Positivity-preserving schemes for Euler equations: sharp and practical CFL conditions

C. Calgaro[a,c], E. Creusé[a,c], T. Goudon[d], Y. Penel[a,b,*]

[a] *INRIA Lille – Nord Europe, EPI SIMPAF, B.P. 70478, 59658 Villeneuve d'Ascq Cedex – France*
[b] *CETMEF, Team ANGE, Ministry of Ecology, Sustainable Development and Energy, LJLL, 4 place Jussieu, 75005 Paris – France*
[c] *Laboratoire Paul Painlevé (UMR 8524), Université Lille 1, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex – France*
[d] *INRIA Sophia-Antipolis – Méditerranée, EPI COFFEE, B.P. 93, 06902 Sophia-Antipolis Cedex – France*

Abstract

When one solves PDEs modelling physical phenomena, it is of great importance to take physical constraints into account. More precisely, numerical schemes have to be designed such that discrete solutions satisfy the same constraints as exact solutions. For instance, the underlying physical assumptions for the Euler equations are the positivity of both density and pressure variables.

We consider in this paper an unstructured vertex-based tesselation in $\mathbb{R}^2$. Given a MUSCL finite volume scheme and given a reconstruction method (including a limiting process), the point is to determine whether the overall scheme ensures the positivity. The present work is issued from seminal papers from Perthame & Shu and Berthon. They proved in different frameworks that under assumptions on the corresponding one-dimensional numerical flux, a suitable CFL condition guarantees that density and pressure remain positive.

We first analyse Berthon's method by presenting the ins and outs. We then propose a more general approach adding non geometric degrees of freedom. This approach includes an optimization procedure in order to make the CFL condition explicit and as less restrictive as possible. The reconstruction method is handled independently by means of τ-limiters and of an additional damping parameter. An algorithm is provided in order to specify the adjustments to make in a preexisting code based on a certain numerical flux. Numerical simulations are carried out to prove the accuracy of the method and its ability to deal with low densities and pressures.

Keywords: Positivity-preserving MUSCL schemes, 2nd-order finite volume method, Euler equations, CFL condition, τ-limiters.

*Corresponding author
Email addresses: caterina.calgaro@math.univ-lille1.fr (C. Calgaro), creuse@math.univ-lille1.fr (E. Creusé), thierry.goudon@inria.fr (T. Goudon), yohan.penel@gmail.com (Y. Penel)

1. Introduction

Partial differential equations are widely used in physics modelling and are thus expected to reproduce actual situations. In particular, most models in fluid dynamics consist in systems of conservation laws whose solutions (must) satisfy some properties: from a physical point of view to make sense (for instance, we may think of the positivity of density or fluid height) and from the point of view of mathematical analysis for the problem to be well-posed.

The transition to industrial codes must take these elements into account in addition to other considerations such as accuracy and computational costs. The design of numerical schemes consists in a balance between these antagonistic criteria since the quest for a more and more accurate solution may increase the number of unknowns and thus the computational time. As for physical constraints which include maximum principle and positivity preservation, it is important to bear in mind that they may ensure discrete problems to be solvable at each time step.

Finite volume methods (FVm) seem to suit quite well to computational fluid dynamics since they stick to the derivation of equations from basic physical principles. In addition, they enable to simulate both classical and weak solutions as well as to handle general unstructured meshes. It is worth underlying that the term FVm may relate to very different approaches [15]. FVm may differ from one another by the location of unknowns (*cell-centered* vs. *vertex-based*); notice that the issue of which one is the "best" is still open [21], especially in the case of irregular meshes where major drawbacks may appear. For instance, in vertex-based approaches, the colocation point is generally not the centroid of the control volume which prevents from identifying mean values to pointwise values and may induce a large discrepency for this kind of meshes. Similarly in the framework of MUSCL techniques and multidirectional slope procedures, cell-centered methods can lead to severe CFL restrictions as well as to a loss of accuracy in the case of mesh distortions (see e.g. [8], particularly Remark 3 page 38).

Other distinctions between MUSCL procedures are for instance the reconstruction step or the numerical flux – see for example [14, 16, 19] for hyperbolic systems of conservation laws.

To illustrate what was introduced above, let us consider the following system of conservation laws:

$$\begin{cases} \partial_t \mathbf{W} + \nabla \cdot \mathbf{F}(\mathbf{W}) = 0, \ (t, \boldsymbol{x}) \in [0, \mathcal{T}] \times \Omega, \\ \mathbf{W}(0, \boldsymbol{x}) = \mathbf{W}_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \end{cases} \tag{1}$$

for some time $\mathcal{T} > 0$ and for a bounded domain $\Omega \subset \mathbb{R}^d$ together with suitable boundary conditions. The question of boundary conditions supplementing such hyperbolic systems is quite delicate and subtle. This is due to the fact that the number and the nature of boundary conditions depend on local properties of

the solution itself and it might change with time (specializing to gas dynamics, it depends whether the flow is sub or supersonic). Hence, the difficulty is two–fold: firstly one has to design physically relevant boundary conditions which guarantees well-posedness to the initial boundary value problem. About this point we refer to [1] where practical criteria are detailed. Secondly the numerical scheme has to treat appropriately these conditions. Here we shall thus work with appropriate in-coming fields.

Then the solution $\mathbf{W}$ would be supposed to lie in a given set of admissible states $\mathcal{W}$. For instance, for the Euler equations, $\mathbf{W}$ includes density, momentum and pressure; $\mathcal{W}$ is then the set of vectors $\mathbf{W}$ for which density and pressure are positive. From a theoretical point of view (well-posedness, relevance), it is particularly important to study the invariance of $\mathcal{W}$ with respect to Syst. (1) [27, 29]: if $\mathbf{W}_0 \in \mathcal{W}$, do we have $\mathbf{W}(t,\cdot) \in \mathcal{W}$ for all $t > 0$?

This question being answered, it is relevant to study its discrete version. The present work deals with vertex-based strategies insofar as we analyze and modify ideas from works restricted to that framework [3, 4, 6, 7]. We thus introduce a vertex-based tesselation of Ω made of triangles whose nodes are denoted by $(M_i)_{1\leq i\leq N_n}$. The control volume Ω_i corresponding to M_i is built joining barycenters of cells having M_i as a vertex – see Fig. 1. $\mathcal{V}(i)$ is the set of neighbours' indices, Γ_{ij} is the interface between Ω_i and Ω_j and $\boldsymbol{n}_{ij}$ the unit normal vector to Γ_{ij} oriented from Ω_i to Ω_j. Moreover, we assume that the mesh is smooth so that:

$$\forall\, i \neq j,\ \Gamma_{ij} \cap [M_i M_j] \neq \emptyset. \tag{H0}$$

If M_i is located on the boundary, the associated control volume Ω_i connects barycenters to the midpoints of the boundary edges. Likewise, the time interval is discretized by means of a sequence of time steps Δt^n. We thus set $t^0 = 0$ and $t^{n+1} = t^n + \Delta t^n$.

To solve (1), we can use a vertex-based finite-volume scheme:

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \Delta t^n \sum_{j\in\mathcal{V}(i)} \frac{|\Gamma_{ij}|}{|\Omega_i|} \boldsymbol{\mathcal{F}}\left(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{n}_{ij}\right),$$

where $\mathbf{W}_i^n$ is a cell-wise constant approximation at time t^n of the solution over Ω_i corresponding to an internal node M_i and $\boldsymbol{\mathcal{F}}$ is the numerical flux – see further for properties satisfied by $\boldsymbol{\mathcal{F}}$. For details about the numerical handling of the boundary conditions, the reader may refer to [6]. The question now reads: if $\mathbf{W}_i^n \in \mathcal{W}$, do we have $\mathbf{W}_i^{n+1} \in \mathcal{W}$? The case of several first-order classical schemes (such as Lax-Friedrichs or Siliciu) is handled in Bouchut's review [5].

With the quest for better accuracy by means of higher order schemes, the question has to be re-addressed. When one uses a MUSCL second-order (in space) scheme as designed by Van Leer [33]:

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \Delta t^n \sum_{j\in\mathcal{V}(i)} \frac{|\Gamma_{ij}|}{|\Omega_i|} \boldsymbol{\mathcal{F}}\left(\mathbf{W}_{ij}^n, \mathbf{W}_{ji}^n, \boldsymbol{n}_{ij}\right), \tag{2}$$

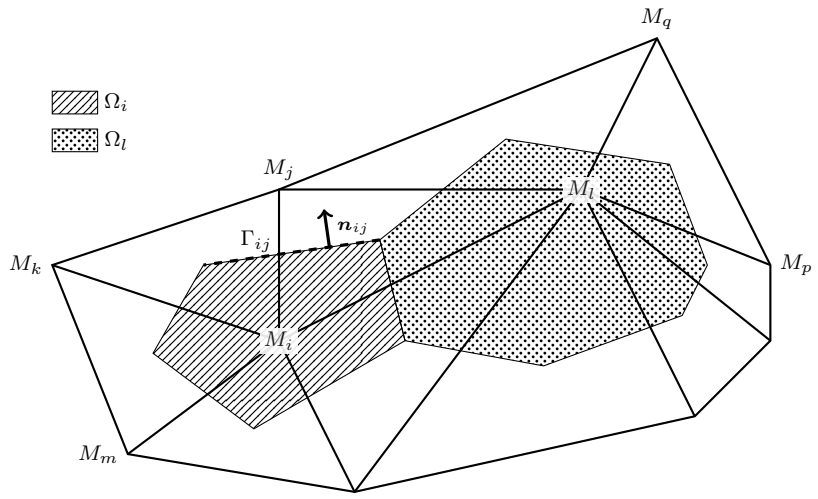


Figure 1: Vertex-based mesh: control volumes Ω_i and Ω_l

where $\mathbf{W}_{ij}^n$ and $\mathbf{W}_{ji}^n$ are reconstructed values of the solution in the vicinity of Γ_{ij}, one has to take care of the reconstruction method in the answering process. A standard Runge-Kutta second-order (in time) extension would be straightforward since it uses a decomposition into two first-order schemes to which the present approach applies.

For scalar conservation laws, discrete maximum principle has been extensively studied. For instance, we may refer to [2] where positive schemes are introduced as well as to [8] for cell-centered MUSCL schemes or to [7] for vertex-based tesselations. The case of systems of conservation laws seems as usual to be more difficult. In the framework of gas dynamics (Euler equations for ideal gas), a partial answer has been yielded by Cournède *et al.* [10]. Khobalatte & Perthame [18] and Estivalezes & Villedieu [13] provided results for kinetic schemes. Then Perthame & Shu [25] (cell-centered), Linde & Roe [20] (convex control volumes) and Berthon [3, 4] (vertex-based) proved that the second-order scheme (2) preserves positivity of density and pressure provided the 1D corresponding numerical flux does. Figuring out whether a 2D scheme preserves positivity thus comes down to studying properties of the 1D scheme. Besides, it is also the case for other kinds of numerical methods. We can for example mention the recent works of Parent applied to the resolution of the multidimensional Euler equations in generalized curvilinear coordinates using the so-called "rule of the positive coefficients" [22, 23].

Moreover, while it is well known that CFL conditions ensure stability from a numerical point of view, they proved that CFL-like estimates for the time steps also seem to be necessary to preserve positivity. However, the latter conditions for the time steps are more restrictive than usual ones. As they result from sufficient conditions, the issue to know whether they are practically necessary

remains open.

Noticing that the procedures in [3, 4, 25] do not seem optimal, we propose a slightly different approach in order to optimize the coefficients involved in their proofs and hence to make the CFL condition less restrictive and fully explicit. In Section 2, we recall Berthon's strategy showing that properties of a 1D numerical flux may imply the positivity-preserving property for a 2D finite-volume scheme. In Section 3, we revisit this approach but without any geometric interpretation in order to improve the computation of the time step. No matter what the approach which is used (geometric or algebraic), a central assumption is that the reconstruction step is "nicely" performed (in the sense that the reconstructed values are physically admissible and that a certain intermediate state to be specified exists in $\mathcal{W}$). This will be the point in Section 4. Section 5 is a summary of §§ 3 and 4: an analysis of the overall process is carried out leading to a practical procedure. The last part is devoted to numerical results for gas dynamics.

2. Berthon's approach for the Euler equations

Noting ρ, $\boldsymbol{u} = (u, v)$ and E the density, velocity and energy of a gas, the Euler equations read:

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0, \\ \partial_t (\rho \boldsymbol{u}) + \nabla \cdot \left(\rho |\boldsymbol{u}|^2 + p\right) = 0, \\ \partial_t (\rho E) + \nabla \cdot \left[(\rho E + p)\boldsymbol{u}\right] = 0, \end{cases} \tag{3}$$

where the pressure is given by the ideal gas equation of state:

$$p = \rho(\gamma - 1)\left(E - \frac{|\boldsymbol{u}|^2}{2}\right),$$

$\gamma \in [1, 3]$ being the ratio of specific heats. Eqs. (3) may be equivalently considered as a system of conservation laws and written like (1) with:

$$\mathbf{W} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{pmatrix} \quad \text{and} \quad \mathbf{F}(\mathbf{W}) = \begin{pmatrix} \rho u & \rho v \\ \rho u^2 + p & \rho uv \\ \rho uv & \rho v^2 + p \\ (\rho E + p)u & (\rho E + p)v \end{pmatrix}. \tag{4}$$

From a physical point of view, the density and pressure variables must be positive which implies that the solution $\mathbf{W}$ of (1)–(4) must lie in the set:

$$\mathcal{W} := \left\{ \mathbf{W} \in \mathbb{R}^4 \ : \ \rho = W_1 > 0 \ \text{ and } \ p = (\gamma - 1)\left[W_4 - \frac{W_2^2 + W_3^2}{2W_1}\right] > 0 \right\}.$$

A partial-order relation may be defined on $\mathbb{R}^4$ to characterize $\mathcal{W}$ [13]. It can also be shown that $p = p(\mathbf{W})$ is homogeneous of degree 1 and a continuous

concave function. Consequently, $\mathcal{W}$ is an open convex cone in $\mathbb{R}^4$. In addition, it has been proven that the Euler equations preserve the positivity of density and pressure [27, 29].

Physically speaking, it is essential that the numerical scheme we wish to use to simulate Syst. (1) with (4) do preserve the invariance of $\mathcal{W}$. Any procedure to investigate this property must rely on the convexity of the set of physical states $\mathcal{W}$. A review has been carried out in [36] where two main approaches are detailed. One approach is based on polynomial reconstructions (for a Discontinuous Galerkin method) and on quadrature integration formulae (which are convex procedures). Constraints then apply to the values of the polynomial at the quadrature nodes – see the series of papers by Zhang et al. [35, 36, 37].

