Staggered Residual Distribution scheme for compressible flow

R. Abgrall and K. Ivanova

Institute of Mathematics, University of Zürich

Wintherhurrestrasse 190, CH 8057 Zürich, Switzerland

November 23, 2021

Abstract

This paper is focused on the approximation of the Euler equation of compressible fluid dynamics on a staggered mesh. To this aim, the flow parameter are described by the velocity, the density and the internal energy. The thermodynamic quantities are described on the elements of the mesh, and this the approximation is only L^2, while the kinematic quantities are globally continuous. The method is general in the sense that the thermodynamical and kinetic parameters are described by arbitrary degree polynomials, in practice the difference between the degrees of the kinematic parameters and the thermodynamical ones is equal to 1. The integration in time is done using a defect correction method. As such, there is no hope that the limit solution, if it exists, will be a weak solution of the problem. In order to guaranty this property, we introduce a general correction method in the spirit of the Lagrangian stagered method described in [1, 2, 3], and we prove a Lax Wendroff theorem. The proof is valid for multidimensional version of the scheme, though all the numerical illustrations, on classical benchmark problems, are all one dimensional because we have an easy access to the exact solution for comparison. We conclude by explaning that the method is general and can be used in a different setting as the specific one used here, for example finite volume, of discontinuous Galerkin methods.

1 Introduction

The Euler equations of fluid dynamics are, in their conservative version,

$$
\begin{cases}
\dfrac{\partial \rho}{\partial t} + \mathrm{div}(\rho \mathbf{u}) = \mathbf{0}, \\[2ex]
\dfrac{\partial \rho \mathbf{u}}{\partial t} + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u} + \mathrm{p}\mathbf{I}) = \mathbf{0}, \\[2ex]
\dfrac{\partial E}{\partial t} + \mathrm{div}((E + \mathrm{p})\mathbf{u}) = \mathbf{0}.
\end{cases}
\tag{1}
$$

As usual, $\rho \geq 0$ is the density, $\mathbf{u}$ is the velocity, $E = e + \frac{1}{2}\rho \mathbf{u}^2$ is the total energy, $e \geq 0$ is the internal energy, and p is the pressure. The system is closed by an equation of state, $p = p(\rho, e)$. The simplest is that of calorically perfect gas where

$$
p = \frac{e}{\gamma - 1}
$$

where the ratio of specific heats γ is a constant.

arXiv:2111.10647v1 [math.NA] 20 Nov 2021

When the solution is smooth, the system (1) can be equivalently written in nonconservative form as:

$$
\begin{cases}
\dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = \mathbf{0}, \\[2mm]
\dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \dfrac{\nabla p}{\rho} = \mathbf{0}, \\[2mm]
\dfrac{\partial e}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{e} + (e + p)\operatorname{div}\mathbf{u} = \mathbf{0}.
\end{cases}
\tag{2}
$$

When the solution is not smooth, the form (2) is meaningless because the differential operators are no longer defined. This is why the form (1) is preferred, in particular in its weak form, see [4]. This fact has a very strong implication for the design of numerical schemes applied to (1): the Lax-Wendroff theorem implies and guaranties that a suitable numerical approximation should be written in term of flux.

However, the form (2) is better suited for engineering purposes, since one has a direct access to the velocity and the internal energy. Hence a rather natural question is to ask how to discretise the Euler equations directly from (2), and still have convergence to the correct weak solutions, at least formally.

One obvious way to achieve this, is to start from a conservation approximation of (1), and by simple algebraic manipulations which amount to multiply the numerical scheme by approximations of

$$
\begin{pmatrix}
1 & 0 & 0 \\
\mathbf{u} & \rho & 0 \\
\dfrac{\mathbf{u}^2}{2} & \rho\mathbf{u} & 1
\end{pmatrix}.
\tag{3}
$$

lead to a scheme directly working on the primitive variable. This "new" scheme is equivalent to the original one.

This is not exactly the question we want to address here. We are interested in designing locally conservative approximations of (2) for which the thermodynamic variables are approximated in L^2 while the velocity is globally continuous. This can be seen as an Eulerian version of the Lagrangian schemes designed in [1, 2] or [3] and the related works by these authors. A similar question has been addressed by Herbin and co-authors, see for example [5, 6, 7] in the finite volume context. In this references, the authors describe a class of numerical schemes where the thermodynamic variables and the velocity are piecewise constant but logically described on a staggered mesh. They show the convergence towards the weak solution. The scheme can be partially implicit, so that in the low Mach number limit the scheme "degenerates" to a Mac-type scheme, see [8]. The schemes are second order in time and space.

In this paper, we describe a different technique which allows to reach an arbitrary order of accuracy, both in time and space. Before describing this method, we shortly review the class of residual distribution schemes that will be the main tool we use, thanks to some reinterpretation. Then we describe the scheme, and explain why it is locally conservative. Last, we show a variant of the Lax-Wendroff theorem that is adapted to our setting.

2   A high-order nonconservative approach

We have in mind a numerical approximations where the variables are piece-wise polynomial in simplex. We also assume that the velocity is globally continuous, in contrast to Discontinuous Galerkin(DG)–like approximations. This constraint is motivated by the *choice* that we want to extend the technique of [1], where a Petrov Galerkin technique, inspired from [9] and the reference therein. If nothing special is done, we need to invert a mass matrix. This can be cumbersome, and even impossible if we want to extend the techniques of [10] because the equivalent of the mass matrix changes at every time step. This is why we use a particular time stepping that we describe now. It relies on series of Euler forward type of discretisation. This is indeed **the** essential point: if one prefers to forget the globally continuous methods, rely on a DG–like

approach, and use a Strong Stability Preserving (SSP) Runge-Kutta approach, one can extend our correction technique and build schemes that converge to a weak solution of the problem, starting from (2).

2.1   Time stepping approach

Consider a hyperbolic system in the form

$$\frac{\partial U}{\partial t} + \mathrm{L(U)} = 0. \tag{4}$$

For time discretization, we want to get a high order accurate approximation. To do so, we will use the Deferred Correction (DeC) approach. The aim of DeC schemes is to avoid implicit methods, without losing the high order of accuracy of a scheme. The high order method that we want to approximate will be denoted by $\mathcal{L}^2$. To use the DeC procedure, we also need another method, which is easy and fast to be solved with low order of accuracy $\mathcal{L}^1$. The DeC algorithm is providing an iterative procedure that wants to approximate the solution of the $\mathcal{L}^2$ scheme U^* in the following way:

$$\mathcal{L}^1\left(U^{(1)}\right) = 0 \tag{5}$$

$$\mathcal{L}^1\left(U^{(k)}\right) = \mathcal{L}^1\left(U^{(k-1)}\right) - \mathcal{L}^2\left(U^{(k-1)}\right), \quad \text{with} \quad k = 2, .., K, \tag{6}$$

where K is the number of iterations that we compute. We need as many iterations as the order of accuracy that we want to reach. We know from [11]:

Proposition 2.1. *Let $\mathcal{L}^1$ and $\mathcal{L}^2$ be two operators defined on R^m, which depend on the discretization scale $\Delta \sim \Delta x \sim \Delta t$, such that*

- *$\mathcal{L}^1$ is coercive with respect to a norm, i.e., $\exists \alpha_1 > 0$ independent of Δ, such that for any U, V we have that*

$$\alpha_1 \big|\big| U - V \big|\big| \leq \big|\big| \mathcal{L}^1(U) - \mathcal{L}^1(V) \big|\big|,$$

- *$\mathcal{L}^1 - \mathcal{L}^2$ is Lipschitz with constant $\alpha_2 > 0$ uniformly with respect to Δ, i.e., for any U, V*

$$\big|\big| \left(\mathcal{L}^1(U) - \mathcal{L}^2(U)\right) - \left(\mathcal{L}^1(V) - \mathcal{L}^2(V)\right) \big|\big| \leq \alpha_2 \Delta \big|\big| U - V \big|\big|.$$

We also assume that there exists a unique U_Δ^* such that $\mathcal{L}^2(U_\Delta^*) = 0$. Then, if $\eta := \dfrac{\alpha_2}{\alpha_1}\Delta < 1$, the DeC is converging to U_Δ^* and after k iterations the error $\big|\big| U^{(k)} - U_\Delta^* \big|\big|$ is smaller than $\eta^k \big|\big| U^{(0)} - U_\Delta^* \big|\big|$.

Let us show how to use this for solving the problem (1) or (2) on domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. For now we forget the question of local conservation and

$$L(U) = \begin{pmatrix} \mathrm{div}(\rho\mathbf{u}) \\ \mathrm{div}(\rho\mathbf{u} \otimes \mathbf{u} + \mathrm{pI}) \\ \mathrm{div}((\mathrm{E} + \mathrm{p})\mathbf{u}) \end{pmatrix}$$

for the conservative form and we define

$$L(U) = \begin{pmatrix} \mathrm{div}(\rho\mathbf{u}) \\ (\mathbf{u} \cdot \nabla)\mathbf{u} + \dfrac{\nabla p}{\rho} \\ \mathbf{u} \cdot \nabla e + (e + p)\mathrm{div}\mathbf{u} \end{pmatrix}$$

for the non–conservative form. We will describe the procedure in two steps. First, we consider the case of a scalar problem, and then we look at (1) or (2). The reason is that in (2) not all the variables play the same role, contrarily to (1), and it is easier to start with a system with one variable. We will proceed by forgetting the question of local conservation.

3

2.2 Scalar case

2.2.1 Trial space

We consider a triangulation of Ω made of non overlapping simplex that are generically denoted by K. We assume the triangulation to be conformal, and define

$$V^h = \{v \in L^2(\Omega) \text{ such that for any } K, v_{|K} \in \mathbb{P}^k(\mathbb{R}^d)\} \subset L^2(\Omega)$$

where as usual, $\mathbb{P}^k(\mathbb{R}^d)$ is the set of polynomials in $\mathbb{R}^d$ of degree less or equal to k. We also define

$$W^h = V^h \cap C^0(\Omega).$$

In each element K, a polynomial is defined by a set of degrees of freedom, for example the Lagrange points. We denote by σ a generic degree of freedom. Here, for reasons that will be more clear later, we expand the polynomials in term of Bézier polynomials.

- One dimensional elements: In the element $K = [x_i, x_{i+1}]$, we consider the barycentric coordinates

$$\lambda_1(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \lambda_2(x) = \frac{x - x_i}{x_{i+1} - x_i} = 1 - \lambda_1.$$

If $\sigma \in K$, the restriction of B_σ is defined in the element as follows, if $\sigma \notin K$, then the Bézier form vanishes. We describe the two families of Bézier form we will need:

  - Linear: The degrees of freedom are the vertices, so

$$\varphi_i^{(1)} = \lambda_1, \qquad \varphi_{i+1}^{(1)} = \lambda_2, \text{ and } \sigma = x_i \text{ or } x_{i+1} \text{ here.}$$

  - Quadratic: the degrees of freedom σ are identified to the vertices i, and the mid-points $i + 1/2$.

$$\varphi_\sigma^{(2)}(x) = \begin{cases} \lambda_1^2 & \text{if } \sigma = x_i \\ 2\lambda_1\lambda_2 & \text{if } \sigma = x_{i+1/2} \\ \lambda_2^2 & \text{if } \sigma = x_{i+1}. \end{cases}$$

- Multidimensional elements: we only describe the 2D cases, with triangles, but similar things are obtained for quadrangles, or 3D simplex. A triangle is made of 3 vertices denoted by 1, 2 and 3. The barycentric coordinates with respect to the vertices $1, 2, 3$ are denotes by Λ_1, Λ_2 and Λ_3.

