mu(i)$, *i.e* $\theta_{ijk} = \omega_{ijk} = 0$ if $k \in \overline{\nu}(i) \setminus \mu(i)$. $\square$

Remark 10 We do not have the uniqueness of coefficients θ_{ijk} and ω_{ijk}. Relations (12) and (13) can be obtained by different ways. For example, we can write any u_k, $k \in \{i\} \cup \mu(i)$ has a convex combination of u_{k_m} and u_{k_M} (see definition 1). It results that formulae (12) and (13) can be written under the form

$$\begin{aligned} u_{ij} - u_j &= \theta_{ijk_m}(u_{k_m} - u_j) + \theta_{ijk_M}(u_{k_M} - u_j), \\ u_{ij} - u_i &= -\omega_{ijk_m}(u_{k_m} - u_j) - \omega_{ijk_M}(u_{k_M} - u_j). \qquad \square \end{aligned}$$

We now establish the L^∞-stability theorems based on the μ-admissible reconstruction. To this end, let $\mathcal{T} \in \mathcal{M}(\alpha)$, $K_i \in \mathcal{T}$, then for any $m \leq M$ we define

$$C_g = \max_{\substack{\alpha,\beta \in [m,M] \\ |n|=1}} \{\partial_1 g(\alpha, \beta, n), -\partial_2 g(\alpha, \beta, n)\}. \tag{16}$$

The following lemma holds.

Lemma 1 (μ-local discrete maximum principle) *Let $\mathcal{R}$ be a μ-admissible reconstruction with respect to the structural parameter α. If $u_k^n \in [m, M]$, $k \in \{i\} \cup \mu(i)$, then u_i^{n+1} defined by relation (10) also belongs to $[m, M]$ under the CFL condition*

$$\Delta t < C_{fl}\, h, \text{ with } C_{fl} = \frac{1}{\#\underline{\nu}(i) \#\mu(i) C_g (C_\theta + C_\omega)}, \tag{17}$$

where constant C_g is given by relation (16) whereas constants C_θ, C_ω only depend on the structural parameter α.

Proof The finite volume scheme (10) with the one-point reconstruction writes:

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij})$$

Using the conservative property, we have

$$\begin{aligned} u_i^{n+1} &= u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \left[\mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}) - \mathcal{F}(u_i^n, u_i^n, n_{ij}) \right], \\ &= u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \left[\partial_1 \mathcal{F}(\widetilde{u}_{ij}^n, \widehat{u}_{ij}^n, n_{ij})(u_{ij}^n - u_i^n) + \partial_2 \mathcal{F}(\widetilde{u}_{ij}^n, \widehat{u}_{ij}^n, n_{ij})(u_{ji}^n - u_i^n) \right], \end{aligned}$$

with $\widetilde{u}_{ij}^n = \lambda_{ij} u_{ij}^n + (1-\lambda_{ij}) u_i^n$ and $\widehat{u}_{ij}^n = \lambda_{ij} u_{ji}^n + (1-\lambda_{ij}) u_i^n$ where $\lambda_{ij} \in]0,1[$.
From relations (12) and (13) we deduce

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \Big[- A_{ij} \sum_{k \in \mu(i)} \omega_{ijk} (u_k^n - u_i^n) + B_{ij} \sum_{k \in \mu(i)} \theta_{jik} (u_k^n - u_i^n) \Big],$$

where we have defined

$$A_{ij} = \partial_1 \mathcal{F}(\widetilde{u}_{ij}^n, \widehat{u}_{ij}^n, n_{ij}), \quad B_{ij} = \partial_2 \mathcal{F}(\widetilde{u}_{ij}^n, \widehat{u}_{ij}^n, n_{ij}).$$

After some algebraic manipulations, we obtain

$$u_i^{n+1} = u_i^n - \Delta t \sum_{k \in \mu(i)} \Big[\sum_{j \in \underline{\nu}(i)} -A_{ij} \frac{|S_{ij}|}{|K_i|} \omega_{ijk} + B_{ij} \frac{|S_{ij}|}{|K_i|} \theta_{jik} \Big] (u_k^n - u_i^n).$$

Setting

$$\Theta_{ik} = \sum_{j \in \underline{\nu}(i)} A_{ij} \frac{|S_{ij}|}{|K_i|} \omega_{ijk} - B_{ij} \frac{|S_{ij}|}{|K_i|} \theta_{jik},$$

thanks to the monotonicity of the numerical flux, we have $0 \leq \Theta_{ik} \leq \dfrac{C_m}{h}$ with

$$C_m = \#\underline{\nu}(i) C_g (C_\theta + C_\omega).$$

We rewrite the relation

$$u_i^{n+1} = (1 - \sum_{k \in \mu(i)} \Delta t \Theta_{ik}) u_i^n + \sum_{k \in \mu(i)} \Delta t \Theta_{ik} u_k^n$$

and we get a convex combination with positive coefficient if we satisfy the CFL condition

$$\frac{\Delta t}{h} \#\mu(i) \#\underline{\nu}(i) C_g (C_\theta + C_\omega) \leq 1.$$

Hence $u_i^{n+1} \in [m, M]$ $\square$

Remark 11 We have the μ-local discrete maximum principle applying lemma 1 with $m = m_{i,\mu}$ and $M = M_{i,\mu}$. □

Remark 12 A sufficient condition to provide the L^∞ stability is that the $\Theta_{i,k}$ coefficients are non negative and uniformly bounded while the admissible reconstruction operator condition is more restrictive. Nevertheless, we shall see in the sequel that such a condition cover a very large class of limiters and reconstructions. □

We now state the main theorems of the section where we focus on two particular cases: $\mu(i) = \overline{\nu}(i)$ and $\mu(i) = \underline{\nu}(i)$.

Theorem 1 ($\overline{\nu}$-global discrete maximum principle) *Let $\mathcal{R}$ be a $\overline{\nu}$-admissible reconstruction with respect to the structural parameter α. We consider the second-order finite volume scheme (10) with the initial condition (3) and set $m = \min\{u^0(x),\, x \in \mathbb{R}^2\}$, $M = \max\{u^0(x),\, x \in \mathbb{R}^2\}$.*
If Δt satisfies the CFL condition

$$\Delta t < \overline{C}_{fl}\, h, \text{ with } \overline{C}_{fl} = \frac{1}{N_{\overline{\nu}} N_{\underline{\nu}} C_g (C_\theta + C_\omega)}, \tag{18}$$

then $u_i^n \in [m, M]$ for all $K_i \in \mathcal{T}$ and $t^n \geq 0$. □

Proof We make the proof by induction. The property holds for $t = t^0$ since the mean values are always between the minimum and the maximum of u^0. Assume now that the property holds at time t^n, lemma 1 says that $u_i^{n+1} \in [m, M]$ for any element $K_i \in \mathcal{T}$ if Δt satisfies the CFL condition (17). By definition $N_{\overline{\nu}} = \max_i(\#\overline{\nu}(i))$ and $N_{\underline{\nu}} = \max_i(\#\underline{\nu}(i))$ hence $\overline{C}_{fl} < C_{fl}$. Consequently, the time step controlled by relation (18) is also controlled by relation (17) thus $u_i^{n+1} \in [m, M]$ for any element $K_i \in \mathcal{T}$. □

In the same way, we have the following theorem.