The other approach has been introduced in [25] and consists in expressing the updated value $\mathbf{W}_i^{n+1}$ as a convex combination of states lying in $\mathcal{W}$. Perthame and Shu use a cell-centered tesselation combined to a monoslope[1] MUSCL procedure. More precisely, the second-order reconstructed values are computed by means of a linear function within the control volume. This results in the existence of positive coefficients ω_{ij} such that:

$$\mathbf{W}_i^n = \sum_{j \in \mathcal{V}(i)} \omega_{ij} \mathbf{W}_{ij}^n. \tag{5}$$

The authors then proved that $\mathbf{W}_i^{n+1}$ satisfies a similar relation with $\mathbf{W}_{ij}^n$ replaced by solutions of one-dimensional first order schemes (see below for more details). Assuming that reconstructed values belong to $\mathcal{W}$ and that the first-order scheme preserves positivity leads to the conclusion. The same trick applies for the 2D MUSCL scheme on cartesian grids detailled in [30] and for convex control volumes in [20].

This argument no longer holds when using a multislope procedure in the reconstruction step. Berthon adapted this idea to vertex-based meshes and multislope reconstructions in 1D [3] and in 2D [4]. We only focus on the 2D case in the sequel. This very general approach deals with the relevant CFL condition to apply given a reconstruction method and a (suitable) numerical flux.

As relations like (5) do not naturally occur in multislope methods, Berthon introduced a new state $\mathbf{W}_i^* \in \mathcal{W}$ so that $\mathbf{W}_i^n$ is a convex combination of states in $\mathcal{W}$ – see (6). But this new variable must be associated to a geometric element. That is why new points m_{ijp} are added on the segment $M_i G_{ijp}$ to build the sub-cell Ω_i^* where G_{ijp} is the barycenter of triangle $M_i M_j M_p$ (see Fig. 2). For the sake of simplicity, we suppose that the ratios $r_{ijp} := M_i m_{ijp}/M_i G_{ijp}$ are equal to r_i within the control volume Ω_i, such that the quadrilaterals $\Omega_{ij} := m_{ijp} m_{ijl} G_{ijl} G_{ijp}$ are trapezoidals.

[1] *Monoslope* means that a unique gradient is computed within the control volume. This gradient is used to reconstruct values at all interfaces. *Multisplope* methods use one gradient per interface.

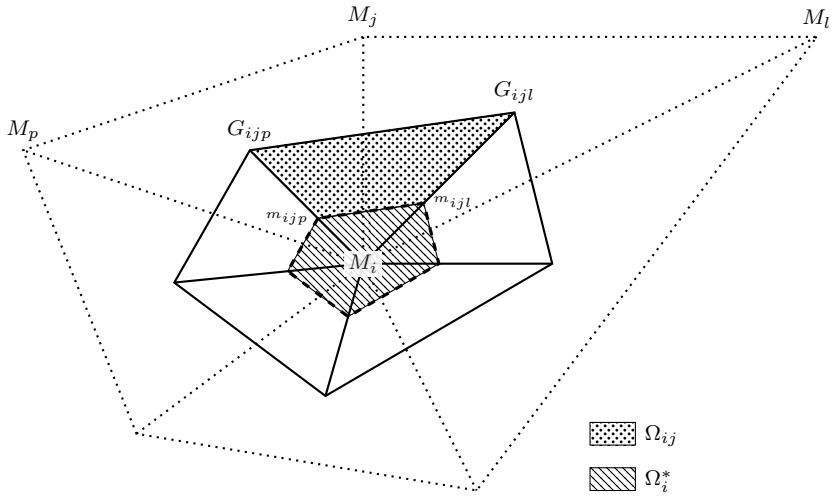


Figure 2: Sub-cells within a control volume

The convex combination reads in [4]:

$$\mathbf{W}_i^n = \frac{|\Omega_i^*|}{|\Omega_i|}\mathbf{W}_i^* + \sum_{j\in\mathcal{V}(i)} \frac{|\Omega_{ij}|}{|\Omega_i|}\mathbf{W}_{ij}^n. \tag{6}$$

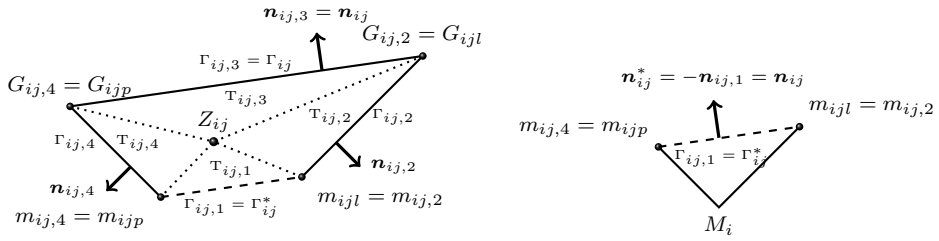
Nevertheless, given reconstructed values $\mathbf{W}_{ij}^n \in \mathcal{W}$ (see Sect. 4) the existence of a state $\mathbf{W}_i^*$ verifying (6) and belonging to $\mathcal{W}$ is not ensured. Indeed, considering the equivalent formulation:

$$\mathbf{W}_i^* = \frac{|\Omega_i|}{|\Omega_i^*|}\mathbf{W}_i^n - \sum_{j\in\mathcal{V}(i)} \frac{|\Omega_{ij}|}{|\Omega_i^*|}\mathbf{W}_{ij}^n, \tag{7}$$

we notice that the right hand side is known and does not necessarily lie in $\mathcal{W}$. The only way to satisfy this requirement is to modify the reconstruction step so that both "starred" density and pressure turn positive. A brief routine is presented in [4]. We propose in Section 4 an algorithm that provides suitable reconstructed values.

Once $\mathbf{W}_i^* \in \mathcal{W}$ is computed, the next step consists in expressing $\mathbf{W}_i^{n+1}$ as a combination similar to (6). Denoting counter-clockwise $\Gamma_{ij,k}$ the edges of Ω_{ij} (see Fig. 3) and Γ_{ij}^* the edges of Ω_i^*, Berthon sets:

$$\begin{cases} \overline{\mathbf{W}}_{ij} = \mathbf{W}_{ij}^n - \Delta t^n \displaystyle\sum_{k=1}^{4} \frac{|\Gamma_{ij,k}|}{|\Omega_{ij}|}\boldsymbol{\mathcal{F}}\left(\mathbf{W}_{ij}^n, \mathbf{W}_{ij,k}^n, \boldsymbol{n}_{ij,k}\right),\ j \in \mathcal{V}(i), & \text{(8a)} \\ \overline{\mathbf{W}}_i^* = \mathbf{W}_i^* - \Delta t^n \displaystyle\sum_{j\in\mathcal{V}(i)} \frac{|\Gamma_{ij}^*|}{|\Omega_i^*|}\boldsymbol{\mathcal{F}}\left(\mathbf{W}_i^*, \mathbf{W}_{ij}^n, \boldsymbol{n}_{ij}^*\right). & \text{(8b)} \end{cases}$$



(a) $\Omega_{ij} = \bigcup_{1 \le k \le 4} \mathrm{T}_{ij,k}$ (b) $\Omega^*_i = \bigcup_{j \in \mathcal{V}(i)} \mathrm{T}^*_{ij}$

Figure 3: Splitting of sub-cells

In accordance with the local indices of the edges, the notations are: $\mathbf{W}^n_{ij,1} = \mathbf{W}^*_i$, $\mathbf{W}^n_{ij,2} = \mathbf{W}^n_{il}$, $\mathbf{W}^n_{ij,3} = \mathbf{W}^n_{ji}$ and $\mathbf{W}^n_{ij,4} = \mathbf{W}^n_{ip}$. Due to (6) and (8), a straightforward calculation shows that (2) also reads:

$$\mathbf{W}^{n+1}_i = \frac{|\Omega^*_i|}{|\Omega_i|}\overline{\mathbf{W}}^*_i + \sum_{j \in \mathcal{V}(i)} \frac{|\Omega_{ij}|}{|\Omega_i|}\overline{\mathbf{W}}_{ij}. \tag{9}$$

Relations (8a) and (8b) correspond to 1st-order schemes applied to local variables $\mathbf{W}^n_{ij}$ and $\mathbf{W}^*_i$. To prove that $\mathbf{W}^{n+1}_i$ belongs to $\mathcal{W}$ by means of a convexity argument, it suffices to show that $\overline{\mathbf{W}}^*_i$ and $\overline{\mathbf{W}}_{ij}$ lie in $\mathcal{W}$. These requirements imply assumptions on the numerical flux $\boldsymbol{\mathcal{F}}$ as it will be shown in the sequel. A numerical flux must satisfy the following standard properties [14]:

- Consistency with the physical flux:

$$\forall\, \mathbf{W} \in \mathcal{W},\ \forall\, \boldsymbol{n} \in \mathbb{S}^2,\ \boldsymbol{\mathcal{F}}(\mathbf{W}, \mathbf{W}, \boldsymbol{n}) = \mathbf{F}(\mathbf{W}) \cdot \boldsymbol{n}; \tag{H1}$$

- Conservativity:

$$\forall\, (\mathbf{W}_l, \mathbf{W}_r, \boldsymbol{n}) \in \mathcal{W}^2 \times \mathbb{S}^2,\ \boldsymbol{\mathcal{F}}(\mathbf{W}_l, \mathbf{W}_r, \boldsymbol{n}) = -\boldsymbol{\mathcal{F}}(\mathbf{W}_r, \mathbf{W}_l, -\boldsymbol{n}); \tag{H2}$$

- Continuity:

$$\boldsymbol{\mathcal{F}} \text{ is locally Lipschitz-continuous.} \tag{H3}$$

In addition to the previous requirements, the proof is based on the invariance of $\mathcal{W}$ wrt the flux. More precisely, Berthon proved that properties of the two-dimensional flux $\boldsymbol{\mathcal{F}}$ may reduce to properties of the corresponding one-dimensional flux in the direction orthogonal to the interface. In the present case, (8a) and (8b) can be decomposed respectively as:

$$\begin{cases} \overline{\mathbf{W}}_{ij} = \displaystyle\sum_{k=1}^{4} \frac{|\mathrm{T}_{ij,k}|}{|\Omega_{ij}|}\overline{\mathbf{W}}_{ij,k}, & \text{(10a)} \\ \overline{\mathbf{W}}_{ij,k} = \\ \qquad \mathbf{W}^n_{ij} - \Delta t^n \dfrac{|\Gamma_{ij,k}|}{|\mathrm{T}_{ij,k}|}\left[\boldsymbol{\mathcal{F}}(\mathbf{W}^n_{ij}, \mathbf{W}^n_{ij,k}, \boldsymbol{n}_{ij,k}) - \boldsymbol{\mathcal{F}}(\mathbf{W}^n_{ij}, \mathbf{W}^n_{ij}, \boldsymbol{n}_{ij,k})\right], & \text{(10b)} \end{cases}$$

and:

$$\begin{cases} \overline{\mathbf{W}}_i^* = \displaystyle\sum_{j\in\mathcal{V}(i)} \frac{|\mathrm{T}_{ij}^*|}{|\Omega_i^*|}\overline{\mathbf{W}}_{ij}^*, & \text{(11a)} \\ \overline{\overline{\mathbf{W}}}_{ij}^* = \mathbf{W}_i^* - \Delta t^n \dfrac{|\Gamma_{ij}^*|}{|\mathrm{T}_{ij}^*|}\left[\boldsymbol{\mathcal{F}}(\mathbf{W}_i^*, \mathbf{W}_{ij}^n, \boldsymbol{n}_{ij}^*) - \boldsymbol{\mathcal{F}}(\mathbf{W}_i^*, \mathbf{W}_i^*, \boldsymbol{n}_{ij}^*)\right]. & \text{(11b)} \end{cases}$$

To derive one-dimensional problems, additional geometric elements were required. A point $Z_{ij} \in \Omega_{ij}$ is introduced in order to build a splitting of Ω_{ij} in four triangles – see Fig. 3a. Its position is left free. Likewise, Ω_i^* is split into $\#\mathcal{V}(i)$ triangles – see Fig. 3b.

The equivalence between (8a) and (10) relies on the Green formula and on the consistency assumption (**H1**):

$$\begin{aligned} \sum_{k=1}^{4} \frac{|\Gamma_{ij,k}|}{|\Omega_{ij}|}\boldsymbol{\mathcal{F}}(\mathbf{W}_{ij}^n, \mathbf{W}_{ij}^n, \boldsymbol{n}_{ij,k}) &= \sum_{k=1}^{4} \frac{|\Gamma_{ij,k}|}{|\Omega_{ij}|}\mathbf{F}(\mathbf{W}_{ij}^n)\cdot\boldsymbol{n}_{ij,k} \\ &= \frac{\mathbf{F}(\mathbf{W}_{ij}^n)}{|\Omega_{ij}|}\cdot\sum_{k=1}^{4}|\Gamma_{ij,k}|\boldsymbol{n}_{ij,k} = \frac{\mathbf{F}(\mathbf{W}_{ij}^n)}{|\Omega_{ij}|}\cdot\int_{\partial\Omega_{ij}}\boldsymbol{n}_{ij}\,\mathrm{d}\sigma \\ &= \frac{1}{|\Omega_{ij}|}\int_{\Omega_{ij}}\nabla\cdot\underbrace{\mathbf{F}(\mathbf{W}_{ij}^n)}_{=constant}\,\mathrm{d}\boldsymbol{x} = \mathbf{0}. \end{aligned}$$

What is interesting in formulations (10) and (11) is that $\overline{\mathbf{W}}_{ij,k}$ and $\overline{\overline{\mathbf{W}}}_{ij}^*$ are now solutions of first-order one-dimensional schemes. To lead to the conclusion, it suffices to have:

$$\forall\,(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{n}) \in \mathcal{W}^2 \times \mathbb{S}^2,\ \exists\,\Delta t > 0,$$
$$\mathbf{W}_1 - \frac{\Delta t}{\ell}\left[\boldsymbol{\mathcal{F}}(\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{n}) - \boldsymbol{\mathcal{F}}(\mathbf{W}_1, \mathbf{W}_1, \boldsymbol{n})\right] \in \mathcal{W}. \quad (12)$$

This is obviously satisfied for Δt small enough since $\mathcal{W}$ is open and $\mathbf{W}_1 \in \mathcal{W}$. The issue reduces to determining how small Δt should be which provides the CFL condition.

