Monoslope and Multislope MUSCL Methods for unstructured meshes

Thierry Buffard, Stéphane Clain

Monoslope and Multislope MUSCL Methods for unstructured meshes

Thierry Buffard [a,*], Stéphane Clain [b]

[a] *Université Clermont–Ferrand II, Laboratoire de Mathématiques, UMR CNRS 6620, 63177 Aubière cedex, France*

[b] *Institut de Mathématiques de Toulouse, UMR CNRS 5219, 118 route de Narbonne, 31062 Toulouse cedex, France*

Abstract

We present new MUSCL techniques associated with cell-centered Finite Volume method on triangular meshes. The first reconstruction consists in calculating a one vectorial slope per control volume by a minimization procedure with respect to a prescribed stability condition. The second technique we propose is based on the computation of three scalar slopes per triangle (one per edges) still respecting some stability condition. The resulting algorithm provides a very simple scheme which is extensible to higher dimensional problems. Numerical approximations have been performed to obtain the convergence order for the advection scalar problem whereas we treat a nonlinear vectorial example, namely the Euler system, to show the capacity of the new MUSCL technique to deal with more complexe situations.

Key words: High-order scheme; Finite Volume; multislope method; unstructured mesh; conservation laws

1 Introduction

Large numerical simulations in industrial framework require efficient but rather simple numerical methods to face the modelling complexity while making easier the implementation. Flexibility is also required to quickly adapt the computation code to new conditions and models. High-resolution methods such

* Corresponding author.
Email addresses: Thierry.Buffard@univ-bpclermont.fr (Thierry Buffard), clain@mip.ups-tlse.fr (Stéphane Clain).

that ENO, WENO or Discontinuous Galerkin methods provide very good accuracy. However, the MUSCL technique is more popular in the industrial context due to its natural simplicity and adaptation capacity to respond to modelling evolutions and complexifications.

Monotone Upstream Scheme for Conservation Law technique (MUSCL technique) has been introduced by Van Leer [17] for one dimensional hyperbolic problems. The main idea is a piecewise linear reconstruction of the solution to achieve higher accurate schemes still preserving the stability (the maximum principle for instance). Initially elaborated for one dimensional scalar problems, the MUSCL technique combined with a conservative scheme had to preserve the Total Variation of the solution. To this end, slopes are limited to prevent spurious oscillations or overshooting of the numerical approximations [15]. A first extention of the MUSCL technique to higher dimensions has been proposed using structured meshes where the MUSCL procedure is applied in each direction [5] but the generalisation of the Total Variation constraint for higher dimensional geometries makes the scheme to be a first order method [9]. To get around this negative result, a new class of positive schemes have been introduced [14] which ensures a local maximum principle.
To handle more flexible refinements and allow discretization of complex bounded domains, new MUSCL methods for unstructured meshes have been considered [11], [6], [1]. A local linear representation is constructed on each element using a gradient prediction which should be limited to prevent oscillations of the numerical solutions [7] (see also [8,12,13] for a mathematical study of the high-order schemes).

The classical MUSCL technique consists in two steps. First, a predicted gradient is computed for each element of the mesh using the neighbouring values. Then the gradient is modified to respect some Maximum Principle or Total Variation Diminishing constraint and provide a vectorial slope on the element. New values are therefore computed on each edge of the element using the linear reconstruction. Finally, an approximation of the flux crossing the interface is performed by employing the two reconstructed values situated on both sides of the edge combined with a monotone numerical flux function. To avoid the predictor-corrector algorithm and obtain some optimal reconstruction, we propose to build the vectorial slope on each element by minimizing a convex functional under stability constraints. The idea is to optimize the slope while respecting the Maximum principle or the Total Variation Diminishing property. We intend in this way to produce the best gradient approximation which respects the stability constraint.

The MUSCL method presented above will be referred to as **monoslope method** since the reconstructed values are obtained using the same vectorial slope on each element. We also introduce a new class of MUSCL method named **multislope method** where we use specific scalar slope for each interface. For a

given element, we consider a set of normalized vectors and we use the neighbouring values to compute the scalar slopes representing an approximation of the directional derivatives. The slopes are modified afterwards to respect some stability constraint and finally, the reconstructed values are computed on each edge using the corrected slopes. The main advantage of the method is that we only deal with one dimensional situations and, as we shall show in the sequel, the scalar slopes are very simple to compute even for higher dimensional geometries.

The remainder of the paper is organized as follows. In Section 2, we introduce the notations we shall used in te sequel to describe the finite volume process on triangular meshes for two-dimensional geometries and we review some classical MUSCL-type methods. In particular, we give a precide description of the Maximum Principle domain and the Total Variation Diminishing domain that we employ to keep the stability condition. Section 3 is devoted to a new monoslope MUSCL method while we describe the multislope MUSCL technique in Section 4. Numerical results are presented for the linear advection problem and the Euler system in Section 5.

2 Second order monoslope MUSCL method.

To illustrate the MUSCL reconstruction, we here introduce the classical advection problem but more complex problems such as nonlinear vectorial systems can of course be considered. Let $\Omega \subset R^2$, be a polygonal open bounded set of $\mathbb{R}^2$, $T > 0$. We denote by $\mathbf{V}(t,x)$ a given $\mathbb{R}^2$ vectorial valued function defined on $Q_T = [0,T] \times \overline{\Omega}$. For $t \in [0,T]$, we set

$$\Gamma^-(t) = \{x \in \partial\Omega; \mathbf{V}(t,x).\mathbf{n}(x) < 0\}, \quad \Gamma^+(t) = \{x \in \partial\Omega; \mathbf{V}(t,x).\mathbf{n}(x) \geq 0\},$$

with $x = (x_1, x_2)$ a generic point of Ω and $\mathbf{n}$ the outwards normal on the boundary $\partial\Omega$.

We consider the advection problem: find $U(t,x)$ a real valued function defined on Q_T such that

$$\begin{aligned} \partial_t U + \nabla.(\mathbf{V}U) &= 0 && \text{in }]0,T[\times\Omega, \\ U(t=0,.) &= U_0(.) && \text{in } \Omega, \\ U(t,.) &= U_b(t,.) && \text{on } \Gamma^-(t), t \in]0,T]. \end{aligned}$$

where U_0 and U_b are given functions.

To deal with the numerical approximation, we introduce the following ingre-

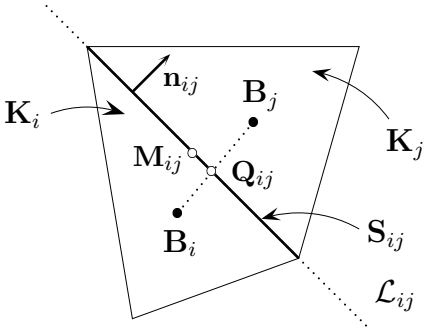

Fig. 1. Notations and conventions of the mesh elements and edges.

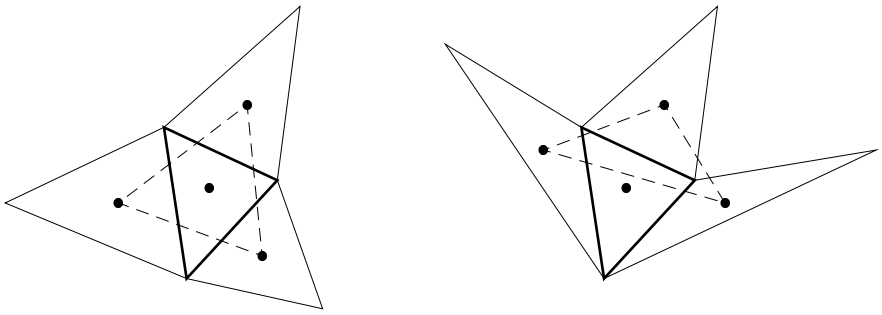

Fig. 2. Configuration satisfying the hypothesis $(\mathcal{H})$ (*on left*). Configuration which not satisfies the hypothesis $(\mathcal{H})$ (*on right*).

dients (see Fig. 1). $\mathcal{T}_h$ is a discretization of Ω with triangles K_i of centroid $\mathbf{B}_i$, $i = 1, \ldots, N$ where N is the number of mesh elements. For a given i, $\nu(i)$ represents the index set of the common edge elements $K_j \in \mathcal{T}_h$, $j \in \nu(i)$ where $S_{ij} = \bar{K}_j \cap \bar{K}_i$ stands for the common edge with midpoint $\mathbf{M}_{ij}$.

We assume furthermore that the mesh satisfies the following hypothesis $(\mathcal{H})$ (see Fig. 2):

$(\mathcal{H})$ *For any $K_i \in \mathcal{T}_h$ such that $|\nu(i)| = 3$, point $\mathbf{B}_i$ is strictly inside the convex set defined by the points $\mathbf{B}_j$, $j \in \nu(i)$.*

If $\mathcal{L}_{ij}$ represents the line containing the edge S_{ij}, point $\mathbf{Q}_{ij}$ is defined as the intersection between the segment $[\mathbf{B}_i, \mathbf{B}_j]$ and the line $\mathcal{L}_{ij}$. Note that $\mathbf{Q}_{ij}$ does not belong *a priori* to S_{ij} but only to $\mathcal{L}_{ij}$. For a given edge S_{ij}, $\mathbf{n}_{ij}$ represents the outward normal of K_i pointing to K_j and $\mathbf{n}_{ji} = -\mathbf{n}_{ij}$.

We will use a cell-centered finite volume method where control volumes are the triangles. The sequence $(t^n)_n$ defines a time discretization of $[0, T]$ with $t^{n+1} = t^n + \Delta t$. Let U_i^n stand for an approximation of the mean value of U at time t^n on the element K_i. The conservative first order finite volume

formulation is given by

$$|K_i|\ U_i^{n+1} = |K_i|\ U_i^n - \Delta t \sum_{j\in\nu(i)} |S_{ij}|\ F_{ij}(U_i^n, U_j^n), \tag{1}$$

where $F_{ij}(U_i, U_j)$ is a numerical flux from K_i to K_j at interface S_{ij}.

For the advection case, classical numerical flux functions are the Lax-Friedrichs flux or the upwind flux:

$$F_{ij}^{LF}(U_i^n, U_j^n) = \frac{1}{2}\left(\mathbf{V}(t^n, \mathbf{B}_i).\mathbf{n}_{ij}\, U_i^n + \mathbf{V}(t^n, \mathbf{B}_j).\mathbf{n}_{ij}\, U_j^n\right) - \lambda(U_j^n - U_i^n),$$
$$F_{ij}^{upwind}(U_i^n, U_j^n) = [\mathbf{V}(t^n, \mathbf{Q}_{ij}).\mathbf{n}_{ij}]^+\, U_i^n + [\mathbf{V}(t^n, \mathbf{Q}_{ij}).\mathbf{n}_{ij}]^-\, U_j^n,$$

where $[.]^+$ represents the positive part and λ is a positive constant to guarantee the scheme stability.

2.1 Classical MUSCL methods

First order schemes give a poor approximation and induce high viscosity effect. A second order scheme provides a better approximation and manages to reduce the viscous smoothing effect in the vicinity of the shocks.

The popular techniques consist in a local linear reconstruction (see [2,8,16]). Assuming that a constant piecewise approximation $U_h^n = (U_i^n)_i$ of U at time t^n is known, we construct a new linear piecewise approximation $\widetilde{U}_h^n$ in the following way

$$\widetilde{U}_i^n(\mathbf{X}) = U_i^n + \mathbf{a}_i.\mathbf{B}_i\mathbf{X}, \quad \mathbf{X} \in K_i, \tag{2}$$

where $\mathbf{B}_i\mathbf{X}$ stands for the vector $\mathbf{X} - \mathbf{B}_i$, $\mathbf{a}_i \in \mathbb{R}^2$ is the vectorial slope we have to construct, $\mathbf{a}_i.\mathbf{B}_i\mathbf{X}$ is the inner product between $\mathbf{B}_i\mathbf{X}$ and $\mathbf{a}_i \in \mathbb{R}^2$.

Remark that such a linear reconstruction satisfies conservation property

$$\int_{K_i} \widetilde{U}_i^n(\mathbf{X})\, dX = |K_i|\ U_i^n,$$

since the centroid point $\mathbf{B}_i$ is chosen as reference point.

Given a point $\mathbf{X}_{ij}$ on the common edge S_{ij}, we set

$$U_{ij}^n = U_i^n + \mathbf{a}_i.\mathbf{B}_i\mathbf{X}_{ij}, \qquad U_{ji}^n = U_j^n + \mathbf{a}_j.\mathbf{B}_j\mathbf{X}_{ij}. \tag{3}$$

We classify this kind of reconstruction as **monoslope method** since we produce values U_{ij}^n on edges S_{ij}, $j \in \nu(i)$ using the same slope: the slope $\mathbf{a}_i$ does not change with subscript j.

Two useful choices for point $\mathbf{X}_{ij}$ are $\mathbf{Q}_{ij}$ or $\mathbf{M}_{ij}$ (see Fig. 1). The first one is natural from a geometrical point of view since it corresponds to the linear interpolation between $\mathbf{B}_i$ and $\mathbf{B}_j$ whereas the second one is natural from the integration point of view since the numerical integration with the midpoint rule is exact for linear functions along the edge S_{ij}.

