VECTORIAL LIMITATION FOR MULTISLOPE MUSCL SCHEMES

ARTHUR TÉTELIN¹, CLÉMENT LE TOUZE¹ AND PHILIPPE VILLEDIEU²

¹ DMPE, ONERA
Université Paris Saclay
F-91123 Palaiseau - France
email: arthur.tetelin@onera.fr, clement.le_touze@onera.fr

² DMPE, ONERA
Université de Toulouse
F-31055 Toulouse - France
email: philippe.villedieu@onera.fr

Key words: Cell-Centered Finite-Volume Methods, Unstructured Meshes, Multislope MUSCL Schemes, Vectorial Slope Limitation, Frame Invariance

Abstract. In finite volume schemes with MUSCL interpolation of scalar variables at the faces of control volumes, a slope limiting function is used in order to prevent non-physical oscillations of the solution. More particularly, these functions are designed to ensure a certain monotonicity criterion at each face of the control volume, criterion which then ensures a stability property of the scheme. For vectorial variables, these slope limiting functions are generally applied componentwise, but this may result in a frame-dependance, as well as a loss of accuracy due to false detection of extrema. In this paper, a new vectorial interpolation method is introduced, which is frame-invariant, second-order accurate and stable in a sense that will be defined.

1 INTRODUCTION

CEDRE software is a numerical code developed by the french institute ONERA to solve multi-physics problems in energetics [1]. Like other CFD codes, it solves hyperbolic systems of conservation equations, which, by focusing only on the convective part, read:

\[ \frac{\partial \mathbf{Q}}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{Q}, \lambda) = 0, \quad (1) \]

where \( \mathbf{Q} \) represents the vector of conserved variables, \( \lambda \) is the velocity vector field, and \( \mathbf{f} \) the physical flux. In CEDRE, the spatial discretization relies on a finite volume cell-centered framework on general unstructured meshes [2], which allows us to write the semi-discretized formulation of equation (1) on a multi-dimensional domain:

\[ \frac{\partial Q_i}{\partial t} = -\frac{1}{|K_i|} \sum_{j \in V(i)} |S_{ij}| \Phi \left( Q_{ij}^+, Q_{ij}^-, n_{ij}, \lambda_{ij} \right), \quad (2) \]
where $K_i$ denotes a cell of the mesh, $S_{ij}$ the face between cells $K_i$ and $K_j$. Notation $|.|$ denote either the length, area or volume measure, depending on the dimension of the object. $\mathcal{V}(i)$ is the face neighborhood of the cell $K_i$, $\Phi$ is the numerical flux and $\mathbf{n}_{ij}$ is a face normal vector pointing from cell $K_i$ to cell $K_j$. Lastly, $Q^n_{ij}$ represents a set of interpolated values at the face center $M_{ij}$ of face $S_{ij}$. With a first order scheme, these interpolated values are equal to the numerical variables $Q_i$ which represents the mean value of $Q$ over the cell $K_i$, located at the cell centroid $B_i$. To reach a higher order of spatial accuracy, CEDRE code relies on a MUSCL approach, on which we will be focusing in this article.

As it is well known by Godunov’s theorem [3], no linear reconstruction can be both high-order accurate and ensure the monotony of the scheme, that is why Van-Leer introduced the MUSCL approach [4–8]. This method consists in evaluating the gradient of the variable, which is then limited by a limitation function [9, 10] in order to ensure the scheme stability. For a one dimensional scheme, this stability is reached through a TVD property as introduced by Harten [11], but in higher dimension, it is known that the TVD property is incompatible with high order accuracy [12]. Since then, many other have studied this issue [13–15] in order to develop a MUSCL method on multi-dimensional meshes. These multi-dimensional MUSCL methods can be grouped in two categories: monoslope methods, in which a single limited gradient is computed for the entire cell [16], and multislope methods, in which a limited directional gradient is computed for each face [17]. The method that we will study here is of multislope type, and has been introduced by Le Touze [18].