It is possible to go further for most schemes by taking into account the rotational invariance of the physical flux in the Euler equations [32, § 3.2.1]. Setting:

$$\mathrm{R}_n = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_n & \sin\theta_n & 0 \\ 0 & -\sin\theta_n & \cos\theta_n & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \theta_n = \arccos(\boldsymbol{n}\cdot\mathbf{e}_1),$$

and:

$$\mathbf{F}_n(\mathbf{W}) = \mathbf{F}(\mathbf{W})\cdot\boldsymbol{n},$$

we have: $\mathbf{F}_n(\mathbf{W}) = \mathrm{R}_n^{-1}\mathbf{F}_{e_1}(\mathrm{R}_n\mathbf{W})$. Notice that $\mathcal{W}$ is also invariant by rotation.

It is thus natural to assume that the numerical flux has the same property, which is the case for most classical fluxes such as Lax-Friedrichs and Godunov schemes:[2] $\boldsymbol{\mathcal{F}}_n(\mathbf{W}_l, \mathbf{W}_r) := \boldsymbol{\mathcal{F}}(\mathbf{W}_l, \mathbf{W}_r, \boldsymbol{n}) = \mathrm{R}_n^{-1}\boldsymbol{\mathcal{F}}_{e_1}(\mathrm{R}_n\mathbf{W}_l, \mathrm{R}_n\mathbf{W}_r)$. Hence we assume that, given $(\mathbf{W}_1, \mathbf{W}_2) \in \mathcal{W}^2$, if $\Delta t > 0$ satisfies

$$\frac{\Delta t}{\ell}\lambda(\mathbf{W}_1, \mathbf{W}_2) \leq \alpha_0, \tag{H4.CFL}$$

then

$$\mathbf{W}_1 - \frac{\Delta t}{\ell}\left[\boldsymbol{\mathcal{F}}_{e_1}(\mathbf{W}_1, \mathbf{W}_2) - \boldsymbol{\mathcal{F}}_{e_1}(\mathbf{W}_1, \mathbf{W}_1)\right] \in \mathcal{W}. \tag{H4}$$

Here $\alpha_0 > 0$ is specific to the flux $\boldsymbol{\mathcal{F}}_{e_1}$. For instance, for the Lax-Friedrichs scheme, $\alpha_0 = 1$ since in the present situation – see (10b) and (11b) – there is no wave arising on the left side. $\lambda(\mathbf{W}_l, \mathbf{W}_r)$ denotes the largest eigenvalue of the Riemann problem with numerical flux $\boldsymbol{\mathcal{F}}_{e_1}$ and initial data $(\mathbf{W}_l, \mathbf{W}_r)$. In the Euler case (4), we have:

$$\lambda(\mathbf{W}_l, \mathbf{W}_r) = \max\{|u_l| + c_l, |u_r| + c_r\},\ c = \sqrt{\frac{\gamma p}{\rho}}.$$

Then (**H4**) implies (12) under the CFL condition (**H4.CFL**) with λ replaced by λ_n:

$$\lambda_n(\mathbf{W}_l, \mathbf{W}_r) = \max\{|\boldsymbol{u}_l \cdot \boldsymbol{n}| + c_l, |\boldsymbol{u}_r \cdot \boldsymbol{n}| + c_r\}.$$

We are now able to conclude. Given a numerical flux $\boldsymbol{\mathcal{F}}$ satisfying (**H1**-**H2**-**H3**-**H4**) and given reconstructed values $\mathbf{W}_{ij}^n \in \mathcal{W}$ such that $\mathbf{W}_i^* \in \mathcal{W}$, we deduce from (10b) and (11b) that $\overline{\mathbf{W}}_{ij,k}$ and $\overline{\mathbf{W}}_{ij}^*$ belong to $\mathcal{W}$ under the following CFL conditions:

$$\Delta t^n \max_{\substack{j\in\mathcal{V}(i)\\ 1\leq k\leq 4}} \frac{|\Gamma_{ij,k}|}{|\mathrm{T}_{ij,k}|}\lambda_{n_{ij,k}}(\mathbf{W}_{ij}^n, \mathbf{W}_{ij,k}^n) \leq \alpha_0, \tag{13a}$$

$$\Delta t^n \max_{j\in\mathcal{V}(i)} \frac{|\Gamma_{ij}^*|}{|\mathrm{T}_{ij}^*|}\lambda_{n_{ij}^*}(\mathbf{W}_i^*, \mathbf{W}_{ij}^n) \leq \alpha_0. \tag{13b}$$

Then, $\overline{\mathbf{W}}_{ij}$ and $\overline{\mathbf{W}}_i^*$ are in $\mathcal{W}$ as (10a) and (11a) are convex combinations. For the same reason, (9) shows that $\mathbf{W}_i^{n+1} \in \mathcal{W}$. To compute explicitly the time step, we have to combine the finite collection of conditions (13) for each node M_i.

In a nutshell, MUSCL scheme (2) guarantees that density and pressure variables remain positive as long as the time step satisfies inequalities (13). In addition to the preservation of positivity, this procedure also guarantees that solutions satisfy entropy inequalities [4]. Indeed, the modifications we present in this paper do not affect Berthon's proof about entropy-satisfying solutions.

[2] For a numerical flux which would not satisfy the rotational invariance, (**H4.CFL**) is replaced by a genuinely multidimensional version.

An important feature of this approach is that the scheme is actually implemented under formulation (2): intermediate variables such as $\overline{\mathbf{W}}_{ij}$ or $\overline{\mathbf{W}}_i^*$ are never computed. They are only useful from a theoretical point of view. Nevertheless, one must emphasize that CFL conditions (13) may lead to severe computational costs. Indeed, two parameters are not specified in [4]: the ratio r_i corresponding to the localization of nodes m_{ijp} and the position of nodes Z_{ij} which influences the values of $|\mathrm{T}_{ij,k}|$. Arbitrary choices for these parameters may significantly decrease Δt^n as it will be shown in § 6. For Z_{ij}, it is easy to prove that its optimal position corresponds either to a segment or to a single node depending on r_i. Indeed, the ratio $|\Gamma_{ij,k}|/|\mathrm{T}_{ij,k}|$ is equal to $2/h_{ij,k}$ where $h_{ij,k}$ is the height associated to Z_{ij} in the triangle $\mathrm{T}_{ij,k}$. The optimization process comes down to maximizing the smallest height in Ω_{ij}. As for r_i, it influences all geometric values as well as the eigenvalues $\lambda_{n_{ij}^*}(\mathbf{W}_i^*, \mathbf{W}_{ij}^n)$ (the state $\mathbf{W}_i^*$ is computed from (7) where the areas $|\Omega_{ij}|$ and $|\Omega_i^*|$ appear).

To avoid this dependance on geometric aspects, we present in the next section a more abstract approach based on Berthon's procedure: we keep the idea of expressing the second-order two-dimensional scheme as a convex combination of first-order one-dimensional schemes but we introduce non geometric coefficients in order to optimize CFL conditions.

3. Algebraic approach

3.1. Deriving a new CFL condition

The core of the method is still the convexity of the physical set $\mathcal{W}$. We keep the same notations for the splitting of Ω_i as the union of Ω_i^* and Ω_{ij} even if these elements will not be used *a posteriori*. Rather than considering the geometry-dependent mean (6), we introduce some coefficients η_{ij} and η_i^* such that:

$$\mathbf{W}_i^n = \eta_i^* \mathbf{W}_i^* + \sum_{j \in \mathcal{V}(i)} \eta_{ij} \mathbf{W}_{ij}^n. \tag{14}$$

We must underline that this very section based on Eq. (14) is not restricted to vertex-based approaches. It can adapt to the cell-centered framework directly, the set $\mathcal{V}(i)$ consisting of exactly three indices.

From now on, we suppose that coefficients $\eta = \{\eta_i^*, \eta_{ij}\}$ are known (see § 5 for more details) and verify:

$$\eta_{ij} \geq 0, \qquad \eta_i^* \geq 0, \qquad \eta_i^* + \sum_{j \in \mathcal{V}(i)} \eta_{ij} = 1. \tag{15}$$

Instead of (8), we update in time $\mathbf{W}_{ij}^n$ and $\mathbf{W}_i^*$ by means of first-order schemes:

$$\begin{cases} \overline{\mathbf{W}}_{ij} = \mathbf{W}_{ij}^n - \Delta t^n \displaystyle\sum_{k=1}^{4} \zeta_{ij,k} \boldsymbol{\mathcal{F}} \left(\mathbf{W}_{ij}^n, \mathbf{W}_{ij,k}^n, \boldsymbol{n}_{ij,k} \right), \ j \in \mathcal{V}(i), \\ \overline{\mathbf{W}}_i^* = \mathbf{W}_i^* - \Delta t^n \displaystyle\sum_{j \in \mathcal{V}(i)} \zeta_{ij}^* \boldsymbol{\mathcal{F}} \left(\mathbf{W}_i^*, \mathbf{W}_{ij}^n, \boldsymbol{n}_{ij}^* \right), \end{cases} \tag{16}$$

where we naturally assume that:

$$\zeta_{ij,k} \geq 0, \tag{17a}$$

$$\sum_{k=1}^{4} \zeta_{ij,k} \boldsymbol{n}_{ij,k} = \mathbf{0}, \tag{17b}$$

$$\zeta^*_{ij} \geq 0, \tag{17c}$$

$$\sum_{j \in \mathcal{V}(i)} \zeta^*_{ij} \boldsymbol{n}^*_{ij} = \mathbf{0}. \tag{17d}$$

Equalities (17b) and (17d) correspond to the preservation of steady states specific to Eqs. (16). To obtain from (2), (14) and (16) the convex combination for $\mathbf{W}_i^{n+1}$:

$$\mathbf{W}_i^{n+1} = \eta^*_i \overline{\mathbf{W}}^*_i + \sum_{j \in \mathcal{V}(i)} \eta_{ij} \overline{\mathbf{W}}_{ij},$$

a straightforward computation shows that it is necessary and sufficient to have (see Fig. 3):

$$\begin{cases} \eta_{ij}\zeta_{ij,1} = \eta^*_i \zeta^*_{ij}, & \text{(18a)} \\ \eta_{ij}\zeta_{ij,2} = \eta_{il}\zeta_{il,4}, & \text{(18b)} \\ \eta_{ij}\zeta_{ij,3} = \dfrac{|\Gamma_{ij}|}{|\Omega_i|}. & \text{(18c)} \end{cases}$$

Next step consists in introducing one-dimensional schemes similarly to (10):

$$\begin{cases} \overline{\mathbf{W}}_{ij} = \displaystyle\sum_{k=1}^{4} \nu_{ij,k} \overline{\mathbf{W}}_{ij,k}, \\ \overline{\mathbf{W}}_{ij,k} = \\ \quad \mathbf{W}^n_{ij} - \Delta t^n \mu_{ij,k} \left[\boldsymbol{\mathcal{F}}(\mathbf{W}^n_{ij}, \mathbf{W}^n_{ij,k}, \boldsymbol{n}_{ij,k}) - \boldsymbol{\mathcal{F}}(\mathbf{W}^n_{ij}, \mathbf{W}^n_{ij}, \boldsymbol{n}_{ij,k}) \right], \end{cases}$$

and:

$$\begin{cases} \overline{\mathbf{W}}^*_i = \displaystyle\sum_{j \in \mathcal{V}(i)} \nu^*_{ij} \overline{\mathbf{W}}^*_{ij}, \\ \overline{\mathbf{W}}^*_{ij} = \mathbf{W}^*_i - \Delta t^n \mu^*_{ij} \left[\boldsymbol{\mathcal{F}}(\mathbf{W}^*_i, \mathbf{W}^n_{ij}, \boldsymbol{n}^*_{ij}) - \boldsymbol{\mathcal{F}}(\mathbf{W}^*_i, \mathbf{W}^*_i, \boldsymbol{n}^*_{ij}) \right]. \end{cases}$$

Given (17b) and (17d), the equivalence with (16) holds iff:

$$\begin{cases} \displaystyle\sum_{k=1}^{4} \nu_{ij,k} = 1, \quad \nu_{ij,k}\mu_{ij,k} = \zeta_{ij,k}, \\ \displaystyle\sum_{j \in \mathcal{V}(i)} \nu^*_{ij} = 1, \quad \nu^*_{ij}\mu^*_{ij} = \zeta^*_{ij}. \end{cases}$$

Unknowns $\nu = \{\nu^*_{ij}, \nu_{ij,k}\}$ can be easily eliminated:

$$\nu_{ij,k} = \frac{\zeta_{ij,k}}{\mu_{ij,k}}, \qquad \nu^*_{ij} = \frac{\zeta^*_{ij}}{\mu^*_{ij}},$$

so that hypotheses on ζ and μ reduce to:

$$\begin{cases} \displaystyle\sum_{k=1}^{4} \frac{\zeta_{ij,k}}{\mu_{ij,k}} = 1, & \text{(19a)} \\ \displaystyle\sum_{j\in\mathcal{V}(i)} \frac{\zeta^*_{ij}}{\mu^*_{ij}} = 1. & \text{(19b)} \end{cases}$$

Consequently, this yields the fact that $\mathbf{W}_i^{n+1} \in \mathcal{W}$ under the assumptions (**H1**-**H2**-**H3**-**H4**) and the CFL conditions:

$$\begin{cases} \Delta t^n \displaystyle\max_{\substack{j\in\mathcal{V}(i)\\1\le k\le 4}} \mu_{ij,k}\lambda_{n_{ij,k}}(\mathbf{W}_{ij}^n, \mathbf{W}_{ij,k}^n) \le \alpha_0, \\ \Delta t^n \displaystyle\max_{j\in\mathcal{V}(i)} \mu^*_{ij}\lambda_{n^*_{ij}}(\mathbf{W}_i^*, \mathbf{W}_{ij}^n) \le \alpha_0. \end{cases}$$

Given that $\boldsymbol{n}^*_{ij} = -\boldsymbol{n}_{ij,1} = \boldsymbol{n}_{ij,3} = \boldsymbol{n}_{ij}$, it is convenient to consider the weaker formulation:

$$\Delta t^n \max_{j\in\mathcal{V}(i)} \left\{ \mu^*_{ij}, \max_{1\le k\le 4} \mu_{ij,k} \right\} \times \bar{\lambda}_i^n \le \alpha_0, \qquad (20)$$

$$\bar{\lambda}_i^n := \max_{\substack{j\in\mathcal{V}(i)\\1\le k\le 4}} \left\{ |\boldsymbol{u}_{ij}^n \cdot \boldsymbol{n}_{ij,k}| + c_{ij}^n, |\boldsymbol{u}_{ij,k}^n \cdot \boldsymbol{n}_{ij,k}| + c_{ij,k}^n \right\}.$$

This CFL condition consists of classical MUSCL eigenvalues together with the ones (for $k = 1$) associated to the new states $\mathbf{W}_i^*$ and which will be investigated in Sect. 5.

