

# STAGGERED SCHEMES FOR COMPRESSIBLE FLOW: A GENERAL CONSTRUCTION

R. ABGRALL\*

**Abstract.** This paper is focused on the approximation of the Euler equations of compressible fluid dynamics on a staggered mesh. With this aim, the flow parameters are described by the velocity, the density and the internal energy. The thermodynamic quantities are described on the elements of the mesh, and thus the approximation is only in  $L^2$ , while the kinematic quantities are globally continuous. The method is general in the sense that the thermodynamic and kinetic parameters are described by an arbitrary degree of polynomials. In practice, the difference between the degrees of the kinematic parameters and the thermodynamic ones is set to 1. The integration in time is done using the forward Euler method but can be extended straightforwardly to higher-order methods. In order to guarantee that the limit solution will be a weak solution of the problem, we introduce a general correction method in the spirit of the Lagrangian staggered method described in [1, 2, 3], and we prove a Lax Wendroff theorem. The proof is valid for multidimensional versions of the scheme, even though most of the numerical illustrations in this work, on classical benchmark problems, are one-dimensional because we have easy access to the exact solution for comparison. We conclude by explaining that the method is general and can be used in different settings, for example, Finite Volume, or discontinuous Galerkin method, not just the specific one presented in this paper.

**1. Introduction.** The Euler equations of fluid dynamics are, formulated in their conservative version,

$$(1) \quad \begin{aligned} \frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) &= \mathbf{0}, \\ \frac{\partial \rho \mathbf{u}}{\partial t} + \operatorname{div}(\rho \mathbf{u} \otimes \mathbf{u} + p \mathbf{I}) &= \mathbf{0}, \\ \frac{\partial E}{\partial t} + \operatorname{div}((E + p)\mathbf{u}) &= \mathbf{0}. \end{aligned}$$

As usual,  $\rho \geq 0$  is the density,  $\mathbf{u}$  is the velocity vector,  $E = e + \frac{1}{2}\rho \mathbf{u}^2$  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 for  $p = p(\rho, e)$ . The simplest one is that of a calorically perfect gas

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

where the ratio of specific heats  $\gamma$  is constant.

When the solution is smooth, the system (1) can be equivalently written in non-conservative form as

$$(2) \quad \begin{aligned} \frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) &= \mathbf{0}, \\ \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{\nabla p}{\rho} &= \mathbf{0}, \\ \frac{\partial e}{\partial t} + \mathbf{u} \cdot \nabla e + (e + p) \operatorname{div} \mathbf{u} &= \mathbf{0}. \end{aligned}$$

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

---

\* Institute of Mathematics, University of Zürich, Zürich, Switzerland. [remi.abgrall@math.uzh.ch](mailto:remi.abgrall@math.uzh.ch)

numerical schemes applied to (1): the Lax-Wendroff theorem implies and guarantees that a suitable numerical approximation should be written in terms of flux.

However, the form (2) is better suited for engineering purposes, since one has direct access to the velocity and the internal energy. Hence a rather natural question is how to discretise the Euler equations directly from (2), and still have convergence to the correct weak solutions, at least formally. In addition to this theoretical question, there are other reasons to use the (2) system, and we list a few of them:

- In the Lagrangian hydrodynamics community (i.e. the US National Laboratories, AWE in the UK, CEA in France and their Russian and Chinese counterparts), it is very common to describe, for certain applications, the equations of compressible fluid dynamics using volume, velocity and specific internal energy. This is, for example, what is done with the Wilkins scheme, see [5], and even in the finite element version of this scheme, see [3]. Variables are represented either by cell or by point values, depending on whether they are intensive (such as velocity) or extensive (such as mass, volume and energy): this implies that thermodynamic variables are in  $L^2$  only, while velocity is globally continuous (except when we need to introduce slip surfaces).

We think it is interesting to understand how the mechanism that makes these schemes conservative can be translated into the Eulerian framework. The technique we develop in this article can be seen as the Eulerian counterpart of what was done in [1] and then [2] in the Lagrangian framework.

- In many multi-physics applications, it is not obvious whether the fully conservative formulation is the most appropriate. MHD equations are a good example. In the finite-volume community, it is customary to describe flow with density, momentum and total energy, i.e. the sum of kinetic energy, thermodynamic energy and magnetic energy. The magnetic field evolution equation is described in conservation form (using Ohm's law and Faraday's equation), and we have the constraint  $\text{div } B = 0$ . However, the natural way to write the magnetic field equation is not using the divergence operator but the curl operator, and the consequence is the preservation of the divergence involution. Merging the magnetic field and the mechanical and thermodynamic energies may be considered somewhat artificial, since we also have an evolution relation for the magnetic field equation. Hence, it may be considered interesting to separate the thermodynamics from the magnetic field. See [6] for an example of this type of approach in the Lagrangian framework. In addition, one way of preserving the structure of the Faraday equation is to use a staggered mesh. This is not necessarily done as in the present paper, but respecting local conservation of total energy may require the same kind of algebraic manipulations as here, see [7].
- When considering a mixture of gases, the most natural variable describing fluid energy is not total energy but internal energy, or even pressure. Consequently, the use of a formulation based on a non-conservative form of the system may offer certain advantages.
- One numerical strategy for simulating incompressible flows is to use a staggered mesh, for reasons of stability. It is well known in the finite volume community that when data are collocated and the Mach number tends towards 0, the behavior of the numerical scheme degrades considerably. This problem has been studied in numerous articles, [8, 9, 10] for example, and the references therein. Several strategies exist, such as preconditioning, but not only. One might expect staggering, even for the compressible case, to be

a good strategy when the Mach number tends towards 0. This was done by Herbin et al. in [11] and by Bijl et al. in [12]. The scheme must be implicit. Here we are not interested in the low Mach effect, but in conservation issues, and the extension of our work to small Mach numbers could be a possibility.

We will be using the expression "locally conservative" for a scheme. By this we mean that for finite volume schemes, discontinuous Galerkin schemes, finite differences and even those using continuous finite elements (see [13] for a *explicit* construction), each degree of freedom can be associated with a (control) volume. We say that a scheme is locally conservative if, for any sub-domain obtained by gathering such volumes, the update of the conservative variable is obtained by the contribution on the boundary of this sub-domain.

One obvious way to write a scheme on the primitive variables is to start from a locally conservative approximation of (1), and by simple algebraic manipulations which amount to multiplying the numerical scheme by approximations of

$$(3) \quad \begin{pmatrix} 1 & 0 & 0 \\ \mathbf{u} & \rho & 0 \\ \frac{\mathbf{u}^2}{2} & \rho\mathbf{u} & 1 \end{pmatrix},$$

we can obtain a scheme directly working on the primitive variables. 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, [14, 15, 16] in the Finite Volume context. In these 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 also be partially implicit, so that in the low Mach number limit the scheme "degenerates" to a Mac-type scheme, see [11]. Their schemes are second-order accurate in time and space.

In this article, we describe a different technique that allows *a priori* to achieve an arbitrary level of precision, both in time and space. This technique is not particularly designed for any specific class of scheme. The main restriction seems to be that the time scheme must be based on a sequence of Euler steps. Examples are the Runge Kutta SSP schemes, or the Defect Correction (DeC) methods in the [17] version. We have chosen to illustrate the technique on the example of residual distribution (RD) schemes where evolution over time is done by DeC, see [17]. This choice is also motivated by the fact that local conservation recovery is simpler for RD schemes (see [13]). Therefore, before describing this method, we briefly review the class of residual distribution schemes that will be the main tool we use, and sketch how dG (and therefore finite-volume) schemes can be reformulated in this framework. We then describe the scheme and explain why it is locally conservative. Finally, we show a variant of the Lax-Wendroff theorem adapted to our framework. Finally, we show how to adapt the method to finite volume schemes and discontinuous Galerkin methods. Numerical examples illustrate the soundness of the approach.