As far as the authors know, while the limitation framework has already been well studied for scalar variables, such as temperature or pressure, limitation methods of vectorial variables such as velocity have received far less attention than their scalar counterpart. On a software such as CEDRE, vectorial limitation was obtained until now by limiting vectors componentwise, but this kind of procedure turns out to be frame-dependant, and a loss of accuracy can also be observed. For the vectorial limitation problem, one can divide solutions found in the literature in three major methods. The first one has been introduced by Luttwak and Falcovitz for a monoslope framework [19–21]. As the stability of a scalar variable is usually defined through a maximum principle, the authors defined vectorial variable stability by using the convex hull of vectors from cell neighborhood, and called it VIP (Vector Image Polygon / Polyhedron). A VIP method consists in the computation of a vector at the time step $n + 1$, followed by its projection on the VIP set if it doesn’t already lie into it. With this process, a vector will satisfy a maximum principle componentwise, whatever basis we choose, which is the natural extension of the scalar maximum principle to vector variables. The second method has been developped by Maire et al. [22,23]. It consists in computing a local basis for vectors to be reconstructed. Then vectors are projected onto this basis, and the vector reconstruction is computed componentwise in this basis. The last method has been developped by Zeng and Scovazzi [24]. It consists in determining an axis on which the vector field will be projected in order to get scalar variables. Then, a simple scalar limiter is computed, in order to get the general vectorial limiter.

The VIP method has the advantage to present the most accurate extension of the scalar maximum principle to the vectorial case. However, its algorithmic cost seems to be important and even prohibitive on large-scale simulations, as we have to compute intersection of vectors with a convex hull for each reconstructed variables. On the other hand, projection methods seems to have a lower computation cost, but the choice of the axis or basis of projection remains
arbitrary, and the vectorial aspect of the variable is dropped by these methods. For all these reasons, we propose a new vectorial limitation method in the context of multislope MUSCL methods. In section 2, we will quickly detail the multislope scalar MUSCL scheme, and we will develop a new way of limiting vectors variables in section 3. Nevertheless, we will only focus here on final results. Intermediate demonstrations and numerical tests will be presented in a forthcoming paper.

2 SCALAR MULTISLOPE MUSCL SCHEME

2.1 General reconstruction

The numerical scheme introduced by Le Touze and al. [18] consists in the computation of various geometrical parameters for each face of each cell of the mesh. All these parameters are computed at the intersection of the line drawn by the forward cells centroids ($B^{i}_{j+1}$) and the axis $B^{i}_{j}M^{ij}$, for a $d$-dimensional mesh. From the intersection of the line drawn by the forward cells centroids ($B^{i}_{j+1}$) and the axis $B^{i}_{j}M^{ij}$, we define the forward point $H^{+}_{ij}$. By a similar way, we define the backward point $H^{-}_{ij}$.

Then, by giving each scalar value ($U^{+}_{ij}$) of the forward cells a weight ($\beta^{+}_{ij}$), one can determine an interpolated value at the forward point: $U_{H^{+}_{ij}} = \sum_{k=1}^{d} \beta^{+}_{ijk} U^{+}_{ijk}$. Similarly, one can write $U_{H^{-}_{ij}} = \sum_{k=1}^{d} \beta^{-}_{ijk} U^{-}_{ijk}$. Hence, along the axis $B^{i}_{j}M^{ij}$, a scalar monodimensional framework has been created. Thus, we can define the forward and backward slopes, as well as their ratio:

$$p^{+}_{ij} = \frac{U^{+}_{H^{+}_{ij}} - U^{+}_{ij}}{\|B^{i}_{j}H^{+}_{ij}\|}, \quad p^{-}_{ij} = \frac{U^{-}_{H^{-}_{ij}} - U^{-}_{ij}}{\|B^{i}_{j}H^{-}_{ij}\|}, \quad r_{ij} = \frac{p^{-}_{ij}}{p^{+}_{ij}}. \quad (3)$$