To obtain a second order method, we then substitute the numerical flux $F_{ij}(U_i^n, U_j^n)$ by $F_{ij}(U_{ij}^n, U_{ji}^n)$ in relation (1) and obtain:

$$|K_i|\ U_i^{n+1} = |K_i|\ U_i^n - \Delta t \sum_{j \in \nu(i)} |S_{ij}|\ F_{ij}(U_{ij}^n, U_{ji}^n). \tag{4}$$

Several slope evaluations have been proposed (see [8], [10], [2] for an exhaustive list), where two leading requirements have to be satisfied:

(C1) the linearly reconstructed function $\widetilde{U}_h$ satisfies $\widetilde{U}_h = U$ if the function U is linear. In this paper, this property will be referred to as linear consistency of the reconstruction;
(C2) the reconstruction has to respect a maximum principle to avoid overshooting leading to a descripency of the numerical approximation.

Remark 1 *In the sequel, we only consider the situation where an element K_i is strictly inside the domain, i.e. the element has no edge on the boundary, otherwise we bring back to a first order scheme setting $\mathbf{a}_i = 0$.*

2.1.1 Gradient methods

Let denote by K_{j_1}, K_{j_2}, K_{j_3} the three adjacent triangles of K_i. We consider the three following hyperplans in the x_1, x_2, U space: hyperplane $\pi_{i,1}$ is defined by the points $\mathbf{B}_i$, $\mathbf{B}_{j_2}$, $\mathbf{B}_{j_3}$ with elevations U_i, U_{j_2}, U_{j_3} and $\pi_{i,2}$, $\pi_{i,3}$ are obtained in the same way. The hyperplane $\pi_{1,2,3}$ is defined by the points $\mathbf{B}_{j_1}$, $\mathbf{B}_{j_2}$, $\mathbf{B}_{j_3}$ with elevations U_{j_1}, U_{j_2}, U_{j_3} (see Fig. 3).

For example, $\pi_{i,1}$ is given by equation

$$(u - U_i^n) = \mathbf{G}_{i,1} \,.\, \mathbf{B}_i\mathbf{X}$$

where $\mathbf{G}_{i,1} \in \mathbb{R}^2$ while $\pi_{1,2,3}$ is given by

$$(u - U_{j_1}^n) = \mathbf{G}_{1,2,3} \,.\, \mathbf{B}_{j_1}\mathbf{X}.$$

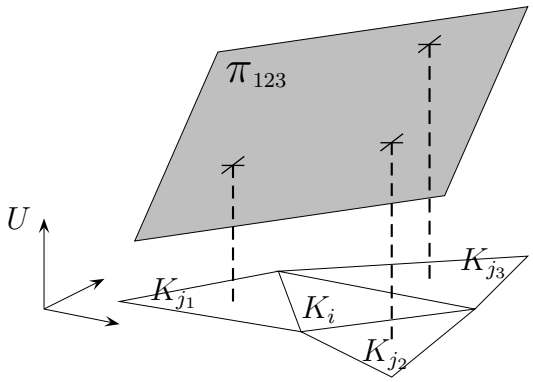

Fig. 3. Plane $\pi_{1,2,3}$ representation.

A first choice consists to take the slope $\mathbf{a}_i = \mathbf{G}_{1,2,3}$ and we obtain a linear consistent reconstruction. Other possible choices use a combinaison of $\mathbf{G}_{i,1}$, $\mathbf{G}_{i,2}$, $\mathbf{G}_{i,3}$, setting

$$\mathbf{a}_i = \sigma(\mathbf{G}_{i,1}, \mathbf{G}_{i,2}, \mathbf{G}_{i,3}).$$

The linear consistency is obtained if and only if $\mathbf{a} = \sigma(\mathbf{a}, \mathbf{a}, \mathbf{a})$ for all $\mathbf{a} \in \mathbb{R}^2$.

2.1.2 Minimization method

In reference [4], the authors consider the hyperplane minimizing the distance with the four points $(\mathbf{B}_i, U_i)$, $(\mathbf{B}_j, U_j)$, $j \in \nu(i)$. One has to seek a vector $\mathbf{G}_{LS}$ using a Least Square Method, that is to say which minimizes the functional

$$E(\mathbf{a}) = \sum_{j \in \nu(i)} \Big(U_j^n - (U_i^n + \mathbf{a}.\mathbf{B}_i\mathbf{B}_j) \Big)^2. \tag{5}$$

Existence and uniqueness of the minimum is obvious since the functional is strictly convex.

Moreover, if U is linear, the four points lie in the same hyperplane and the minimum corresponds to the gradient of U, hence we get the linear consistency of the reconstruction.

2.2 The stability conditions.

Let consider two adjacent triangles K_i and K_j. To avoid numerical artefacts in the vicinity of large gradients (overshooting or spurious oscillations), one imposes that the reconstructed values U_{ij} and U_{ji} on S_{ij} satisfy some stability property. To this end, we introduce the following conditions:

(1) The L^∞ stability condition (Maximum Principle constraint or MP constraint):

$$\min(U_i^n, U_j^n) \leq U_{ij}^n, U_{ji}^n \leq \max(U_i^n, U_j^n). \tag{6}$$

(2) The Total Variation-like condition (TVD constraint):

$$\text{if } U_i^n \leq U_j^n \text{ then } U_i^n \leq U_{ij}^n \leq U_{ji}^n \leq U_j^n. \tag{7}$$

The last condition is named TVD constraint since the property (7) implies the preservation of the BV norm between the initial piecewise constant function and its piecewise linear reconstruction. Moreover, relation (7) implies relation (6) so the Total Variation condition is a subcase of the L^∞ stability condition.

The slope $\mathbf{a}_i$ provided by one of the above methods does not *a priori* satisfy the stability condition. We impose the stability multiplying the slope by a limiter ϕ_i such that the values U_{ij}^n and U_{ji}^n obtained with the new slope $\widetilde{\mathbf{a}}_i = \phi_i \mathbf{a}_i$ satisfy one of the two stability conditions. In particular, if $\phi_i = 0$, we find again the first order scheme.

In the case of a linear solution, a predicted slope process which satisfies condition (C1) provides a slope equal to the function gradient. In this particular case, the limiting procedure has no impact since the predicted slope respects the two stability constraints and one has $\phi_i = 1$. Therefore, it is natural to choose the highest value of $\phi_i \in [0, 1]$ such that the reconstruction statisfies a prescribed stability condition.

2.2.1 The Maximum Principle domain

For a given element K_i, we define the Maximum Principle domain (MP domain) as

$$MP_i = \{\mathbf{a} \in \mathbb{R}^2 \ ; \ \min(U_j^n - U_i^n, 0) \leq \mathbf{a}.\mathbf{B}_i\mathbf{Q}_{ij} \leq \max(U_j^n - U_i^n, 0), \ j \in \nu(i)\}.$$

If $\mathbf{a}_i \in MP_i$ then $U_{ij}^n = U_i^n + \mathbf{a}_i.\mathbf{B}_i\mathbf{Q}_{ij}$ satisfies the stability condition (6) and conversely.

For the sake of simplicity, we introduce a new set of vectors

$$\mathbf{s}_k = \operatorname{sgn}(U_{j_k}^n - U_i^n)\, \mathbf{B}_i\mathbf{Q}_{ij_k}, \quad k = 1, 2, 3,$$

where $\operatorname{sgn}(x) = \begin{cases} 1 \text{ for } x \geq 0 \\ -1 \text{ for } x < 0. \end{cases}$

The MP_i region is now simply given by

$$MP_i = \{\mathbf{a} \in \mathbb{R}^2 \ ; \ 0 \leq \mathbf{a}.\mathbf{s}_k \leq \gamma_k, \ k = 1, 2, 3\}$$

with $\gamma_k = |U_i^n - U_{j_k}^n|$, $k = 1, 2, 3$.

We require that sgn(0) being non zero in order to extend the equivalence:

$$\mathbf{a}.\mathbf{B}_i\mathbf{Q}_{ij_k} = 0 \quad \Longleftrightarrow \quad \mathbf{a}.\mathbf{s}_k = 0,$$

to the particular situation $U_{j_k}^n = U_i^n$.

Hypothesis $(\mathcal{H})$ implies that any couple of the three vectors $\mathbf{s}_k$, $k = 1, 2, 3$ defined a basis of the $\mathbb{R}^2$ space. Therefore we can express one vector from the two others and we have the following unique expansions with non zero coefficients:

$$\mathbf{s}_1 = \alpha_{12} \ \mathbf{s}_2 + \alpha_{13} \ \mathbf{s}_3 \tag{8}$$
$$\mathbf{s}_2 = \alpha_{21} \ \mathbf{s}_1 + \alpha_{23} \ \mathbf{s}_3 \tag{9}$$
$$\mathbf{s}_3 = \alpha_{31} \ \mathbf{s}_1 + \alpha_{32} \ \mathbf{s}_2. \tag{10}$$

A withdraw computation gives the following proposition.

Proposition 2 *We have the relations*

$$\alpha_{lm} \ \alpha_{ml} = 1, \tag{11}$$
$$\alpha_{lk} \ \alpha_{km} = -\alpha_{lm}, \tag{12}$$

for any circular permutation (l, m, k) *of* $(1, 2, 3)$.

PROOF. To check properties (11) and (12), let us consider the decomposition of $\mathbf{s}_1$

$$\mathbf{s}_1 = \alpha_{12} \ \mathbf{s}_2 + \alpha_{13} \ \mathbf{s}_3.$$

Thanks to hypothesis $(\mathcal{H})$, $\mathbf{s}_1$ is neither colinear to $\mathbf{s}_2$ nor to $\mathbf{s}_3$, hence coefficients α_{12} and α_{13} do not vanish. The relation can be rewritten

$$\mathbf{s}_2 = \frac{1}{\alpha_{12}} \ \mathbf{s}_1 - \frac{\alpha_{13}}{\alpha_{12}} \ \mathbf{s}_3 = \alpha_{21} \ \mathbf{s}_1 + \alpha_{23} \ \mathbf{s}_3$$

which gives relations (11) and (12) by identification thanks to the uniqueness of the decomposition. □

We deduce that MP_i domain is only characterized by coefficients α and γ. If at least two of the three γ coefficients vanish, we easily deduce $MP_i = \{(0, 0)\}$.

We now consider the other situations.

Proposition 3 *Assume that one coefficient, say γ_k, vanishes while the two others, say γ_l and γ_m, are not zero. Then we have*

$$MP_i = \{(0,0)\} \quad \Longleftrightarrow \quad \alpha_{lm} < 0.$$

PROOF. Let us first remark that if $\mathbf{a} \in \mathbb{R}^2$ with $\mathbf{a}.\mathbf{s}_k = 0$, then we have:

$$\mathbf{a}.\mathbf{s}_l = \alpha_{lk}\, \mathbf{a}.\mathbf{s}_k + \alpha_{lm}\, \mathbf{a}.\mathbf{s}_m = \alpha_{lm}\, \mathbf{a}.\mathbf{s}_m. \tag{13}$$

$\Leftarrow$) Suppose that $\alpha_{lm} < 0$ and let $\mathbf{a} \in MP_i$.
Since $\gamma_k = 0$, relation (13) is satisfied. From condition $\mathbf{a} \in MP_i$, we have the relations $\mathbf{a}.\mathbf{s}_l \geq 0$ and $\mathbf{a}.\mathbf{s}_m \geq 0$. It follows that $\mathbf{a}.\mathbf{s}_m = \mathbf{a}.\mathbf{s}_l = 0$ since we have $\alpha_{lm} < 0$. Hence $\mathbf{a} = (0,0)$.

$\Rightarrow$) Conversely, suppose that $\alpha_{lm} \geq 0$. Since all the coefficients are non-vanishing, we have $\alpha_{lm} > 0$. We shall now construct a non zero vector of MP_i. To this end, consider $\mathbf{a} \in \mathbb{R}^2$ such that $\mathbf{a}.\mathbf{s}_k = 0$ and $\mathbf{a}.\mathbf{s}_l = \min(\gamma_l, \alpha_{lm}\,\gamma_m) \in]0, \gamma_l]$ .We obtain a non zero vector which satisfies $0 < \mathbf{a}.\mathbf{s}_m = \dfrac{1}{\alpha_{lm}}\, \mathbf{a}.\mathbf{s}_l \leq \gamma_m$, then $\mathbf{a} \in MP_i$. □

Remark 4 *If only one of the γ_k is zero, the MP_i domain is reduced to the null vector or to a segment.*

Proposition 5 *Assume that all the coefficients γ_k are positive, $k = 1, 2, 3$. Then the following assertions are equivalent:*

(i) $\alpha_{12} < 0$ *and* $\alpha_{13} < 0$.
(ii) $\alpha_{21} < 0$ *and* $\alpha_{23} < 0$.
(iii) $\alpha_{31} < 0$ *and* $\alpha_{32} < 0$.
(iv) $MP_i = \{(0,0)\}$.