**2. A first-order nonconservative approach.** We have in mind a numerical approximations where the variables are piecewise 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 is used, inspired from [17] 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 [18] because the equivalent of the mass matrix changes at every time step. This is why a particular time stepping should be preferred, for example, the Deferred Correction (DeC) approach, see Appendix A. 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). This will be explained in section 3.3. Since the novelty of the approach lies in the correction technique, we will focus for simplicity on a single Euler forward step in time.

We consider a hyperbolic system in the form

$$(4) \quad \frac{\partial U}{\partial t} + L(U) = 0$$

on a domain  $\Omega \subset \mathbb{R}^d$ ,  $d = 1, 2, 3$ . For solving (1) or (2) we define

$$(5) \quad L(U) =: \begin{pmatrix} \operatorname{div}(\rho \mathbf{u}) \\ \operatorname{div}(\rho \mathbf{u} \otimes \mathbf{u} + p \mathbf{I}) \\ \operatorname{div}((E + p)\mathbf{u}) \end{pmatrix}$$

for the conservative form and

$$(6) \quad L(U) =: \begin{pmatrix} \operatorname{div}(\rho \mathbf{u}) \\ (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{\nabla p}{\rho} \\ \mathbf{u} \cdot \nabla \mathbf{e} + (e + p) \operatorname{div} \mathbf{u} \end{pmatrix}$$

for the non-conservative form. In what follows, we will describe the procedure for solving the equations 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 of the variables play the same role, contrarily to (1), and it is easier to start with a system with one variable. For now, we will proceed by forgetting the question of local conservation.

## 2.1. Scalar case.

**2.1.1. Trial space.** We consider a triangulation of  $\Omega$  made of non-overlapping simplices that are generically denoted by  $K$ . We assume that the triangulation is conformal, and define

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

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

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

In each element  $K$ , a polynomial is defined by a set of degrees of freedom, for example, the Lagrange points. We denote by  $\sigma$  a generic degree of freedom. Here, for

reasons that will be more clear later on, we expand the polynomials in terms of Bézier polynomials.

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

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

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

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

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

- Quadratic: The degrees of freedom  $\sigma$  are identified with the vertices  $i$ , and the mid-points  $i + \frac{1}{2}$

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

- Multidimensional elements: We only describe the 2D cases, with triangles, but similar things are obtained for quadrangles, or 3D simplices. A triangle is made of three vertices denoted by 1, 2 and 3. The barycentric coordinates with respect to the vertices 1, 2, 3 are denoted by  $\Lambda_1$ ,  $\Lambda_2$  and  $\Lambda_3$ .

- 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 three vertices  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$  as well as the midpoints of the edges:

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

The Bézier polynomials are:

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

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  $\varphi_\sigma^K$  is any of the linear, quadratic (or higher-order) functions 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 notations, we will use the second expansion throughout this paper, depending if we see  $\varphi_\sigma$  per element or more globally.

**2.1.2. Test space.** As we mentioned 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  $\sigma$ : we can identify functions of  $X_h(\Omega)$  that are indexed by the  $\sigma$  and span this space. We denote them by  $\Xi_\sigma$ . For example, in the SUPG method, we define  $\Xi_\sigma$  in each  $K$  by: for  $\mathbf{x} \in K$ ,

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

Here  $h_K$  is the diameter of  $K$  and  $\tau_K$  is a positive matrix. In [18, 19, 20, 21, 22], examples are given, where  $\Xi_\sigma$  depends on the solution, in order to get  $L^\infty$  stability. In all the examples we are considering, the support of  $\Xi_\sigma$  is that of  $\varphi_\sigma$ .

**2.1.3. Description of the time discretisation.** We start by integrating (4) which gives

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

By applying the forward Euler method per simplex, we obtain

$$(7) \quad \int_\Omega \int_{t^n}^{t^{n+1}} \Xi_\sigma \left( \frac{\partial U}{\partial t} + L(U) \right) dt d\mathbf{x} = \int_\Omega \Xi_\sigma (U^{n+1} - U^n) d\mathbf{x} + \Delta t \int_\Omega \Xi_\sigma L(U^n) d\mathbf{x} = 0.$$

The definition of  $\int_K \Xi_\sigma L(u) d\mathbf{x}$  is somewhat formal and we replace it by some approximation  $\Phi_\sigma^K(U)$  which will be defined later. The only constraint is that we have the relation

$$(8) \quad \sum_{\sigma \in K} \Phi_\sigma^K(U) = \Phi^K(U),$$

where the precise definition of  $\Phi^K(U)$  depends on whether we are dealing with the problem (4) with the  $L$  operator in conservation form  $L(U) = \operatorname{div} \mathbf{f}(U)$  as in (5) or in non conservation form  $L(U) = \mathbf{a}(U) \cdot \nabla U$  as in (6). More specifically,

- If  $L$  is in conservation form, we set

$$\Phi^K(U) := \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}} d\gamma,$$

where  $\hat{\mathbf{f}}_{\mathbf{n}}$  is a constant approximation of the flux  $\mathbf{f}$  in the direction  $\mathbf{n}$  (normal to  $\partial K$ ),

- If  $L$  is in non conservation form, we set

$$\Phi^K(U) := \int_K \mathbf{a}(U^h) \cdot \nabla U^h d\mathbf{x}$$

where a quadrature formula is employed. In fact, and in that case, the situation is slightly more complicated, because written as such, there might seem there is no coupling between elements. Since this is a case by case procedure, we give an example in section 4.

We replace the temporal terms by

$$\int_\Omega \varphi_\sigma (U^{n+1} - U^n) d\mathbf{x}$$

and then "lump" the mass matrix, set  $\int_K \varphi_\sigma dx = C_K$  (it does not depend on  $\sigma$ ) and obtain

$$(9) \quad |C_\sigma|(U^{n+1} - U^n) + \Delta t \sum_{K, \sigma \in K} \Phi_\sigma^K(U^n) = 0.$$

We note that this is the reason why we use a Bézier approximation since we are sure that the lumped mass is non zero because it holds

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

This scheme corresponds to the  $\mathcal{L}^1$  operator of the DeC procedure described in appendix A. It will be high order in time and space, provided some conditions described in appendix A are fulfilled. We stick to this, to avoid useless complications, and also because its form is that of an Euler forward method. Hence our discussion becomes valid for any algorithm that can be put in the form (9) with  $\Phi_\sigma^K$  satisfying (8).

Let us give an other example: the discontinuous Galerkin method in the conservative setting. Using basis functions  $\{\varphi^\sigma\}$  that are now see as polynomial in each  $K$  but only in  $L^2$ , we have for any  $\sigma \in K$

$$\int_K \varphi_\sigma \frac{\partial U}{\partial t} d\mathbf{x} - \int_K \nabla \varphi_\sigma \cdot \mathbf{f}(U) d\mathbf{x} + \int_{\partial K} \varphi_\sigma \hat{\mathbf{f}}_{\mathbf{n}} d\gamma = 0$$

and we set

$$\Phi_\sigma^K := - \int_K \nabla \varphi_\sigma \cdot \mathbf{f}(U) + \int_{\partial K} \varphi_\sigma \hat{\mathbf{f}}_{\mathbf{n}} d\gamma.$$

Since in  $K$ ,  $\sum_{\sigma \in K} \varphi_\sigma = 1$ , we have (8). The non conservative setting works *formally* similarly, provided a case by case strategy is again adopted.