To make this approach legitimate, we aim at minimizing the coefficients $\boldsymbol{\mu} = \{\mu^*_{ij}, \mu_{ij,k}\}$ in order to maximize Δt^n. As we shall see, there is a balance between the order of the method and the CFL condition through the coefficients η_{ij}. To better understand the dependency wrt to these coefficients, we solve the system of constraints.

3.2. Solving the constraints

We first notice that constraint (17d) is redundant. Indeed:

$$\begin{aligned}\sum_{j\in\mathcal{V}(i)}\zeta_{ij}^*\boldsymbol{n}_{ij}^* &= -\sum_{j\in\mathcal{V}(i)}\zeta_{ij}^*\boldsymbol{n}_{ij,1} \overset{(18a)}{=} -\frac{1}{\eta_i^*}\eta_{ij}\zeta_{ij,1}\boldsymbol{n}_{ij,1}\\ &\overset{(17b)}{=} \frac{1}{\eta_i^*}\sum_{j\in\mathcal{V}(i)}\eta_{ij}\left[\zeta_{ij,2}\boldsymbol{n}_{ij,2}+\zeta_{ij,3}\boldsymbol{n}_{ij,3}+\zeta_{ij,4}\boldsymbol{n}_{ij,4}\right]\\ &\overset{(18c)}{=} \frac{1}{\eta_i^*}\Bigg[\underbrace{\sum_{j\in\mathcal{V}(i)}\frac{|\Gamma_{ij}|}{|\Omega_i|}\boldsymbol{n}_{ij}}_{=\mathbf{0}\text{ (Green)}}+\sum_{j\in\mathcal{V}(i)}\eta_{ij}\zeta_{ij,2}\boldsymbol{n}_{ij,2}+\underbrace{\sum_{j\in\mathcal{V}(i)}\eta_{ij}\zeta_{ij,4}\boldsymbol{n}_{ij,4}}_{\overset{(18b)}{=}\sum_{p\in\mathcal{V}(i)}\eta_{ip}\zeta_{ip,2}(-\boldsymbol{n}_{ip,2})}\Bigg]=\mathbf{0}.\end{aligned}$$

Another remark is that (17c) is a consequence of (17a) due to (18a). Moreover, from a linear algebra argument, there are only two independent vectors in the set $\{\boldsymbol{n}_{ij,k}\}_{1\le k\le 4}$. In our case, we choose $\boldsymbol{n}_{ij,3}$ and $\boldsymbol{n}_{ij,4}$ as linearly independent vectors. Others are connected by the relations:

$$\sum_{k=1}^{4}|\Gamma_{ij,k}|\boldsymbol{n}_{ij,k}=\mathbf{0},\qquad \boldsymbol{n}_{ij,1}=-\boldsymbol{n}_{ij,3}. \tag{21a}$$

Due to the fact that $m_{ij,2}G_{ij,2}G_{ij,4}m_{ij,4}$ is a trapezoidal, the combination of these equalities yields (see Fig. 3 for notations):

$$\boldsymbol{n}_{ij,2}=-\frac{|\Gamma_{ij}|}{M_iG_{ij,2}}\boldsymbol{n}_{ij,3}-\frac{M_iG_{ij,4}}{M_iG_{ij,2}}\boldsymbol{n}_{ij,4}. \tag{21b}$$

Taking (21) into account, (17b) now reads:

$$\mathbf{0}=\left(-\zeta_{ij,1}-\frac{|\Gamma_{ij}|}{M_iG_{ij,2}}\zeta_{ij,2}+\zeta_{ij,3}\right)\boldsymbol{n}_{ij,3}+\left(-\frac{M_iG_{ij,4}}{M_iG_{ij,2}}\zeta_{ij,2}+\zeta_{ij,4}\right)\boldsymbol{n}_{ij,4}.$$

As the vectors are independent, we obtain:

$$\begin{cases}\zeta_{ij,1}=\zeta_{ij,3}-\dfrac{|\Gamma_{ij}|}{M_iG_{ij,2}}\zeta_{ij,2},\\[2ex] \zeta_{ij,4}=\dfrac{M_iG_{ij,4}}{M_iG_{ij,2}}\zeta_{ij,2},\end{cases}$$

or equivalently, due to (18c):

$$\frac{\eta_{ij}\zeta_{ij,2}}{M_iG_{ij,2}}=\frac{\eta_{ij}\zeta_{ij,4}}{M_iG_{ij,4}}=\frac{1}{|\Omega_i|}-\frac{\eta_{ij}\zeta_{ij,1}}{|\Gamma_{ij}|}. \tag{22}$$

The system of constraints (17a-18a-18b-18c-19a-19b-22) is now solvable. Let $j_0 \in \mathcal{V}(i)$ be a fixed index. Set $X = \zeta_{ij_0,1}$. As $M_i G_{ij,2} = M_i G_{il,4} = M_i G_{ijl}$ and due to (18b), we induce that:

$$\frac{1}{|\Omega_i|} - \frac{\eta_{ij}\zeta_{ij,1}}{|\Gamma_{ij}|} = \frac{1}{|\Omega_i|} - \frac{\eta_{il}\zeta_{il,1}}{|\Gamma_{il}|}.$$

We then iteratively verify that for all $j \in \mathcal{V}(i)$:

$$\zeta_{ij,1} = \frac{|\Gamma_{ij}|}{|\Gamma_{ij_0}|}\frac{\eta_{ij_0}}{\eta_{ij}}X, \qquad \zeta^*_{ij} = \frac{|\Gamma_{ij}|}{|\Gamma_{ij_0}|}\frac{\eta_{ij_0}}{\eta^*_i}X, \qquad \zeta_{ij,3} = \frac{|\Gamma_{ij}|}{\eta_{ij}|\Omega_i|},$$

$$\zeta_{ij,2} = \frac{M_i G_{ijl}}{\eta_{ij}}\left[\frac{1}{|\Omega_i|} - \frac{\eta_{ij_0}}{|\Gamma_{ij_0}|}X\right], \qquad \zeta_{ij,4} = \frac{M_i G_{ijk}}{\eta_{ij}}\left[\frac{1}{|\Omega_i|} - \frac{\eta_{ij_0}}{|\Gamma_{ij_0}|}X\right],$$

with:

$$0 \le X \le X_{max} := \frac{|\Gamma_{ij_0}|}{\eta_{ij_0}|\Omega_i|}.$$

The upper bound for X comes from the positivity of $\zeta_{ij,2}$ and $\zeta_{ij,4}$.

3.3. Optimizing the CFL condition

Minimizing the largest $\mu_{ij,k}$ satisfying (19a) comes down to taking $\mu_{ij,k}$ equal to each other for all k (*cf.* **Lemma 1** in the Appendix), *i.e.*:

$$\mu_{ij,k} \equiv \mu_{ij}(X) = \sum_{k=1}^{4}\zeta_{ij,k} = \frac{|\partial \mathrm{T}_{ij}|}{\eta_{ij}|\Omega_i|} - \frac{X}{|\Gamma_{ij_0}|}\frac{\eta_{ij_0}}{\eta_{ij}}\big(|\partial \mathrm{T}_{ij}| - 2|\Gamma_{ij}|\big).$$

Here T_{ij} denotes the triangle $M_i G_{ijl} G_{ijk}$. As for the condition (19b) associated to Ω^*_i, we similarly take (**Lemma 1**):

$$\mu^*_{ij} \equiv \mu^*_i(X) = \sum_{j\in\mathcal{V}(i)}\zeta^*_{ij} = \frac{|\partial\Omega_i|}{|\Gamma_{ij_0}|}\frac{\eta_{ij_0}}{\eta^*_i}X.$$

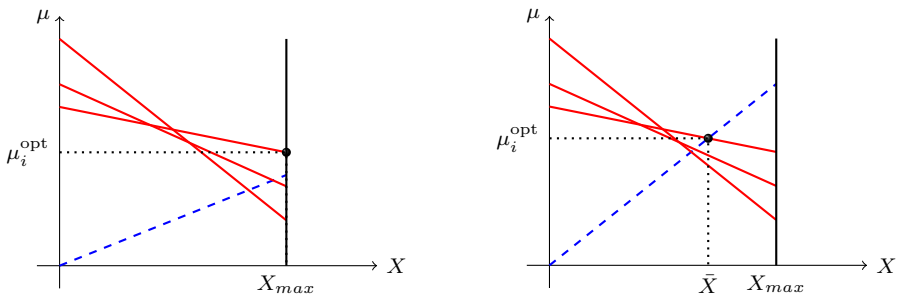
The optimization problem thus reduces to:

$$\mu_i^{\mathrm{opt}} := \min_{0\le X\le X_{max}} \max\left\{\mu^*_i(X), \max_{j\in\mathcal{V}(i)}\mu_{ij}(X)\right\}.$$

As each constraint is linear wrt X and monotone (increasing for μ^*_i and decreasing for μ_{ij}), its solution is (see Fig. 4):

(a). If $\mu^*_i(X_{max}) \le \max_{j\in\mathcal{V}(i)}\mu_{ij}(X_{max})$, $\mu_i^{\mathrm{opt}} = \max_{j\in\mathcal{V}(i)}\mu_{ij}(X_{max})$;

(b). Otherwise, $\mu_i^{\mathrm{opt}} = \mu^*_i(\bar{X})$ where $\bar{X} = \max_{j\in\mathcal{V}(i)}\{X_j : \mu^*_i(X_j) = \mu_{ij}(X_j)\}$.



(a) $\mu_i^*(X_{max}) \leq \max\limits_{j\in\mathcal{V}(i)} \mu_{ij}(X_{max})$ (b) $\mu_i^*(X_{max}) > \max\limits_{j\in\mathcal{V}(i)} \mu_{ij}(X_{max})$

Figure 4: Graphs of μ_i^* (dashed blue line) and μ_{ij} (plain red lines)

Given the expressions of μ_i^* and μ_{ij} with respect to X, this solution reads:

$$\mu_i^{\text{opt}} = \begin{cases} 2\dfrac{|\Gamma_{ij_1}|}{\eta_{ij_1}|\Omega_i|}, & \text{if } \dfrac{|\partial\Omega_i|}{\eta_i^*} \leq 2\dfrac{|\Gamma_{ij_1}|}{\eta_{ij_1}}, \\ \dfrac{|\partial\Omega_i|}{|\Omega_i|}\dfrac{|\partial \mathrm{T}_{ij_2}|}{\eta_{ij_2}|\partial\Omega_i| + \eta_i^*\left(|\partial \mathrm{T}_{ij_2}| - 2|\Gamma_{ij_2}|\right)}, & \text{otherwise,} \end{cases} \tag{23}$$

where $j_1 = \underset{j\in\mathcal{V}(i)}{\text{argmax}} \dfrac{|\Gamma_{ij}|}{\eta_{ij}}$ and $j_2 = \underset{j\in\mathcal{V}(i)}{\text{argmax}} \{X_j : \mu_i^*(X_j) = \mu_{ij}(X_j)\}$.

The optimal choice μ_i^{opt} therefore highly depends on the coefficients η_{ij}. The final step would consist in optimizing μ_i^{opt} wrt all (η_{ij}) such that (15) holds. However, these coefficients also influence the reconstruction procedure. That is why it is necessary first to get interested in the computation of reconstructed values.

4. Reconstruction step

This section is largely independent from the previous one. The overall process in the design of positivity-preserving schemes relies on the convex combination (14) and thus on the existence of an intermediate state such that:

$$\mathbf{W}_i^* = \frac{1}{\eta_i^*}\left[\mathbf{W}_i^n - \sum_{j\in\mathcal{V}(i)} \eta_{ij}\mathbf{W}_{ij}^n\right] \in \mathcal{W}. \tag{24}$$

We assumed previously that the reconstructed values $\mathbf{W}_{ij}^n$ belong to $\mathcal{W}$. In this section, we first explain how to compute these values and secondly how to ensure that $\mathbf{W}_i^* \in \mathcal{W}$.

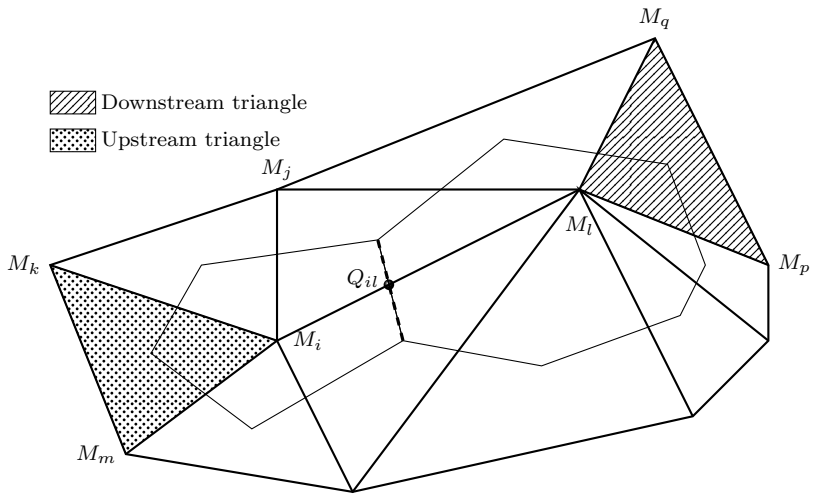


Figure 5: Reconstruction step: upstream and downstream gradients

4.1. Computation of the gradients

Let $\Delta\mathbf{W}_{ij}^n$ be the gradients on each interface of the control volume for the conservative variables, *i.e.* such that $\mathbf{W}_{ij}^n = \mathbf{W}_i^n + \Delta\mathbf{W}_{ij}^n$. There are three issues to address when computing $\Delta\mathbf{W}_{ij}^n$: guaranteeing that $\mathbf{W}_{ij}^n \in \mathcal{W}$, improving the accuracy of the method and avoiding numerical instabilities. It has been proven in the literature that the use of flux/slope limiters may help achieve these purposes [33, 31]. A limiter enables to combine a high-order scheme (where the solution is smooth) and a first-order scheme (elsewhere). Most limiters are functions of consecutive gradients (in one dimension) or upstream/downstream gradients (in higher dimensions) – see Fig. 5.