PROOF. Equivalences between (i), (ii) and (iii) derive from relations (11)-(12). It remains to prove the equivalence between (i) and (iv). To this end, let us assume that assertion (i) holds and let $\mathbf{a} \in MP_i$. One has

$$\mathbf{a}.\mathbf{s}_1 = \alpha_{12}\, \mathbf{a}.\mathbf{s}_2 + \alpha_{13}\, \mathbf{a}.\mathbf{s}_3, \quad \text{with} \quad \alpha_{12}\, \mathbf{a}.\mathbf{s}_2 \leq 0 \quad \text{and} \quad \alpha_{13}\, \mathbf{a}.\mathbf{s}_3 \leq 0. \tag{14}$$

It follows that $\mathbf{a}.\mathbf{s}_1 \leq 0$, hence that $\mathbf{a}.\mathbf{s}_1 = 0$ since $\mathbf{a}.\mathbf{s}_1 \geq 0$. Relation (14) now gives $\mathbf{a}.\mathbf{s}_2 = \mathbf{a}.\mathbf{s}_3 = 0$ and we conclude that $\mathbf{a}$ is the null vector because $\mathbf{s}_1, \mathbf{s}_2$ is a basis.

Conversely, let us assume that (i) does not hold. We shall construct a non zero vector $\mathbf{a}$ such that $\mathbf{a} \in MP_i$.
Since assertion (i) is wrong, we have $\alpha_{12} > 0$ or $\alpha_{13} > 0$. Suppose $\alpha_{12} > 0$ for example. Let $\mathbf{a}$ be the vector of $\mathbb{R}^2$ such that $\mathbf{a}.\mathbf{s}_3 = 0$ and $\mathbf{a}.\mathbf{s}_2 = \min\left(\frac{\gamma_1}{\alpha_{12}}\,,\,\gamma_2\right) \in]0, \gamma_2]$ .We obtain a non zero vector which satisfies $0 < \mathbf{a}.\mathbf{s}_1 = \alpha_{12}\ \mathbf{a}.\mathbf{s}_2 \leq \gamma_1$, then $\mathbf{a} \in MP_i$. □

Under the same assumption as the above proposition, we have the following corollary using relation (12).

Corollary 6 *Assume that all the coefficients γ_k are positive, $k = 1, 2, 3$. Then the MP_i domain is not reduced to the null vector if and only if one of the three following assertions holds*

(i) $\alpha_{12} > 0$ *and* $\alpha_{13} > 0$,
(ii) $\alpha_{21} > 0$ *and* $\alpha_{23} > 0$,
(iii) $\alpha_{31} > 0$ *and* $\alpha_{32} > 0$.

PROOF. ⇒) We first assume that MP_i domain is not reduced to the null vector. From proposition 5, we deduce that $\alpha_{12} \geq 0$ or $\alpha_{13} \geq 0$, hence $\alpha_{12} > 0$ or $\alpha_{13} > 0$ since the coefficients are non zero. If both the coefficients are positive, assertion (i) is right otherwise one of the two coefficients is negative (says $\alpha_{13} < 0$). From relations (11) and (12) we have $\alpha_{21} > 0$ and $\alpha_{23} > 0$ and assertion (ii) holds.

⇐) Conversely, if for example $\alpha_{12} > 0$ and $\alpha_{13} > 0$ then proposition 5 immediately implies that MP_i domain is not reduced to the null vector. □

When MP_i domain is not reduce to the null vector one of the three assertions of corollary 6 holds. In this case, we adopt the following convention:

Convention *We choose the local indexation such that* $\alpha_{31} > 0$ *and* $\alpha_{32} > 0$.

The MP_i domain is a convex polygonal set (see Fig. 4) which consists in the intersection of the three bands limited by the lines

$$d_k = \{\mathbf{a} \in \mathbb{R}^2;\ \mathbf{a}.\mathbf{s}_k = \gamma_k\},\ k = 1, 2, 3, \tag{15}$$
$$\delta_k = \{\mathbf{a} \in \mathbb{R}^2;\ \mathbf{a}.\mathbf{s}_k = 0\},\ k = 1, 2, 3. \tag{16}$$

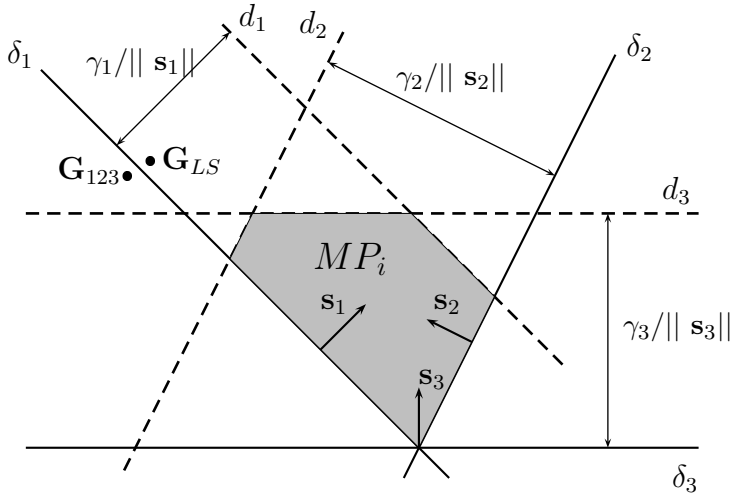

Fig. 4. Maximum Principle domain

2.2.2 The slope limiter

Let $\mathbf{a}_i$ be a predicted gradient obtained, for example, by one of the methods presented in Section 2.1. The Maximum Principle constraint yields that $\mathbf{a}_i$ has to be in the MP_i domain. If not, we reduce the slope by a limiter $\phi_i \in [0,1]$ such that $\widetilde{\mathbf{a}}_i = \phi_i \mathbf{a}_i \in MP_i$. The most classical limiting procedure (see [8], [2]) consists in constructing the three limiters

$$\phi_{i,k} = \begin{cases} \max\left(0, \dfrac{\gamma_k}{\mathbf{a}_i.\mathbf{s}_k}\right) & \text{if } \ \mathbf{a}_i.\mathbf{s}_k \neq 0, \\ 1 & \text{if } \ \mathbf{a}_i.\mathbf{s}_k = 0. \end{cases} \tag{17}$$

Taking $\phi_i = \min(1, \phi_{i,1}, \phi_{i,2}, \phi_{i,3})$, we set $\widetilde{\mathbf{a}}_i = \phi_i \mathbf{a}_i \in MP_i$. If for one local subscript k, $\mathbf{a}_i.\mathbf{s}_k < 0$, the limiter is zero and we obtain a first order method. Numerical experiences indicate that such a phenomena often occurs resulting in a poor approximation accuracy [4]. For example, let us consider a configuration where the predicted slope $\mathbf{G}_{1,2,3}$ is on the left side of line δ_1 while $\mathbf{G}_{LS}$ stands on the right side (see Fig. 4). Applying the limiting procedure (17) yields that $\phi_i > 0$ if we choose $\mathbf{a}_i = \mathbf{G}_{LS}$ whereas $\phi_i = 0$ if we choose $\mathbf{a}_i = \mathbf{G}_{1,2,3}$. In the first case, the resulting slope provides a second order scheme but the second situation reduces to a first order scheme.

To avoid the descrepancy, some authors propose to limit the predicted gradient using a orthogonal projection of point $\mathbf{a}_i$ on the boundary of the MP_i domain (see [10]).

2.2.3 The TVD domain

We now consider the more restrictive TVD constraint (7) using the same framework introduced for the MP constraint. For a given edge S_{ij}, the TVD constraint involves the two slopes $\mathbf{a}_i$ and $\mathbf{a}_j$ which also depends on the neighbouring elements leading to a coupling between all the slopes. To avoid the complex interactions between the slopes, we introduce a more restrictive definition of the TVD constraint such that $\mathbf{a}_i$ is computed independently of the other slopes but only depends on the data of the three neighbouring elements. The requirement on slope $\mathbf{a}_i$ is that reconstructed values U_{ij}^n and U_{ji}^n (with $j \in \nu(i)$) have to satisfy

$$\text{if } U_i^n \leq U_j^n \text{ then } U_i^n \leq U_{ij}^n \leq U_{ij}^{ref} \leq U_{ji}^n \leq U_j^n \tag{18}$$

where U_{ij}^{ref} is the reference value at point $\mathbf{Q}_{ij}$ defined by

$$U_{ij}^{ref} = U_i^n + \frac{|\mathbf{B}_i\mathbf{Q}_{ij}|}{|\mathbf{B}_i\mathbf{B}_j|}\,(U_j^n - U_i^n) = U_j^n + \frac{|\mathbf{B}_j\mathbf{Q}_{ij}|}{|\mathbf{B}_j\mathbf{B}_i|}\,(U_i^n - U_j^n) = U_{ji}^{ref}. \tag{19}$$

We define the TVD_i domain by

$$TVD_i = \{\mathbf{a} \in \mathbb{R}^2;\ \min(U_{ij}^{ref} - U_i^n, 0) \leq \mathbf{a}.\mathbf{B}_i\mathbf{Q}_{ij} \leq \max(U_{ij}^{ref} - U_i^n, 0),\ j \in \nu(i)\}.$$

The TVD_i domain is also characterized by

$$TVD_i = \{\mathbf{a} \in \mathbb{R}^2;\ 0 \leq \mathbf{a}.\mathbf{s}_k \leq \mu_k,\ k = 1, 2, 3\}$$

with $\mu_k = |U_i^n - U_{ij_k}^{ref}|$, $k = 1, 2, 3$.

The TVD_i domain is a convex polygonal set which consists in the intersection of the three bands limited by the lines

$$d_k = \{\mathbf{a} \in \mathbb{R}^2;\ \mathbf{a}.\mathbf{s}_k = \mu_k\},\ k = 1, 2, 3, \tag{20}$$
$$\delta_k = \{\mathbf{a} \in \mathbb{R}^2;\ \mathbf{a}.\mathbf{s}_k = 0\},\ k = 1, 2, 3. \tag{21}$$

To conclude the section, notice that

$$\mu_k = \frac{|\mathbf{B}_i\mathbf{Q}_{i,j_k}|}{|\mathbf{B}_i\mathbf{B}_{j_k}|}\,\gamma_k \leq \gamma_k,$$

hence, we deduce that the TVD_i domain is a subset of the MP_i domain and all the limiting techniques presented for the MP_i domain can directly be adapted to the TVD_i domain using μ_k in place of γ_k.

3 A new monoslope method

All the second order schemes presented above are developed following two steps: first we compute a predicted slope and, secondly, we use a limiting procedure. We propose here a new method where we build the slope in only one procedure in which we optimize the slope under the MP constraint or the TVD constraint.

As we state in the convention presented in subsection 2.2.1, we choose the local indexation such that the coefficients α_{31} and α_{32} are positive.

3.1 Minimization under the TVD constraint

We only present the construction of the optimized slope respecting the TVD constraint. The construction of the optimized slope under the MP constraint can also been considered and adapted.

3.1.1 Problem formulation

Let us consider a triangular control volume K_i. It is clear that if U is a linear function defined by $U(\mathbf{X}) = U_0 + \mathbf{L}.\mathbf{X}$, then $U(\mathbf{Q}_{ij}) = U(\mathbf{B}_i) + \mathbf{L}.\mathbf{B}_i\mathbf{Q}_{ij} = U_{ij}^{ref}$ for all $j \in \nu(i)$. For the general case, we wish to obtain a slope $\mathbf{a}_i$ on K_i for which deviations $U_i + \mathbf{a}_i.\mathbf{B}_i\mathbf{Q}_{ij} - U_{ij}^{ref}$ are as close as possible to 0. Moreover, the slope should provide a reconstruction which respects the stability condition.

We then compute the slope by using a least square method under the TVD constraint on element K_i and the optimization problem reads:

find the slope $\widetilde{\mathbf{a}}_i$ minimizing the functional

$$E_i(\mathbf{a}) = \sum_{j\in\nu(i)} \left(U_{ij}^{ref} - (U_i + \mathbf{a}.\mathbf{B}_i\mathbf{Q}_{ij}) \right)^2 \quad \text{with } \mathbf{a} \in TVD_i. \tag{22}$$

Using the notations introduced in Section 2.2.1, we can rewrite the minimization problem as

$$E_i(\mathbf{a}) = \sum_{k=1,2,3} (\mu_k - \mathbf{a}.\mathbf{s}_k)^2 \tag{23}$$

$$\text{with } \ 0 \le \mathbf{a}.\mathbf{s}_k \le \mu_k, \ k = 1, 2, 3. \tag{24}$$

Remark 7 *We can also consider another minimization problem using the minimization functional (5). If we add now the MP constraint (see [3]), the optimization problem then reads:*

find the slope $\widetilde{\mathbf{a}}_i$ *minimizing the functional*

$$E_i(\mathbf{a}) = \sum_{j \in \nu(i)} \left(U_j - (U_i + \mathbf{a}.\mathbf{B}_i\mathbf{B}_j) \right)^2 \quad \textit{with } \mathbf{a} \in MP_i. \tag{25}$$

Note that problem (22) is not equivalent to problem (25). □

Since the functionnal (23) is strictly convex and the domain defined by (24) is convex and bounded, we get the existence and the uniqueness of the minimum $\widetilde{\mathbf{a}}_i$. With the slope in hand, we build the new predicted values at any given collocation point $\mathbf{X}_{ij}$

$$U_{ij} = U_i + \widetilde{\mathbf{a}}_i \,.\, \mathbf{B}_i\mathbf{X}_{ij}, \quad j \in \nu(i). \tag{26}$$

3.1.2 Computation of the optimal slope

We are now interested in finding the minimum $\widetilde{\mathbf{a}}$ of the functional (23) under constraints (24). To simplify the notations, we skip the index i in this subsection.