**2.2. Case of system (2).** We describe the residuals and develop the method as before for simplicity for the forward Euler method in time. But as mentioned before, it can be extended to higher orders in a straightforward way.

We assume that the computational domain  $\Omega$  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  $\{\varphi_{\sigma_{\mathcal{V}}}\}_{\sigma_{\mathcal{V}} \in D_{\mathcal{V}}}$ , 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 thermodynamic 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 (or linear) Bernstein elements, while the thermodynamic space  $\mathcal{E}$  has a piecewise-linear (or piece-wise constat) basis. The velocity field is approximated by

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

where the  $\varphi_{\sigma_{\mathcal{V}}}$  are the linear/quadratic (or linear) Bézier polynomials, and the density, the pressure and the internal energy, are given by

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

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

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

$$\frac{\partial U}{\partial t} + \mathbf{a}(U) \cdot \nabla U = 0.$$

The only thing to do is to describe how the method of the previous section adapts to this case, and this amounts to describing the general structure residuals, that is how (8) is written. Since the velocity is globally continuous, we write

$$\Phi^{K,\mathbf{u}} = \int_K \left( \mathbf{u} \otimes \mathbf{u} + \frac{\nabla p}{\rho} \right) d\mathbf{x},$$

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

For the density, the evolution equation is in conservation form and we use a numerical flux  $\hat{\mathbf{f}}$ :

$$\Phi^{K,\rho} = \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}} d\gamma.$$

Here, any consistent numerical flow can be used a priori. Of course, the stability of the method depends on this choice, but the conservation properties of the method do not.

Last, for the internal energy, we write

$$\Phi^{K,E} = \int_K \left( \mathbf{u} \cdot \nabla e + (e + p) \operatorname{div} \mathbf{u} \right) d\mathbf{x}$$

and again we use a 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 first-order case, but the extension to the more general case is straightforward. In the appendix B, 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 conservative form.

Nothing has to be done for the density since it is already in conservative 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 denote with  $\psi_K^n$  the value of  $\psi$  at time  $t_n$  at the centroid of  $K$ , and consider the following approximation of  $\psi$  that we still denote by  $\psi$ :

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

Then we consider

$$(10) \quad \begin{aligned} \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} &= \sum_K \psi_K^n \int_K (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} \\ &= \sum_K \psi_K^n \left[ \int_K \rho^{n+1} (\mathbf{u}^{n+1} - \mathbf{u}^n) d\mathbf{x} + \int_K \mathbf{u}^n (\rho^{n+1} - \rho^n) d\mathbf{x} \right]. \end{aligned}$$

Introducing  $\Delta \mathbf{u}_{\sigma_\nu} := \mathbf{u}_{\sigma_\nu}^{n+1} - \mathbf{u}_{\sigma_\nu}^n$  and  $\Delta \rho_{\sigma_\varepsilon} := \rho_{\sigma_\varepsilon}^{n+1} - \rho_{\sigma_\varepsilon}^n$ , we can write

$$\int_K \rho^{n+1} (\mathbf{u}^{n+1} - \mathbf{u}^n) d\mathbf{x} = \sum_{\sigma_\nu \in K} \Delta \mathbf{u}_{\sigma_\nu} \int_K \rho^{n+1} \varphi_{\sigma_\nu} d\mathbf{x}$$

and

$$\int_K \mathbf{u}^n (\rho^{n+1} - \rho^n) d\mathbf{x} = \sum_{\sigma_\varepsilon \in K} \Delta \rho_{\sigma_\varepsilon} \int_K \mathbf{u}^n \varphi_{\sigma_\varepsilon} d\mathbf{x}.$$

Hence, (10) can be rewritten as:

$$(11) \quad \begin{aligned} &\int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} \\ &= \sum_K \psi_K^n \left[ \sum_{\sigma_\nu \in K} \Delta \mathbf{u}_{\sigma_\nu} \int_K \rho^{n+1} \varphi_{\sigma_\nu} d\mathbf{x} + \sum_{\sigma_\varepsilon \in K} \Delta \rho_{\sigma_\varepsilon} \int_K \mathbf{u}^n \varphi_{\sigma_\varepsilon} d\mathbf{x} \right] \\ &= \sum_K \psi_K^n \left[ \sum_{\sigma_\nu \in K} \omega_{\sigma_\nu}^{\rho, n+1, K} |C_{\sigma_\nu}| \Delta \mathbf{u}_{\sigma_\nu} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} |C_{\sigma_\varepsilon}| \Delta \rho_{\sigma_\varepsilon} \right] \\ &= \sum_K \left[ \sum_{\sigma_\nu \in K} \psi_{\sigma_\nu} \omega_{\sigma_\nu}^{\rho, n+1, K} |C_{\sigma_\nu}| \Delta \mathbf{u}_{\sigma_\nu} \right] \\ &\quad + \sum_K \sum_{\sigma_\nu} (\psi_K^n - \psi_{\sigma_\nu}) |C_{\sigma_\nu}| \omega_{\sigma_\nu}^{\rho, n+1} \Delta \mathbf{u}_{\sigma_\nu} + \sum_K \psi_K^n \left[ \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} |C_{\sigma_\varepsilon}| \Delta \rho_{\sigma_\varepsilon} \right] \\ &= \sum_{\sigma_\nu} \psi_{\sigma_\nu}^n \omega_{\sigma_\nu}^{\rho, n+1} |C_{\sigma_\nu}| \Delta \mathbf{u}_{\sigma_\nu} + \sum_K \sum_{\sigma_\nu \in K} (\psi_K^n - \psi_{\sigma_\nu}) |C_{\sigma_\nu}| \omega_{\sigma_\nu}^{\rho, n+1} \Delta \mathbf{u}_{\sigma_\nu} \\ &\quad + \sum_K \psi_K^n \left[ \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} |C_{\sigma_\varepsilon}| \Delta \rho_{\sigma_\varepsilon} \right] \end{aligned}$$

where we have set for simplicity

$$\omega_{\sigma_\nu}^{\rho, n+1, K} := \frac{\int_K \rho^{n+1} \varphi_{\sigma_\nu} d\mathbf{x}}{|C_{\sigma_\nu}|}, \quad \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} := \frac{\int_K \mathbf{u}^n \varphi_{\sigma_\varepsilon} d\mathbf{x}}{|C_{\sigma_\varepsilon}|}$$

and (using that the support of  $\varphi_{\sigma_\nu}$  is the union of the elements that share  $\sigma_\nu$ ),

$$\omega_{\sigma_\nu}^{\rho, n+1} := \frac{\int_\Omega \rho^{n+1} \varphi_{\sigma_\nu} d\mathbf{x}}{|C_{\sigma_\nu}|}$$