More precisely, we use the approach described in [7]. Let M_i and M_l be two nodes (with M_l an internal node). We aim at approaching values of a reconstructed variable ξ at the point Q_{il} by means of the formulae:

$$\xi_{il} = \xi_i + \alpha_{il}\varphi(r_{il})\overline{\Delta\xi}_{il} = \xi_i + \alpha_{il}\left[\phi(r_{il})\overline{\Delta\xi}_{il}^{up} + \big(1-\phi(r_{il})\big)\overline{\Delta\xi}_{il}\right], \tag{25a}$$

$$\xi_{li} = \xi_l - \alpha_{li}\varphi(r_{li})\overline{\Delta\xi}_{il} = \xi_l - \alpha_{li}\left[\phi(r_{li})\overline{\Delta\xi}_{il}^{dn} + \big(1-\phi(r_{li})\big)\overline{\Delta\xi}_{il}\right]. \tag{25b}$$

The notations are – see Fig. 5:

- $\alpha_{il} = \dfrac{M_iQ_{il}}{M_iM_l}$, $\alpha_{li} = 1-\alpha_{il}$;
- $\overline{\Delta\xi}_{il} = \xi_l - \xi_i = \nabla\hat{\xi}_{ijl}\cdot\overrightarrow{M_iM_l}$ where $\hat{\xi}_{ijl}$ is a linear approximation of ξ within $M_iM_jM_l$;

- $\overline{\Delta\xi}_{il}^{up} = \nabla\hat{\xi}_{ikm} \cdot \overrightarrow{M_i M_l}$ where $\hat{\xi}_{ikm}$ is a linear approximation of ξ within the upstream triangle $M_i M_k M_m$; if this triangle does not exist (when M_i is located on the boundary), we set $\overline{\Delta\xi}_{il}^{up} = 0$;
- $\overline{\Delta\xi}_{il}^{dn} = \nabla\hat{\xi}_{lqp} \cdot \overrightarrow{M_i M_l}$ where $\hat{\xi}_{lqp}$ is a linear approximation of ξ within the downstream triangle $M_l M_q M_p$;
- $r_{il} = \dfrac{\overline{\Delta\xi}_{il}^{up}}{\overline{\Delta\xi}_{il}}$ and $r_{li} = \dfrac{\overline{\Delta\xi}_{il}^{dn}}{\overline{\Delta\xi}_{il}}$; if $\overline{\Delta\xi}_{il} = 0$, we set $r_{il} = r_{li} = 0$ so that $\xi_{il} = \xi_{li} = \xi_i = \xi_l$;
- φ is the limiter.

For scalar equations in one dimension, the TVD property requires a slope limiter φ to satisfy estimates [31] like:

$$0 \leq \varphi(r) \leq \min(2r, 2).$$

To reach higher orders, it is necessary to have $\varphi(1) = 1$ and φ as smooth as possible in the neighbourhood of 1 [30], which is equivalent to assuming smoothness for ϕ such that $\varphi(r) = 1 - (1 - r)\phi(r)$. The readers may refer to [26] and [6] about procedures for designing new limiters.

For scalar equations in two dimensions, the same inequality turns out to be sufficient for reconstruction at the middle of the edges even if the TVD concept no longer applies [17]. However, for control volumes as on Fig. 1, interfaces do not cross edges at their middles. That is why limiters have to be adapted. Q-limiters have been introduced for cell-centered meshes [8] and τ-limiters for vertex-based triangulations [7]. These limiters satisfy:

$$0 \leq \varphi(r) \leq \min(\tau r, \tau) \tag{26}$$

where τ is a geometric parameter to be specified. To guarantee that $\xi_{il} > 0$ and $\xi_{li} > 0$ for a limiter φ satisfying (26), it is necessary to take:

$$\tau = \min_{i,l} \frac{1}{\alpha_{il}}. \tag{27}$$

Due to Hyp. **H0**, Q_{il} lies between M_i and M_l which ensures that $\alpha_{il} \in [0, 1]$ and it can be shown [7] that $\tau \in [1, 2]$.

To extend this approach to systems of equations, the first question that arises is about the set of variables to which to apply the limitation procedure. Indeed for the Euler equations, it is possible to limit the conservative variables $(\rho, \rho u, \rho v, \rho E)$ [20], the physical variables (ρ, u, v, p) [9], the characteristic variables [6] or the entropic variables [3]. Berthon carried out numerical simulations in [3] to highlight the influence of this choice. The fact still remains that there is no compelling argument to make the decision. However as far as this study is concerned, it seems more relevant to limit the physical variables insofar as we are designing a method to ensure that two physical variables (ρ, p) remain

positive. Denoting by $\mathbf{U} = (\rho, u, v, p)$, the reconstruction is computed like this:

Algorithm 1

1. Perform the change of variable $\mathbf{U}_i = \kappa(\mathbf{W}_i)$;
2. Compute "physical" gradients $\Delta\mathbf{U}_{ij}$ by applying (25) to each component of $\mathbf{U}_i$;
3. Compute "conservative" gradients $\Delta\mathbf{W}_{ij} = \kappa^{-1}\big(\mathbf{U}_i + \Delta\mathbf{U}_{ij}\big) - \mathbf{W}_i$.

By construction, $\widetilde{\mathbf{W}}_{ij} := \mathbf{W}_i + \Delta\mathbf{W}_{ij}$ lies in $\mathcal{W}$.

4.2. Construction of $\mathbf{W}_i^*$

We then investigate the existence of $\mathbf{W}_i^*$ belonging to $\mathcal{W}$. Due to the facts that $\mathcal{W}$ is open and that $\mathbf{W}_i^n \in \mathcal{W}$, there exist η_{ij} small enough such that (24) holds. However, according to § 3.3 (see (23) for example), the smaller η_{ij}, the worse the CFL condition. That is why another strategy has to be found out. Let us introduce $\beta_i \in [0,1]$ and set:

$$\mathbf{W}_{ij} := \mathbf{W}_i + \beta_i \Delta\mathbf{W}_{ij}. \tag{28}$$

The equality $\mathbf{W}_{ij} = (1-\beta_i)\mathbf{W}_i + \beta_i\widetilde{\mathbf{W}}_{ij}$ shows that $\mathbf{W}_{ij} \in \mathcal{W}$. Due to (15) and (28), Eq. (24) then reads:

$$\mathbf{W}_i^* = \mathbf{W}_i^n - \beta_i^n \sum_{j\in\mathcal{V}(i)} \frac{\eta_{ij}}{\eta_i^*}\Delta\mathbf{W}_{ij}^n = \mathbf{W}_i^n + \frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\Delta\mathbf{W}_i^*,$$
$$\Delta\mathbf{W}_i^* := \frac{-1}{\displaystyle\sum_{k\in\mathcal{V}(i)}\eta_{ik}} \sum_{j\in\mathcal{V}(i)} \eta_{ij}\Delta\mathbf{W}_{ij}. \tag{29}$$

$\Delta\mathbf{W}_i^*$ is (up to the minus sign) an average of the gradients over the cell Ω_i. As above, the facts that $\mathbf{W}_i^n \in \mathcal{W}$ and that $\mathcal{W}$ is open imply that there exists β_i^n small enough so that $\mathbf{W}_i^* \in \mathcal{W}$. The case $\beta_i^n = 1$ corresponds to the second order (except for extrema, where the gradient is zero and the scheme degenerates to order 1 owing to the limiter) while $\beta_i^n = 0$ corresponds to the first order. $\beta_i^n < 1$ thus implies a loss of accuracy compared to the pure second-order MUSCL scheme but turns out to be necessary for guaranteeing positivity. As it will be highlighted in the numerical simulations, $\beta_i^n \neq 1$ only locally.

The requirements $\rho_i^* > 0$ and $p_i^* > 0$ corresponding to $\mathbf{W}_i^* \in \mathcal{W}$ reduce to:[3]

$$\mathcal{P}_1\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right) := 1 + D_i^n\left[\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right] > 0, \tag{30a}$$
$$\mathcal{P}_2\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right) := 1 + B_i^n\left[\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right] + A_i^n\left[\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right]^2 > 0, \tag{30b}$$

[3] $\mathcal{P}_1$ and $\mathcal{P}_2$ are such that $\rho_i^* = \rho_i^n\mathcal{P}_1\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right)$ and $p_i^* = p_i^n\frac{\mathcal{P}_2}{\mathcal{P}_1}\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right)$.

with:

$$A_i^n = \frac{\gamma-1}{\rho_i^n p_i^n}\left[\Delta\rho_i^*\,\Delta(\rho E)_i^* - \frac{|\Delta(\rho\boldsymbol{u})_i^*|^2}{2}\right];$$
$$B_i^n = \frac{\gamma-1}{p_i^n}\left[E_i^n\Delta\rho_i^* + \Delta(\rho E)_i^* - \boldsymbol{u}_i^n\cdot\Delta(\rho\boldsymbol{u})_i^*\right]; \qquad D_i^n = \frac{\Delta\rho_i^*}{\rho_i^n}.$$

Consider the roots of the linear inequality (30a) and of the (at most) quadratic constraint (30b). We must bear in mind that our aim is to take β_i^n as large as possible in $[0,1]$ not to damage accuracy. For the first inequality, we set:

$$\vartheta_i^{(\rho)} := \begin{cases} +\infty, & \text{if } D_i^n \geq 0, \\ -\dfrac{1}{D_i^n}, & \text{otherwise,} \end{cases} \qquad \text{and } \beta_i^{(\rho)} := \min\left\{1, \frac{\eta_i^*}{1-\eta_i^*}\vartheta_i^{(\rho)}\right\}. \tag{31a}$$

As for the other inequality, we set $\delta_i^n = (B_i^n)^2 - 4A_i^n$. This term is clearly positive:[4]

$$\delta_i^n = \left(\frac{\gamma-1}{p_i^n}\right)^2\left[\Delta(\rho E)_i^* - \boldsymbol{u}_i^n\cdot\Delta(\rho\boldsymbol{u})_i^* + \Delta\rho_i^*\left(|\boldsymbol{u}_i^n|^2 - E_i^n\right)\right]^2$$
$$+ 2\frac{\gamma-1}{p_i^n\rho_i^n}\left[\left(\Delta\rho_i^*|\boldsymbol{u}_i^n| - \frac{\boldsymbol{u}_i^n\cdot\Delta(\rho\boldsymbol{u})_i^*}{|\boldsymbol{u}_i^n|}\right)^2 + \left(|\Delta(\rho\boldsymbol{u})_i^*|^2 - \frac{[\boldsymbol{u}_i^n\cdot\Delta(\rho\boldsymbol{u})_i^*]^2}{|\boldsymbol{u}_i^n|^2}\right)\right].$$

Hence we set:

$$\vartheta_i^{(p)} := \begin{cases} +\infty, & \text{if } (A_i^n = 0, B_i^n \geq 0) \text{ or } (A_i^n > 0, B_i^n > 0), \\ -\dfrac{1}{B_i^n}, & \text{if } (A_i^n = 0, B_i^n < 0), \\ -\dfrac{B_i^n + \sqrt{\delta_i^n}}{2A_i^n}, & \text{otherwise,} \end{cases} \tag{31b}$$

and $\beta_i^{(p)} := \min\left\{1, \dfrac{\eta_i^*}{1-\eta_i^*}\vartheta_i^{(p)}\right\}$. To satisfy (30a) and (30b) simultaneously, we then have to take:

$$\beta_i^n \leq \min\left\{\beta_i^{(\rho)}, \beta_i^{(p)}\right\}. \tag{32}$$

The analysis of this result and the actual choice for β_i^n are specified in the next section.

[4] This expression holds if $\boldsymbol{u}_i^n \neq \boldsymbol{0}$. Otherwise, $p_i^n = (\gamma-1)\rho_i^n E_i^n$ and $\delta_i^n = \left\{\left[E_i^n\Delta\rho_i^* - \Delta(\rho E)_i^*\right]^2 + 2E_i^n\left|\Delta(\rho\boldsymbol{u})_i^*\right|^2\right\}/(\rho_i^n E_i^n)^2 > 0$.

5. Practical issues

The final step of our approach consists in choosing coefficients η_{ij} so as to obtain to a balance between the CFL condition (characterized by $\bar{\lambda}_i^n$ and μ_i^{opt}) and the order of the scheme (characterized by β_i^n). A first comment is that it is always possible to introduce $\tilde{\eta}_{ij}$ such that:

$$\eta_{ij} = (1-\eta_i^*)\tilde{\eta}_{ij} \text{ and } \sum_{j\in\mathcal{V}(i)} \tilde{\eta}_{ij} = 1 \quad \Longrightarrow \quad \Delta\mathbf{W}_i^* = -\sum_{j\in\mathcal{V}(i)} \tilde{\eta}_{ij}\Delta\mathbf{W}_{ij}.$$

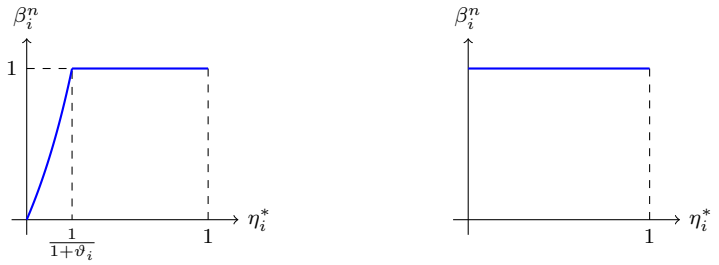
Note that $\Delta\mathbf{W}_i^*$ is independent from η_i^*. The optimal choice would consist in maximizing Δt^n to avoid extra computations (equivalent to minimizing $\bar{\lambda}_i^n$ and μ_i^{opt}) and maximizing β_i^n to improve accuracy. However, as it will be demonstrated in the sequel, these two goals are antagonistic. Moreover, due to numerous nonlinearities, it does not seem achievable to state an optimal choice for the three aforementioned coefficients. That is why we choose to mainly focus on μ_i^{opt}. Indeed, the worst case for β_i^n corresponds to $\beta_i^n = 0$ and the scheme locally degenerates to order 1. An unrelevant choice for η_i^* and $\tilde{\eta}_{ij}$ would make μ_i^{opt} go to ∞ and Δt^n to 0. Hence the priority given to μ_i^{opt} involved in the CFL condition even if we first study the influence of η_i^* on the two other coefficients (parameters $\tilde{\eta}_{ij}$ are assumed to be given).