We first note that $\widetilde{\mathbf{a}}$ is obviously the null vector if $TVD = \{(0,0)\}$. Note that if $\mu_3 = 0$, we have $TVD = \{(0,0)\}$ by the indexation convention.

From now on we make the assumption that μ_1, μ_2 and μ_3 are positive. The case where $\mu_1 = 0$ or $\mu_2 = 0$ will also be treated further.

Proposition 8 *Let* $\overline{\mathbf{a}}$ *be the minimum of* $E(\mathbf{a})$ *without constraint then* $\overline{\mathbf{a}}$ *is inside the triangle* T_{123} *formed by the three lines* d_k, $k = 1, 2, 3$ *defined by relation (20). In particular, if the triangle is not reduced to a point,* $\overline{\mathbf{a}}$ *is strictly inside the triangle.*

PROOF. Let us set $\mathbf{G}_1 = d_2 \cap d_3$ (see Fig. 5). We then have

$$\mathbf{G}_1.\mathbf{s}_2 = \mu_2, \quad \mathbf{G}_1.\mathbf{s}_3 = \mu_3.$$

We define in the same way $\mathbf{G}_2 = d_1 \cap d_3$ and $\mathbf{G}_3 = d_1 \cap d_2$ satisfying

$$\mathbf{G}_2.\mathbf{s}_1 = \mu_1, \ \mathbf{G}_2.\mathbf{s}_3 = \mu_3, \quad \mathbf{G}_3.\mathbf{s}_1 = \mu_1, \ \mathbf{G}_3.\mathbf{s}_2 = \mu_2.$$

If $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ belong to the same line then hyptothesis $(\mathcal{H})$ yields $\mathbf{G}_1 = \mathbf{G}_2 = \mathbf{G}_3 = \mathbf{G}$, thus $\overline{\mathbf{a}} = \mathbf{G}$ since $E(\overline{\mathbf{a}}) = 0$ in this exceptional case.

We now assume that the three points define a non degenerated triangle T_{123} and we seek $\overline{\mathbf{a}} = \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \lambda_3\mathbf{G}_3$ using the barycentric coordinates with $\lambda_1 + \lambda_2 + \lambda_3 = 1$.
Existence and uniqueness of the minimun $\overline{\mathbf{a}}$ is clear since $E(\mathbf{a})$ is strictly convex and $\overline{\mathbf{a}}$ has to satisfy the linear system

$$\sum_{k=1,2,3} (\mu_k - \overline{\mathbf{a}}.\mathbf{s}_k)\mathbf{s}_k = 0. \tag{27}$$

Using the barycentric coordinates property and the definition of $\mathbf{G}_k$, we get

$$\sum_{k=1,2,3} \lambda_k(\mu_k - \mathbf{G}_k.\mathbf{s}_k)\mathbf{s}_k = 0.$$

The inner product between the last relation and vector $\mathbf{G}_1$ gives

$$\lambda_1(\mu_1 - \mathbf{G}_1.\mathbf{s}_1)\mathbf{G}_1.\mathbf{s}_1 + \lambda_2(\mu_2 - \mathbf{G}_2.\mathbf{s}_2)\mu_2 + \lambda_3(\mu_3 - \mathbf{G}_3.\mathbf{s}_3)\mu_3 = 0.$$

Using also vector $\mathbf{G}_2$ and $\mathbf{G}_3$, we obtain

$$\lambda_1(\mu_1 - \mathbf{G}_1.\mathbf{s}_1)\mu_1 + \lambda_2(\mu_2 - \mathbf{G}_2.\mathbf{s}_2)\mathbf{G}_2.\mathbf{s}_2 + \lambda_3(\mu_3 - \mathbf{G}_3.\mathbf{s}_3)\mu_3 = 0,$$

$$\lambda_1(\mu_1 - \mathbf{G}_1.\mathbf{s}_1)\mu_1 + \lambda_2(\mu_2 - \mathbf{G}_2.\mathbf{s}_2)\mu_2 + \lambda_3(\mu_3 - \mathbf{G}_3.\mathbf{s}_3)\mathbf{G}_3.\mathbf{s}_3 = 0.$$

From the three relations, we deduce

$$\lambda_1(\mu_1 - \mathbf{G}_1.\mathbf{s}_1)^2 = \lambda_2(\mu_2 - \mathbf{G}_2.\mathbf{s}_2)^2 = \lambda_3(\mu_3 - \mathbf{G}_3.\mathbf{s}_3)^2 \tag{28}$$

Since triangle T_{123} is not reduced to a point, the quantities $(\mu_k - \mathbf{G}_k.\mathbf{s}_k)^2$ are positive and thus the coordinates λ_k have the same sign. Moreover, the condition $\lambda_1 + \lambda_2 + \lambda_3 = 1$ yields that $\lambda_k > 0$, hence $\overline{\mathbf{a}}$ is strictly inside the triangle. □

Remark 9 *An explicit calculation of coefficients λ_k provides an expression independant of U_i and U_j:*

$$\lambda_1 = \frac{\alpha_{31}^2}{1 + \alpha_{31}^2 + \alpha_{32}^2}, \quad \lambda_2 = \frac{\alpha_{32}^2}{1 + \alpha_{31}^2 + \alpha_{32}^2}, \quad \lambda_3 = \frac{1}{1 + \alpha_{31}^2 + \alpha_{32}^2}.$$

Remark 10 *The exceptional situation where the triangle T_{123} is reduced to a point corresponds to the case where the four points $(\mathbf{B}_i, U_i^n)$, $(\mathbf{B}_j, U_j^n)$, $j = j_1, j_2, j_3$ lie in the same hyperplane of the (x_1, x_2, U) space. In this case, the optimal slope $\widetilde{\mathbf{a}}$ under constraint corresponds to the optimal slope $\overline{\mathbf{a}}$ without*

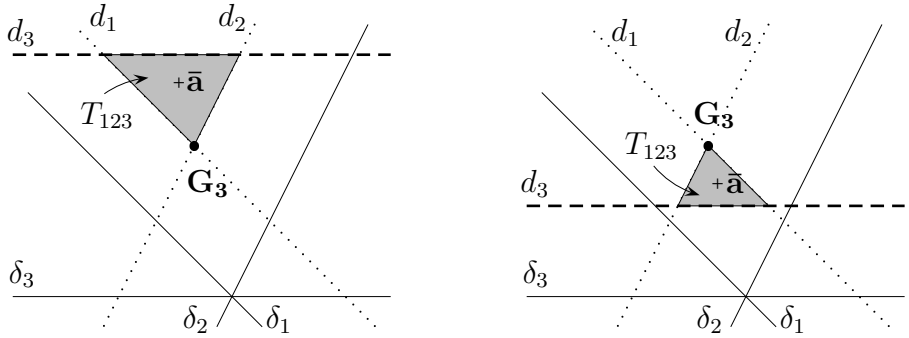

Fig. 5. The triangle T_{123} is above point $\mathbf{G}_3$ (*left*). The triangle T_{123} is under point $\mathbf{G}_3$ (*right*).

constraint and the reconstruction is consistent for linear functions.

Corollary 11 *If triangle T_{123} is not reduced to a point, the minimum without constraint $\overline{\mathbf{a}}$ does not satisfy the TVD constraint. Furthermore the minimum $\widetilde{\mathbf{a}}$ with constraint satisfies, at least, one of the six constraints: $\widetilde{\mathbf{a}}.\mathbf{s}_k = 0$ or $\widetilde{\mathbf{a}}.\mathbf{s}_k = \mu_k$ with $k = 1, 2, 3$, i.e. $\widetilde{\mathbf{a}} \in \partial TVD$.*

PROOF. We notice that $TVD \cap T_{123}$ is reduced to the point $\mathbf{G}_3$ or the segment $[\mathbf{G}_1, \mathbf{G}_2]$ whether d_3 is above ($\mu_3 \geq \mathbf{G}_3.\mathbf{s}_3$, see Fig. 5 *left*) or under ($\mu_3 \leq \mathbf{G}_3.\mathbf{s}_3$, see Fig. 5 *right*) the point $\mathbf{G}_3$. Since $\overline{\mathbf{a}}$ is strictly inside T_{123}, we conclude that $\overline{\mathbf{a}} \notin TVD$.

Finally, if $\widetilde{\mathbf{a}}$ is strictly inside the TVD domain, then no constraint is saturated and we have $\nabla E(\widetilde{\mathbf{a}}) = 0$ thus $\widetilde{\mathbf{a}} = \overline{\mathbf{a}}$ which is not possible since $\overline{\mathbf{a}} \notin TVD$. □

Proposition 12 *The minimum under constraint $\widetilde{\mathbf{a}}$ belongs to d_1, d_2 or d_3.*

PROOF. Let us denote by ${\mathbf{s}_1}^{\perp}$ the orthogonal vector to $\mathbf{s}_1$ such that ${\mathbf{s}_1}^{\perp}.\mathbf{s}_3 > 0$. Since α_{31} and α_{32} are positive, we have also ${\mathbf{s}_1}^{\perp}.\mathbf{s}_2 > 0$. In the other hand, the half-line δ_1 which touches the TVD domain is characterized by $\lambda {\mathbf{s}_1}^{\perp}$ with $\lambda > 0$ and we have

$$E(\lambda {\mathbf{s}_1}^{\perp}) = \mu_1^2 + (\mu_2 - \lambda {\mathbf{s}_1}^{\perp}.\mathbf{s}_2)^2 + (\mu_3 - \lambda {\mathbf{s}_1}^{\perp}.\mathbf{s}_3)^2.$$

Since ${\mathbf{s}_1}^{\perp}.\mathbf{s}_3 > 0$ and ${\mathbf{s}_1}^{\perp}.\mathbf{s}_2 > 0$, we deduce that E decreases as λ increases till it reaches the first of the two intersection points $\delta_1 \cap d_2$ or $\delta_1 \cap d_3$. In conclusion the minimum $\widetilde{\mathbf{a}}$ does not belong to $\delta_1 \cap TVD$ excepted point $\delta_1 \cap d_2$ or $\delta_1 \cap d_3$. The same arguments hold using vectors ${\mathbf{s}_2}^{\perp}$ and the minimum $\widetilde{\mathbf{a}}$ does not belong to $\delta_2 \cap TVD$ excepted points $\delta_2 \cap d_1$ or $\delta_2 \cap d_3$. □

Remark 13 *If $\mu_1 = 0$ and $\mu_2 \mu_3 \neq 0$, the TVD domain is reduced to a segment on line δ_1. The previous proof shows in that case that the minimum $\widetilde{\mathbf{a}}$ is $\delta_1 \cap d_2$*

or $\delta_1 \cap d_3$. The case $\mu_2 = 0$ and $\mu_1\mu_3 \neq 0$ is similar.

We precise the position of the minimum with constraint in the next proposition.

Proposition 14 *Let $\tilde{\mathbf{a}}$ be the minimum with the TVD constraint. Then we have the following alternative:*

i) If d_3 is above point $d_1 \cap d_2$ then $\tilde{\mathbf{a}} = \mathbf{G}_3$.
ii) If d_3 is under point $d_1 \cap d_2$ then $\tilde{\mathbf{a}}$ belongs to d_3.

PROOF. We first study the situation for the line d_1 where we prove that $\tilde{\mathbf{a}}$ does not belong to d_1 exept point $\mathbf{G}_3$. The same argument holds for line d_2. Since $\mathbf{G}_3 = d_1 \cap d_2$, we have $\mathbf{G}_3.\mathbf{s}_1 = \mu_1$ and $\mathbf{G}_3.\mathbf{s}_2 = \mu_2$. Consider now a point $\mathbf{a}$ on the segment $d_1 \cap TVD$. Using the parametrisation

$$\mathbf{a} = \mathbf{G}_3 + \lambda {\mathbf{s}_1}^{\perp}, \tag{29}$$

we obtain

$$E(\mathbf{a}) = E(\mathbf{G}_3 + \lambda {\mathbf{s}_1}^{\perp}) = F(\lambda) = \lambda^2({\mathbf{s}_1}^{\perp}.\mathbf{s}_2)^2 + (\mu_3 - \mathbf{G}_3.\mathbf{s}_3 - \lambda {\mathbf{s}_1}^{\perp}.\mathbf{s}_3)^2. \tag{30}$$

We get a convex parabolic curve and the minimum λ_0 is given by

$$\lambda_0 = \frac{(\mu_3 - \mathbf{G}_3.\mathbf{s}_3){\mathbf{s}_1}^{\perp}.\mathbf{s}_3}{({\mathbf{s}_1}^{\perp}.\mathbf{s}_3)^2 + ({\mathbf{s}_1}^{\perp}.\mathbf{s}_2)^2}.$$

Due to the orientation convention ${\mathbf{s}_1}^{\perp}.\mathbf{s}_3 > 0$, any point $\mathbf{a} \in d_1 \cap TVD$ satisfies $\lambda \leq 0$.