For the velocity, we have:

$$|C_{\sigma_\nu}| (\mathbf{u}_{\sigma_\nu}^{n+1} - \mathbf{u}_{\sigma_\nu}^n) + \Delta t_n \sum_{K, \sigma_\nu \in K} \Phi_{\sigma_\nu, K}^{\mathbf{u}} = 0,$$

where, for the forward Euler scheme (7),

$$\Phi_{\sigma_{\nu},K}^{\mathbf{u}} = \Phi_{\sigma_{\nu},K}^{\mathbf{u}}(U^n).$$

For the density, we have

$$|C_{\sigma_{\varepsilon}}|(\rho_{\sigma_{\varepsilon}}^{n+1} - \rho_{\sigma_{\varepsilon}}^n) + \Delta t_n \sum_{K, \sigma_{\varepsilon} \in K} \Phi_{\sigma_{\varepsilon},K}^{\rho} = 0$$

and we note that the sum reduces to one term, hence

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

where again

$$\Phi_{\sigma_{\varepsilon},K}^{\rho} = \Phi_{\sigma_{\varepsilon},K}^{\rho}(U^n)$$

and  $K$  is *the* element such that  $\sigma_{\varepsilon} \in K$ . Using these relations in (11), we get

$$\begin{aligned} & \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} + \Delta t_n \underbrace{\sum_{\sigma_{\nu}} \psi_{\sigma_{\nu}}^n \omega_{\sigma_{\nu}}^{\rho, n+1} \left[ \sum_{K, \sigma_{\nu} \in K} \Phi_{\sigma_{\nu},K}^{\mathbf{u}} \right]}_I \\ (12) \quad & + \Delta t_n \sum_K \left\{ \sum_{\sigma_{\nu} \in K} (\psi_K^n - \psi_{\sigma_{\nu}}) \omega_{\sigma_{\nu}}^{\rho, n+1, K} \left[ \sum_{K', \sigma_{\nu} \in K' \cap K} \Phi_{\sigma_b, K}^{\mathbf{u}} \right] \right\} \\ & + \Delta t_n \sum_K \psi_K^n \left[ \sum_{\sigma_{\varepsilon} \in K} \omega_{\sigma_{\varepsilon}}^{\mathbf{u}, n, K} \Phi_{\sigma_{\varepsilon}, K}^{\rho} \right] \\ & = 0 \end{aligned}$$

The term  $I$  can be rewritten as

$$\begin{aligned} & \sum_{\sigma_{\nu}} \psi_{\sigma_{\nu}}^n \omega_{\sigma_{\nu}}^{\rho, n+1} \left[ \sum_{K, \sigma_{\nu} \in K} \Phi_{\sigma_{\nu}, K}^{\mathbf{u}} \right] = \sum_K \psi_K^n \sum_{\sigma_{\nu} \in K} \omega_{\sigma_{\nu}}^{\rho, n+1} \Phi_{\sigma_{\nu}, K}^{\mathbf{u}} \\ & + \sum_K \left[ \sum_{\sigma_{\nu} \in K} (\psi_K^n - \psi_{\sigma_{\nu}}) \omega_{\sigma_{\nu}}^{\rho, n+1} \Phi_{\sigma_{\nu}, K}^{\mathbf{u}} \right] \end{aligned}$$

Hence, gathering all together, we get

$$\begin{aligned} & \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} \\ & + \Delta t_n \sum_K \psi_K^n \left[ \sum_{\sigma_{\nu} \in K} \omega_{\sigma_{\nu}}^{\rho, n+1} \Phi_{\sigma_{\nu}, K}^{\mathbf{u}} \sum_{\sigma_{\varepsilon} \in K} \omega_{\sigma_{\varepsilon}}^{\mathbf{u}, n, K} \Phi_{\sigma_{\varepsilon}, K}^{\rho} \right] \\ & + \Delta t_n \sum_K \left[ \sum_{\sigma_{\nu} \in K} (\psi_K^n - \psi_{\sigma_{\nu}}) \omega_{\sigma_{\nu}}^{\rho, n+1} \Phi_{\sigma_{\nu}, K}^{\mathbf{u}} \right] \\ & + \Delta t_n \sum_K \left\{ \sum_{\sigma_{\nu} \in K} (\psi_K^n - \psi_{\sigma_{\nu}}) \omega_{\sigma_{\nu}}^{\rho, n+1, K} \left[ \sum_{K', \sigma_{\nu} \in K' \cap K} \Phi_{\sigma_b, K}^{\mathbf{u}} \right] \right\} = 0 \end{aligned}$$

Thus, we obtain the master equation:

$$(13a) \quad \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} \\ + \Delta t \sum_K \psi_K^n \left[ \sum_{\sigma_\nu \in K} \omega_{\sigma_\nu}^{\rho, n+1} \Phi_{\sigma_\nu, K}^{\mathbf{u}} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} \Phi_{\sigma_\varepsilon, K}^\rho \right] \\ + \Delta t_n \sum_K \left( F_K^{\mathbf{m}} + D_K^{\mathbf{m}} \right) = 0$$

with

$$(13b) \quad F_K^{\mathbf{m}} = \sum_{\sigma_\nu \in K} (\psi_K^n - \psi_{\sigma_\nu}^n) \omega_{\sigma_\nu}^{\rho, n+1} \Phi_{\sigma_\nu, K}^{\mathbf{u}} \\ D_K^{\mathbf{m}} = \sum_{\sigma_\nu \in K} (\psi_K^n - \psi_{\sigma_\nu}^n) \omega_{\sigma_\nu}^{\rho, n+1, K} \left[ \sum_{K', \sigma_\nu \in K' \cap K} \Phi_{\sigma_b, K}^{\mathbf{u}} \right] \\ \omega_{\sigma_\nu}^{\rho, n+1, K} = \frac{\int_K \rho^{n+1} \varphi_{\sigma_\nu} d\mathbf{x}}{|C_{\sigma_\nu}|}, \quad \omega_{\sigma_\nu}^{\rho, n+1} = \sum_{K, \sigma_\nu \in K} \omega_{\sigma_\nu}^{\rho, n+1, K} \\ \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} = \frac{\int_K \mathbf{u}^n \varphi_{\sigma_\varepsilon} d\mathbf{x}}{|C_{\sigma_\varepsilon}|}.$$

Let us now consider the total energy. First, we remark that (with similar notations as before) the following holds:

$$\Delta(\rho \mathbf{u}^2) = \mathbf{u}^{n+1} \cdot \Delta(\rho \mathbf{u}) + \rho^n \mathbf{u}^n \cdot \Delta \mathbf{u}.$$

Combined with

$$\Delta(\rho \mathbf{u}) = \rho^{n+1} \Delta \mathbf{u} + \mathbf{u}^n \Delta \rho$$

we obtain

$$\Delta(\rho \mathbf{u}^2) = (\rho^{n+1} \mathbf{u}^{n+1} + \rho^n \mathbf{u}^n) \cdot \Delta \mathbf{u} + \mathbf{u}^{n+1} \cdot \mathbf{u}^n \Delta \rho.$$

To simplify, we will set

$$\tilde{\mathbf{m}} = \frac{\rho^{n+1} \mathbf{u}^{n+1} + \rho^n \mathbf{u}^n}{2}, \quad \tilde{q}^2 = \mathbf{u}^{n+1} \cdot \mathbf{u}^n.$$

Using these relations, we see that

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

First, we notice that

$$\int_K \Delta e d\mathbf{x} = -\Delta t \sum_{\sigma_\varepsilon \in K} \Phi_{\sigma_\varepsilon, K}^e.$$