  - Linear: the degrees of freedom are the vertices and $\varphi_{\sigma_i} = \Lambda_i$ and $i = 1, 2, 3$.
  - Quadratic: The degrees of freedom are the 3 vertices σ_1, σ_2 and σ_3, and the midpoint of the edges:

$$\sigma_4 = \frac{\sigma_1 + \sigma_2}{2}, \sigma_5 = \frac{\sigma_2 + \sigma_3}{2}, \sigma_6 = \frac{\sigma_2 + \sigma_1}{2}.$$

  The Bézier polynomials are:

$$\varphi_{\sigma_i} = \Lambda_i^2 \text{ for } i = 1, 2, 3, \quad \varphi_{\sigma_4} = 2\Lambda_1\Lambda_2, \quad \varphi_{\sigma_5} = 2\Lambda_3\Lambda_2, \quad \varphi_{\sigma_6} = 2\Lambda_1\Lambda_3.$$

Then, considering $u \in V_h(\Omega)$ or $u \in W_h(\Omega)$, for any K, we expand $u_{|K}$ as

$$u_{|K} = \sum_{\sigma \in K} u_\sigma \varphi_\sigma^K$$

where φ_σ^K is any of the linear, quadratic (or more) function defined above. If $u \in V_h(\Omega)$ then we have the expansion

$$u = \sum_K \sum_{\sigma \in K} u_\sigma^K \varphi_\sigma^K,$$

and if $u \in W_h(\Omega)$, we can expand u as

$$u = \sum_\sigma u_\sigma \varphi_\sigma.$$

With some abuse of language, we will use the second notation throughout this paper, depending if we see φ_σ per element or more globally.

2.2.2 Test space

As we said earlier, we rely on a Petrov-Galerkin approach. This means that the test functions will belong to a finite dimensional subspace $X_h(\Omega)$ of $L^2(\Omega)$ that can also be described by the degrees of freedom σ: we can identify functions of $X_h(\Omega)$ that are indexed by the σ and span this space. We denote them by Ξ_σ. For example, in the SUPG method, we define Ξ_σ in each K by: for $\mathbf{x} \in K$,

$$\Xi_\sigma(\mathbf{x}) = \varphi_\sigma(\mathbf{x}) + h_K \left(\nabla_U L(U) \tau_K \right) \cdot \nabla \varphi_\sigma(\mathbf{x}).$$

Here h_K is the diameter of K and τ_K is a positive matrix. In [10, 12, 13, 14, 15], examples are given where Ξ_σ depends on the solution, in order to get L^∞ stability. In all the examples we are considering, the support of Ξ_σ is that of φ_σ.

2.2.3 Description of the time discretisation

The first step is to define $\mathcal{L}^2$. Let's integrate the system (4):

$$\int_\Omega \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x} = 0, \quad \int_{t^n}^{t^{n+1}} \int_\Omega \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x} = 0.$$

Using the locality of the basis function, we have

$$\int_\Omega \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x} = \sum_{K, \sigma \in K} \int_K \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x},$$

and

$$\int_\Omega \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x} = \sum_{K, \sigma \in K} \int_K \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x},$$

so we will define the restriction of $\mathcal{L}^2$ on K. Then we introduce M sub-time steps between t_n and t_{n+1}: $t^0 = t_n < t^1 < \ldots < t^k < \ldots < t^M = t_{n+1}$ and for any degree of freedom σ the corresponding approximations $U_\sigma^{(l)}$ for $0 \le l \le M$. Of course, $U_\sigma^{(0)} = U_\sigma^n$ is the approximation at time t_n. This being done, we will consider Lagrange interpolation in time and will integrate exactly in time. More details can be found in [9]. We describe here two examples as illustrations.

In the first one, we take $M = 1$, so that there is no intermediate points, and we simply get the Crank Nicholson scheme:

$$\int_K \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(\mathrm{U}) \right) dt \, d\mathbf{x} = \int_K \Xi_\sigma \left(U^{n+1} - U^n \right) dx + \frac{\Delta t}{2} \left(\int_K \Xi_\sigma \mathrm{L}(\mathrm{U}^n) dx + \int_K \Xi_\sigma \mathrm{L}(\mathrm{U}^{n+1}) dx \right) \quad (7)$$

If we look for third order in time, we consider the mid point time step, $t_{n+1/2} = t_n + \frac{\Delta t}{2}$ and get

$$\int_K \int_{t^n}^{t^{n+1/2}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(U)\right) \mathrm{dt}\ \mathrm{dx} = \int_K \Xi_\sigma \left(U^{n+1/2} - U^n\right) dx + \Delta t \left(\frac{5}{24} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^n}) \mathrm{dx} + \frac{1}{3} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^{n+1/2}}) \mathrm{dx}\right.$$

$$\left. - \frac{1}{24} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^{n+1}}) \mathrm{dx}\right)$$

$$\int_K \int_{t^n}^{t^{n+1}} \Xi_\sigma \left(\frac{\partial U}{\partial t} + \mathrm{L}(U)\right) \mathrm{dt}\ \mathrm{dx} = \int_K \Xi_\sigma \left(U^{n+1} - U^n\right) dx + \Delta t \left(\frac{1}{6} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^n}) \mathrm{dx} + \frac{1}{3} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^{n+1/2}}) \mathrm{dx}\right.$$

$$\left. + \frac{1}{6} \int_K \Xi_\sigma \mathrm{L}(\mathrm{U^{n+1}}) \mathrm{dx}\right) \tag{8}$$

The definition of $\int_K \Xi_\sigma \mathrm{L}(\mathrm{u}) \mathrm{dx}$ is somewhat formal, we replace it by some approximation $\Phi_\sigma^K(u)$ to be defined later, the only constraint is that we have the relation

$$\sum_{\sigma \in K} \Phi_\sigma^K(U) = \int_K \widetilde{L(U)}\ d\mathbf{x}. \tag{9}$$

The precise definition of the term $\widetilde{L(U)}$ depends on wether $U \in V_h(\Omega)$ or $U \in W_h(U)$.

- If $U \in W_h$, then $\widetilde{L(U)} = L(U)$ and the integration is done via quadrature formula of order at least equal to the polynomial degree,

- If $U \in V_h(\Omega)$, then U may be discontinuous across the faces of K, we only consider a consistent approximation. If L is in conservation form, then we set

$$\int_K \widetilde{L(U)}\ d\mathbf{x} := \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}}\ d\gamma$$

where $\hat{\mathbf{f}}$ is a numerical flux. If instead, L is not in conservation form, for example

$$L(U) = \sum_{j=1}^d a_j(U) \frac{\partial U}{\partial x_j}$$

and one of the $a_j(U)$ is not a jacobian, we simply set

$$\int_K \widetilde{L(U)}\ d\mathbf{x} := \int_K \left(\sum_{j=1}^d a_j(U) \frac{\partial U}{\partial x_j}\right) d\mathbf{x}$$

which is evaluated by means of quadrature formula.

We set, for (7), $U(t_n) \approx U_0$, $U(t_{n+1}) \approx U_1$ and define $\mathcal{L}^2$ as a vector indexed by the degrees of freedom, and the components are:

$$\mathcal{L}_\sigma^2(U_1, U_0) = \sum_{K, \sigma \in K} \Bigg(\underbrace{\int_K \Xi_\sigma \left(U_1 - U_0\right) dx + \frac{\Delta t}{2} \left(\Phi_\sigma^K(U_0) + \Phi_\sigma^K(U_1)\right)}_{L_{\sigma, K}^2(U1, U2)} \Bigg) \tag{10a}$$

and for (8), $U(t_n) \approx U_0, U(t_{n+1/2}) \approx U_1,\ U(t_{n+1}) \approx U_2$ and define $\mathcal{L}_\sigma^2$ by

$$
\mathcal{L}^2(U_1, U_2, U_0) = \begin{cases} \displaystyle\sum_{K, \sigma \in K} \overbrace{\left(\int_K \Xi_\sigma (U_1 - U_0) dx + \Delta t \left(\frac{5}{24} \Phi_\sigma^K(U_0) + \frac{1}{3} \Phi_\sigma^K(U_1) - \frac{1}{24} \Phi_\sigma^K(U_2) \right) \right)}^{\mathcal{L}_{\sigma,K}^{2,I}} \\ \displaystyle\sum_{K, \sigma \in K} \underbrace{\left(\int_K \Xi_\sigma (U_2 - U_0) dx + \Delta t \left(\frac{1}{6} \Phi_\sigma^K(U_0) + \frac{1}{3} \Phi_\sigma^K(U_1) + \frac{1}{6} \Phi_\sigma^K(U_2) \right) \right)}_{\mathcal{L}_{\sigma,K}^{2,II}} \end{cases} \tag{10b}
$$

So again we have a "vector" of "components" $(\mathcal{L}_{\sigma,K}^{2,I}, \mathcal{L}_{\sigma,K}^{2,II})^T$. The operator $\mathcal{L}^1$ is defined by replacing the temporal terms by

$$
\int_K \varphi(U^{n+\xi} - U^n)\, d\mathbf{x}
$$

and then by "lumping" the mass matrix: this is why we use Bézier approximation since we are sure that the lumped mass is non zero because it is

$$
C_\sigma = \int_\Omega \varphi_\sigma\, d\mathbf{x} > 0.
$$

We set $\int_K \varphi_\sigma dx = C_K$ (it does not depend on σ) and set, for (7)

$$
\mathcal{L}_{\sigma,K}^1(U_1, U_0) = C_K(U_{1,\sigma} - U_{0,\sigma}) + \Delta t \sum_{K, \sigma \in K} \Phi_\sigma^K(U_0) = 0 \tag{11a}
$$

and similarly, for (8),

$$
\mathcal{L}_{\sigma,K}^1(U_1, U_2, U_0) = \begin{pmatrix} C_K(U_{1,\sigma} - U_{0,\sigma}) & + \dfrac{\Delta t}{2} \displaystyle\sum_{K, \sigma \in K} \Phi_\sigma^K(U_0) = 0 \\ C_\sigma^K(U_{2,\sigma} - U_{0,\sigma}) & + \Delta t \displaystyle\sum_{K, \sigma \in K} \Phi_\sigma^K(U_0) = 0 \end{pmatrix} \tag{11b}
$$

The generalisation to formally higher order temporal schemes is obvious. Last, one can show that the two operators satisfies the requirements of the lemma 2.1.

The DeC iteration is then defined by

$$
\mathcal{L}^1(U^{(k+1)}) = \mathcal{L}^1(U^{(k)}) - \mathcal{L}^2(U^{(k)}).
$$

After simplifications, this gives the following iteration: knowing $U^{(k)}$, we get $U^{(k+1)}$ by, for all σ, we have for (7),

$$
C_\sigma(U_{1,\sigma}^{(k+1)} - U_{1,\sigma}^{(k)}) + \sum_{K, \sigma \in K} \left[\int_K \Xi_\sigma(U_1^{(k)} - U_0) dx + \frac{\Delta t}{2} \left(\Phi_\sigma^K(U_0) + \Phi_\sigma^K(U_1^{(k)}) \right) \right] \tag{12}
$$

and for (8),

$$
\begin{aligned}
C_\sigma(U_{1,\sigma}^{(k+1)} - U_{1,\sigma}^{(k)}) + \sum_{K, \sigma \in K} &\left[\int_K \Xi_\sigma(U_1^{(k)} - U_0) dx \right.\\
&\left. + \Delta t \left(\frac{5}{24} \Phi_\sigma^K(U_0) + \frac{1}{3} \Phi_\sigma^K(U_1^{(k)}) - \frac{1}{24} \Phi_\sigma^K(U_2^{(k)}) \right) \right] \\
C_\sigma(U_{2,\sigma}^{(k+1)} - U_{2,\sigma}^{(k)}) + \sum_{K, \sigma \in K} &\left[\int_K \Xi_\sigma(U_2^{(k)} - U_0) dx \right.\\
&\left. + \Delta t \left(\frac{1}{6} \Phi_\sigma^K(U_0) + \frac{1}{3} \Phi_\sigma^K(U_1^{(k)}) + \frac{1}{6} \Phi_\sigma^K(U_2^{(k)}) \right) \right]
\end{aligned} \tag{13}
$$

For now, this is all we need.