Case i)
If d_3 is above $\mathbf{G}_3$, *i.e.* $\mathbf{G}_3.\mathbf{s}_3 < \mu_3$ then $\lambda_0 > 0$ and we deduce that the minimum on the segment $d_1 \cap TVD$ is obtained at point $\lambda = 0$ since λ has to be non positive.

Case ii)
If d_3 is under $\mathbf{G}_3$, *i.e.* $\mathbf{G}_3.\mathbf{s}_3 > \mu_3$ then $\lambda_0 < 0$. On the other hand, the point $\mathbf{G}_2 = d_1 \cap d_3$ corresponds to the parameter ν such that $(\mathbf{G}_3 + \nu {\mathbf{s}_1}^{\perp}).\mathbf{s}_3 = \mu_3$ and we deduce

$$\nu = \frac{(\mu_3 - \mathbf{G}_3.\mathbf{s}_3)}{{\mathbf{s}_1}^{\perp}.\mathbf{s}_3}.$$

We obtain then

$$\frac{\lambda_0}{\nu} = \frac{({\mathbf{s}_1}^{\perp}.\mathbf{s}_3)^2}{({\mathbf{s}_1}^{\perp}.\mathbf{s}_3)^2 + ({\mathbf{s}_1}^{\perp}.\mathbf{s}_2)^2} < 1.$$

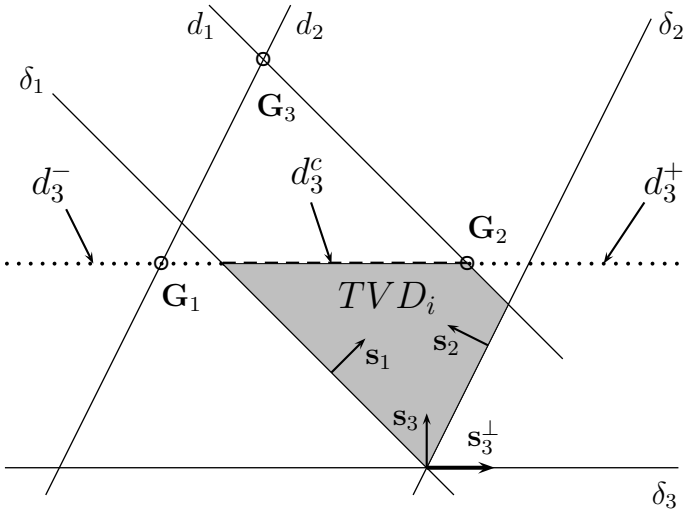

Fig. 6. Partition of the line d_3.

We conclude that $\nu < \lambda_0 < 0$ and the minimun of E on the segment $d_1 \cap TVD$ occurs for $\lambda = \nu$, thus the minimum belongs to d_3. □

The first situation corresponds to the choice $\widetilde{\mathbf{a}} = \mathbf{G}_3$ whereas the following proposition completes the second assertion.

Proposition 15 *Assume that d_3 is under point $\mathbf{G}_3$. Line d_3 is parted into three pieces: d_3^c is the segment $d_3 \cap TVD$, d_3^- is the left part of d_3 with respect to d_3^c while d_3^+ is the right part of d_3 with respect to d_3^c (see Fig. 6).*
Let $\widehat{\mathbf{a}}$ be the minimum of functional $E(\mathbf{a})$ under the constraint $\mathbf{a} \in d_3$. We have the following situations:

- *case 1: if $\widehat{\mathbf{a}} \in d_3^-$ then $\widetilde{\mathbf{a}}$ is the left bound of segment d_3^c.*
- *case 2: if $\widehat{\mathbf{a}} \in d_3^c$ then $\widetilde{\mathbf{a}} = \widehat{\mathbf{a}}$,*
- *case 3: if $\widehat{\mathbf{a}} \in d_3^+$ then $\widetilde{\mathbf{a}}$ is the right bound of segment d_3^c.*

PROOF. Let us denote by ${\mathbf{s}_3}^\perp$ the orthogonal vector to $\mathbf{s}_3$ such that ${\mathbf{s}_3}^\perp$ goes from the left to the right (see Fig. 6). Line d_3 can be parametrized by using a free parameter λ

$$\mathbf{a} = \widehat{\mathbf{a}} + \lambda {\mathbf{s}_3}^\perp. \tag{31}$$

On line d_3, functional E is then given by

$$E(\mathbf{a}) = E(\widehat{\mathbf{a}} + \lambda {\mathbf{s}_3}^\perp) = F(\lambda),$$

where $F(\lambda)$ is a parabolic function, strictly decreasing for $\lambda < 0$ and strictly increasing for $\lambda > 0$.

If $\widehat{\mathbf{a}} \in d_3^c$, then $\widehat{\mathbf{a}} \in TVD$. Since proposition 14 says that $\widetilde{\mathbf{a}} \in d_3$, we deduce that $\widetilde{\mathbf{a}} = \widehat{\mathbf{a}}$.
If $\widehat{\mathbf{a}} \in d_3^+$, function $F(\lambda)$ is a decreasing function for λ such that $\mathbf{a} \in TVD$. Therefore, the minimum is obtained at the right bound of segment d_3^c. On the contrary, if $\widehat{\mathbf{a}} \in d_3^-$, function $F(\lambda)$ is an increasing function for λ such that $\mathbf{a} \in TVD$. Therefore, the minimum is obtained at the left bound of segment d_3^c. □

We conclude this subsection by a summary of optimal slope computation.

- If at least two of the three μ coefficients vanish, then $\widehat{\mathbf{a}} = (0, 0)$.
- Assume that all μ coefficients are positive.
  From proposition 5, $\widehat{\mathbf{a}} = (0, 0)$ if and only if $\alpha_{21} < 0$, $\alpha_{31} < 0$ and $\alpha_{32} < 0$.
  Otherwise, using the convention on the local indexation ($\alpha_{31} > 0$ and $\alpha_{32} > 0$), we derived the following Table from propositions 14-15:

Cases	optimal slope $\widehat{\mathbf{a}}$
case 1: $\psi = \mathbf{G}_3.\mathbf{s}_3 - \mu_3$ $= \alpha_{31}\mu_1 + \alpha_{32}\mu_2 - \mu_3 \leq 0$	$\mathbf{G}_3$ such that $\mathbf{G}_3.s_1 = \mu_1$ and $\mathbf{G}_3.s_2 = \mu_2$
case 2: $\psi > 0$ and $\mu_1 - \dfrac{\psi}{\alpha_{31}^2 + \alpha_{32}^2}\,\alpha_{31} \leq 0$	$\mathbf{P}_1^3$ such that $\mathbf{P}_1^3.s_1 = 0$ and $\mathbf{P}_1^3.s_3 = \mu_3$
case 3: $\psi > 0$ and $0 \leq \mu_1 - \dfrac{\psi}{\alpha_{31}^2 + \alpha_{32}^2}\,\alpha_{31} \leq \dfrac{\mu_3}{\alpha_{31}}$	$\widehat{\mathbf{a}}$ such that $\widehat{\mathbf{a}}.s_i = \mu_i - \dfrac{\psi}{\alpha_{31}^2 + \alpha_{32}^2}\,\alpha_{3i},\ i=1,2$
case 4: $\psi > 0$ and $\dfrac{\mu_3}{\alpha_{31}} \leq \mu_1 - \dfrac{\psi}{\alpha_{31}^2 + \alpha_{32}^2}\,\alpha_{31}$	$\mathbf{P}_2^3$ such that $\mathbf{P}_2^3.s_2 = 0$ and $\mathbf{P}_2^3.s_3 = \mu_3$

Remark 16 *If one of $\mu_k = 0$, says μ_1, and $\mu_2\mu_3 \neq 0$, the previous procedure is modified. From proposition 3 we deduced that $\widehat{\mathbf{a}} = (0, 0)$ if and only if $\alpha_{32} < 0$. Otherwise, using the convention on the local indexation, we now obtain the following expression for the optimal slope:*

if $\alpha_{32}\,\mu_2 \leq \mu_3$, $\widehat{\mathbf{a}} = \mathbf{G}_3$ *such that* $\mathbf{G}_3.s_1 = \mu_1$ *and* $\mathbf{G}_3.s_2 = \mu_2$,

else, $\widehat{\mathbf{a}} = \mathbf{G}_2$ *such that* $\mathbf{G}_2.s_1 = \mu_1$ *and* $\mathbf{G}_2.s_3 = \mu_3$.

The Q method consists in predicting the value U_{ij} using the collocation point $X_{ij} = Q_{ij}$ and we get

$$U_{ij} = U_i + \widetilde{\mathbf{a}}_i \,.\, \mathbf{B}_i \mathbf{Q}_{ij}, \quad j \in \nu(i). \tag{32}$$

The reconstruction is consistent with the linear solutions and satisfy *a priori* the stability constraint whether $\widetilde{\mathbf{a}}_i \in TVD_i$ or $\widetilde{\mathbf{a}}_i \in MP_i$. Nevertheless, the Q method is not optimal. Indeed, flux F_{ij} is an approximation of the exact flux integrated on the edge S_{ij}, therefore numerical integration using the value at the midpoint $\mathbf{M}_{ij}$ provides a better approximation than the value at $\mathbf{Q}_{ij}$. Consequently, we aim to evaluate U_{ij} at point $\mathbf{M}_{ij}$ in place of $\mathbf{Q}_{ij}$ leading to the following M method:

$$U_{ij} = U_i + \widetilde{\mathbf{a}}_i . \mathbf{B}_i \mathbf{M}_{ij}, \quad j \in \nu(i). \tag{33}$$

Note that the reconstruction is still consistent with the linear solutions but does not satisfy *a priori* any stability constraint even if the slope belongs to the TVD or MP domain. Theoritical stability is lost but as we shall show in the numerical test section, the solution remains stable with a better accuracy than the former method using points $\mathbf{Q}_{ij}$.

4 The multislope technique

All the above second order method are based on the linear reconstruction (2) where the slope $\mathbf{a}_i$ computed on element K_i is used to obtain all the reconstructed values U_{ij}, $j \in \nu(i)$. A different approach consists in providing three slopes, one for each edge of the element, such that we satisfy the two following basic conditions:
- the reconstruction is consistent for the linear function U, *i.e.* $U_{ij} = U(\mathbf{X}_{ij})$,
- if we have a local extremum at point $\mathbf{B}_i$, we find again a first order scheme, *i.e.* the slopes vanish.
We call this method a multislope method since each value U_{ij} is obtained using a specific slope for each $j \in \nu(i)$.

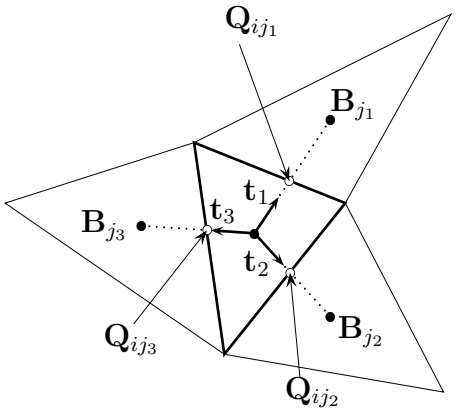

Fig. 7. Definition of vector $\mathbf{t}_k$.

4.1 The fundamental decomposition

We first construct the slopes in each direction. To this end, we introduce the normalized vectors

$$\mathbf{t}_k = \mathbf{t}_{ij_k} = \frac{\mathbf{B}_i\mathbf{B}_{j_k}}{|\mathbf{B}_i\mathbf{B}_{j_k}|},\ k = 1, 2, 3.$$

We have the following proposition.

Proposition 17 *Assume that the mesh satisfies hypothesis ($\mathcal{H}$), then the following decomposition holds:*

$$\mathbf{t}_1 = \beta_{12}\ \mathbf{t}_2 + \beta_{13}\ \mathbf{t}_3 \tag{34}$$
$$\mathbf{t}_2 = \beta_{21}\ \mathbf{t}_1 + \beta_{23}\ \mathbf{t}_3 \tag{35}$$
$$\mathbf{t}_3 = \beta_{31}\ \mathbf{t}_1 + \beta_{32}\ \mathbf{t}_2, \tag{36}$$

with

$$\beta_{ml}\ \beta_{lm} = 1, \tag{37}$$
$$\beta_{ml}\ \beta_{lk} = -\beta_{mk}, \tag{38}$$

for any circular permutation (m, l, k) of $(1, 2, 3)$ and all the coefficients are negative.