Introducing

$$\theta_{\sigma_\nu}^{\mathbf{m},K} = \frac{\int_K \tilde{\mathbf{m}} \varphi_{\sigma_\nu} dx}{|C_{\sigma_\nu}|} \quad \text{and} \quad \theta_{\sigma_\varepsilon}^{q^2,K} = \frac{\int_K \tilde{q}^2 \varphi_{\sigma_\varepsilon} dx}{|C_{\sigma_\varepsilon}|},$$

we get

$$\sum_{\sigma_\nu \in K} \Delta \mathbf{u}_{\sigma_\nu} \cdot \int_K \tilde{\mathbf{m}} \varphi_{\sigma_\nu} dx = -\Delta t \sum_{\sigma_\nu \in K} \theta_{\sigma_\nu}^{\mathbf{m},K} \cdot \left( \sum_{K', \sigma_\nu \in K'} \Phi_{\sigma_\nu, K'}^{\mathbf{u}} \right)$$

and because  $\sigma_\varepsilon$  belongs to a single element we have

$$\sum_{\sigma_\varepsilon \in K} \Delta \rho_{\sigma_\varepsilon} \int_K \tilde{q}^2 \varphi_{\sigma_\varepsilon} dx = -\Delta t \sum_{\sigma_\varepsilon \in K} \theta_{\sigma_\varepsilon}^{q^2,K} \Phi_{\sigma_\varepsilon, K}^\rho.$$

Then proceeding as for the velocity, and introducing

$$\theta_{\sigma_\nu}^{\mathbf{m}} = \frac{\sum_{K, \sigma_\nu \in K} \int_K \tilde{\mathbf{m}} \varphi_{\sigma_\nu} dx}{|C_{\sigma_\nu}|} = \sum_{K, \sigma_\nu \in K} \theta_{\sigma_\nu}^{\mathbf{m},K},$$

we get

$$\begin{aligned} & \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (E^{n+1} - E^n) dx \\ & + \Delta t_n \sum_K \psi_K \left( \sum_{\sigma_\varepsilon \in K} \Phi_{\sigma_\varepsilon, K}^e + \sum_{\sigma_\nu \in K} \theta_{\sigma_\nu}^{\mathbf{m}} \cdot \Phi_{\sigma_\nu, K}^{\mathbf{u}} + \frac{1}{2} \sum_{\sigma_\varepsilon \in K} \theta_{\sigma_\varepsilon}^{q^2,K} \Phi_{\sigma_\varepsilon, K}^\rho \right) \\ (14) \quad & + \Delta t_n \underbrace{\sum_K \left[ \sum_{\sigma_\nu} (\psi_K^n - \psi_{\sigma_\nu}^n) \theta_{\sigma_\nu}^{\mathbf{m},K} \cdot \left\{ \sum_{K', \sigma_\nu \in K' \cap K} \Phi_{\sigma_\nu, K'}^{\mathbf{u}} \right\} \right]}_{D_K^E} \\ & + \Delta t_n \underbrace{\sum_K \sum_{\sigma_\nu \in K} (\psi_{\sigma_\nu} - \psi_K^n) \theta_{\sigma_\nu}^{\mathbf{m},K} \cdot \Phi_{\sigma_\nu, K}^{\mathbf{u}}}_{:= F_K^E} = 0. \end{aligned}$$

As it is customary, we say that a family of meshes is shape regular if there exists  $\alpha > 0$  depending only on this family such that the ratio of the inner and outer diameters of any element of any mesh of this family is greater than  $\alpha$ . We show the following result in Appendix B:

**PROPOSITION 1.** *Assume that the mesh  $\mathcal{T}_h$  is shape regular, we denote by  $h$  the maximum diameter of the element of the mesh. For any  $K$ , the residuals  $\Phi_{\sigma_\varepsilon, K}^\rho$ ,  $\Phi_{\sigma_\varepsilon, K}^e$ ,  $\Phi_{\sigma_\nu, K}^{\mathbf{u}}$  are Lipschitz continuous functions of their arguments, with Lipschitz constant of the form  $C \cdot h$ , where  $C$  only depends on  $\alpha$  and the maximum norm of the solution.*

*Assume that we have a family of meshes  $\mathcal{F} = \{\mathcal{T}_{h_n}\}$  with  $\lim_{n \rightarrow +\infty} h_n = 0$ . We denote by  $(U_{h_n})_{n \geq 0}$  the sequence of functions fulfilling:*

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

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

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

and

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

Here  $\{(\rho_{\sigma_\varepsilon}^n), (\mathbf{u}_{\sigma_\nu}^n), (e_{\sigma_\varepsilon}^n)\}_{n \geq 0, \sigma_\varepsilon, \sigma_\nu}$  are defined by the introduced 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  $\mathbf{u} \in (L^2(\mathbb{R}^d \times [0, T]))^d$ .

We also assume that the residuals satisfy

$$(15) \quad \sum_{\sigma_\nu \in K} \omega_{\sigma_\nu}^{\rho, n+1} \Phi_{\sigma_\nu, K}^{\mathbf{u}} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} \Phi_{\sigma_\varepsilon, K}^\rho = \int_{\partial K} \mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n} \, d\gamma$$

and

$$(16) \quad \sum_{\sigma_\varepsilon \in K} \Phi_{\sigma_\varepsilon, K}^e + \sum_{\sigma_\nu \in K} \theta_{\sigma_\nu}^{\mathbf{m}} \cdot \Phi_{\sigma_\nu, K}^{\mathbf{u}} + \frac{1}{2} \sum_{\sigma_\varepsilon} \theta_{\sigma_\varepsilon}^{q^2, K} \Phi_{\sigma_\varepsilon} = \int_{\partial K} \mathbf{f}^E(U^n) \cdot \mathbf{n} \, d\gamma,$$

where we have set

$$(17) \quad \begin{aligned} \omega_{\sigma_\nu}^{\rho, n+1} &= \frac{\sum_{K, \sigma_\nu \in K} \int_K \rho^{n+1} \varphi_{\sigma_\nu} \, d\mathbf{x}}{|C_{\sigma_\nu}|}, & \omega_{\sigma_\nu}^{\mathbf{u}, n, K} &= \frac{\int_K \mathbf{u}^n \varphi_{\sigma_\varepsilon} \, d\mathbf{x}}{|C_{\sigma_\varepsilon}|}, \\ \tilde{\mathbf{m}} &= \frac{\rho^{n+1} \mathbf{u}^{n+1} + \rho^n \mathbf{u}^n}{2}, & \tilde{q}^2 &= \mathbf{u}^{n+1} \cdot \mathbf{u}^n, \\ \theta_{\sigma_\nu}^{\mathbf{m}} &= \frac{\sum_{K, \sigma_\nu \in K} \int_K \tilde{\mathbf{m}} \varphi_{\sigma_\nu} \, d\mathbf{x}}{|C_{\sigma_\nu}|}, & \theta_{\sigma_\varepsilon}^{q^2, K} &= \frac{\int_K \tilde{q}^2 \varphi_{\sigma_\varepsilon} \, d\mathbf{x}}{|C_{\sigma_\varepsilon}|} \end{aligned}$$

with the assumption that there exists  $C$  independent of  $n$ , such that  $\Delta t \leq Ch$ . In (15) (resp. (16)),  $\mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n}$  (resp.  $\mathbf{f}^E(U^n) \cdot \mathbf{n}$ ) is the momentum component of the normal flux (resp. its energy component).

