Finite volume $L^\infty$-stability for hyperbolic scalar problems
Stephane Clain

To cite this version:
Stephane Clain. Finite volume $L^\infty$-stability for hyperbolic scalar problems. 2010. hal-00467650

HAL Id: hal-00467650
https://hal.archives-ouvertes.fr/hal-00467650
Preprint submitted on 27 Mar 2010

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

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Finite volume $L^\infty$-stability for hyperbolic scalar problems

Stéphane Clain

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 our formalism and provide new proofs of the $L^\infty$ stability.

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

1 Introduction

$L^\infty$ 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 $L^\infty$ 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 lose the maximum principle unless some restrictions (limiting procedures for instance) are employed to recover the MP property.

Important efforts have been realized during the last four 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 been 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 papers 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 $L^\infty$-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 $L^\infty$-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 $K_i, i \in \mathcal{E}_{cl}$ 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 $\nu(i) \subset \mathcal{E}_{cl}$ 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 $\nu(i) \subset \mathcal{E}_{cl}$ contains all the index $j$ of elements $K_j \in \mathcal{T}$ such that $K_i \cap K_j \neq \emptyset$. In other word, $\bigcup_{j \in \nu(i)} K_j$ 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

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

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, $\text{perim}(K_i) = \sum_{j \in \nu(i)} |S_{ij}|$ is the perimeter of element $K_i$. 

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 $\alpha$ defined in the following (see definition 4 for example) whereas $h$ is the mesh size parameter given by
\begin{equation}
    h = \min_{\kappa_i \in \mathcal{T}} \frac{|K_i|}{|S_{ij}|}.
\end{equation}

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

\textit{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
\begin{equation}
    \partial_t u + \partial_{x_1} f_1(u) + \partial_{x_2} f_2(u) = 0,
\end{equation}
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 $K_i \in T$, 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 $K_i \in T$ at time $t^n$ while the initial condition is given by
\[
u_i^0 = \frac{1}{|K_i|} \int_{K_i} u^0(\cdot) \, dx.
\]

The generic first-order explicit finite volume scheme provides an approximation at time $t^{n+1} = t^n + \Delta t$ by
\[
u_i^{n+1} = \nu_i^n - \Delta t \sum_{j \in \mathcal{N}(i)} \frac{|S_{ij}|}{|K_i|} g(u^j_i, u^j_i, n_{ij}).
\]
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 $(f_1, f_2)$:
\[
g(u_i, u_i, n_{ij}) = f_1(u_i)n_{ij,1} + f_2(u_i)n_{ij,2}.
\]

(c) monotony: the numerical flux is monotone:
\[
\partial_1 g(u_i, u_j, n_{ij}) \geq 0, \quad \partial_2 g(u_i, u_j, n_{ij}) \leq 0.
\]

Note that the consistency condition implies the conservation property
\[
\sum_{j \in \mathcal{N}(i)} \frac{|S_{ij}|}{|K_i|} g(u_i, u_i, n_{ij}) = 0.
\]

Remark 2 The flux conservation across the interface is usually satisfied by the numerical flux:
\[
g(u_i, u_j, n_{ij}) = -g(u_j, u_i, n_{ji}).
\]
which implies equivalence
\[
\partial_1 g(u_i, u_j, n_{ij}) \geq 0 \iff \partial_2 g(u_i, u_j, n_{ij}) \leq 0.
\]

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. \qed
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 $\Omega$ with reflection 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 $K_i \in T$ and $u_h$ corresponds to the constant piecewise representation given by

$$u_h = \sum_{i \in E_T} u_i \mathbb{1}_{K_i},$$

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 $T$ we define the one-point reconstruction operator $\mathcal{R}(T)$ by

$$(u_i)_{K_i \in T} \xrightarrow{\mathcal{R}(T)} (u_{ij})_{K_i \in T, j \in \nu(i)},$$

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

$$(u_i)_{K_i \in T} \xrightarrow{\mathcal{R}(T)} (u_{ij,r})_{K_i \in T, j \in \nu(i), r=1,...,R}$$

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

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

(10)

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

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

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

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

\[ u_{i}^{n+1} = u_{i}^{n} - \Delta t \sum_{j \in G(i)} \frac{|S_{ij}|}{|K_{i}|} \sum_{r=1}^{R} \zeta_{r} g(u_{ij,r}^{n}, u_{ji,r}^{n}, n_{ij}). \]  

(11)

where \( \zeta_{r} \) are non-negative convex weight coefficients for the numerical integration of \( f(u) \) 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

\( L^{\infty}\)-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 \( \nu(i) \subset \mu(i) \subset \nu(i) \). To this end, we have to impose some specific conditions on the reconstruction leading to the notion of admissible reconstruction operator.

**Definition 1 (\( \mu \)-local discrete extrema)** Let \( K_{i} \in T \), and \( \mu(i) \) an index set such that \( \nu(i) \subset \mu(i) \subset \nu(i) \). We define the \( \mu \)-local discrete minimum and maximum on the stencil \( \{i\} \cup \mu(i) \) by

\[ m_{i,\mu}^{n} = \min_{j \in \mu(i)} (u_{i}^{n}, u_{j}^{n}), \quad M_{i,\mu}^{n} = \max_{j \in \mu(i)} (u_{i}^{n}, u_{j}^{n}), \]

and we associate two corresponding indexes \( k_{m}, k_{M} \in \{i\} \cup \mu(i) \) such that

\[ u_{k_{m}} = m_{i,\mu}, \quad u_{k_{M}} = M_{i,\mu}. \]

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

a) There exists \( C_{\theta} = C_{\theta}(\alpha) \geq 0 \) and coefficients \( \theta_{ijk} \), \( i \in E_{el}, \ j \in \nu(i), \ k \in \mu(i) \) with

\[ 0 \leq \theta_{ijk} \leq C_{\theta}, \]

such that

\[ u_{ij} - u_{j} = \sum_{k \in \mu(i)} \theta_{ijk} (u_{k} - u_{j}). \]  

(12)
b) There exists $C_\omega = C_\omega(\alpha) \geq 0$ and coefficients $\omega_{ijk}$, $i \in \mathcal{E}_\text{el}$, $j \in \mathcal{G}(i)$, $k \in \mu(i)$ with
\[ 0 \leq \omega_{ijk} \leq C_\omega, \]
such that
\[ u_{ij} - u_i = - \sum_{k \in \mu(i)} \omega_{ijk} (u_k - u_i). \quad \square \tag{13} \]

The following more restrictive definition is also considered.