2.3 Case of system (2).

We describe the residuals, and develop the method first for the case of the second order in time, since it is simpler. The generalisation to higher order is straightforward and done in the paragraph that follows.

We assume that the computational domain Ω is covered by non-overlapping simplices $\{K_j\}_{j\in\mathcal{T}}$. The velocity field $\mathbf{u}$ belongs to a kinematic space $\mathcal{V}$ of finite dimension; it has a basis denoted by $\{\psi_{\sigma_\mathcal{V}}\}_{\sigma_\mathcal{V}\in D_\mathcal{V}}$, where $D_\mathcal{V}$ is the set of kinematic degrees of freedom with the total degrees of freedom given by $\#D_\mathcal{V} = N_\mathcal{V}$. The thermodynamic quantities such as the internal energy, the density and the pressure belong to a thermodynamic space $\mathcal{E}$; this space is also finite dimensional and its basis is $\{\varphi_{\sigma_\mathcal{E}}\}_{\sigma_\mathcal{E}\in D_\mathcal{E}}$. The set $D_\mathcal{E}$ is the set of thermodynamical degrees of freedom with the total degrees of freedom $\#D_\mathcal{E} = N_\mathcal{E}$. The kinematic space $\mathcal{V}$ is formed by the quadratic Bernstein elements, while the thermodynamic space $\mathcal{E}$ has a piecewise-linear basis. The velocity field is approximated by

$$\mathbf{u}(\mathbf{x},t) = \sum_{\sigma_\mathcal{V}\in D_\mathcal{V}} \mathbf{u}_{\sigma_\mathcal{V}}(t)\psi_{\sigma_\mathcal{V}}(\mathbf{x})$$

where the $\psi_{\sigma_\mathcal{V}}$ are the linear/quadratic Bézier, and the density, the pressure and the internal energy, are given by

$$\rho(\mathbf{x},t) = \sum_{\sigma_\mathcal{E}\in D_\mathcal{E}} \rho_{\sigma_\mathcal{E}}(t)\varphi_{\sigma_\mathcal{E}}(\mathbf{x}), \quad p(\mathbf{x},t) = \sum_{\sigma_\mathcal{E}\in D_\mathcal{E}} p_{\sigma_\mathcal{E}}(t)\varphi_{\sigma_\mathcal{E}}(\mathbf{x}), \quad e(\mathbf{x},t) = \sum_{\sigma_\mathcal{E}\in D_\mathcal{E}} e_{\sigma_\mathcal{E}}(t)\varphi_{\sigma_\mathcal{E}}(\mathbf{x}).$$

where the $\varphi_{\sigma_\mathcal{E}}$ are the piecewise constant/linear per elements functions. Note that the degrees of freedom for the velocity are assumed to be globally continuous, so in W_h^d, while the thermodynamical ones are discontinuous across the boundary of elements, so in V_h^2.

We can rewrite the Euler equations (2) in the following form

$$\frac{\partial U}{\partial t} + A \cdot \nabla U = 0.$$

The only thing to do is to describe how the method of the previous section adapt to this case, and this amounts to describing the general structure residuals, that is how (9) is written. Since the velocity is globaly continuous, we write

$$\int_K \widetilde{L(U)} \, d\mathbf{x} = \int_K \left(\mathbf{u}\otimes\mathbf{u} + \frac{\nabla p}{\rho}\right) d\mathbf{x}$$

where $\mathbf{u}\in W_h$ and $p,\rho\in V_h$. Since we are on K, these are simply polynomials, and the integration is carried out by numerical quadrature.

For the density, $L(U) = \mathrm{div}\,(\rho\mathbf{u})$ is in conservation form and we use a numerical flux $\hat{\mathbf{f}}$:

$$\int_K \widetilde{L(U)} \, d\mathbf{x} = \int_{\partial K} \hat{\mathbf{f}}_\mathbf{n} \, d\gamma.$$

Last, for the internal energy, we write

$$\int_K \widetilde{L(U)} \, d\mathbf{x} = \int_K \left(\mathbf{u}\cdot\nabla e + (e+p)\,\mathrm{div}\,\mathbf{u}\right) d\mathbf{x}$$

and again we use quadrature formula.

In the numerical section, we will describe the residuals that we use.

3 A discussion on conservation

3.1 A set of sufficient conditions to achieve convergence to a weak solution

Again, to simplify the notations, we focus on the second order case, but the extension to the more general case is straightforward. In the appendix 6, we show a Lax Wendroff theorem for this type of discretisation.

What we do here is to show how to go from the system in non–conservative form to the one in conservation form.

There is nothing to do for the density because it is already in conservation form and the standard proof [4] applies. There is no need to repeat it here since the proof we give for the momentum and the total energy, modulo some complications, is essentially the same.

Let us first look at the momentum. Considering a test function $\psi \in C_0^1(\mathbb{R}^d \times \mathbb{R})$, we define ψ_K^n the value of ψ at time t_n at the centroid of K, and consider the following approximation of ψ that we still denote by ψ:

$$\psi(\mathbf{x},t) = \sum_K \psi_K^n 1_K \qquad \text{for } t \in [t_n, t_{n+1}[.$$

Then we consider

$$\int_{\mathbb{R}^d} \psi(\mathbf{x},t) \left(\rho^{(p+1)}\mathbf{u}^{(p+1)} - \rho^{(p)}\mathbf{u}^{(p)}\right) d\mathbf{x} = \sum_K \psi_K^n \int_K \left(\rho^{(p+1)}\mathbf{u}^{(p+1)} - \rho^{(p)}\mathbf{u}^{(p)}\right) d\mathbf{x}$$
$$= \sum_K \psi_K^n \left[\int_K \rho^{(p+1)}\left(\mathbf{u}^{(p+1)} - \mathbf{u}^{(p)}\right) d\mathbf{x} + \int_K \mathbf{u}^{(p)}\left(\rho^{(p+1)} - \rho^{(p)}\right) d\mathbf{x} \right]. \tag{14}$$

Then, we can write, introducing $\Delta\mathbf{u}_{\sigma_\mathcal{V}} =: \mathbf{u}_{\sigma_\mathcal{V}}^{(p+1)} - \mathbf{u}_{\sigma_\mathcal{V}}^{(p)}$ and $\Delta\rho_{\sigma_\mathcal{E}} := \rho_{\sigma_\mathcal{E}}^{(p+1)} - \rho_{\sigma_\mathcal{E}}^{(p)}$,

$$\int_K \rho^{(p+1)}\left(\mathbf{u}^{(p+1)} - \mathbf{u}^{(p)}\right) d\mathbf{x} = \sum_{\sigma_\mathcal{V} \in K} \Delta\mathbf{u}_{\sigma_\mathcal{V}} \int_K \rho^{(p+1)}\varphi_{\sigma_\mathcal{V}} \, d\mathbf{x}$$

and

$$\int_K \mathbf{u}^{(p)}\left(\rho^{(p+1)} - \rho^{(p)}\right) d\mathbf{x} = \sum_{\sigma_\mathcal{E} \in K} \Delta\rho_{\sigma_\mathcal{E}} \int_K \mathbf{u}^{(p)}\varphi_{\sigma_\mathcal{E}} \, d\mathbf{x}.$$

Hence, (14) is rewritten as:

$$\int_{\mathbb{R}^d} \psi(\mathbf{x},t) \left(\rho^{(p+1)}\mathbf{u}^{(p+1)} - \rho^{(p)}\mathbf{u}^{(p)}\right) d\mathbf{x} = \sum_K \psi_K^n \left[\sum_{\sigma_\mathcal{V} \in K} \Delta\mathbf{u}_{\sigma_\mathcal{V}} \int_K \rho^{(p+1)}\varphi_{\sigma_\mathcal{E}} \, d\mathbf{x} + \sum_{\sigma_\mathcal{E} \in K} \Delta\rho_{\sigma_\mathcal{E}} \int_K \mathbf{u}^{(p)}\varphi_{\sigma_\mathcal{V}} \, d\mathbf{x} \right]$$
$$= \sum_K \psi_K^n \left[\sum_{\sigma_\mathcal{V} \in K} \omega^{\rho,p+1,K}_{\sigma_\mathcal{V}}|C_{\sigma_\mathcal{V}}|\Delta\mathbf{u}_{\sigma_\mathcal{V}} + \sum_{\sigma_\mathcal{E} \in K} \omega^{\mathbf{u},p,K}_{\sigma_\mathcal{V}}|C_{\sigma_\mathcal{E}}|\Delta\rho_{\sigma_\mathcal{E}} \right]$$
$$\tag{15}$$

where we have set for simplicity

$$\omega^{\rho,p+1,K}_{\sigma_\mathcal{V}} := \frac{\int_K \rho^{(p+1)}\varphi_{\sigma_\mathcal{V}} \, d\mathbf{x}}{|C_{\sigma_\mathcal{V}}|} \text{ and } \omega^{\mathbf{u},p,K}_{\sigma_\mathcal{V}} := \frac{\int_K \mathbf{u}^{(p)}\varphi_{\sigma_\mathcal{E}} \, d\mathbf{x}}{|C_{\sigma_\mathcal{E}}|}.$$

For the velocity, we have:

$$|C_{\sigma_\mathcal{V}}|\left(\mathbf{u}_{\sigma_\mathcal{V}}^{(p+1)} - \mathbf{u}_{\sigma_\mathcal{V}}^{(p)}\right) + \Delta t_n \sum_{K,\sigma_\mathcal{V} \in K} \Phi^{\mathbf{u}}_{\sigma_\mathcal{V},K} = 0$$

where, for the scheme (7),

$$\Phi^{\mathbf{u}}_{\sigma_\mathcal{V},K} = \frac{1}{2}\left(\Phi^{\mathbf{u}}_{\sigma_\mathcal{V},K}(U^{(p)}) + \Phi^{\mathbf{u}}_{\sigma_\mathcal{V},K}(U^{(0)})\right)$$

and for the density, we have

$$|C_{\sigma_\mathcal{E}}|\left(\rho_{\sigma_\mathcal{E}}^{(p+1)} - \rho_{\sigma_\mathcal{E}}^{(p)}\right) + \Delta t_n \sum_{K,\sigma_\mathcal{E} \in K} \Phi^{\rho}_{\sigma_\mathcal{E},K} = 0.$$

9

We note that the sum reduces to one term, hence

$$|C_{\sigma_{\mathcal{E}}}|(\rho_{\sigma_{\mathcal{E}}}^{(p+1)} - \rho_{\sigma_{\mathcal{E}}}^{(p)}) + \Delta t_n \Phi_{\sigma_{\mathcal{E}},K}^{\rho} = 0$$

where again

$$\Phi_{\sigma_{\mathcal{E}}}^{\rho} = \frac{1}{2}\big(\Phi_{\sigma_{\mathcal{V}}}^{\rho}(U^{(p)}) + \Phi_{\sigma_{\mathcal{V}}}^{\rho}(U^{(0)})\big)$$

and K is *the* element such that $\sigma_{\mathcal{E}} \in K$

Using these relations in (15), we get

$$\int_{\mathbb{R}^d} \psi(\mathbf{x},t)\,\big(\rho^{(p+1)}\mathbf{u}^{(p+1)} - \rho^{(p)}\mathbf{u}^{(p)}\big)\,d\mathbf{x}$$

$$= -\Delta t_n \sum_K \psi_K^n \bigg[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \bigg\{ \sum_{K',\sigma_{\mathcal{V}} \in K'} \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}} \bigg\} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho} \bigg] \qquad (16)$$