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 dependency of the residuals with respect to the element.

Given a set of residuals that satisfy (8) also satisfy (15) and (16). In this section, we will show how to slightly modify the original scheme so that the new one will satisfy (8), (15) and (16), and hence if the scheme converges, we have convergence towards a weak solution. To achieve this, following [1, 23, 24], we introduce the correction terms in the residuals. This needs to be done only for the velocity and the internal energy.

Knowing at time  $t_n$  the solution  $(\rho^n, \mathbf{u}^n, e^n)$  we obtain with the forward Euler step  $(\rho^{n+1}, \mathbf{u}^{n+1}, e^{n+1})$ . For this, we first compute  $\rho^{n+1}$  and then perform the update for the velocity and the energy:

#### Momentum.

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

$$(18) \quad \Psi_{\sigma_\nu}^{\mathbf{u}} = \Phi_{\sigma_\nu}^{\mathbf{u}}(U^n) + r_{\sigma_\nu}^{\mathbf{u}}$$

is such that (15) holds true for the new set of residuals, i.e.

$$\sum_{\sigma_V \in K} \omega_{\sigma_V}^{\rho, n+1} r_{\sigma_V}^{\mathbf{u}} = \int_{\partial K} \mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n} \, d\gamma - \left\{ \sum_{\sigma_V \in K} \omega_{\sigma_V}^{\rho, n+1} \Phi_{\sigma_V, K}^{\mathbf{u}} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} \Phi_{\sigma_\varepsilon, K}^\rho \right\}.$$

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

$$\sum_{\sigma_V \in K} \omega_{\sigma_V}^{\rho, n+1} > 0$$

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

$$(19) \quad \left( \sum_{\sigma_V \in K} \omega_{\sigma_V}^{\rho, n+1} \right) r^{\mathbf{u}} = \int_{\partial K} \mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n} \, d\gamma - \left\{ \sum_{\sigma_V \in K} \omega_{\sigma_V}^{\rho, n+1} \Phi_{\sigma_V, K}^{\mathbf{u}} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\varepsilon}^{\mathbf{u}, n, K} \Phi_{\sigma_\varepsilon, K}^\rho \right\}.$$

Once this is known, we can update the velocity and compute  $\mathbf{u}_{\sigma_V}^{n+1}$ .

### Energy.

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

$$\Psi_{\sigma_\varepsilon}^e = \Phi_{\sigma_\varepsilon}^e(U^n) + r_{\sigma_\varepsilon}^e$$

to satisfy (16), we simply need:

$$(20) \quad \sum_{\sigma_\varepsilon \in K} r_{\sigma_\varepsilon}^e = \int_{\partial K} \mathbf{f}^E(U^n) \cdot \mathbf{n} \, d\gamma - \left\{ \sum_{\sigma_\varepsilon \in K} \Phi_{\sigma_\varepsilon, K}^e + \sum_{\sigma_V \in K} \theta_{\sigma_V}^{\mathbf{m}} \cdot \Phi_{\sigma_V, K}^{\mathbf{u}} + \frac{1}{2} \sum_{\sigma_\varepsilon} \theta_{\sigma_\varepsilon}^{q^2, K} \Phi_{\sigma_\varepsilon} \right\}.$$

Since there is no reason to favour one degree of freedom with respect to the other ones, we take  $r_{\sigma_\varepsilon}^e = r^e$ , and again we can explicitly solve the equation and obtain the energy at the new time instance.

**3.3. Modifications for other schemes.** We have presented this conservation recovery method using a class of schemes that might seem a bit narrow. In this section, we want to explain that it is not the case. This can apply to more general schemes, as soon as the update of any variable  $w$  (density, velocity, energy), described by degrees of freedom  $\sigma$  (point values, averages, moments), can be written as:

$$\sum_{K, \sigma \in K} \Phi_\sigma^K.$$

*About accuracy:* The calculations made for the first order in time can be immediately extended to the higher accuracy ones in time. One just has to add a temporal contribution into the new residuals, see for example [17] for more details. In Appendix A we briefly present a straightforward choice to obtain a high order accurate approximation, the Deferred Correction (DeC) approach which was used for the numerical results in Section 4.2.

We also see that the exact form of the residuals is never used, so this can also be extended to any type of residuals, including for high order ones as in [17] or [25]. We also note that we have never used the global continuity of the velocity: instead of

using  $(W_h)^2$  for the velocity, we could have used  $(V_h)^2$  in a discontinuous Galerkin like spirit.

The coefficients of (17) can be computed with any quadrature formula provided that the geometrical location of the quadrature points needed to evaluate the boundary integrals depend only on the faces and not the element, so that the edge contribution will sum up to zero. However, in the calculations we have always used enough points so that the quadrature formula are exact for the polynomial degree that we need. It is only to test global conservation that we have also used non exact quadrature formula.

**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  $K = K_{j+1/2}$ , for  $j \in \mathbb{Z}$ .

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

$$U_{j+1/2}^+ = \lim_{x \rightarrow x_{j+1/2}, x > x_{j+1/2}} (\rho, \rho \mathbf{u}, e + \frac{1}{2} \rho \mathbf{u}^2)(x)$$

$$U_{j+1/2}^- = \lim_{x \rightarrow x_{j+1/2}, x < x_{j+1/2}} (\rho, \rho \mathbf{u}, e + \frac{1}{2} \rho \mathbf{u}^2)(x).$$

Here  $\rho(x)$ ,  $\mathbf{u}(x)$  and  $e(x)$  are obtained from the approximation space. The flux  $\hat{\mathbf{f}}$  has a  $\rho$  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 a solver constructed from an approximate Riemann solver because it appears we need intermediate states in the present description of the residual, see above: this allows to couple the neighbouring cells. This could be the exact solver, we have used the HLLC one. Both provide solutions with the same success, the HLLC is easier to generalise.

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

For simplicity, we reduce the formal time accuracy to second order, and we only need to describe the spatial terms:  $\Phi_{\sigma_\varepsilon}^\rho$  for the density,  $\Phi_{\sigma_\varepsilon}^e$  for the energy and  $\Phi_{\sigma_\varepsilon}^u$  for the velocity. The update of the density is done by the dG scheme:

$$(21) \quad \Phi_{\sigma_\varepsilon}^\rho = - \int_{K_{j+1/2}} \nabla \varphi_{\sigma_\varepsilon} \mathbf{f}^\rho \, d\mathbf{x} + \left( \hat{\mathbf{f}}_{j+1/2}^\rho \varphi_{\sigma_\varepsilon}(x_{j+1/2}) - \hat{\mathbf{f}}_{j-1/2}^\rho \varphi_{\sigma_\varepsilon}(x_{j-1/2}) \right).$$

The update of the velocity is done by considering the centered residual:

$$(22) \quad \rho_K^* \Psi_{\sigma_\nu}^u = \int_K \varphi_{\sigma_\nu} \rho \mathbf{u} \frac{\partial \mathbf{u}}{\partial x} \, d\mathbf{x} - \int_K p \frac{\partial \varphi_{\sigma_\nu}}{\partial x} \, d\mathbf{x} + \int_{\partial K} p^* \varphi_{\sigma_\nu} \, d\gamma,$$

where  $p^*$  is the pressure evaluated at the quadrature points by the Riemann solver (this is why we have chosen HLLC), and  $\rho_K^*$  is the average of the density in  $K$ .