Theorem 2 ($\underline{\nu}$-global discrete maximum principle) *Let $\mathcal{R}$ be a $\underline{\nu}$-admissible reconstruction with respect to the structural parameter α. We consider the second-order finite volume scheme (10) with the initial condition (3) and set $m = \min\{u^0(x),\, x \in \mathbb{R}^2\}$, $M = \max\{u^0(x),\, x \in \mathbb{R}^2\}$.*
If Δt satisfies the CFL condition

$$\Delta t < \underline{C}_{fl}\, h, \text{ with } \underline{C}_{fl} = \frac{1}{N_{\underline{\nu}} N_{\underline{\nu}} C_g (C_\theta + C_\omega)}, \tag{19}$$

then $u_i^n \in [m, M]$ for all $K_i \in \mathcal{T}$ and $t^n \geq 0$. □

Remark 13 Higher-order time discretization schemes based on convex combinations of the explicit Euler time discretization also satisfy the maximum principle under an appropriate CFL condition. For example, the third-order TVD Runge-Kutta scheme writes

$$u_h^{n,1} = \mathcal{H}(u_h^n), \quad u_h^{n,2} = \frac{3}{4}u_h^n + \frac{1}{4}\mathcal{H}(u_h^{n,1}), \quad u_h^{n+1} = \frac{1}{3}u_h^n + \frac{2}{3}\mathcal{H}(u_h^{n,2}).$$

Each step satisfies the maximum principle hence the global scheme satisfy the maximum principle since we employ convex combinations. □

Remark 14 The CFL constant are too restrictive with respect to the practical numerical experiences. Indeed, we have over-estimated the number of non-vanishing coefficients θ_{ijk} and ω_{ijk}. As suggested in remark 10, the number of non vanishing coefficient can be reduced to 2 with respect to the local maximum and the minimum value on the stencil. Consequently, a less restrictive CFL constant would be

$$\Delta t < C_{fl}\ h, \text{ with } C_{fl} = \frac{1}{2N_{\underline{\nu}}C_g(C_\theta + C_\omega)}. \qquad \square \tag{20}$$

3 Application to classical MUSCL methods

The goal of the section is to cast the classical limiting methods into the general form proposed in the previous section. We consider the most useful limiting reconstructions employed in the literature and determine their associated coefficient θ_{ijk} and ω_{ijk} in order to apply the L^∞-stability theorem. One of the first reconstruction operator on unstructured meshes has been proposed by Barth and Jespersen [3] based on the $\underline{\nu}$ stencil. The original method has been developed with triangle but we here present the stability result for more general elements. A recent extension have been developed by Park, Yoon and Kim [18] where, this time, the $\overline{\nu}$ stencil has to be used to provide the L^∞-stability. At this stage, we outline an important remark. We have to distinguish two kinds of points: the **control points** where the limiting procedure is applied and the **collocation points** where the reconstructed values are computed: the goal is to prove the L^∞-stability at the collocation points when using the limiting procedure on the controls points. Subsection 3.1 is dedicated to the L^∞-stability when one employes the nodes as control points while subsection 3.2 deals with the general case when control points generate a convex hull around the cell.

In the remainder part of the section, we shall only consider linear conservative reconstructions on elements K_i of the form

$$\widehat{u}_i(X) = u_i + a_i.B_iX$$

where B_i is the centroid of K_i and $a_i \in \mathbb{R}^2$ is the slope of the reconstruction, usually an approximation of $\nabla u(B_i)$ in cell K_i. Other kinds of reconstruction can also be considered like the multislope MUSCL method [6].

3.1 Limiting process with nodes as control points

Let $\mathcal{T}$ be a mesh of $\mathbb{R}^2$, for any elements $K_i \in \mathcal{T}$, we denote by P_m, $m \in \lambda(i)$ the nodes of element K_i (see figure 1). Let B_i be the centroid of element K_i

characterized by the barycentric coordinates in function of the nodes

$$B_i = \sum_{m \in \lambda(i)} \rho_{im} P_m$$

with $\sum_{m \in \lambda(i)} \rho_{im} = 1$. Note that we do not have *a priori* a unique set of barycentric coordinates for each B_i. For example, we can evaluate the barycentric coordinates using only three nodes and set the other coefficients to zero.

Definition 4 (structural coefficient with the nodes) Let $\alpha > 0$, we say that $\mathcal{T}$ belongs to $\mathcal{M}(\alpha)$, if and only if, for any $K_i \in \mathcal{T}$, there exists a set of barycentric coordinates with respect to the nodes such that

$$\min_{\substack{K_i \in \mathcal{T} \\ m \in \lambda(i)}} \rho_{im} \geq \alpha. \qquad \square \tag{21}$$

Remark 15 Since $\rho_{im} \geq 0$ and $\sum_{m \in \lambda(i)} \rho_{im} = 1$, we must have $\alpha \leq \dfrac{1}{N_\lambda}$ with

$$N_\lambda = \max_i \#\lambda(i).$$

For example we can choose $\rho_{im} = \dfrac{1}{\#\lambda(i)}$ and $\alpha = \dfrac{1}{N_\lambda}$. $\square$

Remark 16 Condition (21) may be relaxed in practice. Indeed, we can also foresee a limiting process where we consider a subset of nodes. In that case, the index set $\lambda(i)$ is reduced to the index set where we intend to evaluate the limiting condition. For example, if the element K_i is not convex, we can use the nodes of the convex hull which is a subset of the nodes (see subsection 3.2).
In the other hand, due to the linearity of the reconstruction the minimum and the maximum of $\widehat{u}_i(X)$ are reached on two distinct nodes (said P_{m_1} and P_{m_2}). The index set can be reduced to the two indexes m_1 and m_2 and a third index $m_3 \in \lambda(i)$ such that P_{m_1}, P_{m_2} and P_{m_3} defines a triangle which contains B_i but such a choice depend on the local linear reconstruction, hence of u_h. $\square$

Limiting procedure is performed at the node points but we have to decide what kind of maximum principle we want to respect. Indeed, the choice of the elements (index subset $\mu(i)$ in definition 1) where we seek the minimum and the maximum has to be fixed. In this paper, we propose a study for two extremes cases $\mu(i) =_u ndernu(i)$ and $\mu(i) = \overline{\nu}(i)$.

Lemma 2 (limiter with the $\underline{\nu}$ stencil) *Let $\mathcal{T} \in \mathcal{M}(\alpha)$, for any $K_i \in \mathcal{T}$, we assume that a_i satisfies the property: for all nodes P_m, $m \in \lambda(i)$ we have (see definition 1)*

$$m_{i,\underline{\nu}} = u_{k_m} \leq u_i + a_i . B_i P_m \leq u_{k_M} = M_{i,\underline{\nu}}. \tag{22}$$

Then there exist coefficients $\theta_{ik}(P_m)$, $\omega_{ik}(P_m)$, $k \in \underline{\nu}(i)$ *and constant value* $C_\omega = C_\omega(\alpha) \geq 0$ *which satisfy the following properties*

$$\widehat{u}_i(P_m) = \sum_{k \in \underline{\nu}(i)} \theta_{ik}(P_m) u_k, \qquad \sum_{k \in \underline{\nu}(i)} \theta_{ik}(P_m) = 1, \tag{23}$$

and

$$\widehat{u}_i(P_m) - u_i = - \sum_{k \in \underline{\nu}(i)} \omega_{ik}(P_m)(u_k - u_i), \quad 0 \leq \omega_{ik}(P_m) \leq C_\omega, \tag{24}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

Proof We give the construction for each class of coefficients.
Coefficients θ_{ik}. Condition (22) yields that

$$u_{k_m} < \widehat{u}_i(P_m) < u_{k_M}$$

hence there exists $\chi = \chi(P_m) \in [0, 1]$ such that

$$\widehat{u}_i(P_m) = \chi u_{k_m} + (1 - \chi) u_{k_M}.$$

We then write $\widehat{u}_i(P_m)$ as a convex combination with $\theta_{ik_m}(P_m) = \chi(P_m)$ and $\theta_{ik_M}(P_m) = 1 - \chi(P_m)$ while the other coefficients are set to zero.
coefficient ω_{ik}. Using the barycentric coordinates, we write

$$0 = \sum_{m' \in \lambda(i)} \rho_{im'} B_i P_{m'}.$$

We distinguish the particular node P_m and we obtain

$$B_i P_m = - \sum_{\substack{m' \in \lambda(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} B_i P_{m'},$$

where $\rho_{im} \geq \alpha > 0$. Since we consider a linear reconstruction at point P_m, we have

$$\widehat{u}_i(P_m) - u_i = a_i . B_i P_m = - \sum_{\substack{m' \in \lambda(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} a_i . B_i P_{m'}.$$