We see that

$$\sum_K \psi_K^n \bigg[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \bigg\{ \sum_{K',\sigma_{\mathcal{V}} \in K'} \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}} \bigg\} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho} \bigg]$$

$$= \sum_{\sigma_{\mathcal{V}} \in K} \underbrace{\bigg[\sum_{K',\sigma_{\mathcal{V}} \in K'} \sum_{K,\sigma_{\mathcal{V}} \in K \cap K'} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \psi_K^n\, \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}} \bigg]}_{I}$$

$$+ \sum_K \sum_{\sigma_{\mathcal{E}} \in K} \psi_K^n \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K}\, \Phi_{\sigma_{\mathcal{E}}}^{\rho}$$

Then, I can be written as:

$$I = \psi_{\sigma_{\mathcal{V}}}^n \sum_{K',\sigma_{\mathcal{V}} \in K'} \bigg\{ \sum_{K,\sigma_{\mathcal{V}} \in K \cap K'} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \bigg\} \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}}$$

$$+ \underbrace{\bigg[\sum_{K',\sigma_{\mathcal{V}} \in K'} \sum_{K,\sigma_{\mathcal{V}} \in K \cap K'} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \big(\psi_K^n - \psi_{\sigma_{\mathcal{V}}}^n\big)\, \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}} \bigg]}_{:=D_{\sigma_{\mathcal{V}}}}$$

so that, setting

$$\omega_{\sigma_{\mathcal{V}}}^{\rho,p+1} := \frac{\sum_{K,\sigma_{\mathcal{V}} \in K} \int_K \rho^{(p+1)} \varphi_{\sigma_{\mathcal{V}}}\,d\mathbf{x}}{|C_{\sigma_{\mathcal{V}}}|},$$

$$\sum_K \psi_K^n \bigg[\sum_{\sigma_{\mathcal{V}}} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \bigg\{ \sum_{K',\sigma_{\mathcal{V}} \in K'} \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}} \bigg\} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho} \bigg] = \sum_K \sum_{\sigma_{\mathcal{V}} \in K} \psi_{\sigma_{\mathcal{V}}}^n \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{V}}} D_{\sigma_{\mathcal{V}}}$$

$$+ \sum_K \psi_K^n \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho}$$

$$= \sum_K \psi_K^n \bigg[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1,K} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho} \bigg]$$

$$+ \sum_K \bigg(\underbrace{\sum_{\sigma_{\mathcal{V}} \in K} (\psi_{\sigma_{\mathcal{V}}}^n - \psi_K^n) \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}}}_{:=F_K} + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}} \bigg)$$

so that we get the master equation:

$$\int_{\mathbb{R}^d} \psi(\mathbf{x},t)\left(\rho^{(p+1)}\mathbf{u}^{(p+1)} - \rho^{(p)}\mathbf{u}^{(p)}\right) d\mathbf{x} + \sum_K \psi_K^n \left[\sum_{\sigma_\mathcal{V} \in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1} \Phi_{\sigma_\mathcal{V},K}^{\mathbf{u}} + \sum_{\sigma_\mathcal{E} \in K} \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K} \Phi_{\sigma_\mathcal{E},K}^{\rho}\right]$$
$$+ \sum_K \left(F_K + \sum_{\sigma_\mathcal{V} \in K} D_{\sigma_\mathcal{V}}\right) = 0 \tag{17a}$$

with

$$F_K = \sum_{\sigma_\mathcal{V} \in K} \left(\psi_{\sigma_\mathcal{V}}^n - \psi_K^n\right) \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K} \Phi_{\sigma_\mathcal{V},K}^{\mathbf{u}}$$

$$D_K = \sum_{K',\sigma_\mathcal{V} \in K'} \left[\sum_{K,\sigma_\mathcal{V} \in K \cap K'} \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K}\left(\psi_K^n - \psi_{\sigma_\mathcal{V}}^n\right) \Phi_{\sigma_\mathcal{V},K'}^{\mathbf{u}}\right] \tag{17b}$$

$$\omega_{\sigma_\mathcal{V}}^{\rho,p+1} = \frac{\sum_{K,\sigma_\mathcal{V} \in K} \int_K \rho^{(p+1)} \varphi_{\sigma_\mathcal{V}}\, d\mathbf{x}}{|C_{\sigma_\mathcal{V}}|}, \qquad \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K} = \frac{\int_K \mathbf{u}^{(p)} \varphi_{\sigma_\mathcal{E}}\, d\mathbf{x}}{|C_{\sigma_\mathcal{E}}|}$$

Then let us look at the total energy. The first thing to do is to remark that (with similar notations as before) that
$$\Delta \rho \mathbf{u}^2 = \mathbf{u}^{(p+1)} \cdot \Delta \rho \mathbf{u} + \rho^{(p)} \mathbf{u}^{(p)} \cdot \Delta \mathbf{u}$$

which combined to
$$\Delta \rho \mathbf{u} = \rho^{(p+1)} \Delta \mathbf{u} + \mathbf{u}^{(p)} \Delta \rho$$

gives:
$$\Delta \rho \mathbf{u}^2 = \left(\rho^{(p+1)} \mathbf{u}^{(p+1)} + \rho^{(p)} \mathbf{u}^{(p)}\right) \cdot \Delta \mathbf{u} + \mathbf{u}^{(p+1)} \cdot \mathbf{u}^{(p)} \Delta \rho.$$

To simplify, we will set
$$\widetilde{\mathbf{m}} = \frac{\rho^{(p+1)} \mathbf{u}^{(p+1)} + \rho^{(p)} \mathbf{u}^{(p)}}{2}, \qquad \widetilde{q^2} = \mathbf{u}^{(p+1)} \cdot \mathbf{u}^{(p)}.$$

Using these relations, we see that

$$\sum_K \psi_K \int_K \Delta E\, d\mathbf{x} = \sum_K \psi_K \left(\int_K \Delta e\, d\mathbf{x} + \int_K \tilde{m} \cdot \Delta \mathbf{u}\, d\mathbf{x} + \frac{1}{2}\int_K \widetilde{q^2} \Delta \rho\, d\mathbf{x}\right)$$
$$= \sum_K \psi_K \left[\int_K \Delta e\, d\mathbf{x} + \sum_{\sigma_\mathcal{V} \in K} \Delta \mathbf{u}_{\sigma_\mathcal{V}} \cdot \int_K \widetilde{\mathbf{m}} \varphi_{\sigma_\mathcal{V}}\, d\mathbf{x}\right.$$
$$\left. + \frac{1}{2}\sum_{\sigma_\mathcal{E} \in K} \Delta \rho_{\sigma_\mathcal{E}} \int_K \widetilde{q^2} \varphi_{\sigma_\mathcal{V}}\, d\mathbf{x}\right]$$

First, we notice that
$$\int_K \Delta e\, d\mathbf{x} = -\Delta t \sum_{\sigma_\mathcal{E} \in K} \Phi_{\sigma_\mathcal{E},K}^e.$$

Then we have, introducing
$$\theta_{\sigma_\mathcal{V}}^{\mathbf{m},K} = \frac{\int_K \widetilde{\mathbf{m}} \varphi_{\sigma_\mathcal{V}}\, dx}{|C_{\sigma_\mathcal{V}}|}, \text{ and } \theta_{\sigma_\mathcal{V}}^{q^2,K} = \frac{\int_K \widetilde{q^2} \varphi_{\sigma_\mathcal{E}}\, dx}{|C_{\sigma_\mathcal{E}}|},$$

we have
$$\sum_{\sigma_\mathcal{V} \in K} \Delta \mathbf{u}_{\sigma_\mathcal{V}} \cdot \int_K \widetilde{\mathbf{m}} \varphi_{\sigma_\mathcal{V}}\, d\mathbf{x} = -\Delta t \sum_{\sigma_\mathcal{V} \in K} \theta_{\sigma_\mathcal{V}}^{\mathbf{m},K} \cdot \left(\sum_{K',\sigma_\mathcal{V} \in K'} \Phi_{\sigma_\mathcal{V},K'}^{\mathbf{u}}\right)$$

11

and

$$\sum_{\sigma_{\mathcal{E}}} \in K \Delta\rho_{\sigma_{\mathcal{E}}} \int_K \widetilde{q}^2 \varphi_{\sigma_{\mathcal{V}}} \, d\mathbf{x} = -\Delta t \sum_{\sigma_{\mathcal{E}} \in K} \theta_{\sigma_{\mathcal{V}}}^{q^2,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho}$$

because $\sigma_{\mathcal{E}}$ belongs to a single element. Then proceeding as for the velocity, and introducing

$$\theta_{\sigma_{\mathcal{V}}}^{\mathbf{m}} = \frac{\sum\limits_{K,\sigma_{\mathcal{V}} \in K} \int_K \widetilde{\mathbf{m}} \varphi_{\sigma_{\mathcal{V}}} \, dx}{|C_{\sigma_{\mathcal{V}}}|} = \sum_{K,\sigma_{\mathcal{V}} \in K} \theta_{\sigma_{\mathcal{V}}}^{\mathbf{m},K},$$

we get

$$
\begin{aligned}
\sum_K \psi_K \int_K \Delta E \, d\mathbf{x} &+ \Delta t_n \sum_K \psi_K \left(\sum_{\sigma_{\mathcal{E}} \in K} \Phi_{\sigma_{\mathcal{E}},K}^e + \sum_{\sigma_{\mathcal{V}} \in K} \theta_{\sigma_{\mathcal{V}}}^{\mathbf{m}} \cdot \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \frac{1}{2} \sum_{\sigma_{\mathcal{E}} \in K} \theta_{\sigma_{\mathcal{V}}}^{q^2,K} \Phi_{\sigma_{\mathcal{E}}} \right) \\
&+ \sum_{\sigma_{\mathcal{V}} \in K} \sum_{K',\sigma_{\mathcal{V}} \in K'} \sum_{K,\sigma_{\mathcal{V}} \in K \cap K'} \underbrace{(\psi_K^n - \psi_{\sigma_{\mathcal{V}}}^n) \theta_{\sigma_{\mathcal{V}}}^{\mathbf{m},K} \cdot \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}}}_{:=D_{\sigma_{\mathcal{V}}}^E} \\
&+ \sum_K \sum_{\sigma_{\mathcal{V}} \in K} \underbrace{(\psi_{\sigma_{\mathcal{V}}} - \psi_K^n) \theta_{\sigma_{\mathcal{V}}}^{\mathbf{m},K} \cdot \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}}}_{:=F_K^E}
\end{aligned}
\tag{18}
$$

We show in the appendix the following result:

Proposition 3.1. *Assume that the mesh $\mathcal{T}_h$ is shape regular[1], and denote by h the maximum diameter of the element of the mesh. If, for any K, for the residuals $\Phi_{\sigma_{\mathcal{E}},K}^{\rho}$, $\Phi_{\sigma_{\mathcal{E}},K}^{e}$, $\Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}}$ are Lipschitz continuous function of their arguments, with Lipschitz constant of the form $C\,h$ where C only depend on α and the maximum norm of the solution.*

Assume that we have a family of meshes $\mathcal{F} = \{\mathcal{T}_{h_n}\}$ with $\lim\limits_{n \to +\infty} h_n = 0$. We consider $(U_{h_n})_{n \geq 0}$ the sequence of functions defined by:

$$\text{if } t \in [t_n, t_{n+1}[, U(\mathbf{x}, t) = (\rho(\mathbf{x}, t_n), \mathbf{u}(\mathbf{x}, t_n), e(\mathbf{x}, t_n))^T$$

with, if K is the element that exists almost every where such that $\mathbf{x} \in K$

$$\rho(\mathbf{x}, t_n) = \sum_{\sigma_{\mathcal{E}} \in K} \rho_{\sigma_{\mathcal{E}}}^n \varphi_{\sigma_{\mathcal{E}}}(\mathbf{x}), \quad e(\mathbf{x}, t_n) = \sum_{\sigma_{\mathcal{E}} \in K} e_{\sigma_{\mathcal{E}}}^n \varphi_{\sigma_{\mathcal{E}}}(\mathbf{x}),$$

and

$$\mathbf{u}(\mathbf{x}, t_n) = \sum_{\sigma_{\mathcal{V}}} \mathbf{u}_{\sigma_{\mathcal{V}}}^n \varphi_{\sigma_{\mathcal{V}}}(\mathbf{x}).$$

Here $\{(\rho_{\sigma_{\mathcal{E}}}^n), (\mathbf{u}_{\sigma_{\mathcal{V}}}^n), (e_{\sigma_{\mathcal{E}}}^n)\}_{n \geq 0, \sigma_{\mathcal{E}}, \sigma_{\mathcal{V}}}$ are defined by the scheme.