Inspired by what is done in the RD context, and by [26], we may need to consider a jump term for the velocity only because it is globally continuous. We have taken

$$(23) \quad J_\sigma^K = \theta_K \beta_K h_K^2 \int_{\partial K} [\nabla \varphi_{\sigma_\nu}] [\nabla \mathbf{u}] d\gamma.$$

$\theta_K$  is a parameter (set to 0.1 in the experiments),  $\beta_K$  is an upper bound of the wave speeds in  $K$  and a Local Lax Friedrich dissipation term

$$(24) \quad D_\sigma^K = \alpha_K (\mathbf{u}_{\sigma_\nu} - \bar{\mathbf{u}})$$

where  $\alpha_K$  is an upper bound of the wave speeds in  $K$  and  $\bar{\mathbf{u}}$  is the arithmetic average of the velocity within  $K$ . In practice, we will take either the residual

$$(25a) \quad \Psi_{\sigma_\nu}^{\mathbf{u}} = \Psi_{\sigma_\nu}^{\mathbf{u}} + D_\sigma^K$$

which will leads to a first order scheme or

$$(25b) \quad \Psi_{\sigma_\nu}^{\mathbf{u}} = \Psi_{\sigma_\nu}^{\mathbf{u}} + J_\sigma^K$$

which will be higher order and stable.

The update of the internal energy is simply done by

$$(26) \quad \Phi_{\sigma_\varepsilon}^e = \int_K \varphi_{\sigma_\varepsilon} \left( \mathbf{u} \cdot \nabla e + (e + p) \operatorname{div} \mathbf{u} \right) d\mathbf{x}$$

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:

- Procedure 1: We use the residuals (21) and (26) for the thermodynamic variables, and for the velocity, we replace  $\Phi_{\sigma_\nu}^{\mathbf{u}}$  by  $(\Phi_{\sigma_\nu}^{\mathbf{u}})^*$  defined by:

1. Compute  $\Phi^{\mathbf{u}} = \sum_{\sigma_\nu} \Phi_{\sigma_\nu}^{\mathbf{u}}$ .
2. If  $\|\Phi^{\mathbf{u}}\| > 0$ , define

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

and

$$(\Phi_{\sigma_\nu}^{\mathbf{u}})^* = \frac{x_{\sigma_\nu}}{\sum_{\sigma_\nu \in K_{j+1/2}} x_{\sigma_\nu}} \Phi^{\mathbf{u}}.$$

3. Else  $(\Phi_{\sigma_\nu}^{\mathbf{u}})^* = 0$ .

- Procedure 2: We do the same as before for  $\Phi_{\sigma_\nu}^{\mathbf{u}}$ ,  $\Phi_{\sigma_\varepsilon}^\rho$  and  $\Phi_{\sigma_\varepsilon}^e$ , where the thermodynamic residuals are now:

$$\Phi_{\sigma_\varepsilon}^\rho + \alpha_K (\rho_{\sigma_\varepsilon} - \bar{\rho}_K) \quad \text{and} \quad \Phi_{\sigma_\varepsilon}^e + \alpha_K (e_{\sigma_\varepsilon} - \bar{e}_K),$$

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

We may also need to add a jump term of the type (23) on all the variables (we have not done this here. We refer the reader to [17] for more details, and in particular why formal accuracy is increased. This procedure will lead, in practice, to a scheme that preserves the positivity of the density and the pressure.

In the experiments where we want to test the accuracy of the method, we will use the combination (21), (25b), (26).

**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 the 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  $\gamma = 1.4$ . All the solutions are displayed with 1000 points: all the examples are shock tube problems where the exact solution is known, so we can show the convergence of the method to the exact solution. This is not needed for other purposes, such as stability or other reasons. We have also computed the solution with a more reasonable number of points, the results are similar to what could be obtained with more classical methods when the correction is activated. We do not show them here to lower the number of plots.

**4.2.1. A smooth case.** The purpose is to test accuracy. We test the accuracy of our scheme (with corrections) on a smooth isentropic flow problem similar to the test case introduced in [27]. The initial data for our test problem is the following:

$$\rho_0(x) = 1 + 0.9 \sin(2\pi x), \quad u_0(x) = 0, \quad p_0(x) = \rho^\gamma(x, 0), \quad x \in [-1, 1].$$

with polytropic index  $\gamma = 3$  and periodic boundary conditions.

The exact density and velocity in this case can be obtained by the method of characteristics and is explicitly given by

$$\rho(x, t) = \frac{1}{2}(\rho_0(x_1) + \rho_0(x_2)), \quad u(x, t) = \sqrt{3}(\rho(x, t) - \rho_0(x_1)),$$

where for each coordinate  $x$  and time  $t$  the values  $x_1$  and  $x_2$  are solutions of the nonlinear equations

$$\begin{aligned} x + \sqrt{3}\rho_0(x_1)t - x_1 &= 0, \\ x - \sqrt{3}\rho_0(x_2)t - x_2 &= 0. \end{aligned}$$

The errors ( $L^1$  only is plotted because all the others shows a similar behaviour) is displayed in figure 1 The errors are somehow between second and first order: the

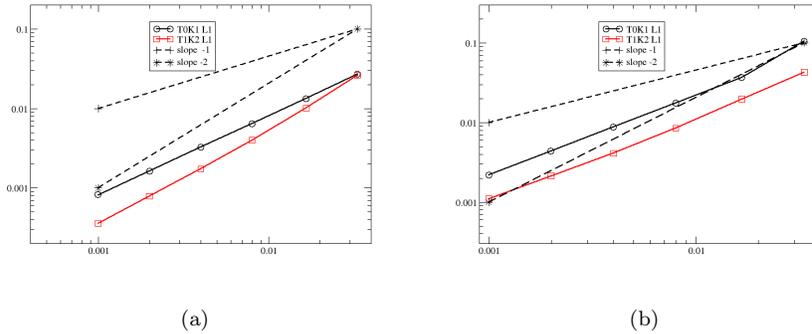


FIG. 1. (a) error on the density, (b) error on the velocity

scheme in time is second order at most in this implementation, this is a choice, we could have used a third order scheme.

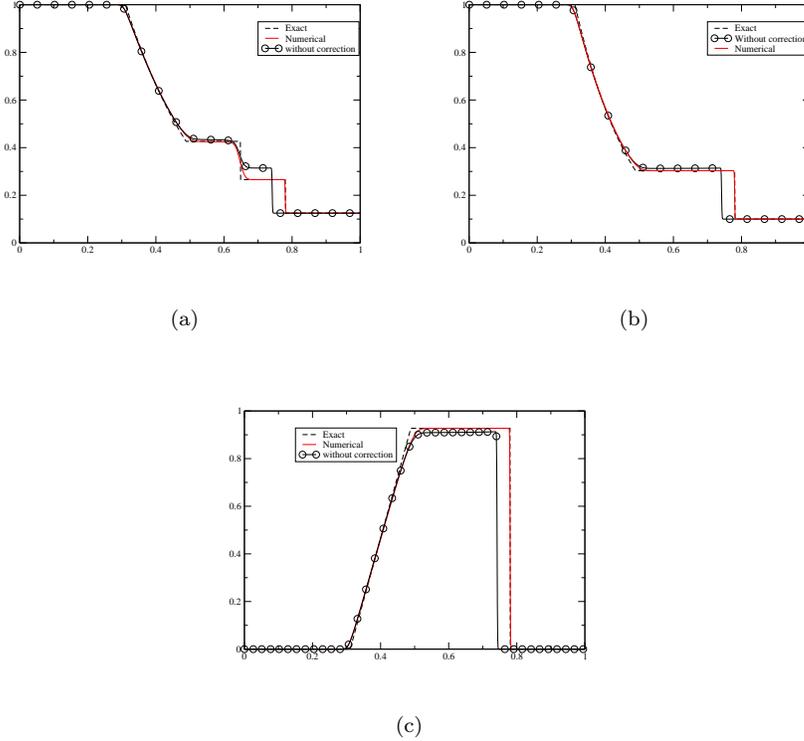


FIG. 2. Solution of the Sod shock tube problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.16$  with  $CFL = 0.4$ . In each figure, the exact, numerical as well as the solution without correction are depicted.