**Definition 3 (convex $\mu$-admissible reconstruction operator)** The reconstruction operator $R = R(T)$ is said convex if definition 2 is satisfied with
\[ \sum_{k \in \mu(i)} \theta_{ijk} = 1. \quad \square \tag{14} \]

**Remark 8** Note that relation (14) implies
\[ u_{ij} = \sum_{k \in \mu(i)} \theta_{ijk} u_k. \quad \square \tag{15} \]
and $C_\theta \leq 1$. \quad \square

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

**Remark 10** We do not have the uniqueness of coefficients $\theta_{ijk}$ and $\omega_{ijk}$. 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 $u_{km}$ and $u_{km}$ (see definition 1). It results that formulae (12) and (13) can be written under the form
\[ u_{ij} - u_j = \theta_{ijk_m} (u_{km} - u_j) + \theta_{ijk_M} (u_{kM} - u_j), \]
\[ u_{ij} - u_i = -\omega_{ijk_m} (u_{km} - u_j) - \omega_{ijk_M} (u_{kM} - u_j). \quad \square \]

We now establish the $L^\infty$-stability theorems based on the $\mu$-admissible reconstruction. To this end, let $T \in \mathcal{M}(\alpha)$, $K_i \in T$, then for any $m \leq M$ we define
\[ C_g = \max_{\alpha, \beta \in [m, M]} \{ \partial_1 g(\alpha, \beta, n), -\partial_2 g(\alpha, \beta, n) \}. \quad \square \tag{16} \]

The following lemma holds.

**Lemma 1 ($\mu$-local discrete maximum principle)** Let $R$ be a $\mu$-admissible reconstruction with respect to the structural parameter $\alpha$. If $u_k^n \in [m, M]$, $k \in \{i\} \cup \mu(i)$, then $u_k^{n+1}$ defined by relation (10) also belongs to $[m, M]$ under the CFL condition
\[ \Delta t < C_{fl} h, \quad \text{with} \quad C_{fl} = \frac{1}{\#\mathcal{P}(i) \#\mu(i) C_{\theta} + C_\omega}, \quad \square \tag{17} \]
where constant $C_{\theta}$ is given by relation (16) whereas constants $C_{\theta}$, $C_\omega$ only depend on the structural parameter $\alpha$. 

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

\[ u_{i}^{n+1} = u_{i}^{n} - \Delta t \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} F(u_{ij}, u_{ji}, n_{ij}) \]

Using the conservative property, we have

\[ u_{i}^{n+1} = u_{i}^{n} - \Delta t \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} \left[ F(u_{ij}^{n}, u_{ji}^{n}, n_{ij}) - F(u_{ij}^{n}, u_{ji}^{n}, n_{ij}) \right], \]

\[ = u_{i}^{n} - \Delta t \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} \left[ \partial_{1} F(\tilde{u}_{ij}^{n}, \tilde{u}_{ij}^{n}, n_{ij})(u_{ij}^{n} - u_{ij}^{n}) + \partial_{2} F(\tilde{u}_{ij}^{n}, \tilde{u}_{ij}^{n}, n_{ij})(u_{ji}^{n} - u_{ji}^{n}) \right], \]

with \( \tilde{u}_{ij}^{n} = \lambda_{ij} u_{ij}^{n} + (1 - \lambda_{ij}) u_{ji}^{n} \) and \( \tilde{u}_{ij}^{n} = \lambda_{ij} u_{ij}^{n} + (1 - \lambda_{ij}) u_{ji}^{n} \) where \( \lambda_{ij} \in [0, 1] \).

From relations (12) and (13) we deduce

\[ u_{i}^{n+1} = u_{i}^{n} - \Delta t \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} \left[ - A_{ij} \sum_{k \in \mu(i)} \omega_{ijk}(u_{k}^{n} - u_{i}^{n}) + B_{ij} \sum_{k \in \mu(j)} \theta_{ijk}(u_{k}^{n} - u_{i}^{n}) \right], \]

where we have defined

\[ A_{ij} = \partial_{1} F(\tilde{u}_{ij}^{n}, \tilde{u}_{ij}^{n}, n_{ij}), \quad B_{ij} = \partial_{2} F(\tilde{u}_{ij}^{n}, \tilde{u}_{ij}^{n}, n_{ij}). \]

After some algebraic manipulations, we obtain

\[ u_{i}^{n+1} = u_{i}^{n} - \Delta t \sum_{k \in \mu(i)} \left[ \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} \omega_{ijk} - B_{ij} \frac{|S_{ij}|}{|K_{i}|} \theta_{ijk} \right] (u_{k}^{n} - u_{i}^{n}). \]

Setting

\[ \Theta_{ik} = \sum_{j \in \mathcal{E}(i)} \frac{|S_{ij}|}{|K_{i}|} \omega_{ijk} - B_{ij} \frac{|S_{ij}|}{|K_{i}|} \theta_{ijk}, \]

thanks to the monotonicity of the numerical flux, we have \( 0 \leq \Theta_{ik} \leq \frac{C_{m}}{h} \) with

\[ C_{m} = \# \mathcal{E}(i) C_{\theta}(C_{\theta} + C_{\omega}). \]

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

\[ \frac{\Delta t}{h} \# \mu(i) \# \mathcal{E}(i) C_{\theta}(C_{\theta} + C_{\omega}) \leq 1. \]

Hence \( u_{i}^{n+1} \in [m, M] \).
Remark 11 We have the $\mu$-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 $L^\infty$ 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: $\mu(i) = \nu(i)$ and $\mu(i) = \nu(i) - \mu(i)$.