We assume that the density, velocity and internal energy are uniformly bounded and that a subsequence converges in L^2 towards $(\rho, \mathbf{u}, e)$ where $\rho, e \in L^2(\mathbb{R}^d \times [0, T])$ and $\mathbf{u} \in \left(L^2(\mathbb{R}^d \times [0, T])\right)^d$.

We also assume that the residuals satisfy

$$\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,p+1} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},p,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho} = \int_K (\mathbf{m}^{(p)} - \mathbf{m}^{(0)}) \, d\mathbf{x} + \Delta t_n \int_{\partial K} \frac{\mathbf{f}^{\mathbf{m}}(U^{(p)}) + \mathbf{f}^{\mathbf{m}}(U^{(0)})}{2} \cdot \mathbf{n} \, d\gamma \tag{19}$$

and

$$\sum_{\sigma_{\mathcal{E}} \in K} \Phi_{\sigma_{\mathcal{E}},K}^e + \sum_{\sigma_{\mathcal{V}} \in K} \theta_{\sigma_{\mathcal{V}}}^{\mathbf{m}} \cdot \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \frac{1}{2} \sum_{\sigma_{\mathcal{V}}} \theta_{\sigma_{\mathcal{V}}}^{q^2,K} \Phi_{\sigma_{\mathcal{E}}} = \int_K \left(E^{(p)} - E^{(0)}\right) d\mathbf{x} + \Delta t_n \int_{\partial K} \frac{\mathbf{f}^E(U^{(p)}) + \mathbf{f}^E(U^{(0)})}{2} \cdot \mathbf{n} \, d\gamma$$

$$\tag{20}$$

---

[1]i.e. the ratio of the inner diameter and the outer diameter of any element of the mesh is larger than an uniform strictly positive constant α.

where we have set:

$$\omega_{\sigma_\mathcal{V}}^{\rho,p+1} = \frac{\sum\limits_{K,\sigma_\mathcal{V}\in K} \int_K \rho^{(p+1)}\varphi_{\sigma_\mathcal{V}}\,d\mathbf{x}}{|C_{\sigma_\mathcal{V}}|}, \qquad \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K} = \frac{\int_K \mathbf{u}^{(p)}\varphi_{\sigma_\mathcal{E}}\,d\mathbf{x}}{|C_{\sigma_\mathcal{E}}|}$$

$$\widetilde{\mathbf{m}} = \frac{\rho^{(p+1)}\mathbf{u}^{(p+1)} + \rho^{(p)}\mathbf{u}^{(p)}}{2}, \qquad \widetilde{q^2} = \mathbf{u}^{(p+1)}\cdot\mathbf{u}^{(p)} \tag{21}$$

$$\theta_{\sigma_\mathcal{V}}^{\mathbf{m}} = \frac{\sum\limits_{K,\sigma_\mathcal{V}\in K} \int_K \widetilde{\mathbf{m}}\varphi_{\sigma_\mathcal{V}}\,dx}{|C_{\sigma_\mathcal{V}}|}, \qquad \theta_{\sigma_\mathcal{V}}^{q^2,K} = \frac{\int_K \widetilde{q^2}\varphi_{\sigma_\mathcal{E}}\,dx}{|C_{\sigma_\mathcal{E}}|}$$

with the assumption that there exists C independant of n, such that $\Delta t \leq Ch$. Then $V = (\rho, \rho\mathbf{u}, e + \frac{1}{2}\rho\mathbf{u}^2)$ is a weak solution of the problem.

3.2   How to achieve discrete conservation.

Since there is no ambiguity, we drop the dependancy of the residuals with respect to the element.

Given a set of residual that satisfy (9) also satisfy (19) and (20). In this section, we will show how to slightly modify the original scheme so that the new one will satisfy (9), (19) and (20), an hence if the scheme converges, we have convergence towards a weak solution. To achieve this, following [1, 16, 17], we introduce the correction terms in the residuals. This needs to be done only for the velocity and the internal energy.

Knowing at time t_n $(\rho^n, \mathbf{u}^n, e^n)$, we start the DeC iteration to get $(\rho^{n+1}, \mathbf{u}^{n+1}, e^{n+1})$. For this, knowing at iteration p, $(\rho^{(p)}, \mathbf{u}^{(p)}, e^{(p)})$, we first compute $\rho^{(p+1)}$.

Momentum.

We introduce a correction $r_{\sigma,K}^{\mathbf{u}}$ so that

$$\Psi_{\sigma_V}^{\mathbf{u}} = \frac{\Phi_{\sigma_V}^{\mathbf{u}}(U^{(p)}) + \Phi_{\sigma_V}^{\mathbf{u}}(U^n)}{2} + r_{\sigma_\mathcal{V}}^{\mathbf{u}} \tag{22}$$

is such that (19) holds true for the new set of residuals, i.e.

$$\sum_{\sigma_\mathcal{V}\in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1} r_{\sigma_\mathcal{V}}^{\mathbf{u}} = \int_{\partial K} \frac{\mathbf{f}^{\mathbf{m}}(U^{(p)}) + \mathbf{f}^{\mathbf{m}}(U^{(0)})}{2}\cdot\mathbf{n}\,d\gamma - \left\{ \sum_{\sigma_\mathcal{V}\in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1}\Phi_{\sigma_\mathcal{V},K}^{\mathbf{u}} + \sum_{\sigma_\mathcal{E}\in K} \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K}\Phi_{\sigma_\mathcal{E},K}^{\rho} \right\}$$

There is no reason to have a different value of $r_{\sigma_\mathcal{V}}^{\mathbf{u}}$ unless possible special needs, so we set $r_{\sigma_\mathcal{V}}^{\mathbf{u}} = r^{\mathbf{u}}$, and since a priori

$$\sum_{\sigma_\mathcal{V}\in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1} > 0$$

we get a unique value of $r^{\mathbf{u}}$ defined by

$$\left(\sum_{\sigma_\mathcal{V}\in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1} \right) r^{\mathbf{u}} = \int_{\partial K} \frac{\mathbf{f}^{\mathbf{m}}(U^{(p)}) + \mathbf{f}^{\mathbf{m}}(U^{(0)})}{2}\cdot\mathbf{n}\,d\gamma - \left\{ \sum_{\sigma_\mathcal{V}\in K} \omega_{\sigma_\mathcal{V}}^{\rho,p+1}\Phi_{\sigma_\mathcal{V},K}^{\mathbf{u}} + \sum_{\sigma_\mathcal{E}\in K} \omega_{\sigma_\mathcal{V}}^{\mathbf{u},p,K}\cdot\Phi_{\sigma_\mathcal{E},K}^{\rho} \right\} \tag{23}$$

Once this is known, we can update the velocity and compute $\mathbf{u}_{\sigma_\mathcal{V}}^{(p+1)}$.

Energy.

Now we know $\rho^{(p)}$, $\rho^{(p+1)}$, $\mathbf{u}^{(p)}$, $\mathbf{u}^{(p+1)}$ and $e^{(p)}$, and have the *updated* residuals for the velocity (there is no change for the density). Again we introduce a correction on the energy, $r_{\sigma_\mathcal{E}}^{e}$, and for the residual

$$\Psi_{\sigma_\mathcal{E}} = \frac{\Phi_{\sigma_\mathcal{E}}^{e}(U^{(p)} + \Phi_{\sigma_\mathcal{E}}^{e}U^{(0)})}{2} + r_{\sigma_\mathcal{E}}^{e}$$

13

to satisfy (20), we simply need:

$$\sum_{\sigma_\mathcal{E} \in K} r^e_{\sigma_\mathcal{E}} = \int_{\partial K} \frac{\mathbf{f}^E(U^{(p)}) + \mathbf{f}^E(U^{(0)})}{2} \cdot \mathbf{n} \, d\gamma - \left\{ \sum_{\sigma_\mathcal{E} \in K} \Phi^e_{\sigma_\mathcal{E},K} + \sum_{\sigma_\mathcal{V} \in K} \theta^{\mathbf{m}}_{\sigma_\mathcal{V}} \cdot \Phi^{\mathbf{u}}_{\sigma_\mathcal{V},K} + \frac{1}{2} \sum_{\sigma_\mathcal{V}} \theta^{q^2,K}_{\sigma_\mathcal{V}} \Phi_{\sigma_\mathcal{E}} \right\} \quad (24)$$

Since there is no reason to favour one degree of freedom with respect to the other ones, we take $r^e_{\sigma_\mathcal{E}} = r^e$, and again we can explicitly solve.

Remark 3.2.

1. *The calculations made for the second order in time can be immediately extended to the higher accuracy ones in time. The only thing is to modify the half sums like*

$$\frac{\Phi^\rho_{\sigma_\mathcal{E}}(U^{(p)}) + \Phi^\rho_{\sigma_\mathcal{E}}(U^n)}{2}$$

   that comes from (12) *into the corresponding terms of* (13).

2. *We also see that the exact form of the residuals are never used, so this can also be extended to any type of residuals, including for high order ones as in [9]. We also note that we have never used the global continuity of the velocity: instead of using W^2_h for the velocity, we could have used V^2_h in a Discontinuous Galerkin like spirit.*

3. *The coefficients of* (21) *must be computed exactly. Since the density and velocity are polynoms, this means we need to use quadrature formula with enough accuracy.*

4 Some numerical results

In this section, we want to illustrate the previous results, and show that the method is effective. We are not claiming that these are the optimal ones, they can be seen more as a proof of concept. It is enough to describe what is done on $K = K_{j+1/2}$, whatever $j \in \mathbb{Z}$.

4.1 Actual schemes

In the following, $\hat{\mathbf{f}}_{j+1/2}$ is a numerical flux evaluated between the states

$$U^+_{j+1/2} = \lim_{x \to x_{j+1/2}, x > x_{j+1/2}} \left(\rho, \rho\mathbf{u}, e + \frac{1}{2}\rho\mathbf{u}^2 \right)(x) \text{ and } U^-_{j+1/2} = \lim_{x \to x_{j+1/2}, x < x_{j+1/2}} \left(\rho, \rho\mathbf{u}, e + \frac{1}{2}\rho\mathbf{u}^2 \right)(x)$$

Here $\rho(x), \mathbf{u}(x)$ and $e(x)$ are obtained from the approximation space. The flux $\hat{\mathbf{f}}$ has a ρ component, a $\mathbf{m}-$ component and a total energy component, they are denoted by $\hat{\mathbf{f}}^\rho$, $\hat{\mathbf{f}}^m$ and $\hat{\mathbf{f}}^E$. Note that $\mathbf{u}$ is continuous. In the numerical experiments, we will consider the exact solver, because it appears we need intermediate states, see above.