This gives us the reconstructed scalar value at $M^{ij}$:

$$U_{ij} = U^{+}_{ij} + \|B^{i}_{j}M^{ij}\|\varphi_{ij}(r_{ij})p^{+}_{ij}, \quad (4)$$
where $\varphi$ denotes the limiting function, ensuring a $L^\infty$ stability property under a CFL condition with a second order accuracy where the solution is smooth. One of the drawbacks of this method is that the ratio $r_{ij}$ cannot be used directly in a vectorial framework, as slopes will also be vectors. One way to use it is to define complex slopes in the complex plane spanned by the slopes vectors. But actually, we will follow another approach that simplifies the stability proof, even in the scalar case, and thereafter makes the vectorial extension more natural. To do so, we recast the reconstruction (4) so that the slope ratio $r_{ij}$ disappears.

### 2.2 Limitation function

We will focus here on the limitation function. We consider here a special case of equation (1), that is the scalar advection equation, as it is easier to study its behaviour and its stability. This equation reads:

$$\frac{\partial u}{\partial t} + \nabla \cdot (\lambda u) = 0,$$

where $u$ is the advected scalar. In the scalar case, the limitation function is shaped in order to achieve a second order accuracy on smooth area, and to limit the solution around discontinuities. Particularly, the limitation part of the limiter is designed so that the discretized scalar advection scheme with an explicit Euler scheme as time-discretization, namely

$$U^{n+1}_i = U_i - \frac{\Delta t}{K_i} \sum_{j \in V(i)} |S_{ij}| \Phi(U_{ij}, U_{ji}, n_{ij}, \lambda_{ij}),$$

achieve a maximum principle under a CFL condition, where $U$ denotes here the discrete scalar solution of equation (5). From [18], the conditions on the shape of the limiting function read as follows:

$$0 \leq \varphi(r_{ij}) \leq \min(\eta^+_{ij}, \eta^-_{ij}),$$

where $\eta^+_{ij}$ and $\eta^-_{ij}$ are geometrical parameters: $\eta^+_{ij} = \frac{\|B_i H^+_i\|}{\|B_i M^+_ij\|}$ and $\eta^-_{ij} = \frac{\|B_i H^-_i\|}{\|B_i M^-_ij\|}$. The second-order accuracy is reached with another inequality:

$$\min(1, r) \leq \varphi(r) \leq \max(1, r).$$

This inequality means that the reconstructed scalar should be a convex combination of forward and backward slopes, which can be written as a $\kappa$-scheme [10, 25]

$$U_{ij} = U_i + \|B_i M_{ij}\| \left( \frac{1 + \kappa}{2} p^+_{ij} + \frac{1 - \kappa}{2} p^-_{ij} \right),$$

where $\kappa$ is a real parameter between $-1$ and 1. Taking now into account the monotonicity constraints, it means that $\kappa$ can no longer be a constant but must depend on the specific slopes for the current face, which leads to write:

$$U_{ij} = U_i + \|B_i M_{ij}\| \left( \frac{1 + \kappa_{ij}}{2} p^+_{ij} + \frac{1 - \kappa_{ij}}{2} p^-_{ij} \right),$$

where $\kappa_{ij} = \kappa(p^+_{ij}, p^-_{ij}) = \kappa(r_{ij})$ is the local value accounting for both second-order accuracy and monotonicity constraints (see Figure 2). This formulation of the reconstructed scalar is
far more convenient as the former one (4) for a vectorial framework as it doesn’t use directly the slope ratio $r_{ij}$. This is why we will use it to define our vectorial method in section 3. To get a limited reconstruction, $\kappa_{ij}$ will therefore be defined as a function, whose maximum and minimum admissible values depend on the limitation inequalities (7). One can easily rewrite them as an inequality on the reconstructed variable:

$$\min(U_i, U_{H_i}^+) \leq U_{ij} \leq \max(U_i, U_{H_i}^+)$$

$$\min(U_i, U_{H_i}^-) \leq U_{ij}^f \leq \max(U_i, U_{H_i}^-)$$

where we have defined $U_{ij}^f = 2U_i - U_{ij}$. If the scheme were monodimensional along the axis $B_i M_{ij}$, and if the variable $U$ had been supposed linear on the cell $(M_{ij}, M_{ij}^f)$ (in blue in Figure 1), then $U_{ij}^f$ would have been the real reconstructed value at $M_{ij}^f$. For an unstructured mesh, the point $M_{ij}^f$ doesn’t play any role and is not explicitely defined but one can still define the hypothetical value of the discrete solution at this point, that’s why we call it the fictitious reconstruction. One can see from (11) that even if this value doesn’t need to be computed in practice, we still need to define it in order to establish the scheme stability property as inequalities on both $U_{ij}$ and $U_{ij}^f$ are required.

### Figure 2
Combination of the second-order zone in green and the monotonicity constraints in red to get the stability zone for the limiter, and equivalence in terms of $\kappa$ values.

### 3 VECTORIAL MULTISLOPE MUSCL SCHEME

#### 3.1 Stability formula

The vectorial MUSCL scheme derives naturally from the stability formula of the vectorial version of the fully discretised scheme (6):

$$V_{i}^{n+1} = V_{i}^{n} - \frac{\Delta t}{|K_i|} \sum_{j \in V(i)} |S_{ij}| \Phi(V_{ij}, V_{ij}),$$

(12)
which is the discrete formulation of the vectorial advection equation:

$$\frac{\partial v}{\partial t} + (\lambda \cdot \nabla)v = 0,$$  \hspace{1cm} (13)

where \(v\) is the exact vectorial solution of (13), and \(V\) represents its vectorial discrete solution.

In order to get the stability formula, one has to make some assumption on the numerical flux function \(\Phi\). We will firstly assume that it is consistent:

$$\forall V \in \mathbb{R}^d, \quad \Phi_{ij}(V, V) = (\lambda \cdot n_{ij})V_{ij}. \hspace{1cm} (14)$$

Secondly, we assume that there exists two scalar parameters \(A_{ij}\) and \(B_{ij}\) such that

$$A_{ij}(V_{ij} - V_i) = \Phi_{ij}(V_{ij}, V_i) - \Phi_{ij}(V_i, V_i),$$

$$B_{ij}(V_{ji} - V_i) = \Phi_{ij}(V_{ij}, V_{ji}) - \Phi_{ij}(V_{ij}, V_i). \hspace{1cm} (15)$$

One can notice that if the numerical flux is the classical upwind numerical flux:

$$\Phi_{ij}(U, V) = \max(\lambda \cdot n_{ij}, 0)U + \min(\lambda \cdot n_{ij}, 0)V, \quad \forall U, V \in \mathbb{R}^d, \hspace{1cm} (16)$$

then these scalar parameters \(A_{ij}\) and \(B_{ij}\) exist and are equal to:

$$A_{ij} = \max(\lambda \cdot n_{ij}, 0), \quad B_{ij} = \min(\lambda \cdot n_{ij}, 0). \hspace{1cm} (17)$$

If we introduce the fictitious vectorial reconstruction \(V_{ij}^f = 2V_i - V_{ij}\), and two other parameters:

$$\nu_{ij}^+ = \Delta t \frac{|S_{ij}|}{|K_i|} A_{ij}, \quad \nu_{ij}^- = -\Delta t \frac{|S_{ij}|}{|K_i|} B_{ij}, \hspace{1cm} (18)$$

we get:

$$V_i^{n+1} = V_i \left(1 - \sum_{j \in V(i)} \left(\nu_{ij}^- + \nu_{ij}^+\right)\right) + \sum_{j \in V(i)} \nu_{ij}^- V_{ji} + \sum_{j \in V(i)} \nu_{ij}^+ V_{ij}^f. \hspace{1cm} (19)$$