Theorem 1 ($\nu$-global discrete maximum principle) Let $\mathcal{R}$ be a $\nu$-admissible reconstruction with respect to the structural parameter $\alpha$. 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 $\Delta t$ satisfies the CFL condition
$$\Delta t < C_{\text{fl}} h, \quad \text{with} \quad C_{\text{fl}} = \frac{1}{N_{\nu} N_{\nu} C_g (C_\theta + C_\omega)},$$

then $u^n_{i} \in [m, M]$ for all $K_i \in T$ 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 $u^{n+1}_{i} \in [m, M]$ for any element $K_i \in T$ if $\Delta t$ satisfies the CFL condition (17). By definition $N_{\nu} = \max_i (\#\nu(i))$ and $N_{\nu} = \max_i (\#\nu(i))$ hence $C_{\text{fl}} < C_{\text{fl}}$. Consequently, the time step controlled by relation (18) is also controlled by relation (17) thus $u^{n+1}_{i} \in [m, M]$ for any element $K_i \in T$. □

In the same way, we have the following theorem.

Theorem 2 ($\nu$-global discrete maximum principle) Let $\mathcal{R}$ be a $\nu$-admissible reconstruction with respect to the structural parameter $\alpha$. 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 $\Delta t$ satisfies the CFL condition
$$\Delta t < C_{\text{fl}} h, \quad \text{with} \quad C_{\text{fl}} = \frac{1}{N_{\nu} N_{\nu} C_g (C_\theta + C_\omega)},$$

then $u^n_{i} \in [m, M]$ for all $K_i \in T$ 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
$$u^{n+1}_h = H(u^n_h), \quad u^{n+1}_h = \frac{3}{4} u^n_h + \frac{1}{4} H(u^n_h), \quad u^{n+1}_h = \frac{1}{3} u^n_h + \frac{2}{3} H(u^{n+1}_h).$$
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 $\theta_{ijk}$ and $\omega_{ijk}$. 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

$$\Delta t < C_{fl} h, \quad \text{with} \quad C_{fl} = \frac{1}{2N_{\nu}C_{g}(C_{\theta} + C_{\omega})}. \quad \square$$

(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 $\theta_{ijk}$ and $\omega_{ijk}$ in order to apply the $L^\infty$-stability theorem. One of the first reconstruction operator on unstructured meshes has been proposed by Barth and Jespersen [3] based on the $\nu$ 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 $\nu$ stencil has to be used to provide the $L^\infty$-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 $L^\infty$-stability at the collocation points when using the limiting procedure on the controls points. Subsection 3.1 is dedicated to the $L^\infty$-stability when one employs 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

$$\tilde{u}_i(X) = u_i + a_i B_i X$$

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 $T$ be a mesh of $\mathbb{R}^2$, for any elements $K_i \in T$, 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 \( T \) belongs to \( \mathcal{M}(\alpha) \), if and only if, for any \( K_i \in T \), there exists a set of barycentric coordinates with respect to the nodes such that

\[ \min_{K_i \in T} \rho_{im} \geq \alpha. \]  \( \Box \) \hspace{1cm} (21)

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

\[ N\lambda = \max_i \#\lambda(i). \]

For example we can choose \( \rho_{im} = \frac{1}{\#\lambda(i)} \) and \( \alpha = \frac{1}{N\lambda} \). \( \Box \)

**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 \( \tilde{u}_i(X) \) 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 \( u_h \). \( \Box \)

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 \( \mu(i) = u \) under\( nu \) and \( \mu(i) = v(i) \).

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

\[ m_i,\nu = u_{km} \leq a_i + a_i B_i P_m \leq u_{kM} = M_i,\nu. \]  \( \Box \) \hspace{1cm} (22)
Then there exist coefficients \( \theta_{ik}(P_m) \), \( \omega_{ik}(P_m) \), \( k \in \mathcal{U}(i) \) and constant value \( C_\omega = C_\omega(\alpha) \geq 0 \) which satisfy the following properties

\[
\tilde{u}_i(P_m) = \sum_{k \in \mathcal{U}(i)} \theta_{ik}(P_m) u_k, \quad \sum_{k \in \mathcal{U}(i)} \theta_{ik}(P_m) = 1, \quad (23)
\]

and

\[
\tilde{u}_i(P_m) - u_i = - \sum_{k \in \mathcal{U}(i)} \omega_{ik}(P_m) (u_k - u_i), \quad 0 \leq \omega_{ik}(P_m) \leq C_\omega, \quad (24)
\]

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

**Proof** We give the construction for each class of coefficients.

**Coefficients** \( \theta_{ik} \). Condition (22) yields that

\[
u_k m < \tilde{u}_i(P_m) < u_k M\]

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

\[
\tilde{u}_i(P_m) = \chi u_k m + (1 - \chi) u_k M.
\]

We then write \( \tilde{u}_i(P_m) \) as a convex combination with \( \theta_{ikm}(P_m) = \chi(P_m) \) and \( \theta_{ikM}(P_m) = 1 - \chi(P_m) \) while the other coefficients are set to zero.