We approximate the thermodynamical variables by polynomials of degree r in each interval $K_{j+1/2}$ and the velocity by a continuous approximation which is polynomial of degree $r+1$ in each interval $K_{j+1/2}$. We denote the approximation by $K(r+1)T(r)$. The time discretisation will use the DeC formulation explained before. For that reason, in each interval we expand the thermodynamical and kinetic function using Bézier polynomials. The reason is that the integrals of the basis functions are always positive.

For simplicity, we reduces the formal time accuracy to second order, and we only need to describe the spatial term: $\Phi^\rho_{\sigma_\mathcal{E}}$ for the density, $\Phi^e_{\sigma_\mathcal{E}}$ for the energy and $\Phi^{\mathbf{u}}_{\sigma_\mathcal{V}}$ for the velocity. The update of the density is done by the DG scheme:

$$\Phi^\rho_{\sigma_\mathcal{E}} = -\int_{K_{j+1/2}} \nabla\varphi_{\sigma_\mathcal{E}} \mathbf{f}^\rho \, d\mathbf{x} + \left(\hat{\mathbf{f}}^\rho_{j+1/2} \varphi_{\sigma_\mathcal{E}}(x_{j+1/2}) - \hat{\mathbf{f}}^\rho_{j-1/2} \varphi_{\sigma_\mathcal{E}}(x_{j-1/2}) \right) \quad (25)$$

The update of the velocity is done by:

$$\rho_K^\star \, \Phi_{\sigma_\mathcal{V}}^{\mathbf{u}} = \int_K \varphi_{\sigma_\mathcal{V}} \rho \mathbf{u} \frac{\partial \mathbf{u}}{\partial x} \, d\mathbf{x} - \int_K p \frac{\partial \varphi_{\sigma_\mathcal{V}}}{\partial x} \, d\mathbf{x} + \int_{\partial K} p^\star \varphi_{\sigma_\mathcal{V}} \, d\gamma + \alpha_K \rho_K^\star \big(\mathbf{u}_{\sigma_\mathcal{V}} - \bar{\mathbf{u}} \big) \tag{26}$$

where $p^\star$ is the pressure evaluated at the quadrature points by the Riemann solver (this is why we have chosen HLLC), and $\rho_K^\star$ is the average of the density in K, α_K is an upper bound of the wave speeds in K, $\bar{\mathbf{u}}$ is the arithmetic average of the velocity within K.

The update of the internal energy is done by (after integration by part of the ue_x term) by

$$\Phi_{\sigma_\mathcal{E}}^e = \int_K \left(p \frac{\partial \mathbf{u}}{\partial x} \varphi_{\sigma_\mathcal{E}} - \mathbf{u} e \frac{\partial \varphi_{\sigma_\mathcal{E}}}{\partial x} \right) d\mathbf{x} + \int_{\partial K} \mathbf{u} e^\star \varphi_{\sigma_\mathcal{E}} \, d\gamma \tag{27}$$

where again $e^\star$ is given by the Riemann solver.

The schemes, even with the Euler forward time stepping, have no chance to be positivity preserving, and we note that the update of the velocity will be at most first order in time. Hence, inspired by the residual distribution schemes, we upgrade formal accuracy in two possible ways:

1. Procedure 1: We use the residuals (25) and (27) for the thermodynamical variables, and for the velocity, we replace $\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}}$ by $\big(\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}} \big)^\star$ defined by:

   (a) Compute $\Phi^{\mathbf{u}} = \sum\limits_{\sigma_\mathcal{V}} \Phi_{\sigma_\mathcal{V}}^{\mathbf{u}}$

   (b) If $\| \Phi^{\mathbf{u}} \| > 0$, define

   $$x_{\sigma_\mathcal{V}} = \max \left(\frac{\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}}}{\Phi^{\mathbf{u}}}, 0 \right)$$

   $$\big(\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}} \big)^\star = \frac{x_{\sigma_\mathcal{V}}}{\sum\limits_{\sigma_\mathcal{V} \in K_{j+1/2}} x_{\sigma_\mathcal{V}}} \Phi^{\mathbf{u}}$$

   (c) Else $\big(\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}} \big)^\star = 0$

2. Procedure 2: we do the same as before for $\Phi_{\sigma_\mathcal{V}}^{\mathbf{u}}$, $\Phi_{\sigma_\mathcal{E}}^\rho$ and $\Phi_{\sigma_\mathcal{V}}^e$ where the thermodynamical residual are now:

   $$\Phi_{\sigma_\mathcal{E}}^\rho + \alpha_K \big(\rho_{\sigma_\mathcal{E}} - \bar{\rho}_K \big) \text{ and } \Phi_{\sigma_\mathcal{E}}^e + \alpha_K \big(e_{\sigma_\mathcal{E}} - \bar{e}_K \big)$$

   where $\bar{\rho}_K$ (resp. $\bar{e}_K$) are the arithmetic average of the density DOFS (resp. internal energy) in K.

The reader can refer to [18] for more details, and in particular why formal accuracy is increased.

4.2 Results

Here we solve a series of shock tube problems to assess the accuracy and robustness of the proposed RD staggered scheme . For numerical experiments of this section we will use the ideal EOS for gas, linking the pressure, the internal energy and the density: $p = (\gamma - 1) \rho e$, where $\gamma = 1.4$.

4.2.1 The Sod shock tube problem.

The Sod shock tube is a common one dimensional Riemann problem for the illustration of the interesting behavior of numerical solutions to hyperbolic Euler equations of gas dynamics. The structure of the solution involves three distinct waves of a left rarefaction wave, a contact discontinuity, and a right shock wave. This test case uses to determine if a scheme recovers properly discrete Rankine-Hugoniot relations on the shock. If we put the initial discontinuity at $x = 0.5$ in the domain $[0, 1]$, the initial data for this problem is given as follows:
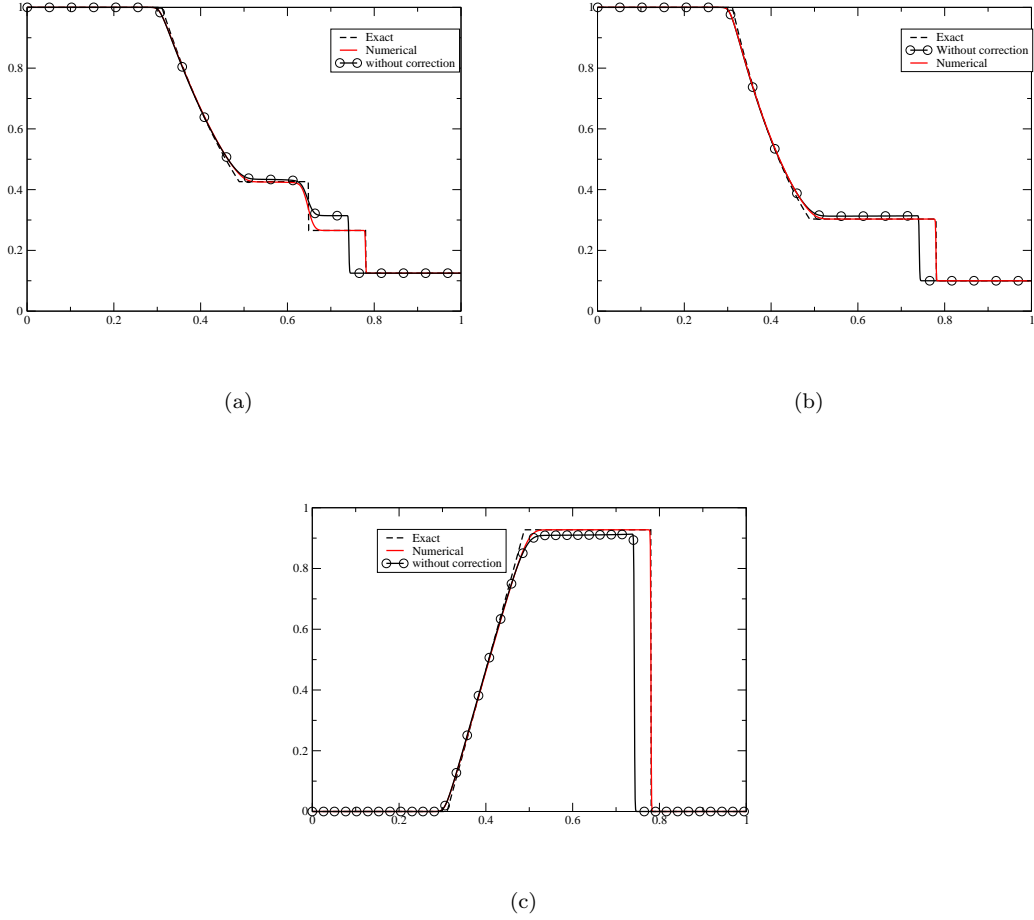
(a)

(b)

(c)

Figure 1: Solution of the Sod shock tube problem for density, velocity and pressure at $T = 0.16, CFL = 0.4$. Exact, numerical and without correction.

$$(\rho_0, u_0, p_0) = \begin{cases} (1.0, 0.0, 1.0), & \text{if} \quad x < 0.5, \\ (0.125, 0.0, 0.1), & \text{if} \quad x > 0.5. \end{cases} \tag{28}$$

In Figure 1, profiles of density, velocity, pressure and internal energy are depicted with a solution of reference in line for a mesh containing 1000 cells. We also have plotted the solution obtained without any correction. Both have been obtained with the T0K1 scheme, and first order in time. We see that the uncorrected solution is completely off, as expected, but also that the correction we have defined provides an accurate approximation of all three distinct waves.

In figure 2, we show the results obtained by the first order (K1T0) and second order (K2T1) in time and space schemes. There are some improvement. The point is more to show that the correction is also effective when we use higher degree polynomial representations.
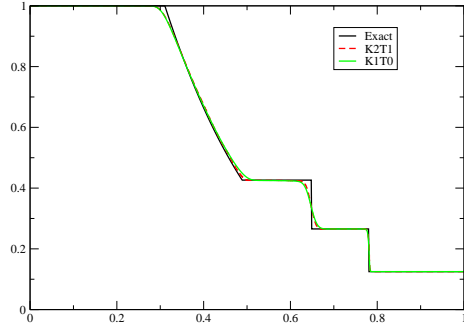
(a)
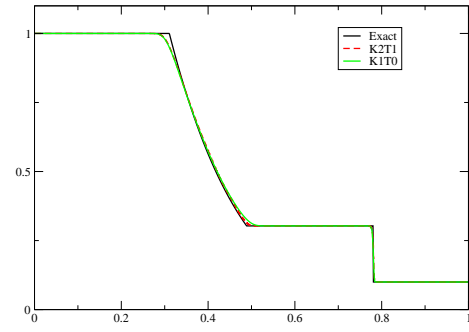
(b)
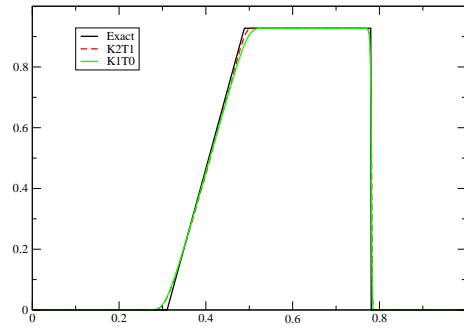
(c)

Figure 2: Solution of the Sod shock tube problem at $T = 0.16, CFL = 0.4..$ Exact, K2T1 and K1T0 solutions, 1000 mesh points.