Condition (22) yields

$$u_{k_m} - u_i < a_i . B_i P_{m'} < u_{k_M} - u_i, \qquad \forall m' \in \lambda(i),$$

then

$$- \sum_{\substack{m' \in \lambda(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} (u_{k_M} - u_i) < \widehat{u}_i(P_m) - u_i < - \sum_{\substack{m' \in \lambda(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} (u_{k_m} - u_i),$$

hence

$$-\frac{1-\rho_{im}}{\rho_{im}}(u_{k_M}-u_i) < \widehat{u}_i(P_m) - u_i < -\frac{1-\rho_{im}}{\rho_{im}}(u_{k_m}-u_i).$$

There exists $\chi(P_m) \in [0,1]$ such that

$$\widehat{u}_i(P_m) - u_i = -\widetilde{\chi}\left(\frac{1-\rho_{im}}{\rho_{im}}\right)(u_{k_m}-u_i) - (1-\widetilde{\chi})\left(\frac{1-\rho_{im}}{\rho_{im}}\right)(u_{k_M}-u_i).$$

Relation (24) holds if we choose

$$\omega_{ik_m}(P_m) = \widetilde{\chi}(P_m)\left(\frac{1-\rho_{im}}{\rho_{im}}\right), \qquad \omega_{ik_M}(P_m) = (1-\widetilde{\chi}(P_m))\left(\frac{1-\rho_{im}}{\rho_{im}}\right),$$

and the other coefficients are set to zero. Since $1 \geq \rho_{im} \geq \alpha$, we have the estimate $C_\omega(\alpha) = \dfrac{1}{\alpha}$. □

Remark 17 Note that a_i do not have to be consistent with the gradient. We only use the stability condition and the linearity of the reconstruction. □

As we outline in the beginning of the section, control points and collocation points where reconstruction is performed may be different. For example, one can control the limiter with the nodes and shall compute the reconstructed values on the side midpoint $M_{ij} \in S_{ij}$. The following proposition says how we can choose the collocation points keep preserving the L^∞-stability.

Proposition 1 *For any $X \in K_i$, there exist coefficients $\theta_{ik}(X)$, $\omega_{ik}(X)$, $k \in \underline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties*

$$\widehat{u}_i(X) = \sum_{k\in\underline{\nu}(i)} \theta_{ik}(X)u_k, \qquad \sum_{k\in\underline{\nu}(i)} \theta_{ik}(X) = 1, \tag{25}$$

and

$$\widehat{u}_i(X) - u_i = -\sum_{k\in\underline{\nu}(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega, \tag{26}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

Proof We write $X \in K_i$ as a convex combination of the nodes with non-negative coefficients $\zeta_m(X)$, $m \in \lambda(i)$

$$X = \sum_{m\in\lambda(i)} \zeta_m(X)P_m, \qquad \sum_{m\in\lambda(i)} \zeta_m(X) = 1.$$

Thanks to the linearity of the reconstruction, we have

$$\widehat{u}_i(X) = \sum_{m\in\lambda(i)} \zeta_m(X)\widehat{u}_i(P_m).$$

Consequently, relations (25) and (26) are satisfied with

$$\theta_{ik}(X) = \sum_{m\in\lambda(i)} \zeta_m(X)\theta_{ik}(P_m), \qquad \omega_{ik}(X) = \sum_{m\in\lambda(i)} \zeta_m(X)\omega_{ik}(P_m).$$

Moreover, we have

$$\begin{aligned}\sum_{k\in\underline{\nu}(i)} \theta_{ik}(X) &= \sum_{k\in\underline{\nu}(i)} \sum_{m\in\lambda(i)} \zeta_m(X)\theta_{ik}(P_m),\\ &= \sum_{m\in\lambda(i)} \zeta_m(X) \sum_{k\in\underline{\nu}(i)} \theta_{ik}(P_m),\\ &= \sum_{m\in\lambda(i)} \zeta_m(X) = 1.\end{aligned}$$

In the other hand we have

$$\omega_{ik}(X) = \sum_{m\in\lambda(i)} \zeta_m(X)\omega_{ik}(P_m) \leq \sum_{m\in\lambda(i)} \zeta_m(X)C_\omega(\alpha) \leq C_\omega(\alpha),$$

hence estimates $C_\omega(\alpha) = \dfrac{1}{\alpha}$ still holds. □

We now treat the case where we use the $\overline{\nu}$ index set to control the reconstruction. Since arguments are very similar, we just give the results.

Lemma 3 (limiter with the $\overline{\nu}$ stencil) *Let $\mathcal{T} \in \mathcal{M}(\alpha)$, for any $K_i \in \mathcal{T}$, we assume that a_i satisfies the property: for all nodes P_m, $m \in \lambda(i)$ we have (see definition 1)*

$$m_{i,\overline{\nu}} = u_{k_m} \leq u_i + a_i.B_iP_m \leq u_{k_M} = M_{i,\overline{\nu}}. \tag{27}$$

Then there exist coefficients $\theta_{ik}(P_m)$, $\omega_{ik}(P_m)$, $k \in \overline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties

$$\widehat{u}_i(P_m) = \sum_{k\in\overline{\nu}(i)} \theta_{ik}(P_m)u_k, \qquad \sum_{k\in\overline{\nu}(i)} \theta_{ik}(P_m) = 1, \tag{28}$$

and

$$\widehat{u}_i(P_m) - u_i = -\sum_{k\in\overline{\nu}(i)} \omega_{ik}(P_m)(u_k - u_i), \quad 0 \leq \omega_{ik}(P_m) \leq C_\omega, \tag{29}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

The following proposition also holds.

Proposition 2 *For any $X \in K_i$, there exist coefficients $\theta_{ik}(X)$, $\omega_{ik}(X)$, $k \in \overline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties*

$$\widehat{u}_i(X) = \sum_{k\in\overline{\nu}(i)} \theta_{ik}(X)u_k, \qquad \sum_{k\in\overline{\nu}(i)} \theta_{ik}(X) = 1, \tag{30}$$

and

$$\widehat{u}_i(X) - u_i = -\sum_{k\in\overline{\nu}(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega, \tag{31}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

3.1.1 Barth-Jespersen limiter on nodes [AIAA 1989]

In [3], Barth and Jespersen propose a MUSCL reconstruction where the limiting procedure is carried out on the nodes using the index set $\underline{\nu}(i)$ to enforce the maximum principle. The numerical routine is summarized by the following steps.

1. On each element K_i a predicted gradient $\widehat{a}_i$ is computed with the values on the neighbouring element K_j, $j \in \underline{\nu}(i)$.
2. The limiting procedure is applied to compute a new slope $a_i = \phi_i \widehat{a}_i$ where the limiting coefficient $\phi_i \in [0,1]$ is evaluated such that relation (22) at the nodes points.
3. The reconstructed values on side S_{ij} are given by

$$u_{ij} = u_i + a_i.B_iX_{ij}, \qquad u_{ji} = u_j + a_j.B_jX_{ij}, \tag{32}$$

   where X_{ij} are given collocation points on sides S_{ij}, $j \in \underline{\nu}(i)$.
4. The update approximation u_h^{n+1} is evaluated with relation (10) and a monotone numerical flux.

We sum-up the stability property in the following proposition.

Proposition 3 (L^∞-stability for the Barth-Jespersen recontruction) *Assume that the mesh $\mathcal{T}$ belongs to the class $\mathcal{M}(\alpha)$, then the maximum principle holds under the CFL condition (20) with $C_\theta = 1$ and $C_\omega = \frac{1}{\alpha}$.*

Proof Since condition (22) is satisfied, proposition 1 holds for any $X_{ij} \in K_i$. Setting

$$\theta_{ijk} = \theta_{ik}(X_{ij}), \qquad \omega_{ijk} = \omega_{ik}(X_{ij}),$$

then we have define a convex $\underline{\nu}$-admissible reconstruction operator (see definition 3) with $C_\theta = 1$ and $C_\omega = \frac{1}{\alpha}$. Theorem 2 gives the L^∞-stability under the CFL condition (19). □

3.1.2 The Park-Yoon-Kim limiter on nodes

In [18], Park, Yoon and Kim, propose an extension of the original Barth-Jesperson limiter we summarize in the following way.