**Coefficient** \( \omega_{ik} \). 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

\[
B_i P_m = - \sum_{m' \in \lambda(i) \setminus m} \frac{\rho_{im'}}{\rho_{im}} B_i P_{m'},
\]

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

\[
\tilde{u}_i(P_m) - u_i = a_i B_i P_m = \sum_{m' \in \lambda(i) \setminus m} \frac{\rho_{im'}}{\rho_{im}} a_i B_i P_{m'}.
\]

Condition (22) yields

\[
u_k m - u_i < a_i B_i P_{m'} < u_k M - u_i, \quad \forall m' \in \lambda(i),
\]

then

\[
- \sum_{m' \in \lambda(i) \setminus m} \frac{\rho_{im'}}{\rho_{im}} (u_k m - u_i) < \tilde{u}_i(P_m) - u_i < - \sum_{m' \in \lambda(i) \setminus m} \frac{\rho_{im'}}{\rho_{im}} (u_k m - u_i),
\]
hence  
\[-\frac{1-\rho_{im}}{\rho_{im}}(u_{k_{ml}} - u_i) < \tilde{u}_i(P_m) - u_i < -\frac{1-\rho_{im}}{\rho_{im}}(u_{k_{ml}} - u_i).\]

There exists \(\chi(P_m) \in [0, 1]\) such that
\[\tilde{u}_i(P_m) - u_i = -\tilde{\chi}\left(1 - \frac{\rho_{im}}{\rho_{im}}\right)(u_{k_{ml}} - u_i) - (1 - \tilde{\chi})\left(1 - \frac{\rho_{im}}{\rho_{im}}\right)(u_{k_{ml}} - u_i).\]

Relation (24) holds if we choose
\[\omega_{ik_{ml}}(P_m) = \frac{1-\rho_{im}}{\rho_{im}}, \quad \omega_{ik}(P_m) = (1 - \tilde{\chi}(P_m))\left(1 - \frac{\rho_{im}}{\rho_{im}}\right),\]

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

**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 \(L^\infty\)-stability.

**Proposition 1** For any \(X \in K_i\), there exist coefficients \(\theta_{ik}(X), \omega_{ik}(X), k \in \nu(i)\) and constant value \(C_{\omega} = C_{\omega}(\alpha) \geq 0\) which satisfy the following properties
\[\tilde{u}_i(X) = \sum_{k \in \nu(i)} \theta_{ik}(X)u_k, \quad \sum_{k \in \nu(i)} \theta_{ik}(X) = 1, \tag{25}\]
and
\[\tilde{u}_i(X) - u_i = -\sum_{k \in \nu(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_{\omega}, \tag{26}\]
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 \(\zeta_m(X), m \in \lambda(i)\)
\[X = \sum_{m \in \lambda(i)} \zeta_m(X)P_m, \quad \sum_{m \in \lambda(i)} \zeta_m(X) = 1.\]

Thanks to the linearity of the reconstruction, we have
\[\tilde{u}_i(X) = \sum_{m \in \lambda(i)} \zeta_m(X)\tilde{u}_i(P_m).\]
Consequently, relations (25) and (26) are satisfied with
\[
\theta_{ik}(X) = \sum_{m \in \lambda(i)} \zeta_m(X) \theta_{ik}(P_m), \quad \omega_{ik}(X) = \sum_{m \in \lambda(i)} \zeta_m(X) \omega_{ik}(P_m).
\]
Moreover, we have
\[
\sum_{k \in \Omega(i)} \theta_{ik}(X) = \sum_{k \in \Omega(i)} \sum_{m \in \lambda(i)} \zeta_m(X) \theta_{ik}(P_m),
\]
\[
= \sum_{m \in \lambda(i)} \zeta_m(X) \sum_{k \in \Omega(i)} \theta_{ik}(P_m),
\]
\[
= \sum_{m \in \lambda(i)} \zeta_m(X) = 1.
\]
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 $C_\omega(\alpha) = \frac{1}{\alpha}$ still holds. \[\square\]

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

**Lemma 3 (limiter with the $\mathfrak{p}$ stencil)** Let $T \in \mathcal{M}(\alpha)$, for any $K_i \in T$, we assume that $a_i$ satisfies the property: for all nodes $P_m$, $m \in \lambda(i)$ we have (see definition 1)
\[
m_{i, \mathfrak{p}} = u_{k,m} \leq u_i + a_i B_i P_m \leq u_{k,m} = M_{i, \mathfrak{p}}.
\]
Then there exist coefficients $\theta_{ik}(P_m)$, $\omega_{ik}(P_m)$, $k \in \mathfrak{p}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties
\[
\bar{u}_i(P_m) = \sum_{k \in \mathfrak{p}(i)} \theta_{ik}(P_m) u_k, \quad \sum_{k \in \mathfrak{p}(i)} \theta_{ik}(P_m) = 1,
\]
and
\[
\bar{u}_i(P_m) - u_i = - \sum_{k \in \mathfrak{p}(i)} \omega_{ik}(P_m)(u_k - u_i), \quad 0 \leq \omega_{ik}(P_m) \leq C_\omega,
\]
with $C_\omega(\alpha) = \frac{1}{\alpha}$. \[\square\]
The following proposition also holds.

**Proposition 2** For any $X \in K_i$, there exist coefficients $\theta_{ik}(X)$, $\omega_{ik}(X)$, $k \in \mathfrak{p}(i)$ and constant value $C_\omega = C_\omega(\alpha) \geq 0$ which satisfy the following properties
\[
\bar{u}_i(X) = \sum_{k \in \mathfrak{p}(i)} \theta_{ik}(X) u_k, \quad \sum_{k \in \mathfrak{p}(i)} \theta_{ik}(X) = 1,
\]
and
\[
\bar{u}_i(X) - u_i = - \sum_{k \in \mathfrak{p}(i)} \omega_{ik}(X)(u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega,
\]
with $C_\omega(\alpha) = \frac{1}{\alpha}$. \[\square\]
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 $\nu(i)$ to enforce the maximum principle. The numerical routine is summarized by the following steps.