In Sect. 4, we detailed ① how to obtain reconstructed values in $\mathcal{W}$ by means of τ-limiters and ② how to ensure $\mathbf{W}_i^* \in \mathcal{W}$ thanks to the use of a damping coefficient β_i^n. The latter parameter must satisfy the constraint (32). A first comment is that if $\beta_i^{(\rho)} < 1$ (resp. $\beta_i^{(p)} < 1$), we cannot take $\beta_i^n = \beta_i^{(\rho)}$ (resp. $\beta_i^n = \beta_i^{(p)}$) since this would imply $\rho_i^* = 0$ (resp. $p_i^* = 0$) and $\mathbf{W}_i^* \notin \mathcal{W}$. Secondly, it is important to mention where $\mathbf{W}_i^*$ is involved from a practical point of view. Although this additional variable is used to prove positivity, it is not part of the numerical scheme (2). It only determines the CFL condition (20) through the term $|\boldsymbol{u}_{ij,1}^n \cdot \boldsymbol{n}_{ij,1}| + c_{ij,1}^n = \left|\frac{(\rho\boldsymbol{u})_i^*}{\rho_i^*} \cdot \boldsymbol{n}_{ij}\right| + c_i^*$. To prevent $c_i^* = \sqrt{\gamma p_i^*/\rho_i^*}$ (and so $\bar{\lambda}_i^n$) from growing drastically, it seems better to have β_i^n close to $\beta_i^{(p)}$ than to $\beta_i^{(\rho)}$. More precisely, we finally take:

$$\beta_i^n := \min\left\{1, \frac{\eta_i^*}{1-\eta_i^*}\sigma_\rho\vartheta_i^{(\rho)}, \frac{\eta_i^*}{1-\eta_i^*}\sigma_p\vartheta_i^{(p)}\right\}, \tag{33}$$

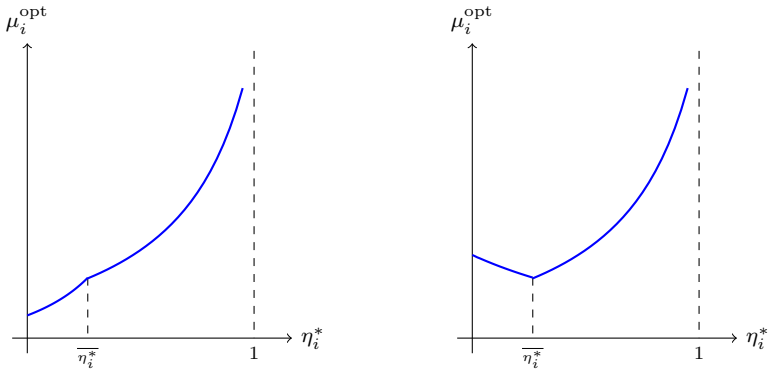
with $\sigma_\rho < \sigma_p < 1$. Fig. 6 shows how β_i^n evolves depending on η_i^*. Hence it is tempting to take η_i^* large enough as it would ensure $\beta_i^n = 1$. But as we shall see later, large η_i^* must be avoided.

To go back to $\bar{\lambda}_i^n$, we sketch out the evolution of the "starred" eigenvalue



(a) $\vartheta_i < +\infty$ (b) $\vartheta_i = +\infty$

Figure 6: Plot of $\beta_i^n(\eta_i^*)$ depending on $\vartheta_i := \min\{\sigma_\rho \vartheta_i^{(\rho)}, \sigma_p \vartheta_i^{(p)}\}$



(a) $\varepsilon_i < 0$ (b) $\varepsilon_i \geq 0$

Figure 7: Plot of $\mu_i^{\text{opt}}(\eta_i^*)$ depending on ε_i

wrt η_i^*:

$$|\boldsymbol{u}_i^* \cdot \boldsymbol{n}_{ij}| + c_i^* = \left| \frac{(\rho \boldsymbol{u})_i^*}{\rho_i^*} \cdot \boldsymbol{n}_{ij} \right| + c_i^* =$$
$$\frac{1}{\mathcal{P}_1\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right)} \left[\left| \boldsymbol{u}_i^n \cdot \boldsymbol{n}_{ij} + \frac{1-\eta_i^*}{\eta_i^*}\beta_i^n \frac{\Delta(\rho \boldsymbol{u})_i^*}{\rho_i^n} \cdot \boldsymbol{n}_{ij} \right| + c_i^n \sqrt{\mathcal{P}_2\left(\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n\right)} \right]. \tag{34}$$

This eigenvalue may be predominant in $\bar{\lambda}_i^n$ in (20). We first notice that other eigenvalues $|\boldsymbol{u}_{ij,k}^n \cdot \boldsymbol{n}_{ij,k}| + c_{ij,k}^n$, $k \neq 1$, remain bounded no matter what η_i^* (they only depend on $\beta_i^n \in [0,1]$). We then deal with two cases: if $\vartheta_i := \min\{\sigma_\rho \vartheta_i^{(\rho)}, \sigma_p \vartheta_i^{(p)}\} < \infty$, then the term $\frac{1-\eta_i^*}{\eta_i^*}\beta_i^n$ belongs to $[0, \vartheta_i]$. The rhs in (34) is thus bounded.

However, when $\vartheta_i^{(\rho)} = \vartheta_i^{(p)} = +\infty$, then $\beta_i^n = 1$, $1/\mathcal{P}_1(\frac{1-\eta_i^*}{\eta_i^*}) \leq 1$ and the rhs in (34) goes to $+\infty$ when $\eta_i^* \to 0$. Once more, large η_i^* seem more suitable.

We now investigate the tuning of μ_i^{opt} in the CFL condition (20). The resulting time step is such that the scheme does preserve positivity of density and pressure. As mentioned above, we expect μ_i^{opt} to be as small as possible. The problem reads:

$$\min_{\substack{(\tilde{\eta}_{ij}) \in \mathbb{R}_+^{\#\mathcal{V}(i)} \\ \sum_j \tilde{\eta}_{ij} = 1}} \min_{0 \leq \eta_i^* \leq 1} \mu_i^{\text{opt}}(\eta_i^*, \tilde{\eta}_{ij}). \tag{$\mathcal{P}$}$$

With the η_i^*-parameterization, (23) reads:

$$\mu_i^{\text{opt}}(\eta_i^*, \tilde{\eta}_{ij}) = \begin{cases} \dfrac{2}{(1-\eta_i^*)|\Omega_i|} \max\limits_{j \in \mathcal{V}(i)} \dfrac{|\Gamma_{ij}|}{\tilde{\eta}_{ij}}, & \text{if } \eta_i^* \geq \overline{\eta_i^*}, \\ \dfrac{|\partial\Omega_i|}{|\Omega_i|} \left[\min\limits_{j \in \mathcal{V}(i)} \left\{ \eta_i^* \left(1 - \dfrac{2|\Gamma_{ij}|}{|\partial \mathrm{T}_{ij}|} - \dfrac{\tilde{\eta}_{ij}|\partial\Omega_i|}{|\partial \mathrm{T}_{ij}|} \right) + \dfrac{\tilde{\eta}_{ij}|\partial\Omega_i|}{|\partial \mathrm{T}_{ij}|} \right\} \right]^{-1}, \end{cases} \tag{35}$$

where:

$$\overline{\eta_i^*}(\tilde{\eta}_{ij}) := \frac{|\partial\Omega_i|}{|\partial\Omega_i| + 2 \max\limits_{j \in \mathcal{V}(i)} \dfrac{|\Gamma_{ij}|}{\tilde{\eta}_{ij}}}.$$

We still denote by j_2 the index for which the minimum is reached in the second case in (35) and we set:

$$\varepsilon_i := 1 - \frac{2|\Gamma_{ij_2}|}{|\partial \mathrm{T}_{ij_2}|} - \frac{\tilde{\eta}_{ij_2}|\partial\Omega_i|}{|\partial \mathrm{T}_{ij_2}|}$$

which is the coefficient in front of η_i^* in (35). The profile of μ_i^{opt} is pictured on Fig. 7 depending on the sign of ε_i.

According to these figures, the minimum is thus reached either for $\eta_i^* = 0$ or $\eta_i^* = \overline{\eta_i^*}$. In view of what was shown above, the former case must not be chosen as $\bar{\lambda}_i^n$ may tend to $+\infty$. Therefore, a reasonable choice seems to be $\eta_i^* = \overline{\eta_i^*}$. In that case, μ_i^{opt} reads:

$$\mu_i^{\text{opt}} = \frac{|\partial\Omega_i| + 2 \max\limits_{j\in\mathcal{V}(i)} \dfrac{|\Gamma_{ij}|}{\tilde{\eta}_{ij}}}{|\Omega_i|}. \tag{36}$$

Hence:

$$(\mathcal{P}) \leq \min_{\substack{(\tilde{\eta}_{ij})\in\mathbb{R}_+^{\#\mathcal{V}(i)} \\ \sum_j \tilde{\eta}_{ij}=1}} \mu_i^{\text{opt}} \left(\overline{\eta_i^*}(\tilde{\eta}_{ij}), \tilde{\eta}_{ij}\right).$$

Applying **Lemma 1**, we obtain the exact solution for the right optimization problem which is:

$$\tilde{\eta}_{ij} = \frac{|\Gamma_{ij}|}{|\partial\Omega_i|}, \qquad \eta_i^* = \frac{1}{3}, \qquad \mu_i^{\text{opt}} = 3\frac{|\partial\Omega_i|}{|\Omega_i|}, \tag{37a}$$

$$\beta_i^n = \min\left\{1, \frac{\sigma_\rho}{2}\vartheta_i^{(\rho)}, \frac{\sigma_p}{2}\vartheta_i^{(p)}\right\}. \tag{37b}$$

The most striking feature of this approach is that η_i^* is independent from the cell Ω_i and this enables to save computational time. Moreover, it is far away from 0 to avoid the issues mentioned previously about β_i^n and $\bar{\lambda}_i^n$. Eventually, numerical simulations go to show that η_i^* is large enough to provide order 2 almost everywhere (see § 6).

To conclude, we should underline that coefficients (37a) are optimized from the point of view of the CFL conditions. Other choices are possible depending on the feature one wants to focus on (accuracy, computational time, ...): these coefficients remain user-tuned.

We see on Fig. 8 the influence of this choice on μ_i^{opt} and thus on the computation of the time step through (20) for a single control volume. Our choice (37a) pictured in dashed red line on Fig. 8 provides a significantly small lower bound for the CFL condition compared to the other coefficients implemented:

Case 1. $\tilde{\eta}_{ij} = \dfrac{1}{\#\mathcal{V}(i)}$;

Case 2. $\tilde{\eta}_{ij} = \dfrac{|\partial \mathrm{T}_{ij}|}{|\Omega_i|}$;

Case 3. $\tilde{\eta}_{ij} = \dfrac{|\Gamma_{ij}|}{|\partial\Omega_i|}$ for various η_i^*.

Case 2. corresponds to coefficients used by Berthon [4] ($\eta_i^* = r_i^2$). Although η_i^* is finally tuned by the user, we notice that an arbitrary choice may have dramatic consequences on the computational time.

We end up this section with a practical algorithm. Given a numerical code solving the Euler equations, here are the few modifications to make:

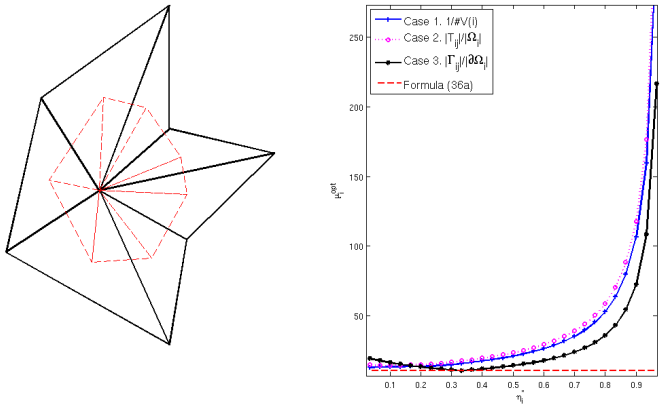


Figure 8: Plot of $\mu_i^{\text{opt}}(\eta_i^*)$ (right) for a single control volume (left)

Algorithm 2

1. *Iteration 0*
   (a) Tune coefficients $\tilde{\eta}_{ij}$;
   (b) Compute $\eta_i^* = \overline{\eta_i^*}$ and deduce μ_i^{opt} from (36);
2. *For every iteration n and every vertex i*
   (a) Compute $\Delta\mathbf{W}_{ij}^n$ according to **Algorithm 1**;
   (b) Compute $\Delta\mathbf{W}_i^*$ given by (29);
   (c) Compute $\vartheta_i^{(\rho)}$ and $\vartheta_i^{(p)}$ from (31); deduce β_i^n from (33);
   (d) Evaluate $\mathbf{W}_{ij}^n$ from (28), $\mathbf{W}_i^*$ from (29) and $\bar{\lambda}_i^n$ from (20);
   (e) Compute Δt^n from (20);
   (f) Update $\mathbf{W}_i^{n+1}$ by (2).

More precisely, there is an initial step that replaces the computation of the smallest height. The very difference with classical codes consists in computing steps 2.(b) and 2.(c). They correspond to an additional loop but nevertheless do not produce prohibitive extra computational cost. It is the price to pay to ensure that no matter how close to vaccum the physical solution may be, the numerical solution remains admissible.

6. Numerical results

This section deals with numerical simulations on benchmarks in order to highlight the effects of the method. Our approach is very general and adapts to many numerical fluxes provided they satisfy (**H1**)-(**H2**)-(**H3**) which are very classical and (**H4**). This is the latter requirement which may fail. The most

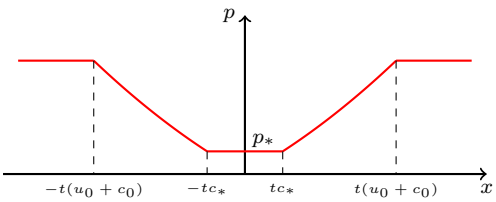


Figure 9: Profile of the pressure at time t

famous (and maybe the most commonly used) scheme that does not work is the Roe scheme [12]. This defect has been corrected in the HLL family (see for instance the HLLE scheme [11]). Entropy fixes have also been derived to avoid non-physical (*i.e.* not entropy-satisfying) solutions. Some of them may be reinterpreted as positive schemes – see *e.g.* [24].