1. On each element K_i a predicted gradient $\widehat{a}_i$ is computed with the values on the neighbouring element K_j, $j \in \underline{\nu}(i)$.
2. For each node P_m, $m \in \lambda(i)$, we denote by $\kappa(m)$ the index set of all the elements which contain node P_m and we set

$$u_{min,m} = \min_{j \in \kappa(m)} \{u_j\}, \qquad u_{max,m} = \max_{j \in \kappa(m)} \{u_j\}.$$

   We then define the coefficients $\phi_{im} \in [0,1]$, $m \in \lambda(i)$ such that

$$u_{min,m} \leq u_i + \phi_{im} a_i.B_iP_m \leq u_{max,m}. \tag{33}$$

3. We determine the new slope $a_i = \phi_i \widehat{a}_i$ with $\phi_i = \min_{m \in \lambda(i)} \phi_{im}$.
4. The reconstructed values u_{ij} and u_{ji} on side S_{ij} are computed by relation (32).
5. The update approximation u_h^{n+1} is evaluated with relation (10) and a monotone numerical flux.

Proposition 4 (L^∞-stability for the Park-Yoon-Kim recontruction) *Assume that the mesh $\mathcal{T}$ belongs to the class $\mathcal{M}(\alpha)$, assumption of lemma 3 is fulfilled and the maximum principle holds under the CFL condition (20) with $C_\theta = 1$ and $C_\omega = \frac{1}{\alpha}$.*

Proof Since $\bigcup_{m \in \lambda(i)} \kappa(m) = \overline{\nu}(i)$, then condition (33) yields

$$m_{i,\overline{\nu}} \leq u_i + a_i.B_iP_m \leq M_{i,\overline{\nu}}, \quad \forall m \in \lambda(i),$$

which corresponds to assumption (27) of lemma 3. Setting

$$\theta_{ijk} = \theta_{ik}(X_{ij}), \qquad \omega_{ijk} = \omega_{ik}(X_{ij}),$$

then we have define a convex $\overline{\nu}$-admissible reconstruction operator with $C_\omega = \frac{1}{\alpha}$. Theorem 1 gives the L^∞-stability under the CFL condition (18). □

3.2 Limiting process with a general convex hull

In the previous subsection, limiting process is achieved using the nodes as control points. A generalization consists in limiting the reconstruction with other control points than P_m, $m \in \lambda(i)$. For any cell K_i, we associate a set of control points $C_{im} \in \mathbb{R}^2$ with $m \in \delta(i)$ where $\delta(i)$ is a local index set. Note that the control point is not necessary a characteristic point of the mesh (node, centroid). Since we intend to limit the slope using the control points, we denote by ρ_{ij} the barycentric coordinates of B_i with respect to the control points C_{im}:

$$B_i = \sum_{m \in \delta(i)} \rho_{im} C_{im} \tag{34}$$

with $\sum_{m \in \delta(i)} \rho_{im} = 1$. Note that we do not have *a priori* a unique set of barycentric coordinate for each B_i. We now introduce a new definition of class mesh $\mathcal{M}(\alpha)$

Definition 5 (structural coefficient: general case) Let $\alpha > 0$, $\mathcal{T}$ belongs to $\mathcal{M}(\alpha)$, if and only if, there exists a set of barycentric coordinates such that

$$\min_{\substack{K_i \in \mathcal{T} \\ m \in \delta(i)}} \rho_{im} \geq \alpha. \qquad \square \tag{35}$$

It is important to note that the structural coefficient depend on the control points where we intend to limit the slope so the class $\mathcal{M}(\alpha)$ may change in function of the control points. For example, a particular interesting choice is $C_{ij} = M_{ij}$ the S_{ij} midpoint with $\delta(i) = \underline{\nu}(i)$.

Remark 18 A precise definition of class $\mathcal{M}(\alpha)$ should require the complete list of the control points as parameters like $\mathcal{M}(\alpha; C_{mi}, K_i \in \mathcal{T}, m \in \delta(i))$ but we omit to mention it for the sake of simplicity. □

We now prove the admissibility of the reconstruction with the following lemma.

Lemma 4 (general limiter with the $\underline{\nu}$ stencil) *Let $\mathcal{T} \in \mathcal{M}(\alpha)$ characterised by definition 4. For any $K_i \in \mathcal{T}$, we assume that a_i satisfies the property:*

$$m_{i,\underline{\nu}} = u_{k_m} \leq u_i + a_i.B_iC_{im} \leq u_{k_M} = M_{i,\underline{\nu}}, \quad \forall m \in \delta(i). \tag{36}$$

Then there exist coefficients $\theta_{ik}(C_{im}) \geq 0$, $\omega_{ik}(C_{im}) \geq 0$, $k \in \underline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties

$$\widehat{u}_i(C_{im}) = \sum_{k \in \underline{\nu}(i)} \theta_{ik}(C_{im}) u_k, \quad \sum_{k \in \underline{\nu}} \theta_{ik}(C_{im}) = 1, \tag{37}$$

and

$$\widehat{u}_i(C_{im}) - u_i = -\sum_{k \in \underline{\nu}(i)} \omega_{ik}(C_{im})(u_k - u_i), \quad 0 \leq \omega_{ijk}(C_{im}) \leq C_\omega. \tag{38}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

Proof We use the same technique given in the proof of lemma 2.
COEFFICIENTS $\theta_{ik}(C_{im})$. Condition (36) yields that for any $j \in \underline{\nu}(i)$

$$u_{k_m} < \widehat{u}_i(C_{im}) < u_{k_M}$$

then there exists $\chi = \chi(C_{im}) \in [0, 1]$ such that

$$\widehat{u}_i(C_{im}) = \chi u_{k_m} + (1 - \chi) u_{k_M}.$$

Hence $\widehat{u}_i(C_{im})$ is a convex combination with $\theta_{ik_m}(C_{im}) = \chi(C_{im})$, $\theta_{ik_M}(C_{im}) = 1 - \chi(C_{im})$ and the other coefficients set to zero.
COEFFICIENT $\omega_{ik}(C_{im})$. From relation(34) we write

$$0 = \sum_{j' \in \delta(i)} \rho_{im'} B_i C_{im'}.$$

We distinguish the particular control point C_{im} and we obtain

$$B_iC_{im} = -\sum_{\substack{m' \in \delta(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} B_iC_{im'},$$

where $\rho_{im} \geq \alpha > 0$. The reconstruction at point C_{im} satisfies the following equality

$$\widehat{u}_i(C_{im}) - u_i = a_i.B_iC_{im} = -\sum_{\substack{m' \in \delta(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}} a.B_iC_{im'}.$$

Condition (36) then yields

$$u_{k_m} - u_i < a.B_iC_{im'} < u_{k_M} - u_i, \qquad \forall m' \in \delta(i),$$

then

$$-\sum_{\substack{m' \in \delta(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}}(u_{k_M} - u_i) < \widehat{u}_i(C_{im}) - u_i < -\sum_{\substack{m' \in \delta(i) \\ m' \neq m}} \frac{\rho_{im'}}{\rho_{im}}(u_{k_m} - u_i),$$

hence

$$-\frac{1-\rho_{im}}{\rho_{im}}(u_{k_M} - u_i) < \widehat{u}_i(C_{im}) - u_i < -\frac{1-\rho_{im}}{\rho_{im}}(u_{k_m} - u_i).$$

Consequently, we can exhibit coefficient $\chi(C_{im}) \in [0,1]$ such that

$$\widehat{u}_i(C_{im}) - u_i = -\widetilde{\chi}\left(\frac{1-\rho_{im}}{\rho_{im}}\right)(u_{k_m} - u_i) - (1-\widetilde{\chi})\left(\frac{1-\rho_{im}}{\rho_{im}}\right)(u_{k_M} - u_i).$$

Relation (38) holds if we choose

$$\omega_{ik_m}(C_{im}) = \widetilde{\chi}(C_{im})\left(\frac{1-\rho_{im}}{\rho_{im}}\right), \qquad \omega_{ik_M}(C_{im}) = (1-\widetilde{\chi}(C_{im}))\left(\frac{1-\rho_{im}}{\rho_{im}}\right),$$

and the other coefficients set to zero. Since $1 \geq \rho_{im} \geq \alpha$, we have $C_\omega(\alpha) \leq \dfrac{1}{\alpha}$. □

We now denote by $\mathcal{C}_i$ the convex hull using point C_{im}, *i.e.* $X \in \mathcal{C}_i$ if there exists a convex combination of C_{im} such that

$$X = \sum_{m \in \delta(i)} \zeta_m(X) C_{im} \text{ with } \zeta_m(X) \geq 0 \text{ and } \sum_{m \in \delta(i)} \zeta_m(X) = 1.$$

As a consequence, the following proposition shows that the reconstruction preserves the L^∞-stability if one chooses the collocation points in the convex hull. The proof is similar to the one given in proposition 1.