1. On each element $K_i$, a predicted gradient $\hat{a}_i$ is computed with the values on the neighbouring element $K_j$, $j \in \nu(i)$.
2. The limiting procedure is applied to compute a new slope $a_i = \phi_i \hat{a}_i$ 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
   \[ u_{ij} = u_i + a_i B_i X_{ij}, \quad u_{ji} = u_j + a_j B_j X_{ij}, \]
   where $X_{ij}$ are given collocation points on sides $S_{ij}$, $j \in \nu(i)$.
4. The update approximation $u^n + 1$ 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 reconstruction)** Assume that the mesh $T$ belongs to the class $M(\alpha)$, 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 $\theta_{ijk} = \theta_{ik}(X_{ij})$, $\omega_{ijk} = \omega_{ik}(X_{ij})$,

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

3.1.2 The Park-Yoon-Kim limiter on nodes

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

1. On each element $K_i$, a predicted gradient $\hat{a}_i$ is computed with the values on the neighbouring element $K_j$, $j \in \nu(i)$.
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
   \[ u_{\text{min},m} = \min_{j \in \kappa(m)} \{ u_j \}, \quad u_{\text{max},m} = \max_{j \in \kappa(m)} \{ u_j \}. \]

We then define the coefficients $\phi_{im} \in [0, 1]$, $m \in \lambda(i)$ such that
   \[ u_{\text{min},m} \leq u_i + \phi_{im} a_i B_i P_m \leq u_{\text{max},m}. \]
3. We determine the new slope $a_i = \phi_i \tilde{a}_i$ with $\phi_i = \min_{m \in \lambda(i)} \phi_m$.

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** ($L^\infty$-stability for the Park-Yoon-Kim reconstruction) Assume that the mesh $T$ 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 $\bigcup_{m \in \lambda(i)} \kappa(m) = \mathcal{P}(i)$, then condition (33) yields

$$m_i, \mathcal{P} \leq u_i + a_i B_i P_m \leq M_i, \mathcal{P}, \quad \forall m \in \lambda(i),$$

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

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

then we have a convex $\mathcal{P}$-admissible reconstruction operator with $C^\omega = \frac{1}{\alpha}$. Theorem 1 gives the $L^\infty$-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 $\rho_{ij}$ the barycentric coordinates of $B_i$ with respect to the control points $C_{im}$:

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

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 $\alpha > 0$, $T$ belongs to $\mathcal{M}(\alpha)$, if and only if, there exists a set of barycentric coordinates such that

$$\min_{\kappa_i \in \mathcal{T}} \min_{m \in \delta(i)} \rho_{im} \geq \alpha. \quad (35)$$
It is important to note that the structural coefficient depend on the control points where we intend to limit the slope so the class \( M(\alpha) \) may change in function of the control points. For example, a particular interesting choice is \( C_{ij} = M_{ij} \) the \( S_{ij} \) midpoint with \( \delta(i) = \nu(i) \).

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

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

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

\[
m_{k} \leq u_{km} \leq a_i, B_i C_{im} \leq u_{kM} = M_{k} \quad \forall m \in \delta(i).
\]  

(36)

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

\[
\tilde{u}_i(C_{im}) = \sum_{k \in \nu(i)} \theta_{ik}(C_{im}) u_k, \sum_{k \in \nu(i)} \theta_{ik}(C_{im}) = 1,
\]  

(37)

and

\[
\tilde{u}_i(C_{im}) - u_i = - \sum_{k \in \nu(i)} \omega_{ik}(C_{im})(u_k - u_i), \quad 0 \leq \omega_{ik}(C_{im}) \leq C_\omega.
\]  

(38)

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

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

\[
u_{km} < \tilde{u}_i(C_{im}) < u_{kM}
\]

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

\[
\tilde{u}_i(C_{im}) = \chi u_{km} + (1 - \chi) u_{kM}.
\]

Hence \( \tilde{u}_i(C_{im}) \) is a convex combination with \( \theta_{ikm}(C_{im}) = \chi(C_{im}), \theta_{ikM}(C_{im}) = 1 - \chi(C_{im}) \) and the other coefficients set to zero. \n
**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

\[
B_i C_{im} = - \sum_{m' \neq m} \frac{\rho_{im}}{\rho_{im'}} B_i C_{im'}.
\]
where \( \rho_{im} \geq \alpha > 0 \). The reconstruction at point \( C_{im} \) satisfies the following equality
\[
\tilde{u}_i(C_{im}) - u_i = a_i B_i C_{im} = - \sum_{m' \in \delta(i)} \frac{\rho_{im'}}{\rho_{im}} a_i B_i C_{im'}.
\]
Condition (36) then yields
\[
 u_{k_m} - u_i < a_i B_i C_{im'} < u_{k_M} - u_i, \quad \forall m' \in \delta(i),
\]
then
\[
- \sum_{m' \in \delta(i)} \frac{\rho_{im'}}{\rho_{im}} (u_{k_m} - u_i) \leq \tilde{u}_i(C_{im}) - u_i < \sum_{m' \in \delta(i)} \frac{\rho_{im'}}{\rho_{im}} (u_{k_m} - u_i),
\]

hence
\[
- \frac{1 - \rho_{im}}{\rho_{im}} (u_{k_m} - u_i) \leq \tilde{u}_i(C_{im}) - u_i < - \frac{1 - \rho_{im}}{\rho_{im}} (u_{k_m} - u_i).
\]