In order to get a convex combination, we have to assume that the numerical flux is monotonous:

$$A_{ij} \geq 0, \quad B_{ij} \leq 0, \hspace{1cm} (20)$$

which means that the coefficients \(\nu_{ij}^+\) and \(\nu_{ij}^-\) are both positive. From this point, one can define the following CFL condition

$$1 - \sum_{j \in V(i)} \left(\nu_{ij}^- + \nu_{ij}^+\right) \geq 0, \hspace{1cm} (21)$$

which ensures that the formula (19) is a convex combination of \(V_i\), \(V_{ji}\) and \(V_{ij}^f\). In order to get the vectorial scheme stability, we still have to choose the way we will limit the reconstruction \(V_{ij}\) and its fictitious counterpart \(V_{ij}^f\) in the same time.
3.2 Monotonicity criterion

From the formula (19), one can observe that with the right choice of limitation for the reconstruction $V_{ij}$, we can get almost any kind of control over $V_{i}^{n+1}$. In the scalar case, a monotonicity criterion is applied on the reconstructed variable in order that $U_{i}^{n+1}$ achieve a maximum principle. On the vectorial case, the control we want over $V_{i}^{n+1}$ should ideally be the vectorial extension of this maximum principle. This vectorial extension corresponds in fact to the VIP area from Luttwak and Falcovitz [19–21], that is the convex hull of vectors from a given cell neighborhood. Hence, if we write our reconstruction under a $\kappa$-scheme form:

$$V_{ij} = V_{i} + \|B_{i}M_{ij}\| \left( \frac{1 + \kappa_{ij}}{2} p_{ij}^{+} + \frac{1 - \kappa_{ij}}{2} p_{ij}^{-} \right),$$

(22)

we only have to find an interval $I = [\kappa^{-}, \kappa^{+}]$ such that for any value of $\kappa_{ij}$ within $I$, we have that both $V_{ij}$ and its associated fictitious $V_{ij}^{f} = 2V_{i} - V_{ij}$ lie into the convex hull of the neighborhood vectors. Nevertheless, this is the best method only from a theoretical point of view. Indeed, as already explained in the introduction, the algorithmic cost of both the convex hull construction and the “lying in the Convex Hull” test may be prohibitive in practice. For this reason, we introduce here an alternative method to the convex hull. This alternative limitation takes the form of a truncated circular sector area around each reconstructed vector $V_{ij}$ and its associated fictitious reconstruction $V_{ij}^{f}$ (see Figure 3). This method ensures some kind of control over the norm and the direction of $V_{i}^{n+1}$, as illustrated in Figure 4. To summarize, this area is constructed as follows:

1. $V_{ij}$ and its fictitious counterpart have both to satisfy an upper bound on their norms:

$$\|V_{ij}\| \leq \max(\|V_{H_{ij}}^{+}\|, \|V_{i}\|), \quad \|V_{ij}^{f}\| \leq \max(\|V_{H_{ij}}^{-}\|, \|V_{i}\|).$$

(23)

This condition ensures that the vector norm $\|V_{i}^{n+1}\|$ is bounded. It is theoretically possible to build a stable numerical scheme only with this condition in the sense that the norm won’t tend toward infinity. But this stability area is much bigger than the well designed convex hull previously mentionned. Especially, if all vectors point roughly to the same direction, we can assume that $V_{i}^{n+1}$ should also point to this direction. With only the norm condition, such a control on the vector direction can not be guaranteed. That’s why we introduce the next condition.