Proposition 5 *Let us assume that the assumptions of lemma 4 are satisfied. For any $X \in \mathcal{C}_i$, there exist coefficients $\theta_{ik}(X) \geq 0$, $\omega_{ik}(X) \geq 0$, $k \in \underline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties*

$$\widehat{u}_i(X) = \sum_{k \in \underline{\nu}(i)} \theta_{ik}(X) u_k, \qquad \sum_{k \in \underline{\nu}(i)} \theta_{ik}(X) = 1, \tag{39}$$

and

$$\widehat{u}_i(X) - u_i = -\sum_{k \in \underline{\nu}(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega, \tag{40}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

We now consider the case when the maximum principle is based on the values defined by the $\overline{\nu}(i)$ index set.

Lemma 5 (general limiter with the $\overline{\nu}$ stencil) *Let $\mathcal{T} \in \mathcal{M}(\alpha)$ characterised by definition 4. For any $K_i \in \mathcal{T}$, we assume that a_i satisfies the property:*

$$m_{i,\overline{\nu}} = u_{k_m} \leq u_i + a_i.B_iC_{im} \leq u_{k_M} = M_{i,\overline{\nu}}, \qquad \forall m \in \delta(i). \tag{41}$$

Then there exist coefficients $\theta_{ik}(C_{im}) \geq 0$, $\omega_{ik}(C_{im}) \geq 0$, $k \in \overline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties

$$\widehat{u}_i(C_{im}) = \sum_{k \in \overline{\nu}(i)} \theta_{ijk}C_{im})u_k, \quad \sum_{k \in \overline{\nu}} \theta_{ik}(C_{im}) = 1, \tag{42}$$

and

$$\widehat{u}_i(C_{im}) - u_i = -\sum_{k \in \overline{\nu}(i)} \omega_{ik}(C_{im})(u_k - u_i), \quad 0 \leq \omega_{ik}(C_{im}) \leq C_\omega. \tag{43}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

As a consequence, the following proposition holds.

Proposition 6 *Let us assume that the assumptions of lemma 5 are satisfied. For any $X \in \mathcal{C}_i$, there exist coefficients $\theta_{ik}(X) \geq 0$, $\omega_{ik}(X) \geq 0$, $k \in \overline{\nu}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties*

$$\widehat{u}_i(X) = \sum_{k \in \overline{\nu}(i)} \theta_{ik}(X)u_k, \quad \sum_{k \in \overline{\nu}(i)} \theta_{ik}(X) = 1, \tag{44}$$

and

$$\widehat{u}_i(X) - u_i = -\sum_{k \in \overline{\nu}(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega, \tag{45}$$

with $C_\omega(\alpha) = \frac{1}{\alpha}$. □

3.2.1 The Barth limiter on side midpoints [VKI03]

In [1], Barth proposes a MUSCL reconstruction where the limiting procedure and the reconstruction are carried out at the same points: the control points and the colocation points are identical. Here, we only consider the useful situation with the side midpoints M_{ij} but the stability result holds for any point X_{ij} on side S_{ij}. The numerical routine is summarized by the following steps.

1. On each element K_i a predicted gradient $\widehat{a}_i$ is computed with the other values on the neighbouring element K_j, $j \in \underline{\nu}(i)$.
2. A limiting procedure is applied to compute a new slope $a_i = \phi_i\widehat{a}_i$ with $\phi_i \in [0, 1]$ such that for all M_{ij}, $j \in \underline{\nu}(i)$

$$m_{i,\underline{\nu}} = u_{k_m} \leq u_i + a_i.B_iM_{ij} \leq u_{k_M} = M_{i,\underline{\nu}}. \tag{46}$$

3. The reconstructed values are given by

$$u_{ij} = u_i + a_i.B_iM_{ij}, \qquad u_{ji} = u_j + a_j.B_jM_{ij}.$$

4. The update approximation u_h^{n+1} is evaluated with relation (10) and a monotone numerical flux.

Proposition 7 (L^∞-stability for the Barth-VKI03 recontruction) *Assume that the mesh $\mathcal{T}$ belongs to the class $\mathcal{M}(\alpha)$ associated to the control points M_{ij}, $j \in \underline{\nu}(i)$. Then the maximum principle holds under the CFL condition (19) with $C_\theta = 1$ and $C_\omega = \frac{1}{\alpha}$.*

Proof Condition (36) is satisfied with $\delta(i) = \underline{\nu}(i)$ on points M_{ij} hence proposition 5 holds. Setting

$$\theta_{ijk} = \theta_{ik}(M_{ij}), \qquad \omega_{ijk} = \omega_{ik}(M_{ij}),$$

then we have a convex $\underline{\nu}$-admissible reconstruction operator with $C_\theta = 1$ and $C_\omega = \frac{1}{\alpha}$. Theorem 2 gives the L^∞-stability under the CFL condition (19). □

Remark 19 Since we use the midpoints of the segment, we have $\alpha = \frac{1}{3}$ and $C_\theta = 1$, $C_\omega = 3$. □

Remark 20 A more restrictive condition is also considered in [1]:

$$\min(u_i, u_j) \leq u_i + a_i.B_iM_{ij} \leq \max(u_i, u_j).$$

Since $u_{k_m} \leq \min(u_i, u_j)$ and $\max(u_i, u_j) \leq u_{k_M}$, it results that condition (46) is also satisfied. □

4 Positivity preserving of the density

Maximum principle is satisfied for scalar autonomous hyperbolic problem *i.e* when the flux only depends on the state variable. However, such a property does not hold any longer when the operator depends on the space variable. For example, the minimum and the maximum of the density with a non free-divergence velocity are not preserved. However, for physical meaningful, the positivity would be preserved and the numerical scheme have to reproduce such a property.

To a great extent, numerical approximations for the Euler system are meaningful from a physical point of view if both density and pressure are non-negative. Numerical fluxes have been designed such that the first-order scheme preserve the density and pressure positivity (see for example [10]) while Perthame and Shu in 1996 [19], Linde and Roe in 1997 [17] prove that second-order schemes based on a linear reconstruction are also positivity preserving. The surprising point is that no limiting procedure is required to achieve the positivity preserving property. In this section, we highlight the link between relations (12)-(13) and the positivity preserving property where we focus on the scalar advection problem which is of practical importance.

4.1 Positivity preserving numerical flux

We consider the following linear advection advection problem

$$\partial_t u + \nabla.(V(x,t)u) = 0,\ x \in \mathbb{R}^2, t > 0, \tag{47}$$

with the initial condition $u(.,t=0) = u^0$. We assume that the velocity $V : \mathbb{R}^2 \times [0,+\infty[\to \mathbb{R}^2$ is a continuous bounded function for the sake of simplicity. If u^0 is a positive real value function, hence the solution has to be positive, therefore, it is convenient that the numerical scheme also preserves the positivity.
The advection problem casts in the generic non autonomous scalar hyperbolic problem

$$\partial_t u + \partial_{x_1} f_1(u;x,t) + \partial_{x_2} f_2(u;x,t) = 0 \tag{48}$$

and we consider the following generic finite volume scheme

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \nu(i)} \frac{|S_{ij}|}{|K_i|} \mathcal{F}(u_i^n, u_j^n, n_{ij}; X_{ij}, t^n), \tag{49}$$

where X_{ij} is a colocation point on side S_{ij}.