Consequently, we can exhibit coefficient \( \chi(C_{im}) \in [0, 1] \) such that
\[
\tilde{u}_i(C_{im}) - u_i = -\chi \left( \frac{1 - \rho_{im}}{\rho_{im}} \right) (u_{k_m} - u_i) - \chi (1 - \chi) \left( \frac{1 - \rho_{im}}{\rho_{im}} \right) (u_{k_M} - u_i).
\]
Relation (38) holds if we choose
\[
\omega_{ik_m}(C_{im}) = -\chi(C_{im}) \left( \frac{1 - \rho_{im}}{\rho_{im}} \right), \quad \omega_{ik_M}(C_{im}) = (1 - \chi(C_{im})) \left( \frac{1 - \rho_{im}}{\rho_{im}} \right),
\]
and the other coefficients set to zero. Since \( 1 \geq \rho_{im} \geq \alpha \), we have \( C_\omega(\alpha) \leq \frac{1}{\alpha} \). □

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 \( L^\infty \)-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 \( \theta_{ik}(X) \geq 0 \), \( \omega_{ik}(X) \geq 0 \), \( k \in \nu(i) \) and constant value \( C_\omega = C_\omega(\alpha) \geq 0 \) which satisfy the following properties
\[
\tilde{u}_i(X) = \sum_{k \in \nu(i)} \theta_{ik}(X) u_k, \quad \sum_{k \in \nu(i)} \theta_{ik}(X) = 1, \quad (39)
\]
and
\[
\tilde{u}_i(X) - u_i = - \sum_{k \in \nu(i)} \omega_{ik}(X) (u_k - u_i), \quad 0 \leq \omega_{ik}(X) \leq C_\omega, \quad (40)
\]
with \( C_\omega(\alpha) = \frac{1}{\alpha} \). □
We now consider the case when the maximum principle is based on the values defined by the $\nu(i)$ index set.

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

$$m_i, \nu = u_{km} \leq u_i + a_i \cdot B_i C_{im} \leq u_{km} = M_i, \nu, \quad \forall m \in \delta(i). \quad (41)$$

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

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

and

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

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

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

and

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

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

$$m_{i, \nu} = u_{km} \leq u_i + a_i \cdot B_i M_{ij} \leq u_{km} = M_{i, \nu}. \quad (46)$$
3. The reconstructed values are given by
\[ u_{ij} = u_i + a_i B_i M_{ij}, \quad u_{ji} = u_j + a_j B_j M_{ij}. \]

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

**Proposition 7 (\( L^\infty \)-stability for the Barth-VKI03 reconstruction)** Assume that the mesh \( T \) belongs to the class \( M(\alpha) \) associated to the control points \( M_{ij}, j \in \mathcal{V}(i) \). 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 \( \delta(i) = \mathcal{V}(i) \) on points \( M_{ij} \) hence proposition 5 holds. Setting \( \theta_{ijk} = \theta_{ik}(M_{ij}), \quad \omega_{ijk} = \omega_{ik}(M_{ij}) \),

then we have a convex \( \mathcal{V} \)-admissible reconstruction operator with \( C_\theta = 1 \) and \( C_\omega = \frac{1}{\alpha} \). Theorem 2 gives the \( L^\infty \)-stability under the CFL condition (19). \( \square \)

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

**Remark 20** A more restrictive condition is also considered in [1]:
\[ \min(u_i, u_j) \leq u_i + a_i B_i M_{ij} \leq \max(u_i, u_j). \]
Since \( u_{km} \leq \min(u_i, u_j) \) and \( \max(u_i, u_j) \leq u_{km} \), it results that condition (46) is also satisfied. \( \square \)

### 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 problem
\[
\partial_t u + \nabla . (V(x, t) u) = 0, \quad x \in \mathbb{R}^2, \quad t > 0,
\]  
(47)

with the initial condition \( u(., t = 0) = u^0 \). We assume that the velocity \( V : \mathbb{R}^2 \times [0, +\infty[ \rightarrow \mathbb{R}^2 \) 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
\[
\partial_t u + \partial_{x_1} f_1(u; x, t) + \partial_{x_2} f_2(u; x, t) = 0
\]  
(48)

and we consider the following generic finite volume scheme
\[
u^{n+1}_i = u^n_i - \Delta t \sum_{j \in \Omega(i)} \frac{|S_{ij}|}{K_i} \mathcal{F}(u^n_i, u^n_j, n_{ij}; X_{ij}, t^n),
\]  
(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
\[
\mathcal{F}(u, n_{ij}; x, t^n) = F(u; x, t^n).n_{ij} = f_1(u; x, t^n).n_{ij,1} + f_2(u; x, t^n).n_{ij,2}.
\]

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

as long as \( \lambda \in [0, \lambda_0] \). \( \square \)

**Remark 21** Definition yields that for any \( t \geq 0, n \in S^2 \) and \( u > 0 \), the physical flux satisfies
\[
\nu \geq \lambda \left[ F(u; x, t) - F(u; y, t) \right].n
\]

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 \( \lambda_0 = \frac{1}{2\|V\|_{L^\infty}} \). \( \square \)
We now show that the two classical numerical fluxes for the advection problem are positivity preserving

\[ F_{\text{up}}(u, u_j, n; x, t) = [V(x, t)\cdot n(x)]^+_u + [V(x, t)\cdot n(x)]^-_u, \]

\[ F_{\text{LF}}(u, u_j, n; x, t) = V(x, t)\cdot n(x)\frac{u_i + u_j}{2} - \nu(u_j - u_i), \]

where \([u]^+ = \max(0, u), [u]^- = \min(0, u)\) and \(\nu > 0\).