Other schemes that do satisfy (**H4**) are for instance the Godunov method, the Lax-Friedrichs flux (and consequently the Rusanov flux) – see [25, Appendix] or [5, § 2.4.2] – and the Siliciu relaxation scheme – see [4] or [5, § 2.4.4].

We must underline that in many situations (**H4**) is not required if the numerical state is far enough from the critical region ($\rho = 0, p = 0$) and if gradients are not too steep. That is why Roe scheme provides admissible results in many cases. What we state in this paper is that our approach including reconstruction (with τ-limiters and β_i^n) and relevant CFL conditions guarantees the positivity of density and pression no matter what the data.

We present in the sequel some numerical simulations with the Rusanov flux. Despite diffusive effects, it yields accurate results. As far as the limiter is concerned, we use a modified Van Leer limiter [8] so that (26) holds:

$$\varphi(r) = \begin{cases} 0, & \text{if } r < 0, \\ \dfrac{\tau r}{1+(\tau-1)r}, & \text{if } 0 \leq r \leq 1, \\ \dfrac{\tau r}{\tau - 1 + r}, & \text{if } r > 1, \end{cases}$$

with τ defined by (27). See Tab. 1 for concrete values.

6.1. 123 problem

The first problem is a one-dimensional Riemann problem for the Euler equations (3) which is known to be one of the most suitable tests to assess the robustness of a scheme. Indeed, the so-called 123 problem consists of two rarefaction waves where the intermediate state $\mathbf{U}_\#$ is close to vacuum ($\rho \ll 1$, $p \ll 1$). It is presented in [32, § 6.4]. The left and right initial physical states are $\mathbf{U}_l = (\rho_0, -u_0, 0, p_0)$ and $\mathbf{U}_r = (\rho_0, u_0, 0, p_0)$. The profile of the pressure at time t is pictured on Fig. 9 (the density has a similar profile).

Mesh	$\min h_k$	$1/\mu_i^{\text{opt}}$	τ
Structured	0.0025	0.00026	1.50
Unstructured	0.0022	0.00017	1.24

Table 1: Coefficients involved in CFL conditions and in reconstruction step

The intermediate state $\mathbf{U}_\#$ is given by:

$$\rho_\# = \rho_0 \left(1 - \frac{\gamma-1}{2}\frac{u_0}{c_0}\right)^{\frac{2}{\gamma-1}}, \; u_\# = 0, \; v_\# = 0, \; p_\# = p_0 \left(1 - \frac{\gamma-1}{2}\frac{u_0}{c_0}\right)^{\frac{2\gamma}{\gamma-1}}.$$

For instance, for $\gamma = 1.4$ and the original data ($\rho_0 = 1, u_0 = 2, p_0 = 0.4$), then the initial value in conservative variables is $\mathbf{W}_r = (1, 2, 0, 3)$ and:

$$\rho_\# \approx 0.0219 \quad \text{and} \quad p_\# \approx 0.0019.$$

But it is possible to tune the initial data to get closer to vacuum. In particular, the critical set of data is given by $u_0 = \frac{2c_0}{\gamma-1}$ for which $\rho_\# = 0$ and $p_\# = 0$. To prove the robustness of the procedure, we take:

$$\rho_0 = 1, \; p_0 = 0.4, \; u_0 = 3.74 < \frac{2c_0}{\gamma-1} \approx 3.742.$$

The corresponding intermediate states are about the machine epsilon. The exact density remains positive and so must do the numerical solution under the CFL condition (20).

This 1D problem is simulated with a 2D code. We consider the 2D domain $\Omega = [-0.5, 0.5] \times [0, 0.25]$ for which two meshes are considered: a cartesian grid (Fig. 10a – 8180 nodes, 15920 triangles) and an unstructured grid with a straight interface generated by Triangle Mesh Generator [28] (Fig. 10b – 8064 nodes, 15799 triangles).

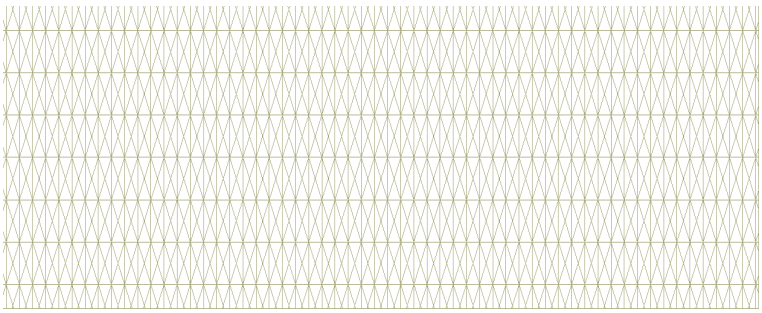
Geometric parameters involved in CFL conditions (smallest height over the whole tesselation for order 1, $(\mu_i^{\text{opt}})^{-1}$ for order 2) are specified in Tab. 1. To ensure positivity of density and pressure, the present study requires (according to the figures) to divide the time step by about 10 (wrt the 1st-order CFL condition) which is equivalent to processing 10 times more iterations. This is in accordance with a similar study devoted to the scalar case [7, Eq. (40)].

The Riemann problem corresponds to the initial condition:

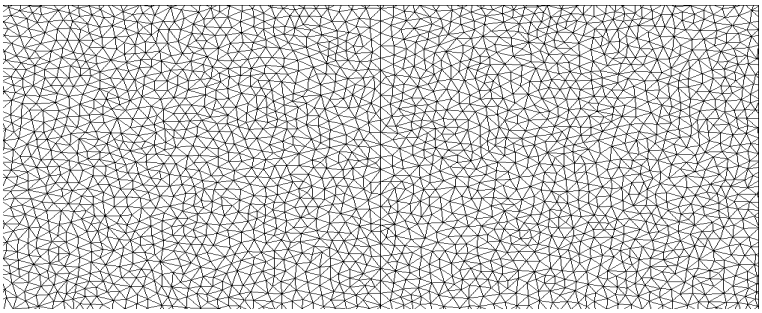
$$\mathbf{W}_0(x, y) = \begin{cases} \mathbf{W}_l, & \text{if } x < 0, \\ \mathbf{W}_r, & \text{if } x > 0. \end{cases}$$

All the results in the sequel are presented over the domain $[0, 0.5] \times \{0.125\}$ insofar as the solution is symmetric (wrt to $x = 0$) and invariant wrt y.

The two specific issues related to this test are about the positivity of density and pressure and the profile of energy (*cf.* [32, Fig. 6.14] or [9, Fig. 14] where



(a) Structured tesselation



(b) Unstructured mesh

Figure 10: Grids for 123 problem

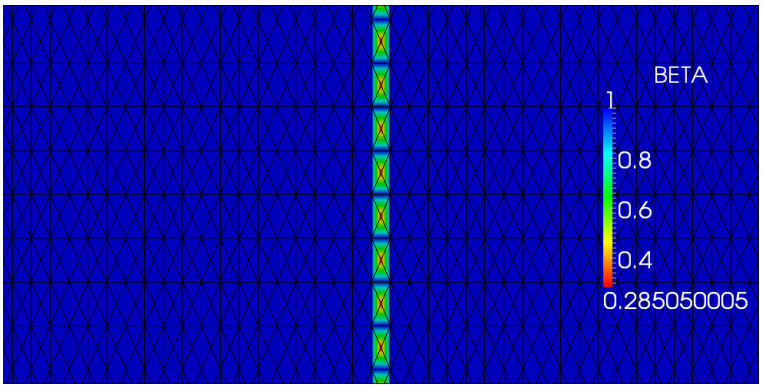


Figure 11: Location of vertices where $\beta_i^n \neq 1$

there is a noticeable overshoot in the vicinity of 0). Typically, **the 2nd-order scheme** (2) **without the β-procedure fails to preserve the positivity** of p which becomes negative at iteration 13 on both kinds of grid no matter what the time step. The use of the coefficient β_i^n (and of the additional state $\mathbf{W}_i^*$) is necessary to ensure that $\mathbf{W}_i^{n+1} \in \mathcal{W}$ as expected. In many non critical cases, the classical version of the FV scheme (2) does not suffer from positivity defects. However, this case proves that the additional numerical tools we present in this paper help to cure this issue. Coefficient β_i^n may be interpreted as a new class of limiters devoted to the preservation of positivity.

More precisely, this coefficient is "activated" (in the sense that $\beta_i^n < 1$) only for a few iterations (up to iteration 19 for the structured mesh, up to iteration 110 for the unstructured grid) and in a narrow area (see FIG. 11 for the structured mesh) consisted of the vertices which belong to the interface. It can be accounted for by the steep gradients on the density variable – this can be inferred from (31a) that the greater $\Delta\rho_i^*$, the lower β_i^n. A similar analysis may be carried out for the pressure variable but expressions for $\beta_i^{(p)}$ is much more intricate.

FIG. 12 shows the comparison (for the density variable) between the 1st-order scheme together with the standard CFL condition

$$\Delta t^n \max_i(|\boldsymbol{u}_i^n| + c_i^n) \leq \alpha_0 \min_i h_i$$

and the modified 2nd-order scheme (*i.e.* including the β_i^n coefficient) – referred to as "order β" in the legend – processed on the structured grid. The latter scheme provides more accurate results as expected.

Results at a larger time (β_i^n is no longer activated) are depicted on FIG. 13 for the energy and on FIGS. 14 for the density. On FIG. 13, we notice that energy is correctly handled (no overshoot) in the vicinity of 0 no matter what the order of the scheme. However, results are not accurate due to the fact that computing $E_\#$ involves the ratio $p_\#/\rho_\#$ while these two variables are close to 0.

FIG. 14a shows that 1st- and β-order versions of the scheme satisfy the positivity constraint (unlike the 2nd-order scheme which fails as mentioned above). The structured solutions obviously seem better than their unstructured counterparts. This clear advantage seems in accordance with the fact that this test is one-dimensional in essence. Indeed, if we pay attention to two specific areas (namely where the solution is less smooth at the origin of the rarefaction wave – FIG. 14b and where it gets close to 0 – FIG. 14c), we remark different qualitative evolutions. In particular, in the vicinity of 0, order β (unstr) yields less accurate results than order 1 (unstr). The conclusion is reversed in the structured case. Moreover, although $\beta_i^n = 1$ everywhere (corresponding to order 2), solutions are close to order 1. The first iterations are thus predominant in this test.

6.2. Mach 3 wind tunnel case

The second problem is a genuine 2D case presented in [34]. It consists in an inward Mach 3 flow within a singular domain. More precisely, the domain is a rectangle which is 3 meter long and 1 meter high with a step (2.4m $\times$

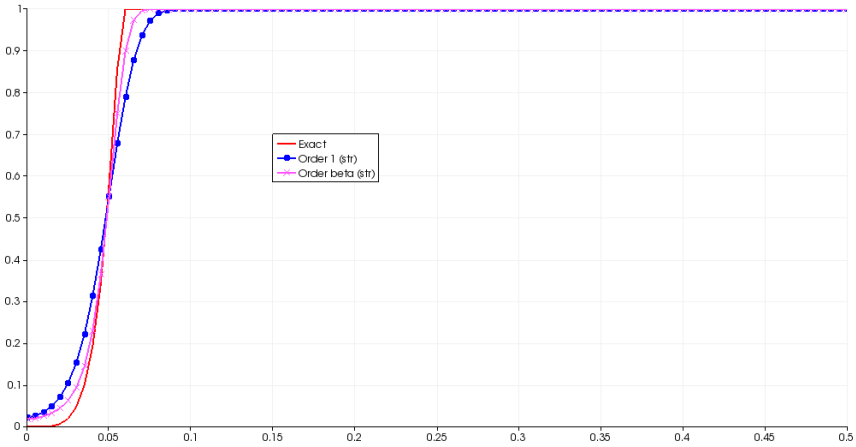


Figure 12: Density plot at time 0.0125: 1st-order and modified 2nd-order schemes on a structured grid

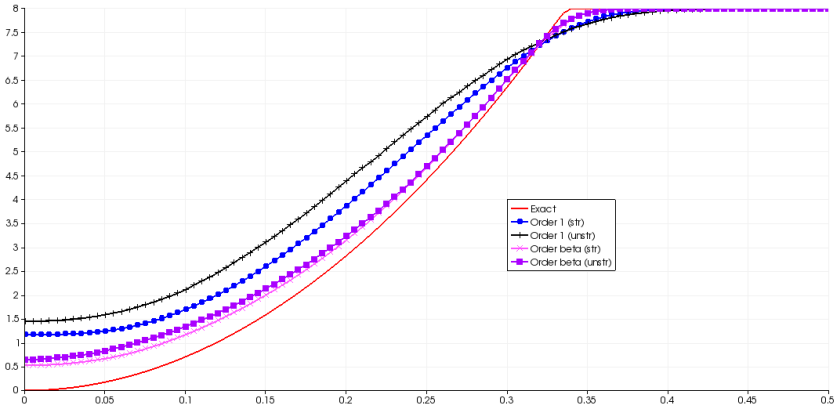
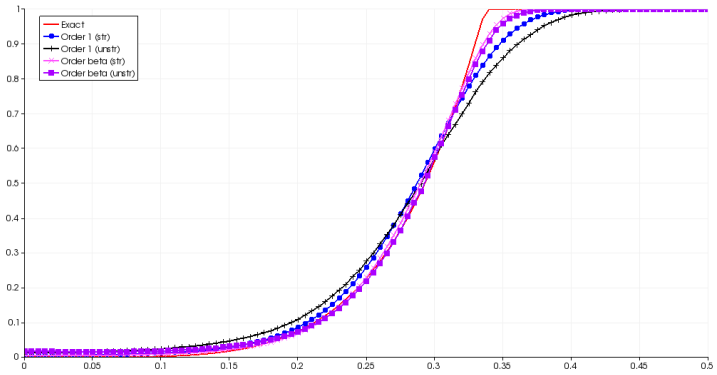
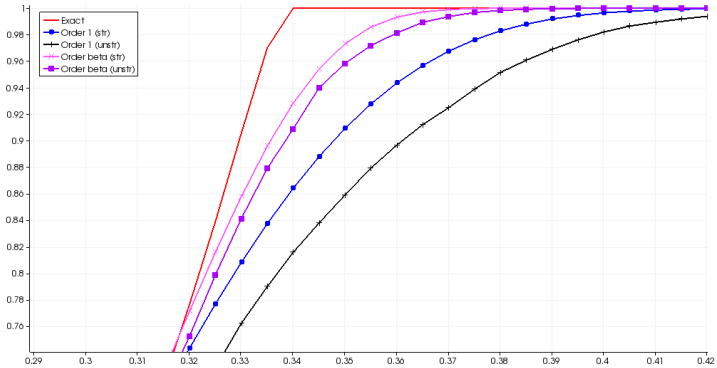