PROOF. Hypothesis ($\mathcal{H}$) reads

$$\mathbf{B}_i = \sum_{k=1,2,3} \rho_k\ \mathbf{B}_{j_k}$$

with $\rho_k > 0$ and $\rho_1 + \rho_2 + \rho_3 = 1$. We then deduce

$$0 = \sum_{k=1,2,3} \rho_k \ \mathbf{B}_i\mathbf{B}_{j_k} = \sum_{k=1,2,3} \rho_k \ |\mathbf{B}_i\mathbf{B}_{j_k}| \ \mathbf{t}_k.$$

Since $\rho_k|\mathbf{B}_i\mathbf{B}_{j_k}| > 0$, we conclude that all the coefficients β_{ij} are negative. Relations (37)-(38) are proved as in proposition 2. □

4.2 Multislope method with the $\mathbf{Q}_{ij}$ points

To build the multislope method, two slope sets are introduced. We define the downstream slopes with respect to point $\mathbf{B}_i$ in direction $\mathbf{t}_{ij_k}$ by

$$p^+_{ij_k} = \frac{U^n_{j_k} - U^n_i}{|\mathbf{B}_i\mathbf{B}_{j_k}|}, \quad k = 1, 2, 3, \tag{39}$$

and we define the upstream slopes by

$$\begin{aligned}
p^-_{ij_1} &= \beta_{12} \ p^+_{ij_2} + \beta_{13} \ p^+_{ij_3}, \\
p^-_{ij_2} &= \beta_{21} \ p^+_{ij_1} + \beta_{23} \ p^+_{ij_3}, \\
p^-_{ij_3} &= \beta_{31} \ p^+_{ij_1} + \beta_{32} \ p^+_{ij_2}.
\end{aligned}$$

Note that the downstream slopes $p^+_{ij_k}$ correspond to an approximation of the directional derivatives in the $\mathbf{t}_{ij_k}$ directions. We now give a general definition of a limiter to provide L^∞ stability for the reconstruction.

Definition 18 *A function $(p, q) \to \theta(p, q)$ is a limiter if it satisfies the properties*

$$\theta(p, p) = p, \quad \forall p \in \mathbb{R}, \tag{40}$$
$$\theta(p, q) = 0, \quad \forall p, q \in \mathbb{R} \text{ with } pq \leq 0, \tag{41}$$
$$\theta(p, q) = \theta(q, p), \quad \forall p, q \in \mathbb{R}. \tag{42}$$

For example the minmod limiter

$$\begin{cases}
\theta(p, q) = 0 & pq \leq 0, \\
\theta(p, q) = \min(p, q) & p \geq 0, \ q \geq 0, \\
\theta(p, q) = \max(p, q) & p \leq 0, \ q \leq 0,
\end{cases}$$

satisfies the properties. Other limiters like Van-Leer's limiter, superbee limiter also satisfy the properties (40)-(42).

Let us define the limited slopes in the $\mathbf{t}_{ij}$ direction by

$$p_{ij} = \theta(p_{ij}^+, p_{ij}^-), \quad j \in \nu(i). \tag{43}$$

The multislope method reads

$$U_{ij} = U_i + p_{ij}\ |\mathbf{B}_i\mathbf{Q}_{ij}|, \quad j \in \nu(i). \tag{44}$$

Proposition 19 *Assume that the mesh satisfies hypothesis ($\mathcal{H}$). Then the reconstruction is consistent for the linear solution and we have a first order scheme at the extrema.*

PROOF. To prove the first assertion, let us consider a linear function $U(\mathbf{X}) = U_0 + \mathbf{L}.\mathbf{X}$. The downstream slope is given by

$$p_{ij_k}^+ = \frac{\mathbf{L}.\mathbf{B}_i\mathbf{B}_{j_k}}{|\mathbf{B}_i\mathbf{B}_{j_k}|} = \mathbf{L}.\mathbf{t}_k$$

and the linearity of function U yields

$$\begin{aligned} p_{ij_1}^- &= \beta_{12}\ p_{ij_2}^+ + \beta_{13}\ p_{ij_3}^+ \\ &= \beta_{12}\ \mathbf{L}.\mathbf{t}_2 + \beta_{13}\ \mathbf{L}.\mathbf{t}_3 \\ &= \mathbf{L}.(\beta_{12}\,\mathbf{t}_2 + \beta_{13}\,\mathbf{t}_3) \\ &= \mathbf{L}.\mathbf{t}_1 = p_{ij_1}^+. \end{aligned}$$

We conclude from property 40 that $p_{ij} = p_{ij}^+$ and finally we get $U_{ij} = U(\mathbf{Q}_{ij})$.

To prove the second assertion, let assume that U_i is a local minimum. All the slopes p_{ij}^+ are non negative since $U_j \geq U_i$, $j \in \nu(i)$. Under hypothesis ($\mathcal{H}$), coefficients β_{ij} are negative hence p_{ij}^- are non positive. In consequence, property 41 yields $p_{ij} = 0$ and the scheme is reduced to a first order one. □

Remark 20 *The particular choice of the minmod limiter provides a TVD reconstruction in each segment $[\mathbf{B}_i, \mathbf{B}_j]$. Indeed, if $U_i \leq U_j$, we have $U_i \leq U_{ij} \leq U_{ij}^{ref} \leq U_{ji} \leq U_j$ (see (19) for definition of U_{ij}^{ref}).*

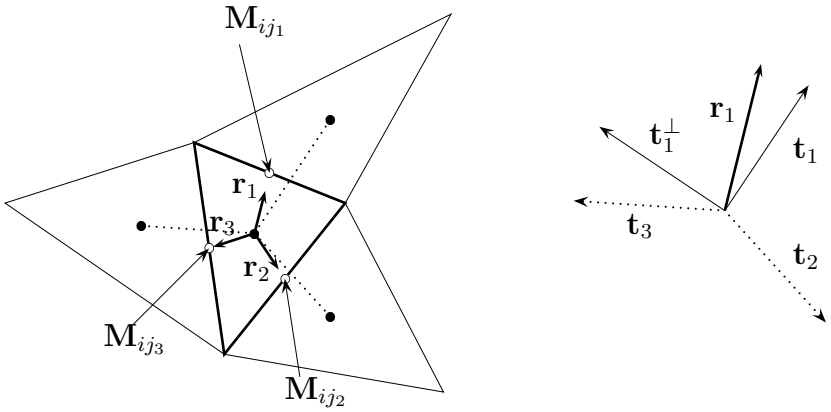

Fig. 8. Vector $\mathbf{r}_k$ (left). Decompositions of vector $\mathbf{r}_1$ in the basis $\mathbf{t}_1$, $\mathbf{t}_1^\perp$ and vector $\mathbf{t}_1^\perp$ in the basis $\mathbf{t}_2$, $\mathbf{t}_3$ (right).

4.3 Multislope method with the M_{ij} points

The numerical flux F_{ij} is an approximation of the exact flux integrated on edge S_{ij}. Numerical integration using midpoint $\mathbf{M}_{ij}$ for the quadrature formula provides a second order approximation. Therefore, better accuracy shall be obtained using $\mathbf{M}_{ij}$ in place of $\mathbf{Q}_{ij}$. We then consider a new set of vectors (Fig. 8 left),

$$\mathbf{r}_k = \mathbf{r}_{ij_k} = \frac{\mathbf{B}_i\mathbf{M}_{ij_k}}{|\mathbf{B}_i\mathbf{M}_{ij_k}|},\ k = 1, 2, 3.$$

As in the previous section, we have the following proposition.

Proposition 21 *Assume that the triangle $K \in \mathcal{T}_h$ is not reduced to a segment. Then the non zero coefficients of the following unique expansions*

$$\mathbf{r}_1 = \delta_{12}\ \mathbf{r}_2 + \delta_{13}\ \mathbf{r}_3 \tag{45}$$
$$\mathbf{r}_2 = \delta_{21}\ \mathbf{r}_1 + \delta_{23}\ \mathbf{r}_3 \tag{46}$$
$$\mathbf{r}_3 = \delta_{31}\ \mathbf{r}_1 + \delta_{32}\ \mathbf{r}_2, \tag{47}$$

satisfy

$$\delta_{ml}\ \delta_{lm} = 1, \tag{48}$$
$$\delta_{ml}\ \delta_{lk} = -\delta_{mk}, \tag{49}$$

for any circular permutation (m, l, k) of $(1, 2, 3)$. Furthermore, since $\mathbf{B}_i$ is strictly inside the triangle $(\mathbf{M}_{ij})_{j\in\nu(i)}$, all the coefficients are negative.

4.3.1 Decomposition of $\mathbf{r}$

Natural directions to compute the slopes are $\mathbf{t}_m = \mathbf{t}_{ij_m}$, $m = 1, 2, 3$ since basic informations (*i.e.* the values of U_i) are given at the centroids. To compute new interpolated values at points $\mathbf{M}_{ij}$, one has to decompose $\mathbf{r}_k$ with respect to the set $(\mathbf{t}_m)_{m=1,2,3}$. Non uniqueness of the decomposition is clear so we propose a decomposition such that we recorver the Q method when M_{ij} and Q_{ij} coincide.

Let $\mathbf{t}_k^\perp$ denote a normalized orthogonal vector to $\mathbf{t}_k$. On the one hand, we consider the unique decomposition of $\mathbf{t}_k^\perp$ in the basis $\{\mathbf{t}_m, \quad m \neq k\}$ (Fig. 8 right)

$$\mathbf{t}_1^\perp = \eta_{12}\ \mathbf{t}_2 + \eta_{13}\ \mathbf{t}_3, \tag{50}$$
$$\mathbf{t}_2^\perp = \eta_{21}\ \mathbf{t}_1 + \eta_{23}\ \mathbf{t}_3, \tag{51}$$
$$\mathbf{t}_3^\perp = \eta_{31}\ \mathbf{t}_1 + \eta_{32}\ \mathbf{t}_2. \tag{52}$$

On the other hand, we decompose $\mathbf{r}_k$ as

$$\mathbf{r}_k = (\mathbf{r}_k.\mathbf{t}_k)\,\mathbf{t}_k + (\mathbf{r}_k.\mathbf{t}_k^\perp)\,\mathbf{t}_k^\perp. \tag{53}$$

We get the decomposition of $\mathbf{r}_k$ thanks to relations (50)-(53):

$$\mathbf{r}_k = \sum_{m=1,2,3} \xi_{km}\ \mathbf{t}_m, \tag{54}$$

with

$$\xi_{kk} = \mathbf{r}_k.\mathbf{t}_k, \quad \xi_{km} = (\mathbf{r}_k.\mathbf{t}_k^\perp)\,\eta_{km}, \quad m \neq k.$$

This decomposition satisfies the property:

$$\text{if } \ \mathbf{r}_k = \mathbf{t}_k \ \text{ then } \ \xi_{kk} = 1 \ \text{ and } \ \xi_{km} = 0,\ m \neq k.$$

4.3.2 Construction of the slopes

We first define the downstream slopes q_{ij}^+ as

$$q_{ij_k}^+ = \sum_{m=1,2,3} \xi_{km}\ p_{ij_m}^+, \quad k = 1, 2, 3. \tag{55}$$

Then we define the upstream slopes

$$q_{ij_1}^- = \delta_{12}\ q_{ij_2}^+ + \delta_{13}\ q_{ij_3}^+,$$

$$q^-_{ij_2} = \delta_{21}\ q^+_{ij_1} + \delta_{23}\ q^+_{ij_3},$$
$$q^-_{ij_3} = \delta_{31}\ q^+_{ij_1} + \delta_{32}\ q^+_{ij_2}.$$

We compute the slopes q_{ij} using the limiter function

$$q_{ij} = \theta(q^+_{ij}, q^-_{ij}), \quad j \in \nu(i). \tag{56}$$

We finally define the reconstruction with

$$U_{ij} = U_i + q_{ij}\ |\mathbf{B}_i\mathbf{M}_{ij}|. \tag{57}$$

Proposition 22 *The reconstruction is consistent for linear functions.*

PROOF. Let us consider a function $U(\mathbf{X}) = U_0 + \mathbf{L}.\mathbf{X}$ with $\mathbf{L} \in \mathbb{R}^2$. By construction, we have $p^+_{ij_k} = \mathbf{L}.\mathbf{t}_k$. Relation (55) implies that

$$q^+_{ij_k} = \mathbf{L}.\sum_{m=1,2,3} (\xi_{km}\,\mathbf{t}_m) = \mathbf{L}.\mathbf{r}_k.$$

Hence we deduce that

$$q^+_{ij_k} = \mathbf{L}.\mathbf{r}_k = \frac{U(\mathbf{M}_{ij_k}) - U(\mathbf{B}_i)}{|\mathbf{B}_i\mathbf{M}_{ij_k}|}.$$

On the other hand, we write for example with $k = 1$

$$\begin{aligned} q^-_{ij_1} &= \delta_{12}\ q^+_{ij_2} + \delta_{13}\ q^+_{ij_3} \\ &= \mathbf{L}.(\delta_{12}\,\mathbf{r}_2 + \delta_{13}\,\mathbf{r}_3) \\ &= \mathbf{L}.\mathbf{r}_1 = q^+_{ij_1}. \end{aligned}$$

Thanks to property 40, we deduce that $q_{ij_k} = q^+_{ij_k}$ and thus $U_{ij} = U(\mathbf{M}_{ij})$ for all $j \in \nu(i)$. □