We first assume that the numerical flux satisfies the consistency property for any $x \in \mathbb{R}^2$ and $n \in S^2$ given by

$$\mathcal{F}(u,u,n_{ij};x,t^n) = F(u;x,t^n).n_{ij} = f_1(u;x,t^n).n_{ij,1} + f_2(u;x,t^n).n_{ij,2}.$$

and the numerical flux conservation

$$\mathcal{F}(u_i,u_j,n_{ij};x,t^n) + \mathcal{F}(u_j,u_i,n_{ji};x,t^n) = 0.$$

Moreover,the numerical flux has to be positivity preserving in the following sense.

Definition 6 The numerical flux is positivity preserving if there exists $\lambda_0 > 0$ such that for any $u_i, u_j > 0$, for any $n \in S^2$, $x, y \in \mathbb{R}^2$ and $t \geq 0$, we have

$$u_i - \lambda \left[\mathcal{F}(u_i,u_j,n;x,t) - \mathcal{F}(u_i,u_i,n;y,t)\right] \geq 0. \tag{50}$$

as long as $\lambda \in [0,\lambda_0]$. □

Remark 21 Definition yields that for any $t \geq 0$, $n \in S^2$ and $u > 0$, the physical flux satisfies

$$u \geq \lambda \left[F(u;x,t) - F(u;y,t)\right].n$$

for $\lambda \in [0,\lambda_0]$. Choosing the normal vector n colinear to $F(u;x,t) - F(u;y,t)$ and we get

$$|F(u;x,t) - F(u;y,t)| \leq \frac{u}{\lambda_0}.$$

Such a property is satisfied by the advection equation since we have

$$|V(x,t)u - V(y,t)u| \leq 2u\|V\|_{L^\infty},$$

hence we choose $\lambda_0 = \dfrac{1}{2\|V\|_{L^\infty}}$. □

We now show that the two classical numerical fluxes for the advection problem are positivity preserving

$$\mathcal{F}_{up}(u_i, u_j, n; x, t) = [V(x,t).n(x)]^+ u_i + [V(x,t).n(x)]^- u_j, \tag{51}$$

$$\mathcal{F}_{LF}(u_i, u_j, n; x, t) = V(x,t).n(x)\frac{u_i + u_j}{2} - \nu(u_j - u_i), \tag{52}$$

where $[u]^+ = \max(0, u)$, $[u]^- = \min(0, u)$ and $\nu > 0$.
We first recall the equalities

$$[u]^+ + [u]^- = u, \qquad [u]^+ - [u]^- = |u|.$$

We now check the positivity preserving property of the fluxes. For the upwind (or splitting) flux (51), we write

$$\begin{aligned}
\widetilde{u}_i &= u_i - \lambda \left[\mathcal{F}_{up}(u_i, u_j, n; x, t) - \mathcal{F}_{up}(u_i, u_i, n; y, t)\right], \\
&= u_i - \lambda \left[[V(x,t).n(x)]^+ u_i + [V(x,t).n(x)]^- u_j - V(y,t).n(y) u_i\right], \\
&= \Big(1 + \lambda [V(x,t).n(x)]^+ - \lambda V(y,t).n(y)\Big) u_i - \lambda [V(x,t).n(x)]^- u_j.
\end{aligned}$$

Hence, $\widetilde{u}_i$ is positive since we have combination of positive values with non negative coefficients (one of them still positive) under the condition

$$\lambda \leq \lambda_0 < \frac{1}{2\|V\|_{L^\infty}}. \tag{53}$$

For the Lax-Friedrisch flux, we write

$$\begin{aligned}
\widetilde{u}_i &= u_i - \lambda \left[\mathcal{F}_{LF}(u_i, u_j, n; x, t) - \mathcal{F}_{LF}(u_i, u_i, n; y, t)\right], \\
&= u_i - \lambda \left[V(x,t).n(x)\frac{u_i + u_j}{2} - \nu(u_j - u_i) - V(y,t).n(y) u_i\right], \\
&= u_i \left(1 - \lambda\nu - \lambda\frac{V(x,t).n(x)}{2} + \lambda V(y,t).n(y)\right) + u_j \lambda \left(\nu - \frac{V(x,t).n(x)}{2}\right).
\end{aligned}$$

Assume that

$$\nu = \frac{\|V\|_{L^\infty}}{2}, \qquad \lambda \leq \lambda_0 < \frac{1}{2\|V\|_{L^\infty}}, \tag{54}$$

then the Lax-Friedrisch flux (52) is positivity preserving.

4.2 Positivity preserving: first-order scheme

We prove the positivity preserving property for the first-order scheme (49).

Proposition 8 (positivity preservation: first-order case) *Let $\mathcal{T}$ be a conform mesh and assume that the approximate solution u_h^n is positive. If the numerical flux is positivity preserving then u_i^{n+1}, $K_i \in \mathcal{T}$ given by relation (49) are positive real values under the CFL condition*

$$\Delta t \leq \frac{\lambda_0}{N_{\underline{\nu}}} h. \quad \square \tag{55}$$

Proof To prove the positivity, we follow the ideas of [19] and [17]. Let B_i be the centroid of element K_i, one has

$$\sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \mathcal{F}(u_i^n, u_i^n, n_{ij}; B_i, t^n) = F(u_i^n; B_i, t^n). \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}| n_{ij}}{|K_i|} = 0$$

because B_i does not depend on the j index. Relation (49) then yields

$$\begin{aligned} u_i^{n+1} &= u_i - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|S_{ij}|}{|K_i|} \Big(\mathcal{F}(u_i^n, u_j^n, n_{ij}; X_{ij}, t^n) - \mathcal{F}(u_i^n, u_i^n, n_{ij}; B_i, t^n) \Big), \\ &= \frac{1}{\mathrm{perim}(K_i)} \sum_{j \in \underline{\nu}(i)} |S_{ij}| \left[u_i - \Delta t \frac{\mathrm{perim}(K_i)}{|K_i|} \Big(\mathcal{F}(u_i^n, u_j^n, n_{ij}; X_{ij}, t^n) - \mathcal{F}(u_i^n, u_i^n, n_{ij}; B_i, t^n) \Big) \right]. \end{aligned}$$

Since the numerical flux is positivity preserving the quantities

$$\widetilde{u}_{ij} = u_i - \Delta t \frac{\mathrm{perim}(K_i)}{|K_i|} \Big(\mathcal{F}(u_i^n, u_j^n, n_{ij}; X_{ij}, t^n) - \mathcal{F}(u_i^n, u_i^n, n_{ij}; B_i, t^n) \Big)$$

are positive as long as

$$\Delta t \frac{\mathrm{perim}(K_i)}{|K_i|} \leq \lambda_0. \tag{56}$$

In the other hand, definition of the perimeter yields

$$\mathrm{perim}(K_i) \leq \#\underline{\nu}(i) \max_{j \in \underline{\nu}(i)} |S_{ij}| \leq N_{\underline{\nu}} \max_{j \in \underline{\nu}(i)} |S_{ij}|,$$

then we have

$$\frac{|K_i|}{\mathrm{perim}(K_i)} \geq \frac{|K_i|}{N_{\underline{\nu}} \max_{j \in \underline{\nu}(i)} |S_{ij}|} \geq \frac{h}{N_{\underline{\nu}}}.$$

Consequently, relation (56) is satisfied if one choose the Δt such that the CFL condition (55) holds. $\square$

For example, let us consider the non free-divergence velocity advection problem. A first-order scheme based on one of the two numerical fluxes proposed above is positivity preserving under the CFL condition

$$\Delta t \leq \frac{h}{2 N_{\underline{\nu}} \|V\|_{L^\infty}}. \quad \square \tag{57}$$

4.3 Positivity preserving: second-order scheme

We now investigate the positivity preserving property when a second-order scheme is employed with a convex μ-admissible reconstruction. Assume that we have a positive real values approximation u_h^n at time t^n, the reconstruction operator $\mathcal{R}$ provides new values u_{ij}^n and u_{ji}^n on both side of the element edges S_{ij}. Assuming that $\mathcal{R}$ is a convex μ-admissible reconstruction then relation (15) holds and u_{ij}^n is positive since it is obtained as a convex combination of positive real values. We plug the reconstructed values into the generic finite volume scheme to provide relation (10). The goal is now to prove that u_i^{n+1} is a positive real value.