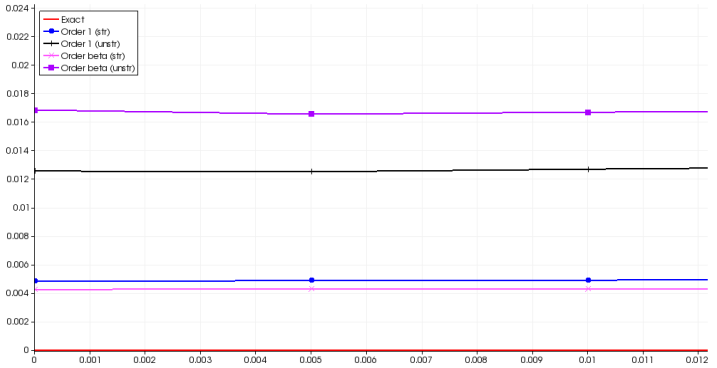
Figure 13: Energy plot at time 0.075



(a) Global view



(b) Magnification: right singularity



(c) Magnification: close-to-vacuum area

Figure 14: Density plot at time 0.075

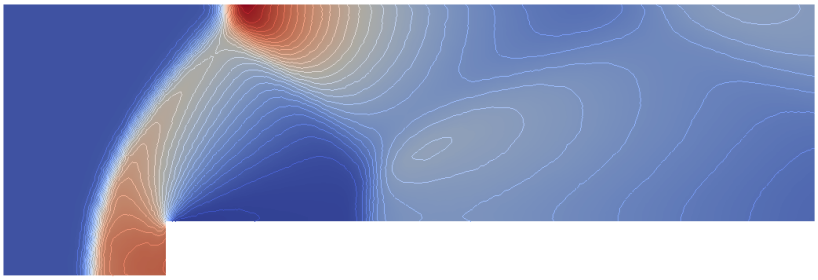
0.2m). Results are classicaly given at time 4 although the steady state is not yet reached. No exact solution is known but according to the extensive number of results in the literature ([6, 4, 34] for instance) the solution may consist of a front wave before the step and then two reflections. From a numerical point of view, two main difficulties are about the handling of the singularity (the edge of the step) and the location of reflections.

Two unstructured meshes are considered (resp. made of 20291 – Mesh 1 – and 80441 nodes – Mesh 2). Using the 1st-order scheme (Fig. 15a) does not enable to locate accurately the first reflection (the second one is not even detectable) and the front wave is blurred. The 2nd-order scheme provides better results (Fig. 15b). Indeed, we recover the two reflections and the front wave is clearer. A striking point is that in this simulation, β_i^n is constantly equal to 1 which means that the pure 2nd-order MUSCL scheme is able to tackle this problem. Decreasing the value of η_i^* (which is the scalar parameter in § 5 and set to $1/3$ in all computations) to 0.1 leads to the activation of β_i^n in the vicinity of the front wave and close to the corner. This remark goes to show that the choice $\eta_i^* = 1/3$ not only improves the computational time, but also preserves accuracy. Furthermore, we see on Fig. 15c that the addition of numerical tools (coefficient β_i^n in the reconstruction, modification of the CFL condition) does not damage the convergence of the scheme: when refining the mesh, we obtain more accurate results with clearer waves.

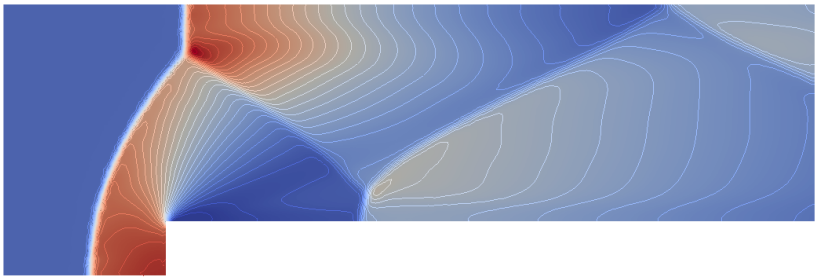
7. Conclusion

We carried out a theoretical study of whether MUSCL schemes for the Euler equations ensure positivity of density and pressure. This is a generalization of seminal works from Perthame & Shu and Berthon. On the one hand, we classicaly rewrote a standard MUSCL scheme as a convex combination of 1st-order 1D schemes and we introduced abstract coefficients allowing for an optimization of the CFL condition (which is thus explicit). This process assumes properties of the numerical flux which are satisfied by many classical schemes but not Roe scheme. It heavily relies on the 2D geometry (and could be extended to 3D) since it consists of one more step compared to dimension 1. On the other hand, to take advantage of an additional state as detailled in Berthon [4], we compute a damping coefficient β_i^n whose role consists in maintaining updated values in the set of physically admissible states. We gave simple and practical directions to modify existing codes in order to guarantee the robustness of the scheme.

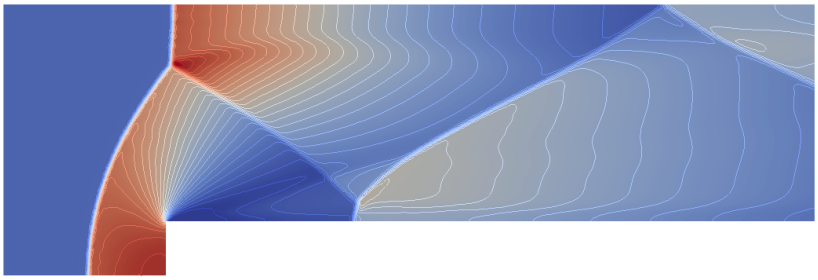
Choices have been made to tune the parameters we introduced in this study in order to find out a balance between accuracy and computational costs. Although we favoured the latter one through the optimization of the CFL condition, our choice turned out to be relevant about accuracy as well. To be more precise about the CFL condition, it is practically about 10 times more constraining than standard ones in accordance with a similar scalar study [7]. We should however bear in mind that our results come from sufficient conditions. Simulations have been performed with the Rusanov flux which is known to be



(a) Order 1 on Mesh 1 (20291 nodes, 39964 triangles)



(b) Order β on Mesh 1



(c) Order β on Mesh 2 (80441 nodes, 159628 triangles)

Figure 15: Density plot at time 4.0: 30 contours from 0.7 to 6.5

diffusive. As our approach can adapt to any scheme, we may obtain more striking results with less diffusive fluxes. Numerical simulations also showed that the coefficient β_i^n is particularly useful when the solution is close to the boundary of the physical set together with steep gradients.

This study was devoted to the Euler equations for ideal gas. A first extension would include the handling of other equations of state. This would only modify the computation of β_i^n. Future work might also deal with other physical models. The process first requires to identify the physical constraints (provided the resulting set is convex) and then to adapt the computation of β_i^n and μ_i^{opt}.

Appendix

Lemma 1

Let $\boldsymbol{\vartheta} \in \mathbb{R}^d$ with positive components, $\boldsymbol{\vartheta} \neq \mathbf{0}$. The optimization problem:

$$\mathcal{X}_{\vartheta} := \max_{\boldsymbol{\chi} \in \mathcal{E}_{\vartheta}} \min_{1 \leq j \leq d} \chi_j$$

with

$$\mathcal{E}_{\vartheta} = \{\boldsymbol{\chi} \in \mathbb{R}^d,\, \chi_j \geq 0,\, \boldsymbol{\chi} \cdot \boldsymbol{\vartheta} = 1\}$$

is solvable with $\mathcal{X}_{\vartheta} = |\boldsymbol{\vartheta}|_1^{-1}$ and a maximizer is given by $\overline{\chi}_j = |\boldsymbol{\vartheta}|_1^{-1}$ for all j.

Proof. For any $\boldsymbol{\chi} \in \mathcal{E}_{\vartheta}$, we observe that:

$$1 = \boldsymbol{\chi} \cdot \boldsymbol{\vartheta} \geq |\boldsymbol{\vartheta}|_1 \min_{1 \leq j \leq d} \chi_j.$$

Thus $\mathcal{X}_{\vartheta} \leq |\boldsymbol{\vartheta}|_1^{-1}$. Let $\overline{\boldsymbol{\chi}}$ be defined by $\overline{\chi}_j = |\boldsymbol{\vartheta}|_1^{-1}$. The vector $\overline{\boldsymbol{\chi}}$ obviously satisfies $\overline{\boldsymbol{\chi}} \in \mathcal{E}_{\vartheta}$ and $\min_{1 \leq j \leq d} \overline{\chi}_j = |\boldsymbol{\vartheta}|_1^{-1}$. Hence we conclude that $\mathcal{X}_{\vartheta} = |\boldsymbol{\vartheta}|_1^{-1}$ and the maximum is reached fo the constant vector $\overline{\boldsymbol{\chi}}$. ■

Acknowledgements

The authors are grateful to the referees for their helpful comments and suggestions which made the reading of technical parts more pleasant.

References

[1] S. Aubert, S. Benzoni & J.-F. Coulombel, *Boundary conditions for Euler equations*, **AIAA J.**, 41(1), 56–63, (2003).

[2] T.J. Barth & M. Ohlberger, *Finite volume methods: foundation and analysis*, **Encyclopedia of Computational Mechanics**, John Wiley & Sons Ltd, (2004).

[3] C. Berthon, *Stability of the MUSCL schemes for the Euler equations*, **Comm. Math. Sciences**, 3(2), 133–157 (2005).

[4] C. Berthon, *Robustness of MUSCL schemes for 2D unstructured meshes*, **J. Comput. Phys.**, 218(2), 495–509 (2006).

[5] F. Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*, **Birkhauser**, (2004).

[6] Ch.-H. Bruneau & P. Rasetarinera, *A finite volume method with efficient limiters for solving conservation laws*, **J. Comput. Fluid Dyn.**, 6(1), 23–38 (1997).

[7] C. Calgaro, E. Chane-Kane, E. Creusé & T. Goudon, *L^∞-stability of vertex-based MUSCL finite volume schemes on unstructured grids; simulation of incompressible flows with high density ratios*, **J. Comput. Phys.**, 229(17), 6027–6046 (2010).

[8] S. Clain & V. Clauzon, *L^∞-stability of the MUSCL methods*, **Numer. Math.**, 116(1), 31–64 (2010).

[9] S. Clain, D. Rochette & R. Touzani, *A multislope MUSCL method on unstructured meshes applied to compressible Euler equations for axisymmetric swirling flows*, **J. Comput. Phys.**, 229(13), 4884–4906 (2010).

[10] P.-H. Cournède, C. Debiez, A. Dervieux, *A positive MUSCL scheme for triangulations*, **INRIA Tech. Report**, 3465, (1998).

[11] B. Einfeldt, *On Godunov-type methods for gas dynamics*, **SIAM J. Numer. Anal.**, 25(2), 294–318 (1988).

[12] B. Einfeldt, C.D. Munz, P.L. Roe & B. Sjogreen, *On Godunov-type methods near low densities*, **J. Comput. Phys.**, 92(2), 273–295 (1991).

[13] J.L. Estivalezes & P. Villedieu, *High-order positivity-preserving kinetic schemes for the compressible Euler equations*, **SIAM J. Numer. Anal.**, 33(5), 2050–2067 (1996).

[14] R. Eymard, T. Gallouet & R. Herbin, *Finite volume methods*, **Handbook of Numerical Analysis**, 7, 713–1018 (2000).

[15] E. Godlewski & P.-A. Raviart, *Hyperbolic systems of conservation laws*, **Ellipses Math. Appl.**, 3/4 (1991).

[16] E. Godlewski & P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, Applied Mathematical Sciences, 118, **Springer-Verlag** (1996).

[17] J.B. Goodman & R.J. LeVeque, *On the accuracy of stable schemes for 2 D scalar conservation laws*, **Math. Comput.**, 45(171), 15–21 (1985).

[18] B. Khobalatte & B. Perthame, *Maximum-principle on the entropy and second-order kinetic schemes*, **Math. Comput.**, 62(205), 119–132 (1994).

[19] R.J. LeVeque, *Finite volume methods for hyperbolic problems*, **Cambridge Univ. Press**, 31, (2002).

[20] T. Linde & P.L. Roe, *Robust Euler codes*, **13th AIAA CFD Conf.**, AIAA-97-2098, (1997).

[21] D. Mavriplis, *Unstructured Mesh Discretizations and Solvers for Computational Aerodynamics*, **18th AIAA CFD Conf.**, AIAA-2007-3955, (2007).

[22] B. Parent, *Positivity-preserving flux-limited method for compressible fluid flow*, **Comput. Fluids**, 44, 238–247, (2011).

[23] B. Parent, *Positivity-preserving high-resolution schemes for systems of conservation laws*, **J. Comput. Phys.**, 231(1), 173–189, (2012).

[24] M. Pelanti, L. Quartapelle & L. Vivegano, *A review of entropy fixes as applied to Roe's linearization*, private communication.

[25] B. Perthame & C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, **Numer. Math.**, 73, 119–130 (1996).

[26] S. Piperno & S. Depeyre, *Criteria for the design of limiters yielding efficient high resolution TVD schemes*, **Comput. Fluids**, 27(2), 183–197 (1998).

[27] D. Serre, *Systems of conservation laws. 2. Geometric structures, oscillations, and initial-boundary value problems*, **Cambridge University Press**, (2000).

[28] J. R. Shewchuk, *Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator*, **Lect. Notes Comput. Sc.**, 1148, 203–222 (1996).

[29] J. Smoller, *Shock waves and reaction-diffusion equations*, **Springer Verlag**, (1994).

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

[31] P. Sweby, *High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws*, **SIAM J. Numer. Anal.**, 21(5), 995–1011 (1984).

[32] E. Toro, *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*, **Springer Verlag** (2009).

[33] B. Van Leer, *Toward the ultimate conservative difference scheme V: A second order sequel of Godunov's methods*, **J. Comput. Phys.**, 32, 101–136 (1979).

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

[35] X. Zhang & C.-W. Shu, *On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, **J. Comput. Phys.**, 229, 8918–8934 (2010).

[36] X. Zhang & C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order schemes for conservation laws: Survey and new developments*, **Proc. R. Soc. A.**, 467(2134), 2752–2776 (2011).

[37] X. Zhang, Y. Xia & C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, **J. Sci. Comput.**, DOI:10.1007/s10915-011-9472-8 (2011).