Remark 23 *Degeneration to first order scheme is not guaranteed by the reconstruction at point M if U_i is a local extremum. Indeed, since the point $\mathbf{B}_i$ is strictly inside the triangle with vertices $\mathbf{M}_{ij}$, $j \in \nu(i)$, all the coefficients δ_{km} are negative. Therefore if all the slopes $q^+_{ij_k}$ have the same sign, we deduce that $q^-_{ij_k} q^+_{ij_k} < 0$ then $q_{ij_k} = 0$ thanks to relation (41). But if all the slopes p^+_{ij} have the same sign, the slopes $q^+_{ij_k}$ given by relations (55) do not have a priori the same sign, hence the slope q_{ij} might be non zero.*

5 Numerical tests

5.1 Tests with linear problems

We present numerical tests for the advection and the rotation problems. Computations have been performed with the six following schemes:

S1	first order scheme (Eq. (1))
S2	the gradient scheme (Section 2.1.1) with TVD limiter (Section 2.2.3)
S3	optimized monoslope scheme (Section 3.1) with point $\mathbf{Q}_{ij}$ (Eq. (32))
S4	optimized monoslope scheme (Section 3.1) with point $\mathbf{M}_{ij}$ (Eq. (33))
S5	multislope scheme with point $\mathbf{Q}_{ij}$ (Section 4.2)
S6	multislope scheme with point $\mathbf{M}_{ij}$ (Section 4.3)

Let Ω be the unit square. To evaluate the method accuracy, we consider four meshes $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ with $N_0 = 228$, $N_1 = 840$, $N_2 = 3300$, $N_3 = 13340$ elements using the Delaunay algorithm to obtain real unstructured meshes. The characteristic length is defined by

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

and we compute $h_0 = 73.2\,E{-}3$, $h_1 = 40.2\,E{-}3$, $h_2 = 20.0\,E{-}3$, $h_3 = 9.9\,E{-}3$ respectively.

We use the forward Euler scheme to update the solution at each time step and we define the characteristic parameter

$$CFL = \frac{|\mathbf{V}|\Delta t}{h}.$$

The time step Δt is adapted to provide stability and we calculate Δt setting the CFL coefficient number. We choose the CFL value in order to obtain the smallest L^1 error. For the present tests, stability is obtained with a CFL parameter lower than 0.6. Moreover, we have taken smaller CFL values up to 0.05 in the regular case to get better accuracy since we deal with a first order scheme in time. A second order scheme in time (Heun scheme) has also been tested with larger CFL values but it appears in our simulations that the computational cost is equivalent since the second order scheme in time requires to compute an intermediate solution for each time step. The performance of

the second order method in time is not significative from a computational point of view for the situations considered here and only first order scheme in time will be employed in the sequel.

We assume that the error estimations converge asymptotically as

$$\|U_h(T,.) - U(T,.)\|_{L^1} \approx Ch^\alpha \tag{58}$$

where α is the scheme order while C is a constant. With low order scheme (*i.e.* $\alpha < 1$), constant C value is crucial to evaluate the scheme accuracy while its influence is less important for higher order schemes (*i.e.* $\alpha > 1$).

In this subsection, we consider the advection problem where an inital compact support function is moved with a constant velocity or by rotation. Two initial functions are used: a regular initial function U_r

$$U_r(x_1,x_2) = \frac{1}{2}(\cos(5\pi r)+1) \text{ if } r < \frac{1}{5}, \quad U_r(x_1,x_2) = 0 \text{ if } r > \frac{1}{5}$$

and an irregular initial function U_d

$$U_d(x_1,x_2) = 1 \text{ if } r < \frac{1}{5}, \quad U_d(x_1,x_2) = 0 \text{ if } r > \frac{1}{5}.$$

Computational experiences have been performed using the six schemes and the upwind flux where the velocity is computed at the midpoint of the edges.

- The advection problem with constant velocity: we take $\mathbf{V} = (0.5, 0.5)$ and the initial functions U_r and U_d are centered at point $(0.25, 0.25)$ ($r = \sqrt{(x_1 - \frac{1}{4})^2 + (x_2 - \frac{1}{4})^2}$). Computational experiences have been performed till the final time $t_f = 1\,s$ for which exact solution is the initial function centered at point $(0.75, 0.75)$.
- The rotation problem: we take $\mathbf{V} = (0.5 - x_2, x_1 - 0.5)$. The test consists in a half-rotation around point $(0.5, 0.5)$ of the initial function U_r or U_d centered at point $(0.75, 0.5)$ ($r = \sqrt{(x_1 - \frac{3}{4})^2 + (x_2 - \frac{1}{2})^2}$). At the final time $t_f = \pi\,s$, the solution is the initial function centered at point $(0.25, 0.5)$.

5.1.1 The advection problem with an initial regular condition

Table 1 lists the errors in the L^1 and L^∞ norms between the exact solution and the approximation at time $t = 1.0$ using a CFL value to provide the smaller L^1 error. Fig. 9 shows the L^1 and L^∞ error curves in function of the mesh characteristic parameter h. Assuming that the L^1 errors satisfy asymptotically

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	2.444e-02	2.404e-02	1.854e-02	2.002e-02	1.825e-02	1.736e-02
$\mathcal{M}_1$	1.902e-02	1.207e-02	8.503e-03	7.839e-03	1.001e-02	8.887e-03
$\mathcal{M}_2$	1.162e-02	7.531e-03	4.417e-03	3.698e-03	4.666e-03	3.966e-03
$\mathcal{M}_3$	6.587e-03	3.667e-03	1.770e-03	1.440e-03	1.947e-03	1.636e-03

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	4.566e-01	3.334e-01	4.712e-01	3.828e-01	4.096e-01	4.278e-01
$\mathcal{M}_1$	3.887e-01	3.213e-01	2.105e-01	1.943e-01	2.372e-01	2.454e-01
$\mathcal{M}_2$	2.496e-01	2.107e-01	1.362e-01	1.298e-01	1.607e-01	1.504e-01
$\mathcal{M}_3$	1.390e-01	1.012e-01	6.638e-02	6.876e-02	7.894e-02	8.127e-02

Table 1
The advection test with the regular function. Error in the L^1 norm (upper table) and the L^∞ norm (lower table)

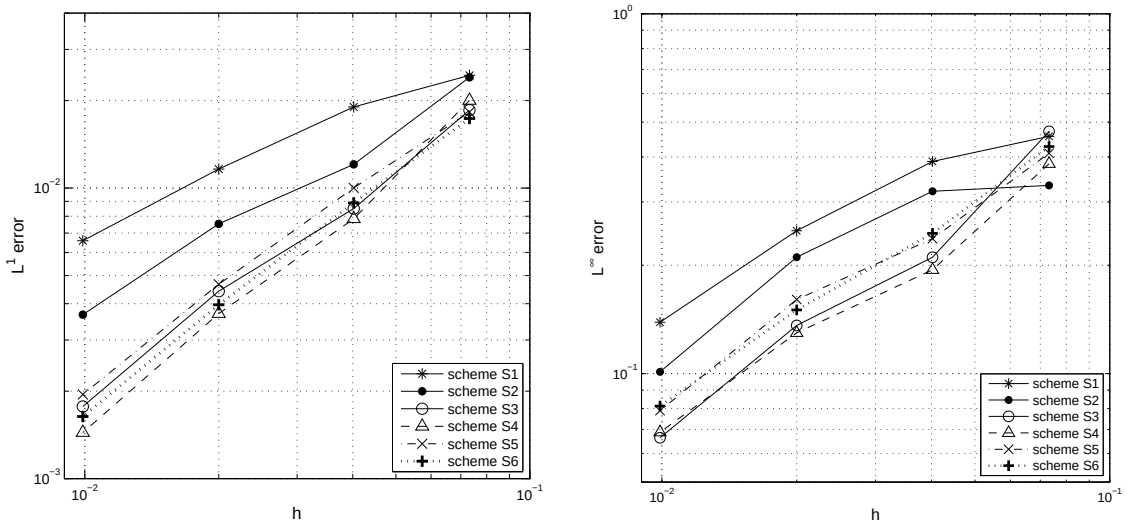

Fig. 9. The advection test with the regular function. Errors in the L^1 norm (left) and the L^∞ norm (right) versus mesh parameter h for the six schemes.

relation (58), coefficients C and α are computed using the last three points and presented in Table 2.

The first order scheme S_1 is clearly the worth in comparison with the others. Scheme $S2$ presents a better convergence order but the rough limiter (17) ensures the scheme to provide convergence order similar to that of schemes $S3$-$S6$. The last four methods have similar convergence order but we note the following facts:

scheme	S1	S2	S3	S4	S5	S6
α	0.76	0.85	1.12	1.21	1.17	1.21
C	0.22	0.19	0.32	0.39	0.44	0.44

Table 2
Convergence order with the L^1 norm for the advection test with the regular function.

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	7.730e-02	6.070e-02	5.432e-02	4.771e-02	6.339e-02	5.046e-02
$\mathcal{M}_1$	6.837e-02	5.143e-02	3.943e-02	3.249e-02	3.870e-02	3.413e-02
$\mathcal{M}_2$	4.961e-02	3.103e-02	2.410e-02	2.189e-02	2.706e-02	2.393e-02
$\mathcal{M}_3$	3.409e-02	1.987e-02	1.406e-02	1.261e-02	1.563e-02	1.367e-02

Table 3
The advection test with the discontinuous function. Error in the L^1 norm for the six schemes.

scheme	S1	S2	S3	S4	S5	S6
α	0.50	0.68	0.74	0.68	0.65	0.65
C	0.34	0.45	0.42	0.29	0.32	0.29

Table 4
Convergence order with the L^1 norm for the advection test with the discontinuous function.

- Schemes using point M give slightly better convergence rates than schemes using point Q.
- For a given point (Q or M), monoslope schemes $S3$ and $S4$ give slightly better convergence rates than the multislope schemes $S5$ and $S6$.

5.1.2 The advection problem with an initial discontinuous function

We present in Table 3 the L^1 errors between the exact solution and the approximation at time $t = 1.0$. Fig. 10 shows the convergence rate plotting the L^1 error in function of the mesh characteristic parameter h and Table 4 gives coefficients C and α.

We obtain the same accuracy classification where $S1$ provides the worst convergence, $S2$ is an intermediate situation and the last four schemes give the best convergence orders. For the discontinuous case, schemes $S3$-$S6$ have the same order of convergence but the constant are differents and the multislope methods $S5$ and $S6$ have a better constant C than the monoslope methods $S3$ and $S4$.

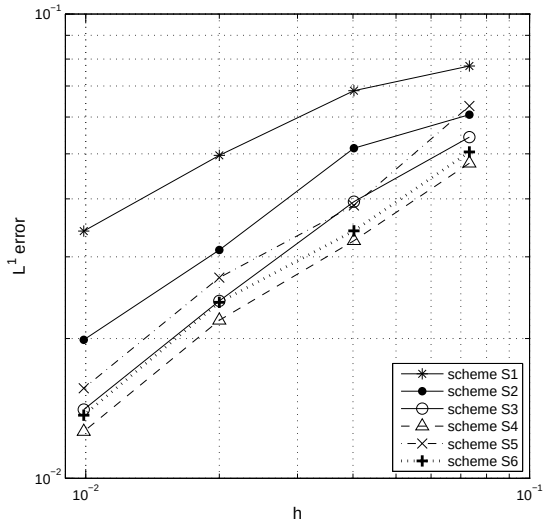

Fig. 10. The advection test with the discontinuous function. Error in the L^1 norm versus mesh parameter h for the six schemes.