Lemma 6 *Let ζ_{ij}^n, $K_i \in \mathcal{T}$, $j \in \underline{\nu}(i)$ be positive real coefficients such that*

$$\zeta_{ij}^n \geq \zeta_0 > 0 \tag{58}$$

and define

$$C_i = \sum_{j \in \underline{\nu}(i)} \Big(\zeta_{ij}^n |K_i| u_{ij} - \Delta t |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n) \Big),$$

where $\mathcal{F}$ is a positivity preserving numerical flux. Then C_i is a positive real value under the CFL condition

$$\Delta t \leq \frac{\zeta_0 \lambda_0}{N_{\underline{\nu}}} h. \quad \square \tag{59}$$

Proof Following [19], we consider a partition of K_i with triangles K_{ij} where S_{ij} is a side of K_{ij} and B_i a node of K_{ij} (see figure 2).

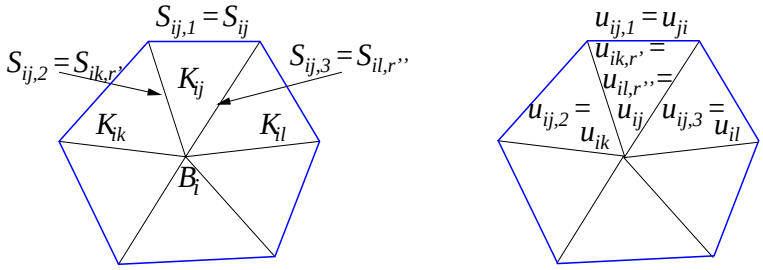

Fig. 2 Mesh subdivision notations. The cell K_i is subdivised in triangles K_{ij} with sides $S_{ij,r}$, $r = 1, 2, 3$ (left). We consider u_{ij} as a mean value of u on cell K_{ij} while $u_{ij,r}$ are the mean values on the other sides (right).

Let denote by $\rho_{ij} = \frac{|K_{ij}|}{|K_i|}$, we have $0 < \rho_{ij} < 1$ and we rewrite C_i under the form

$$C_i = \sum_{j \in \underline{\nu}(i)} \Big(\frac{\zeta_{ij}}{\rho_{ij}} |K_{ij}| u_{ij}^n - \Delta t |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n) \Big).$$

Let denote by K_{ik} and K_{il} the two adjacent sub-triangles (see figure 2 left), we set $S_{ij,2} = K_{ij} \cap K_{ik}$, $S_{ij,3} = K_{ij} \cap K_{il}$ and $S_{ij,1} = S_{ij}$ the common side between K_i and K_{ij}. The reconstructed value u_{ij}^n can be interpreted as the mean value on the sub-triangle K_{ij} and, in the same way, we define $u_{ij,1}^n = u_{ji}^n$, $u_{ij,2}^n = u_{ik}^n$, $u_{ij,3}^n = u_{il}^n$ (see figure 2 right). Note that there exist $r', r'' \in \{2,3\}$ such that $S_{ij,2} = S_{ik,r'}$, $S_{ij,3} = S_{ik,r''}$ and we have $u_{ik,r'}^n = u_{il,r''}^n = u_{ij}^n$ (index $r = 1$ always correspond to values outside of K_i). Finally, we denote by $n_{ij,r}$ the K_{ij} normal outwards unit vector and $M_{ij,r}$ the midpoint of sides $S_{ij,r}$, $r = 1, 2, 3$. The flux conservation property yields

$$\begin{aligned}
&\mathcal{F}(u_{ij}^n, u_{ij,2}^n, n_{ij,2}; M_{ij,2}, t^n) + \mathcal{F}(u_{ik}^n, u_{ik,r'}^n, n_{ik,r'}; M_{ik,r'}, t^n) = 0,\\
&\mathcal{F}(u_{ij}^n, u_{ij,3}^n, n_{ij,3}; M_{ij,3}, t^n) + \mathcal{F}(u_{il}^n, u_{il,r''}^n, n_{il,r''}; M_{il,r''}, t^n) = 0
\end{aligned}$$

From the conservation relations, it results that

$$C_i = \sum_{j \in \underline{\nu}(i)} \left(\frac{\zeta_{ij}^n}{\rho_{ij}} |K_{ij}| u_{ij}^n - \Delta t \sum_{r=1,2,3} |S_{ij,r}| \mathcal{F}(u_{ij}^n, u_{ij,r}^n, n_{ij,r}; M_{ij,r}, t^n) \right).$$

Let B_{ij} be the centroid of triangle K_{ij}, we have

$$\sum_{r=1,2,3} |S_{ij,r}| \mathcal{F}(u_{ij}^n, u_{ij}^n, n_{ij,r}; B_{ij}, t^n) = 0$$

and we rewrite the C_i expression under the form

$$\begin{aligned}
C_i = &\sum_{j \in \underline{\nu}(i)} \left(\frac{\zeta_{ij}^n}{\rho_{ij}} |K_{ij}| u_{ij} - \Delta t \sum_{r=1,2,3} |S_{ij,r}| \right.\\
&\left. \Big[\mathcal{F}(u_{ij}^n, u_{ij,r}^n, n_{ij,r}; M_{ij,r}, t^n) - \mathcal{F}(u_{ij}^n, u_{ij}^n, n_{ij,r}; B_{ij}, t^n) \Big] \right),\\
= &\sum_{j \in \underline{\nu}(i)} \frac{|K_{ij}|}{\mathrm{perim}(K_{ij})} \frac{\zeta_{ij}^n}{\rho_{ij}} \sum_{r=1,2,3} |S_{ij,r}| \left(u_{ij} - \Delta t \frac{\mathrm{perim}(K_{ij})}{|K_{ij}|} \frac{\rho_{ij}}{\zeta_{ij}^n} \right.\\
&\left. \Big[\mathcal{F}(u_{ij}^n, u_{ij,r}^n, n_{ij,r}; M_{ij,r}, t^n) - \mathcal{F}(u_{ij}^n, u_{ij}^n, n_{ij,r}; B_{ij}, t^n) \Big] \right)
\end{aligned}$$

From the positivity preserving property of the numerical flux (50), each term of the $\sum_{r=1,2,3}$ summation is positive if, under condition (58), the time step Δt satisfies the condition

$$\Delta t \, \mathrm{perim}(K_{ij}) \frac{\rho_{ij}}{|K_{ij}|} = \Delta t \frac{\mathrm{perim}(K_{ij})}{|K_i|} \leq \lambda_0 \zeta_0.$$

Since $\mathrm{perim}(K_i) \geq \mathrm{perim}(K_{ij})$, we deduce the more restrictive condition:

$$\Delta t \leq \lambda_0 \zeta_0 \frac{|K_i|}{\mathrm{perim}(K_i)}. \tag{60}$$

Noting that

$$\mathrm{perim}(K_i) \leq \#\underline{\nu}(i) \max_{j\in\underline{\nu}(i)} |S_{ij}| \leq N_{\underline{\nu}} \max_{j\in\underline{\nu}(i)} |S_{ij}|,$$

we deduce that relation (60) is satisfied if (59) is satisfied. hence C_i is positive. □

We sum up the positivity preserving property in the following proposition

Proposition 9 *Let $\mathcal{T}$ belong to $\mathcal{M}(\alpha)$ and $\mathcal{R}$ be a convex μ-admissible reconstruction. Then u_h^{n+1} is a positive real value function under the CFL condition*

$$\Delta t \leq \frac{\lambda_0}{1 + N_\mu C_\omega(\alpha)} \frac{h}{N_{\underline{\nu}}^2}. \quad \square \tag{61}$$