<div align="center">(a)   (b)   (c)</div>
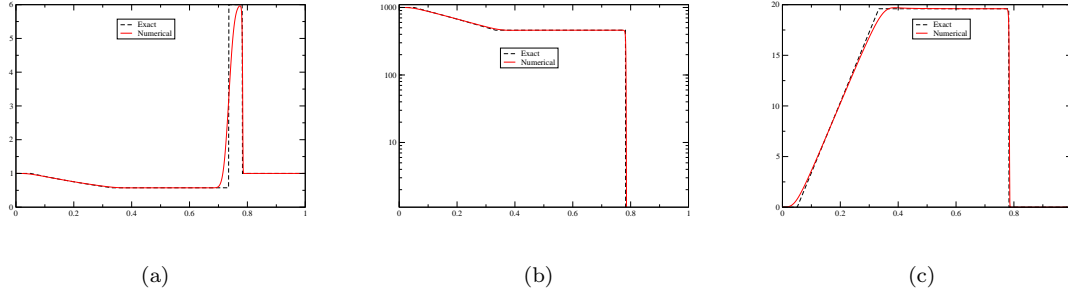
Figure 3: Solution of the strong shock problem at $T = 0.012, CFL = 0.4$.
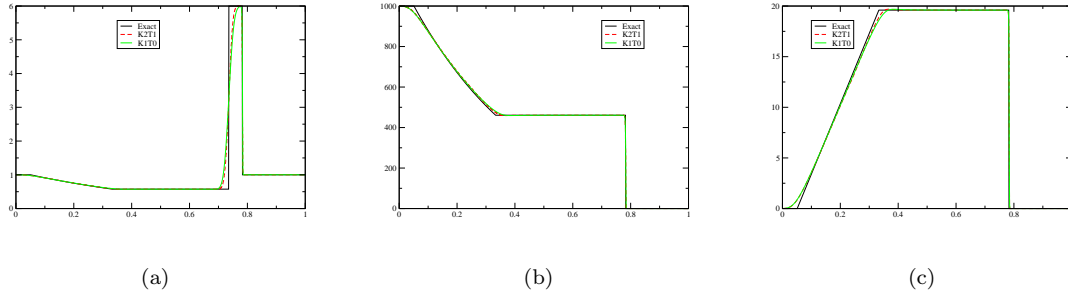

<div align="center">(a)   (b)   (c)</div>

Figure 4: Solution of the strong shock problem at $T = 0.012, CFL = 0.4$ The K2T1 and K2T0 solutions are displayed..

4.2.2   Strong shock.

The next test problem containing a left rarefaction wave, a contact discontinuity, and a strong right shock wave. This test-case highlights the robustness of the numerical methods for fluid dynamics. The initial data, again in the domain $[0, 1]$, are:

$$(\rho_0, u_0, p_0) = \begin{cases} (1.0, 0.0, 1000.0), & \text{if} \quad x < 0.5, \\[2mm] (1.0, 0.0, 0.01), & \text{if} \quad x > 0.5. \end{cases} \tag{29}$$

In Figure 3, profiles of density, velocity, pressure and internal energy are depicted with a solution of reference in line for a mesh containing 1000 cells. It indicates that first-order scheme can accurately resolve strong shocks. As before, the results of figure 4 shows that the correction is effective

4.2.3   123-problem.

For the next test, called the 123 problem, the solution consists of a left rarefaction wave, a contact discontinuity and a right rarefaction wave. Two rarefaction waves are traveling in opposite directions. A low-density and low-pressure region is generated in between. The initial data for this problem is given as follows:

$$(\rho_0, u_0, p_0) = \begin{cases} (1.0, -2.0, 0.4), & if \quad 0.0 \leq x < 0.5, \\ (1.0, 2.0, 0.4), & if \quad 0.5 < x < 1, \end{cases} \tag{30}$$

<div align="center">18</div>

(a)   (b)   (c)

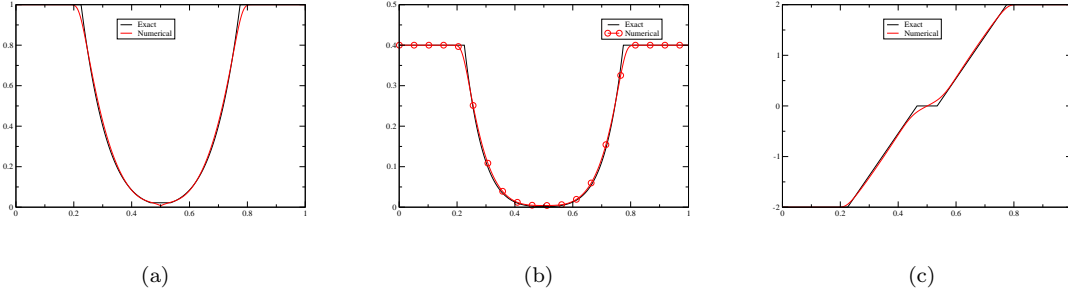Figure 5: The solution for density, velocity and pressure for the 123-problem at $T = 0.15, CFL = 0.4$.

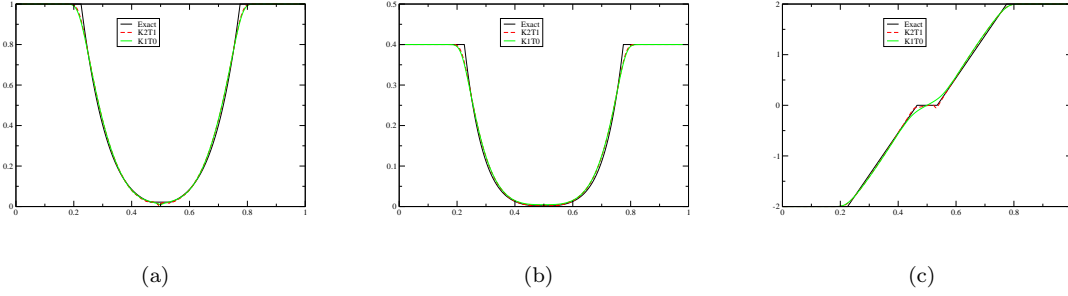(a)   (b)   (c)

Figure 6: The solution for density, velocity and pressure for the 123-problem at $T = 0.15, CFL = 0.4$. K2T1 and K1T0 schemes.

The results for the first-order scheme are shown in Figure 5 with a solution of reference in line for a mesh containing 1000 cells.

4.2.4 Severe test case

The solution of the next test case consists of three strong discontinuities travelling to the right. The initial data consists of two constant states:

$$(\rho_0, u_0, p_0) = \begin{cases} (5.99924, 19.5975, 460.894), & \text{if} \quad 0.0 \leq x < 0.8, \\ (5.992420, -6.19633, 46.0950), & \text{if} \quad 0.8 < x \leq 1.0. \end{cases} \tag{31}$$

This is one of the two test cases designed in [19] which correspond to wave interaction in Collela and Woodward bast waves test case. The exact and numerical solutions are found in the spatial domain $0 \leq x \leq 1$. The numerical solution is computed with 1000 cells and the chosen Courant number coefficient is 0.1. Boundary conditions are transmissive. The results are shown in Figure 7.

5   Conclusions

In this paper we have proposed a staggered higher-order Residual Distribution (RD) scheme for compressible flow. The design of the scheme via the coupling of the RD formulation with a Deferred Correction (DeC) type method allows to reach an arbitrary order of accuracy both in time and space. One of the contributions

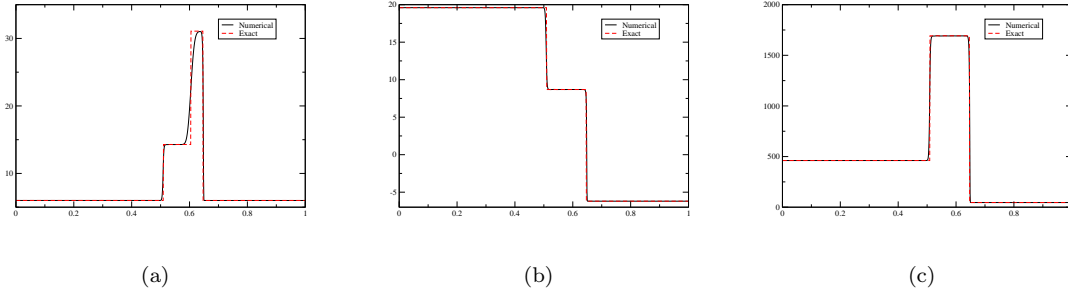<div style="text-align:center">(a)   (b)   (c)</div>

Figure 7: Numerical and exact solutions are compared at the output time 0.012 units for density, velocity and pressure at $T = 0.012, CFL = 0.1$, 1000 cells.

of this paper is to show how one can discretise a non-conservative version of the Euler equation of gas dynamics in Eulerian form and guarantee that the correct weak solutions are recovered. A series of shock tube problems considered in this paper show the accuracy and robustness of the proposed numerical scheme. The RD scheme we have developped provides an accurate numerical approximation and correction we have defined is effective. In addition, the scheme is parameter-free and does not require any artificial viscosity.

Let us write a series of remarks to end this paper.

1. Though the numerical examples are all one dimensional (because here one can compute the exact solution for comparison), the description of the correction introduced in section 3.2, as well as the conditions introduces in proposition 3.1 are formulated for general elements.

2. The Residual distribution formalism introduced here is not restrictive. In [20], it is shown that any classical scheme (finite volume, finite element, discontinuous Galerkin) can be rewritten equivalently in distribution form. If one approximates (for example) the velocity equation with another method, it is certainly possible to write the contribution at element level, as here, and then to rewrite the scheme in the semi-discrete form $\mathcal{L}^{(2)} = 0$ of (10a) (if second order accuracy in time is chosen), or of the type (12), or more general for higher in time approximation. Then, the key fact is to write the local conservation property, not at the level of faces between elements, but on the element themselves: this is what is behind the proof of proposition 3.1, thus corrections of the form (23) and (24) can be written. What is not guaranteed is that the modified scheme will still be stable. In all our experience, we have not see any degradation of the stability condition. We have used this type of correction in other context, see e.g [17, 16, 21], and the conclusions are the same. This is not a proof however.

Further investigations of high order residual distribution schemes and applications to different mathematical models will be considered in forthcoming works.

Acknowledgements.

6 Proof of proposition 3.1.

We first show some estimates for scalar functions (the system case is identical), and then we use it to show the proposition 3.1. We start by some notations: $\mathbb{R}^d$ is subdivided into non overlapping elements,

$$\mathbb{R}^d = \cup K$$

and the mesh is supposed to be conformal (because of the global continuity of the velocity). The parameter h will be the maximum of the diameter of the K. We assume that the partition is shape regular, i.e. the ratio between the inner and outer diameter of the elements is bounded from above and bellow. In $\mathbb{R}^d$, we have functional description of the density, the velocity and the energy: we call them ρ_h, $\mathbf{u}_h$ and e_h to refer they are defined from $\mathbb{R}^d = \cup K$.

Let $T > 0$ and a time discretisation $0 < t_1 < \ldots < t_n < \ldots < t_N \leq T$ of $[0,T]$. We define $\Delta t_n = t_{n+1} - t_n$ and $\Delta t = \max_n \Delta t_n$. We are given the sequences $\{u_h^p\}^{p=0\ldots N}$ where u_h^p belongs to V^h or W^h (see section 2.2). They are defined from degrees of freedom that are again denoted by σ. We can define a function u_Δ by:

$$\text{if } (\mathbf{x},t) \in \Omega \times [t_n, t_{n+1}[, \text{ then } u_\Delta(\mathbf{x},t) = u_h^n(\mathbf{x}).$$

The set of these functions is denoted by X_Δ and is equipped with the L^∞ and L^2 norms.