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	7.861e-03	4.802e-03	4.294e-03	4.651e-03	4.859e-03	4.362e-03
$\mathcal{M}_1$	5.928e-03	2.905e-03	2.022e-03	2.199e-03	2.386e-03	2.085e-03
$\mathcal{M}_2$	3.953e-03	1.664e-03	9.484e-04	8.802e-04	10.30e-04	8.065e-04
$\mathcal{M}_3$	2.373e-03	8.889e-04	4.526e-04	3.819e-04	4.996e-04	3.632e-04

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	12.62e-02	8.927e-02	8.131e-02	8.949e-02	9.186e-02	7.531e-02
$\mathcal{M}_1$	11.36e-02	6.917e-02	5.231e-02	5.910e-02	6.083e-02	5.549e-02
$\mathcal{M}_2$	8.648e-02	4.815e-02	2.915e-02	3.477e-02	3.398e-02	3.456e-02
$\mathcal{M}_3$	5.334e-02	2.759e-02	1.561e-02	1.735e-02	1.685e-02	1.716e-02

Table 5
The rotation test with the regular function. Error in the L^1 norm (upper table) and the L^∞ norm (lower table)

5.1.3 The rotation problem

We first deal with the regular case using function U_r as an initial condition. Table 5 lists the errors in L^1 and L^∞ norms between the exact solution and the approximation at time $t = \pi$ while Fig. 11 shows the L^1 and L^∞ error curves in function of the mesh characteristic parameter h. Moreover, coefficients C and α are presented in Table 6. As in the advection case, schemes $S1$ and $S2$ give the lower rates of convergence while the last four schemes $S3$-$S6$

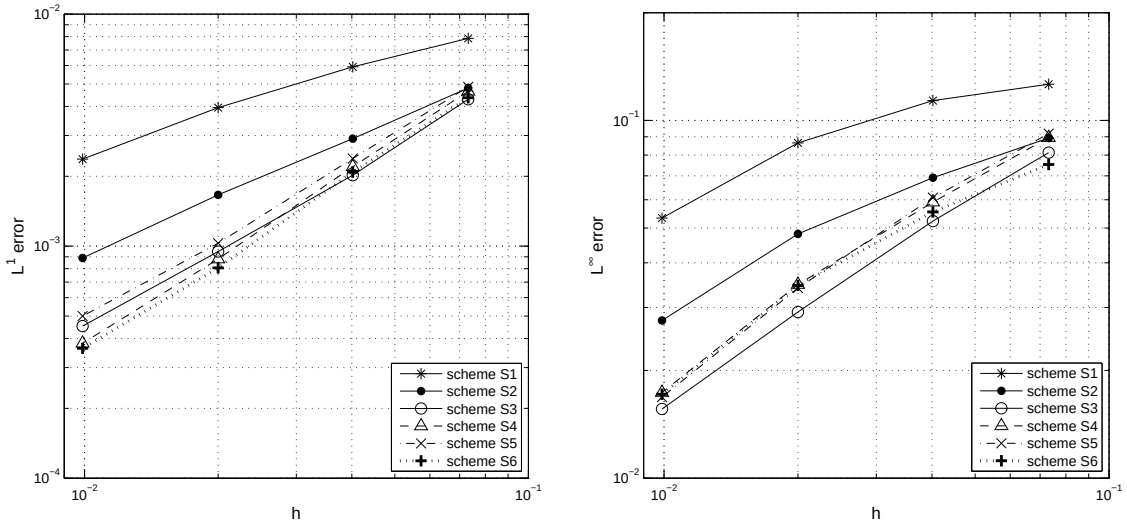

Fig. 11. The rotation test with the regular function. Errors in the L^1 norm (left) and the L^∞ norm (right) versus mesh parameter h for the six schemes.

scheme	S1	S2	S3	S4	S5	S6
α	0.65	0.85	1.07	1.25	1.12	1.25
C	0.05	0.04	0.06	0.12	0.08	0.11

Table 6
Convergence order with the L^1 norm for the rotation test with the regular function.

	S1	S2	S3	S4	S5	S6
$\mathcal{M}_0$	5.634e-02	3.792e-02	3.297e-02	3.493e-02	3.534e-02	3.444e-02
$\mathcal{M}_1$	4.757e-02	3.135e-02	2.486e-02	2.853e-02	2.822e-02	2.522e-02
$\mathcal{M}_2$	3.571e-02	2.194e-02	1.623e-02	1.725e-02	1.793e-02	1.605e-02
$\mathcal{M}_3$	2.566e-02	1.469e-02	1.058e-02	1.105e-02	1.182e-02	1.062e-02

Table 7
The rotation test with the discontinuous function. Error in the L^1 norm for the six schemes.

produce the best convergence order. We have not notice significant difference with the advection problem and the simulations confirm the remarks made in Section 5.1.1.

The second set of tests concerns the irregular case using the discontinuous function U_d as an initial condition. Errors in L^1 norm at time $t_f = \pi$ are given in Table 7 and plotted in Fig. 12. Coefficient C and α are computed and presented in Table 8.

As in the translation tests, scheme $S3$-$S6$ are more efficient but in this example

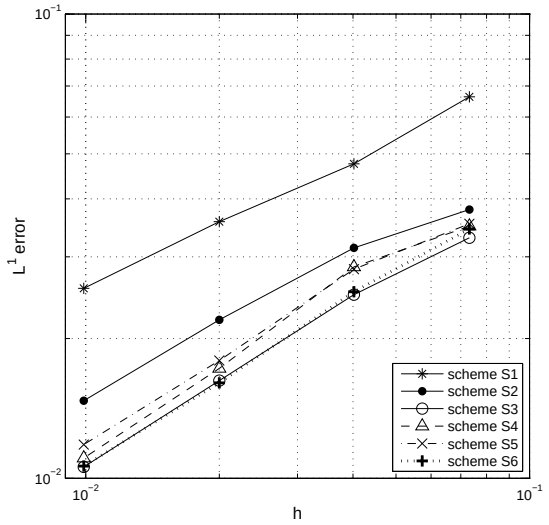

Fig. 12. The rotation test with the discontinuous function. Error in the L^1 norm versus mesh parameter h for the six schemes.

scheme	S1	S2	S3	S4	S5	S6
α	0.44	0.54	0.61	0.68	0.62	0.62
C	0.20	0.18	0.18	0.25	0.21	0.18

Table 8
Convergence order with the L^1 norm for the rotation test with the discontinuous function.

the monoslope scheme provides the best convergence order with the smallest constant.

5.2 The forward facing step problem for the Euler system

We now experiment the MUSCL reconstructions in the nonlinear vectorial framework of the Euler system considering the forward facing step problem (see [18] for the details). We use the HLLC flux to compute the numerical flux (see [16]) while the reconstruction is performed using the density ρ, the velocity (u, v) and the pressure P variables. A right-going supersonic Mach 3 flow is reflected by a 0.2 length unit step. Steady-state is assumed to be reached at $t = 4$ and we compare the solutions computed with three different methods: the first order scheme (run 1), the second order scheme with optimized monoslope at point Q (run 2) and the second order scheme with multislope at point Q (run 3). We use an unstructured Delaunay mesh of 18000 elements and compute the solution till $t_f = 4$.

Simulations have been also done using the M point but numerical artefacts

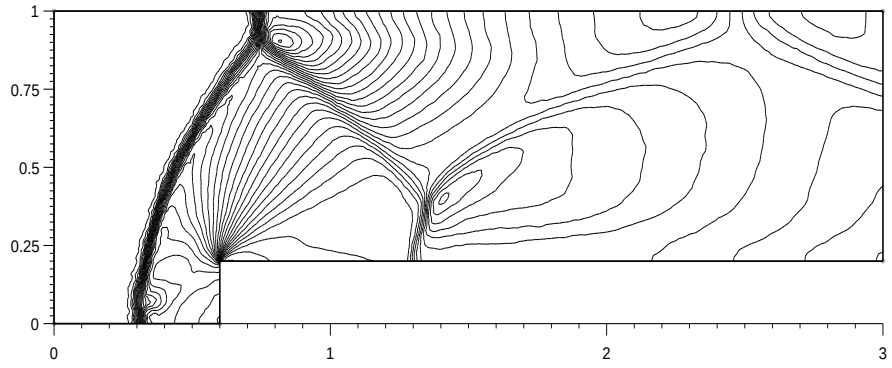

Fig. 13. Run 1: first order scheme.

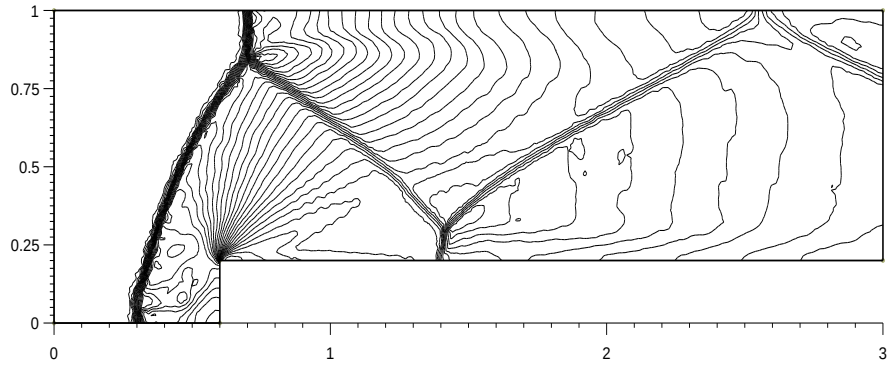

Fig. 14. Run 2: second order optimized monoslope scheme at point Q.

appear leading to negative pressure values. The multislope method at point M is not stable enough to pass critical tests such that the forward facing step problem.

In Figs. 13-15, we represent the density isovalues from 0 to 8 with a step of 0.2. Run 2 and 3 are very similar (second order methods) while the shocks suffer an important diffusion in run 1 (first order method). The second order methods are similar and provide a better resolution of the shock interfaces. The main advantage of the multislope method here is its simplicity and computational cost in comparison with the monoslope method.

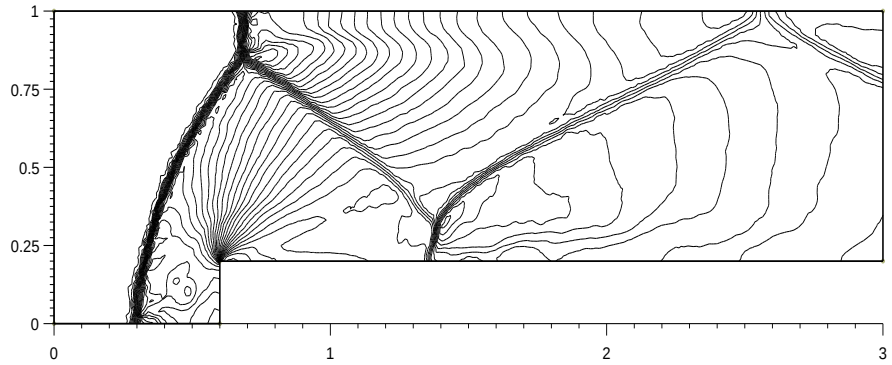

Fig. 15. Run 3: second order multislope scheme at point Q.

6 Conclusions

In this paper, new MUSCL methods have been presented in the context of cell-centered Finite Volume method where control volumes are triangles. First an enhancement of the monoslope MUSCL methods that is to say with one vectorial slope per control volume has been introduced. It is based on a minimization under constraint corresponding to the desired stability condition. Afterwards a new MUSCL method approach has been proposed. It consists in computing three scalar slopes per triangle following the three directions given by the neighbouring triangles.

The methods achieve a better accuracy in comparison with the classical gradient method with a rougth limiter. The convergence rate of both the methods are similar but in the multislope case, time consumption is reduced and the implementation corresponds to a one dimensional MUSCL method in each direction. The generalisation for the three-dimensional situation is straightforward for the multislope where we are concerned with four directions and very few modifications have to be done to adapt the method. For the minimization monoslope MUSCL method, some complementary studies should be considered to provide the three-dimensional extension.

From a numerical point of view, the choice of the edge midpoint M to compute the interpolate values U_{ij} in comparaison with point Q brings higher accuracy but scheme is less stable since the maximum principle is not respected.

References

[1] T. J. Barth, D. C. Jespersen, The design and application of upwind schemes on unstructured meshes, AIAA Report 89-0366, 1989.

[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. Buffard, Analyse de quelques méthodes de volumes finis non structurés pour la résolution des équations d'Euler, Thèse de doctorat de l'Université Paris 6, France, 1993.

[4] P. Chevrier, H. Galley, A Van Leer finite volume scheme for the Euler equations on unstructured meshes, RAIRO Modèl. Math. Anal. Numér. 27 (2) (1993) 183–201.

[5] P. Colella, Multidimensional upwind methods for hyperbolic conservation laws, J. Comput. Phys. 87 (1) (1990) 171–200.

[6] J.-A. Désidéri, A. Dervieux, Compressible flow solvers using unstructured grids, Von Karman Inst. Fluid Dynamics, Lecture Series 1988-05, 1988.

[7] L.-J. Durlofsky, B. Engquist, S. Osher, Triangle based adaptative stencils for the solution of hyperbolic conservation laws, J. Comput. Phys. 98 (1) (1992) 64–73.

[8] E. Godlewski, P. A. Raviart, Numerical approximation of hyperbolic systems of conservation laws, Applied Math. Sc., 118, Springer-Verlag, New York, 1996.

[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] M. E. Hubbard, Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids, J. Comput. Phys. 155 (1) (1999) 54–74.

[11] A. Jameson, D. Mavriplis, Finite volume solution of the two-dimensional Euler equations on a regular triangular mesh, AIAA J. 24 (4) (1986) 611–618.

[12] D. Kröner, Numerical schemes for conservation laws, Wiley Teubner, Series Advances in Numerical Mathematics, John Wiley & Sons Ltd, 1997.

[13] D. Kröner, S. Noelle, M. Rokyta, Convergence of higher order upwind finite volume schemes on unstructured grids for scalar conservation laws in several space dimensions, Numer. Math., 71 (4) (1995) 527–560.

[14] S. P. Spekreijse, Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws, Math. Comp. 49 (179) (1987) 135–155.

[15] P. K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, SIAM J. Numer. Anal. 21 (5) (1984) 995–1011.

[16] E. F. Toro, Riemann solvers and numerical methods for fluid dynamics. A practical introduction, Springer-Verlag, Berlin, 1997.

[17] 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.

[18] P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks, J. Comput. Phys. 54 (1) (1984) 115–173.