We first recall the equalities

\[ [u]^+ + [u]^- = u, \quad [u]^+ - [u]^- = |u|. \]

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

\[ \tilde{u}_i = u_i - \lambda \left[ F_{\text{up}}(u_i, u_j, n; x, t) - F_{\text{up}}(u_i, u_i, n; y, t) \right], \]

\[ = u_i - \lambda \left[ V(x, t)\cdot n(x)^+_u + [V(x, t)\cdot n(x)]^-_u - V(y, t)\cdot n(y)u_i \right], \]

\[ = \left( 1 + \lambda[V(x, t)\cdot n(x)]^+_u - \lambda V(y, t)\cdot n(y) \right)u_i - \lambda [V(x, t)\cdot n(x)]^-_u. \]

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

\[ \lambda \leq \lambda_0 < \frac{1}{2\|V\|_{L^\infty}}. \]

For the Lax-Friedrichs flux, we write

\[ \tilde{u}_i = u_i - \lambda \left[ F_{\text{LF}}(u_i, u_j, n; x, t) - F_{\text{LF}}(u_i, u_i, n; y, t) \right], \]

\[ = u_i - \lambda \left[ V(x, t)\cdot n(x)\frac{u_i + u_j}{2} - \nu(u_j - u_i) - V(y, t)\cdot n(y)u_i \right], \]

\[ = u_i \left( 1 - \lambda\nu - \lambda V(x, t)\cdot n(x) \right) + \nu V(y, t)\cdot n(y) + u_j \lambda \left( \nu - \frac{V(x, t)\cdot n(x)}{2} \right). \]

Assume that

\[ \nu = \frac{\|V\|_{L^\infty}}{2}, \quad \lambda \leq \lambda_0 < \frac{1}{2\|V\|_{L^\infty}}, \]

then the Lax-Friedrichs 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 $T$ be a conform mesh and assume that the approximate solution $u^n_i$ is positive. If the numerical flux is positivity preserving then $u^{n+1}_i$, $K_i \in T$ given by relation (49) are positive real values under the CFL condition

$$\Delta t \leq \frac{\lambda_0}{N_{\nu} h}. \quad (55)$$

**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

$$X_j \in \nu(i) \cap S_{ij} |_{K_i} = 0$$

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

$$u^{n+1}_i = u_i - \Delta t \sum_{j \in \nu(i)} \frac{|S_{ij}|}{|K_i|} \left( F(u^n_i, u^n_j, n_{ij}; X_{ij}, t^n) - F(u^n_i, u^n_i, n_{ij}; B_i, t^n) \right),$$

$$= \frac{1}{\text{perim}(K_i)} \sum_{j \in \nu(i)} |S_{ij}| \left[ u_i - \Delta t \frac{\text{perim}(K_i)}{|K_i|} \left( F(u^n_i, u^n_j, n_{ij}; X_{ij}, t^n) - F(u^n_i, u^n_i, n_{ij}; B_i, t^n) \right) \right].$$

Since the numerical flux is positivity preserving the quantities

$$\tilde{u}_{ij} = u_i - \Delta t \frac{\text{perim}(K_i)}{|K_i|} \left( F(u^n_i, u^n_j, n_{ij}; X_{ij}, t^n) - F(u^n_i, u^n_i, n_{ij}; B_i, t^n) \right)$$

are positive as long as

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

In the other hand, definition of the perimeter yields

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

then we have

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

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

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

$$\Delta t \leq \frac{h}{2N_{\nu} \|V\|_{L^\infty}}. \quad (57)$$
4.3 Positivity preserving: second-order scheme

We now investigate the positivity preserving property when a second-order scheme is employed with a convex \( \mu \)-admissible reconstruction. Assume that we have a positive real values approximation \( u^n_h \) at time \( t^n \), the reconstruction operator \( R \) provides new values \( u^n_{ij} \) and \( u^n_{ji} \) on both side of the element edges \( S_{ij} \). Assuming that \( R \) is a convex \( \mu \)-admissible reconstruction then relation (15) holds and \( u^n_{ij} \) 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 \( u^{n+1}_i \) is a positive real value.

**Lemma 6** Let \( \varepsilon_{ij}^n \) be positive real coefficients such that
\[
\varepsilon_{ij}^n \geq \varepsilon_0 > 0
\]
and define
\[
C_i = \sum_{j \in \mathcal{E}(i)} \left( \varepsilon_{ij}^n |K_{ij}| u_{ij} - \Delta t |S_{ij}| F(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n) \right),
\]
where \( F \) is a positivity preserving numerical flux. Then \( C_i \) is a positive real value under the CFL condition
\[
\Delta t \leq \frac{\varepsilon_0 \lambda_0 h}{N_u}.
\]

**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).

Let denote by \( \rho_{ij} = \frac{|K_{ij}|}{|K_i|} \), we have \( 0 < \rho_{ij} < 1 \) and we rewrite \( C_i \) under the form
\[
C_i = \sum_{j \in \mathcal{E}(i)} \left( \varepsilon_{ij}^n |K_{ij}| u_{ij} - \Delta t |S_{ij}| F(u_{ij}^n, u_{ji}^n, n_{ij}; M_{ij}, t^n) \right).
\]
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 $u^n_{ij}$ can be interpreted as the mean value on the sub-triangle $K_{ij}$ and, in the same way, we define $u^n_{ij,1} = u^n_{ij}$, $u^n_{ij,2} = u^n_{ik}$, $u^n_{ij,3} = u^n_{il}$ (see figure 2 right). Note that there exist $r, r' \in \{2, 3\}$ such that $S_{ij,2} = S_{ik,r}$, $S_{ij,3} = S_{il,r'}$ and we have $u^n_{ik,r} = u^n_{il,r'} = u^n_{ij}$ (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

$$
F(u^n_{ij}, u^n_{ij,2}, n_{ij,2}; M_{ij,2}, t^n) + F(u^n_{ik}, u^n_{ik,r}, n_{ik,r}; M_{ik,r}, t^n) = 0,
$$

$$
F(u^n_{ij}, u^n_{ij,3}, n_{ij,3}; M_{ij,3}, t^n) + F(u^n_{il}, u^n_{il,r'}, n_{il,r'}; M_{il,r'}, t^n) = 0
$$