We have the following lemma

Lemma 6.1. *Let $T > 0$, $\{t_n\}_{n=0,\ldots,N}$ an increasing subdivision of $[0,T]$, $\mathcal{Q}$ a compact of $\mathbb{R}^d$. Let $(u_\Delta)_h$ a sequence of functions of X_Δ defined on $\mathbb{R}^d \times \mathbb{R}^+$. We assume that there exists $C \in \mathbb{R}$ independent of Δ and Δt, and $\mathbf{u} \in L^2_{loc}(\Omega \times [0,T])$ such that*

$$\sup_\Delta \sup_{\mathbf{x},t} |u_\Delta(\mathbf{x},t)| \leq C \quad \text{and} \quad \lim_{\Delta,\Delta t \to 0} |u_\Delta - u|_{L^2(\Omega \times [0,T])} = 0.$$

Then, if $\overline{(u_h^n)}_K$ is the average of u_h^n in K

$$\lim_{h \to 0, \Delta t \to 0} \left(\sum_{n=0}^N \Delta t_n \sum_{K \subset \mathcal{Q}} |K| \sum_{\sigma \in K} \left|\left|(u_h)_\sigma - \overline{(u_h)}_K)\right|\right| \right) = 0. \tag{32}$$

Proof. The proof is inspired from [22] and can be found in [10]. $\qquad\square$

Then we can prove proposition 3.1. We will do it for the momentum, the proof is similar for the energy. We proceed the proof in several lemma.

Lemma 6.2. *Under the conditions of proposition 3.1, for any $\varphi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ we have*

$$\lim_{\Delta t \to 0, \Delta \to 0} \sum_{n=0}^\infty \int_{\mathbb{R}^d} \varphi_h \left(\rho_h^{n+1} \mathbf{u}_h^{n+1} - \rho_h^n \mathbf{u}_h^n \right) d\mathbf{x} = -\int_{\mathbb{R} \times \mathbb{R}^+} \frac{\partial \varphi}{\partial t} u \, dx dt + \int_{\mathbb{R}} \varphi(x,0) u_0 \, dx dt.$$

where

$$\varphi_h(x,t_n) = \sum_K \varphi(x_K, t_k) 1_K, \text{ and } \varphi_h(x,t) = \varphi(x,t_n) \text{ for } t \in [t_n, t_{n+1}[.$$

Proof. This is the classical lemma. $\qquad\square$

Proof of proposition 3.1. Since $\rho^{(M+1)} = \rho^{n+1}$, $\mathbf{u}^{(M+1)} = \mathbf{u}^{n+1}$, $\rho^{(0)} = \rho^n$ and $\mathbf{u}^{(0)} = \mathbf{u}^n$ per definition, we get

$$\rho_h^{n+1} \mathbf{u}_h^{n+1} - \rho_h^n \mathbf{u}_h^n = \sum_{k=0}^M \left(\rho_h^{(p+1)} \mathbf{u}_h^{(p+1)} - \rho_h^{(p)} \mathbf{u}_h^{(p)} \right).$$

To simplify the notations, we will assume $M = 1$, as in the numerical examples, but the reasoning is general.

We start from (17a) written for $\mathbf{u}^{(0)} = \mathbf{u}^n$ and $\mathbf{u}^{(1)}$:

$$\int_{\mathbb{R}^d} \psi(\mathbf{x}, t) \left(\rho^{n+1}\mathbf{u}^{n+1} - \rho^n\mathbf{u}^n\right) d\mathbf{x} + \sum_K \psi_K^n \left[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,2} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},1,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho}\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n, \mathbf{u}^{(1)}) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n, \mathbf{u}^{(1)})\right)$$

$$+ \sum_K \psi_K^n \left[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,1} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},0,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho}\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n)\right) = 0,$$

that is using the assumptions of proposition 3.1,

$$\int_{\mathbb{R}^d} \psi(\mathbf{x}, t) \left(\rho^{n+1}\mathbf{u}^{n+1} - \rho^n\mathbf{u}^n\right) d\mathbf{x} + \sum_K \psi_K^n \left[\int_K \left(\rho^{(1)}\mathbf{u}^{(1)} - \rho^n\mathbf{u}^n\right) d\mathbf{x} + \Delta t_n \int_{\partial K} \mathbf{f_m}\left(\frac{U^{(1)} + U^n}{2}\right) \cdot \mathbf{n} \, d\gamma\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n, \mathbf{u}^{(1)}) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n, \mathbf{u}^{(1)})\right)$$

$$+ \sum_K \psi_K^n \left[\Delta t_n \int_{\partial K} \mathbf{f_m}(U^n) \cdot \mathbf{n} \, d\gamma\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n)\right) = 0,$$

Since, again using (17a) for $p = 0$, we have

$$\int_{\mathbb{R}^d} \psi(\mathbf{x}, t) \left(\rho^{(1)}\mathbf{u}^{(1)} - \rho^n\mathbf{u}^n\right) d\mathbf{x} + \sum_K \psi_K^n \left[\sum_{\sigma_{\mathcal{V}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\rho,1} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}} + \sum_{\sigma_{\mathcal{E}} \in K} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},0,K} \Phi_{\sigma_{\mathcal{E}},K}^{\rho}\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n)\right) = 0$$

we get

$$\int_{\mathbb{R}^d} \psi(\mathbf{x}, t) \left(\rho^{n+1}\mathbf{u}^{n+1} - \rho^n\mathbf{u}^n\right) d\mathbf{x} + \sum_K \psi_K^n \left[\Delta t_n \int_{\partial K} \mathbf{f_m}\left(\frac{U^{(1)} + U^n}{2}\right) \cdot \mathbf{n} \, d\gamma\right]$$

$$+ \sum_K \left(F_K(\mathbf{u}^n, \mathbf{u}^{(1)}) + \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n, \mathbf{u}^{(1)})\right) = 0.$$

From (17b), we see that

$$F_K(U^{(1)}, U^n) = \sum_{\sigma_{\mathcal{V}} \in K} \psi_{\sigma_{\mathcal{V}}}^n - \psi_K^n)\omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},2,K} \Phi_{\sigma_{\mathcal{V}},K}^{\mathbf{u}}$$

$$D_K(U^{(1)}, U^n) = \sum_{K', \sigma_{\mathcal{V}} \in K'} \left[\sum_{K, \sigma_{\mathcal{V}} \in K \cap K'} \omega_{\sigma_{\mathcal{V}}}^{\mathbf{u},2,K} \left(\psi_K^n - \psi_{\sigma_{\mathcal{V}}}^n\right) \Phi_{\sigma_{\mathcal{V}},K'}^{\mathbf{u}}\right]$$

so that, since $\psi_K^n - \psi_{\sigma_{\mathcal{V}}}^n = O(h)$, using the estimates of lemma 6.1, we have

$$\lim_{\Delta t, h \to 0} \sum_K \left(F_K(\mathbf{u}^n, \mathbf{u}^{(1)}) = 0 \text{ and } \lim_{\Delta t, h \to 0} \sum_{\sigma_{\mathcal{V}} \in K} D_{\sigma_{\mathcal{V}}}(\mathbf{u}^n, \mathbf{u}^{(1)}) = 0\right.$$

because the mesh is shape regular and $\Delta t/h$ is bounded. Last, using the same technique as in [10], and thanks again to the lemma 6.1, we see that

$$\lim_{\Delta t, h \to 0} \sum_K \psi_K^n \left[\Delta t_n \int_{\partial K} \mathbf{f_m}(\frac{U^{(1)} + U^n}{2}) \cdot \mathbf{n} \, d\gamma = \int_{\mathbb{R}^+} \int_{\mathbb{R}^d} \nabla_\mathbf{x} \psi(\mathbf{x}, t) \mathbf{f}_m(U) \, d\mathbf{x}. \right.$$

The convergence result for the energy is done with exactly the same method. Then ends the proof of proposition 3.1. $\qquad\square$

References

[1] R. Abgrall and S. Tokareva. Staggered grid residual distribution scheme for lagrangian hydrodynamics. *SIAM SISC*, 39(5):A2345–A2364, 2017. see also `https://hal.inria.fr/hal-01327473`.

[2] Rémi Abgrall, Konstantin Lipnikov, Nathaniel Morgan, and Svetlana Tokareva. Multidimensional staggered grid residual distribution scheme for Lagrangian hydrodynamics. *SIAM J. Sci. Comput.*, 42(1):A343–A370, 2020.

[3] Veselin A. Dobrev, Tzanio V. Kolev, and Robert N. Rieben. High-order curvilinear finite element methods for Lagrangian hydrodynamics. *SIAM J. Sci. Comput.*, 34(5):B606–B641, 2012.

[4] E. Godlewski and P.A. Raviart. *Hyperbolic systems of conservations laws, I.* Ellipse, 1991.

[5] Laura Gastaldo, Raphaèle Herbin, Jean-Claude Latché, and Nicolas Therme. A MUSCL-type segregated–explicit staggered scheme for the Euler equations. *Comput. & Fluids*, 175:91–110, 2018.

[6] Raphaèle Herbin, Jean-Claude Latché, and Chady Zaza. A cell-centred pressure-correction scheme for the compressible Euler equations. *IMA J. Numer. Anal.*, 40(3):1792–1837, 2020.

[7] R. Herbin, J.-C. Latché, S. Minjeaud, and N. Therme. Conservativity and weak consistency of a class of staggered finite volume methods for the Euler equations. *Math. Comp.*, 90(329):1155–1177, 2021.

[8] R. Herbin, J.-C. Latché, and K. Saleh. Low Mach number limit of some staggered schemes for compressible barotropic flows. *Math. Comp.*, 90(329):1039–1087, 2021.

[9] R. Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *Journal of Scientific Computing*, 73(2-3), 2017.

[10] R. Abgrall and P. L. Roe. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.*, 19(1-3):3–36, 2003.

[11] Rémi Abgrall and Davide Torlo. High order asymptotic preserving deferred correction implicit-explicit schemes for kinetic models. *SIAM J. Sci. Comput.*, 42(3):B816–B845, 2020.

[12] Rémi Abgrall. Essentially non oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys.*, 214(2):773–808, 2006.

[13] R. Abgrall, A. Larat, and M. Ricchiuto. Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid meshes. *J. Comput. Phys.*, 230(11):4103–4136, 2011.

[14] D de Santis R. Abgrall, M. Ricchiuto. High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids. *SIAM J. Scientific Computing*, 36(3):A955–A983, 2014.

[15] R. Abgrall and D. de Santis. Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible navier-stokes equations. *Journal of Computational Physics*, 283:329–359, 2015.

[16] R. Abgrall, P. Bacigaluppi, and S. Tokareva. A high-order nonconservative approach for hyperbolic equations in fluid dynamics. *Computers and Fluids*, 2018.

[17] R. Abgrall. A general framework to construct schemes satisfying additional conservation relations, application to entropy conservative and entropy dissipative schemes. *J. Comput. Phys*, 372(1), 2018.

[18] R. Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *J. Sci. Comput.*, 73(2-3):461–494, 2017.

[19] Eleuterio F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, third edition, 2009. A practical introduction.

[20] R. Abgrall. The notion of conservation for residual distribution schemes (or fluctuation splitting schemes), with some applications. *Commun. Appl. Math. Comput.*, 2(3):341–368, 2020.

[21] R. Abgrall, Ph. Öffner, and H. Ranocha. Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous galerkin schemes, 2021.

[22] D. Kröner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-d. *East-West J. Numer. Math.*, 4(4):279–292, 1996.