2. $V_{ij}$ has to lie in an angular sector defined by $V_{i}$ and $V_{H_{ij}}^{+}$ on the one hand, and its fictitious counterpart $V_{ij}^{f}$ has to lie on the angular sector defined by $V_{i}$ and $V_{H_{ij}}^{-}$ on the other hand. This condition ensures that $V_{ij}$ and its associated fictitious $V_{ij}^{f}$ won’t oscillate, which implies an angular stability for $V_{i}^{n+1}$. We have to notice that this condition can occur only for two dimensional vectors. Indeed, in a three dimensional framework, $V_{ij}$ won’t automatically lie in the same plane defined $V_{i}$ and $V_{H_{ij}}^{+}$. A similar problematic occurs for the fictitious reconstruction. It means that the angular sector has to be redefined for vector of dimension 3, which won’t be done here.
3. Both previous conditions are enough to ensure a vectorial stability. But if all vectors are colinear, the current limitation won’t tend toward the scalar case limitation, as in the scalar case, a lower bound is also required on the reconstruction $U_{ij}$ as shown in formula (11). If we impose a lower bound on the vectors norm:

$$\|V_{ij}\| \geq \min(\|V_{H_{ij}}\|, \|\bar{V}_i\|), \quad \|V_{ij}^f\| \geq \min(\|V_{H_{ij}}\|, \|\bar{V}_i\|),$$

(24)

the scheme won’t be satisfying because it will prevent to use an admissible reconstruction for smooth vector fields. This is why we decided not to impose a minimum constraint on the norm, but to truncate the angular sector defined by the two previous bounds. For the angular sector defined for $V_{ij}$, we draw the line going from $\min(\|V_i\|, \|V_{H_{ij}}\|) \frac{V_{H_{ij}}}{\|V_{H_{ij}}\|}$ to $\min(\|V_i\|, \|V_{H_{ij}}\|) \|V_{H_{ij}}\|$. A similar line is drawn for the angular sector associated with the fictitious reconstruction $V_{ij}^f$.

**Figure 3.** Shape of the vectorial stability area. Red dashed line represents the truncated circular sector. Green lines represent the set of all second-order accurate reconstructions. If $V_{ij}$ and $V_{ij}^f$ lie both on the given line but within the red area, then the reconstruction is admissible. Here, $V_{ij}$ lies outside the truncated circular sector. The reconstruction $V_{ij}$ is thus not accepted.

Hence, the reconstruction and limitation process is as follows:

1. We compute the two limit values $\kappa_-$ and $\kappa_+$ such that $V_{ij}$ lies on the boundary of its truncated circular sector. It means that for any value $\kappa$ lying inside the interval $I = [\kappa_-, \kappa_+]$, the reconstructed vector $V_{ij}$ associated with $\kappa$ will lie inside its stability area.

2. We compute the two limit values $\kappa_{ij}^f$ and $\kappa_{ij}^f$ such that $V_{ij}^f$ lies on the boundary of its truncated circular sector. It means that for any value $\kappa$ lying inside the interval $I^f = [\kappa_{ij}^f, \kappa_{ij}^f]$, the fictitious reconstructed vector $V_{ij}^f$ associated with $\kappa$ will lie inside its stability area.
3. We compute the intersection of the two previous intervals, $I \cap I_f$. If this intersection is empty, then it means that we are on a kind of vectorial extremum. As in the scalar case, a vectorial extremum means that the method degenerates at first order, and $V_{ij} = V_i$.

4. If $I \cap I_f$ is not empty, we choose by any kind of way a value $\kappa$ included in $I \cap I_f$. One can for instance choose the closest value to a unique $\kappa_0$ previously chosen, or we can choose the value which maximizes or minimizes the norm $\|V_i - V_{ij}\|$ on the interval $I \cap I_f$. One can also choose for example the mean value of $I \cap I_f$. Different choices of $\kappa = f(I \cap I_f)$ imply different kinds of limitation functions, and eventually different interpolation schemes.

4 CONCLUSION

In this paper, we have presented a new way of limiting vectors in the framework of multislope MUSCL schemes for finite volume methods. This limiting method consists in requiring that the reconstruction vector and its fictitious counterpart lie in a truncated circular sector, which ensures a control over the norm and the direction of $V_i^{n+1}$. Nevertheless, we didn’t presented detailed calculus, and extension of the angular bound to three-dimensional vectors has not been presented here. Moreover, results for this method are only theoretical so far, as no numerical tests have been presented in this paper. All these aspects will be addressed in a forthcoming paper.

REFERENCES