From the conservation relations, it results that

$$
C_i = \sum_{j \in \mathbb{G}(i)} \left( \frac{c^n_{ij}}{\rho_{ij}} |K_{ij}| u^n_{ij} - \Delta t \sum_{r=1,2,3} |S_{ij,r}| F(u^n_{ij}, u^n_{ij,r}, n_{ij,r}; M_{ij,r}, t^n) \right).
$$

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

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

and we rewrite the $C_i$ expression under the form

$$
C_i = \sum_{j \in \mathbb{G}(i)} \left( \frac{c^n_{ij}}{\rho_{ij}} |K_{ij}| u^n_{ij} - \Delta t \sum_{r=1,2,3} |S_{ij,r}| \right)
\left[ F(u^n_{ij}, u^n_{ij,r}, n_{ij,r}; M_{ij,r}, t^n) - F(u^n_{ij}, u^n_{ij,r}, n_{ij,r}; B_{ij}, t^n) \right),
$$

$$
= \sum_{j \in \mathbb{G}(i)} \left( \frac{|K_{ij}|}{\rho_{ij}} \sum_{r=1,2,3} |S_{ij,r}| \right)
\left[ F(u^n_{ij}, u^n_{ij,r}, n_{ij,r}; M_{ij,r}, t^n) - F(u^n_{ij}, u^n_{ij,r}, n_{ij,r}; B_{ij}, t^n) \right).
$$

From the positivity preserving property of the numerical flux (50), each term of the sum is positive if, under condition (58), the time step $\Delta t$ satisfies the condition

$$
\Delta t \frac{\rho_{ij}}{|K_{ij}|} \leq \Delta t \frac{\text{perim}(K_{ij})}{|K_{ij}|} \leq \lambda_0 \zeta_0.
$$

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

$$
\Delta t \leq \lambda_0 \zeta_0 \frac{|K_{ij}|}{\text{perim}(K_{ij})}.
$$

(60)
Noting that
\[ \text{perim}(K_i) \leq \#\nu(i) \max_{j \in L(i)} |S_{ij}| \leq N_L \max_{j \in L(i)} |S_{ij}|, \]
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 \( T \) belong to \( \mathcal{M}(\alpha) \) and \( \mathcal{R} \) be a convex \( \mu \)-admissible reconstruction. Then \( u^n_{K_i} \) is a positive real value function under the CFL condition
\[
\Delta t \leq \frac{\lambda_0}{1 + N\mu C_\omega(\alpha)} \frac{h}{N_L^2} \quad □
\]

**Proof** The convex \( \mu \)-admissible reconstruction assumption says that the equalities
\[
u^n_i - u^n_i + \sum_{k \in \mu(i)} \omega^n_{i,j,k} (u^n_k - u^n_i) = 0, \quad j \in \nu(i)
\]
hold for all \( K_i \in T \). Let \( \zeta^n_{ij}, j \in \nu(i) \) be positive coefficients we shall define ahead, we write the second-order scheme under the following form
\[
|K_i|u^n_{K_i} = |K_i|u^n_i - \Delta t \sum_{j \in \nu(i)} |S_{ij}| \mathcal{F}(u^n_{ij}, u^n_{i}, n_{ij}; M_{ij}, t^n),
\]
\[
= |K_i|u^n_i + \sum_{j \in \nu(i)} \zeta^n_{ij} [K_i] \left( u^n_j - u^n_i + \sum_{k \in \mu(i)} \omega^n_{i,j,k} (u^n_k - u^n_i) \right)
\]
\[
- \Delta t \sum_{j \in \nu(i)} |S_{ij}| \mathcal{F}(u^n_{ij}, u^n_{i}, n_{ij}; M_{ij}, t^n),
\]
\[
= |K_i|u^n_i - |K_i|u^n_i \sum_{j \in \nu(i)} \zeta^n_{ij} (1 + \sum_{k \in \mu(i)} \omega^n_{i,j,k}) + \sum_{k \in \mu(i)} \sum_{j \in \nu(i)} \omega^n_{i,j,k} |K_i|u^n_k
\]
\[
+ \sum_{j \in \nu(i)} \zeta^n_{ij} |K_i|u^n_{ij} - \Delta t \sum_{j \in \nu(i)} |S_{ij}| \mathcal{F}(u^n_{ij}, u^n_{i}, n_{ij}; M_{ij}, t^n),
\]
\[
= A_i + B_i + C_i,
\]
with
\[
A_i := |K_i|u^n_i \left( 1 - \sum_{j \in \nu(i)} \zeta^n_{ij} (1 + \sum_{k \in \mu(i)} \omega^n_{i,j,k}) \right), \quad B_i := \sum_{k \in \mu(i)} \sum_{j \in \nu(i)} \omega^n_{i,j,k} |K_i|u^n_k,
\]
and
\[
C_i := \sum_{j \in \nu(i)} \left[ \zeta^n_{ij} |K_i|u^n_{ij} - \Delta t |S_{ij}| \mathcal{F}(u^n_{ij}, u^n_{i}, n_{ij}; M_{ij}, t^n) \right].
\]
Expression \( A_i \) is non negative if we choose \( \zeta_{ij} \) such that
\[
\sum_{j \in \nu(i)} \zeta^n_{ij} (1 + \sum_{k \in \mu(i)} \omega^n_{i,j,k}) = 1, \quad (62)
\]
which is achieved with

$$\zeta_{ij}^n = \frac{1}{\# \mathcal{I}(i)} \frac{1}{1 + \sum_{k \in \mathcal{P}(i)} \omega_{ijk}^n}.$$ 

Since \( \omega_{ijk}^n \leq C_\omega(\alpha) \) we have the estimate

$$\sum_{k \in \mathcal{P}(i)} \omega_{ijk}^n \leq N_\mu C_\omega$$

and condition (62) is achieved with

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

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 \( \Delta t \) satisfies the CFL condition (61). \( \square \)

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.