Proof The convex μ-admissible reconstruction assumption says that the equalities

$$u_{ij}^n - u_i^n + \sum_{k\in\mu(i)} \omega_{ijk}^n (u_k^n - u_i^n) = 0, \quad j \in \underline{\nu}(i)$$

hold for all $K_i \in \mathcal{T}$. Let ζ_{ij}^n, $j \in \underline{\nu}(i)$ be positive coefficients we shall define ahead, we write the second-order scheme under the following form

$$\begin{aligned}
|K_i| u_i^{n+1} &= |K_i| u_i^n - \Delta t \sum_{j\in\underline{\nu}(i)} |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n), \\
&= |K_i| u_i^n + \sum_{j\in\underline{\nu}(i)} \zeta_{ij}^n |K_i| \left(u_{ij}^n - u_i^n + \sum_{k\in\mu(i)} \omega_{ijk}^n (u_k^n - u_i^n) \right) \\
&\quad - \Delta t \sum_{j\in\underline{\nu}(i)} |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n), \\
&= |K_i| u_i^n - |K_i| u_i^n \sum_{j\in\underline{\nu}(i)} \zeta_{ij}^n (1 + \sum_{k\in\mu(i)} \omega_{ijk}^n) + \sum_{k\in\mu(i)} \sum_{j\in\underline{\nu}(i)} \omega_{ijk}^n |K_i| u_k^n \\
&\quad + \sum_{j\in\underline{\nu}(i)} \zeta_{ij}^n |K_i| u_{ij}^n - \Delta t \sum_{j\in\underline{\nu}(i)} |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n), \\
&= A_i + B_i + C_i,
\end{aligned}$$

with

$$A_i := |K_i| u_i^n \left(1 - \sum_{j\in\underline{\nu}(i)} \zeta_{ij}^n (1 + \sum_{k\in\mu(i)} \omega_{ijk}^n) \right), \; B_i := \sum_{k\in\mu(i)} \sum_{j\in\underline{\nu}(i)} \omega_{ijk}^n |K_i| u_k,$$

and

$$C_i := \sum_{j\in\underline{\nu}(i)} \left[\zeta_{ij}^n |K_i| u_{ij}^n - \Delta t |S_{ij}| \mathcal{F}(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n) \right].$$

Expression A_i is non negative if we choose ζ_{ij} such that

$$\sum_{j\in\underline{\nu}(i)} \zeta_{ij}^n (1 + \sum_{k\in\mu(i)} \omega_{ijk}^n) = 1, \tag{62}$$

which is achieved with

$$\zeta_{ij}^n = \frac{1}{\#\underline{\nu}(i)} \frac{1}{1 + \sum_{k\in\mu(i)} \omega_{ijk}^n}.$$

Since $\omega_{ijk}^n \leq C_\omega(\alpha)$ we have the estimate $\displaystyle\sum_{k\in\mu(i)} \omega_{ijk}^n \leq N_\mu C_\omega$ and condition (62) is achieved with

$$\zeta_{ij}^n \geq \zeta_0 := \frac{1}{N_{\underline{\nu}}} \frac{1}{1 + N_\mu C_\omega(\alpha)}. \tag{63}$$

Clearly, term B is non negative since we have a combination of positive real value numbers with non-negative coefficients. At last, lemma 6 yields that C is positive under the CFL condition (59). Thank to the estimate (63), we deduce that positivity is preserved if Δt satisfies the CFL condition (61). □

5 Conclusions

High order methods for scalar autonomous hyperbolic problems require limiting procedures to preserve the maximal principal property. To achieve such an issue on each element K_i of the mesh, the algorithm is decomposed in several steps.

- First, a local maximum principle has to be defined using neighboured elements characterized by the index set $\mu_{mp}(i)$.
- The second step concerns the local polynomial construction using the mean values situated in K_i and K_j, $j \in \mu_{po}(i)$.
- The third step is the limitation procedure where the predicted polynomial coefficients are modified to respect the maximum principle at some control points characterized by the index set $\mu_{cp}(i)$.
- At last, numerical flux on the sides of K_i is evaluated using colocation points indexed by $\mu_{co}(i)$.

Such a limiting procedure involves four index sets leading to a complicated analysis of the maximum principle preservation. To simplify the study, we propose a generic characterization of the reconstructions based on two fundamental properties. We then prove that we recover a positive coefficient schemes (incremental scheme with non negative coefficients in fact) which immediately gives the maximum principle. We show that the popular MUSCL methods cast in our formalism whatever the choice of the reconstruction or the limitation are (different control and colocation points).
We have also highlighted the connection between the two fundamental properties and the positivity preserving property for non free-divergence velocity advection problem. Such a result provides a condition to achieve the positivity preservation of the density for the classical Euler system (isentropic, real gas) but also for the water height variable of the shallow-water problem.

References

1. T. J. Barth, Numerical methods for conservation laws on structured and unstructured meshes, VKI March 2003 Lectures Series.
2. T. J. Barth, M. Ohlberger, Finite volume methods: foundation and analysis, Volume 1, chapter 15, Encyclopedia of Computational Mechanics, John Wiley & Sons Ltd, (2004).
3. T. J. Barth, D. C. Jespersen, The design and application of upwind schemes on unstructured meshes, AIAA Report 89-0366, 1989.
4. T. Buffard, S. Clain, Monoslope and Multislope MUSCL Methods for unstructured meshes, J. Comput. Phys. ??? (2010) ???-???
5. J. P. Boris, D. L. Book, Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works, J. Comput. Phys. 11 (1973) 38–69.
6. S. Clain, V. Clauzon, L^∞ stability of the MUSCL methods, Numer. Math. Vol.??? (2010) ???–???.
7. C. M. Dafermos, Hyperbolic conservation laws in continuum physics, Springer-Verlag, Berlin, Heidelberg, New-York, 2000.
8. E.F. Toro, V.A. Titarev, Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws, J. Comput. Phys. 227 (2008) 3971–4001.
9. J. B. Goodman, R. J. LeVeque, On the accuracy of stable schemes for 2D scalar conservation laws, Math. Comp. 45 (171) (1985) 15–21.
10. J. Gressier, P. Villedieu, J.-M. Moschetta, Positivity of flux vector splitting schemes, J. Comput. Phys. 155 (1999) 199–220.
11. A. Harten, High resolution schemes for Hyperbolic conservation laws, J. Comput. Phys. 49 (1983) 357–393.
12. A. Harten, On a class of high resolution total variation stable finite difference schemes, SIAM J. Num. Anal. 21 (1984) 1–21.
13. M. E. Hubbard, Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids, J. Comput. Phys. 155 (1) (1999) 54–74.
14. A. Jameson, Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence in transonic and hypersonic flows, AIAA paper No 3359 (1993).
15. A. Jameson, P. D. Lax, Condition for the construction of multipoint variation diminishing difference schemes, App. Numer. Math. Vol. 2, issue 3-5 (1986) 335–345.
16. R. J. LeVeque Numerical methods for conservation laws, Birkhaüser Verlag, Basel, 1992.
17. T. Linde, P. L. Roe, Robust Euler Codes, AIAA report 97-2098, 1997
18. J. S. Park,S.-H. Yoon, C. Kim, Multi-dimensional limiting process for hyperbolic conservation laws on unstructured grids, J. Comput. Phys. 229 (2010) 788–812.
19. B. Perthame, C. W. Shu, On positivity preserving finite volume schemes for Euler equations, Numer. Math. 73 (1996) 119–130.
20. Chi-Wang Shu, High order weighted essentially nonoscillatory schemes for convection dominated problems, SIAM Review, Vol. 51, No 1 (2009) 82–126.
21. C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing scheme, J. Comput. Phys., 77 (1988) 439–471.
22. S. P. Spekreijse, Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws, Math. of Comp. 49 (179) (1987) 135–155.
23. J. Smoller, Shock wawes and raction diffusion equations, Springer-Verlag, New-York, 1983.
24. P. K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, SIAM J. Numer. Anal. 21 (5) (1984) 995–1011.
25. E.F. Toro, V.A. Titarev, Derivative Riemann solvers for system of conservation laws and ADER method, J. Comput. Phys. 212 (2006) 150–165.
26. B. Van Leer, Towards the ultimate conservative difference schemes V. A second-order sequel to Godunov's method, J. Comput. Phys. 32 (1) (1979) 101–136.
27. X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. ??? (2010) ???–???.