**4.2.2. 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: a left rarefaction wave, a contact discontinuity, and a right shock wave. This test case is used 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:

$$(27) \quad (\rho_0, u_0, p_0) = \begin{cases} (1.0, 0.0, 1.0), & \text{if } x < 0.5, \\ (0.125, 0.0, 0.1), & \text{if } x > 0.5. \end{cases}$$

In Figure 2, profiles of density, velocity and pressure are depicted with a reference solution for a mesh containing 1000 cells. We also have plotted the solution obtained without any correction. Both have been obtained with the TOK1 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.

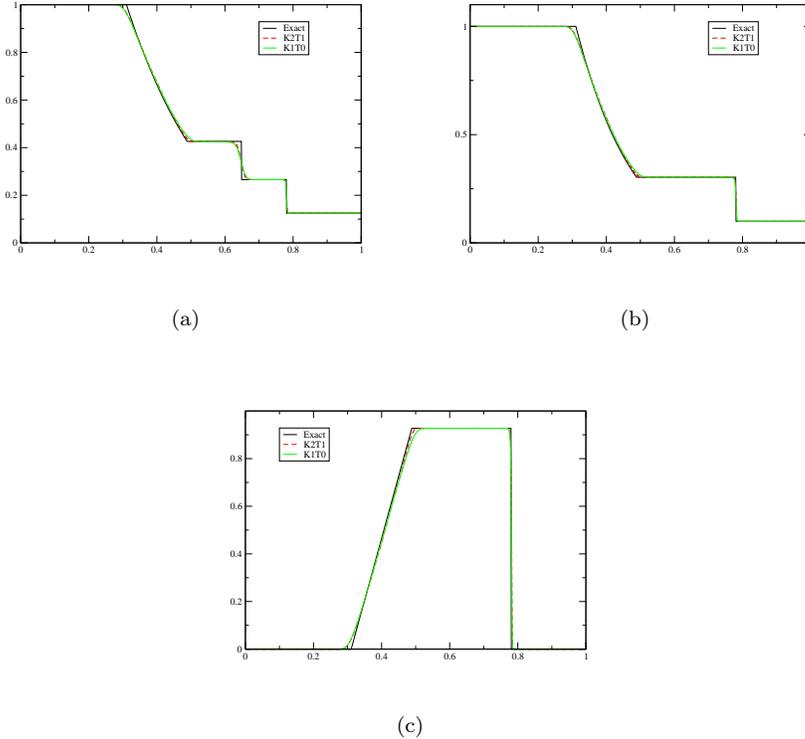


FIG. 3. Solution of the Sod shock tube problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.16$  with  $CFL = 0.4$ . In each figure, the exact solution and the solutions obtained with  $K2T1$  and  $K1T0$  on a mesh with 1000 points are depicted.

In Figure 3, we show the results obtained by the first order ( $K1T0$ ) and second order ( $K2T1$ ) in time and space schemes. One can notice some improvements, but this example mostly shows that the correction is also effective when we use representations with polynomials of higher degree.

**4.2.3. Strong shock.** The next test problem contains 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:

$$(28) \quad (\rho_0, u_0, p_0) = \begin{cases} (1.0, 0.0, 1000.0), & \text{if } x < 0.5, \\ (1.0, 0.0, 0.01), & \text{if } x > 0.5. \end{cases}$$

In Figure 4, profiles of density, velocity and pressure are depicted with a reference solution for a mesh containing 1000 cells. It indicates that the first-order scheme can accurately resolve strong shocks. As before, the results of Figure 5 show that the correction is effective by comparing the results of the first order ( $K1T0$ ) and second order ( $K2T1$ ) in time and space schemes.

**4.2.4. 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

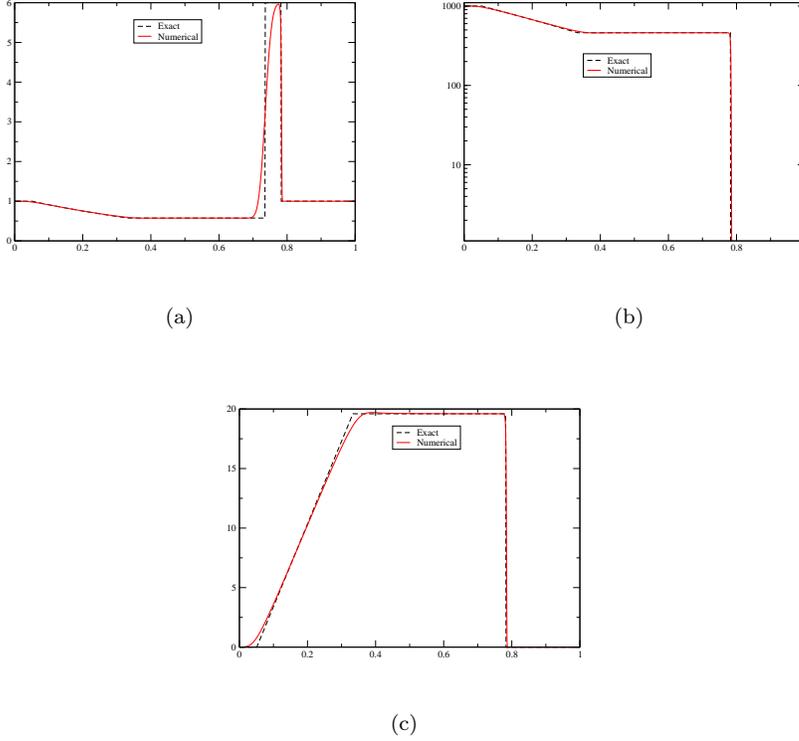


FIG. 4. *Exact and numerical solutions of the strong shock problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.012$  with  $CFL = 0.4$ .*

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:

$$(29) \quad (\rho_0, u_0, p_0) = \begin{cases} (1.0, -2.0, 0.4), & \text{if } 0.0 \leq x < 0.5, \\ (1.0, 2.0, 0.4), & \text{if } 0.5 < x < 1.0. \end{cases}$$

The results for the first-order scheme are depicted in Figure 6 with a reference solution on a mesh containing 1000 cells. In Figure 7 we show again that the correction is effective by comparing the results of the first order (K1T0) and second order (K2T1) in time and space schemes.

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

$$(30) \quad (\rho_0, u_0, p_0) = \begin{cases} (5.99924, 19.5975, 460.894), & \text{if } 0.0 \leq x < 0.8, \\ (5.992420, -6.19633, 46.0950), & \text{if } 0.8 < x \leq 1.0. \end{cases}$$

This is one of the two test cases designed in [28] which correspond to wave interaction in Collela and Woodward blast 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

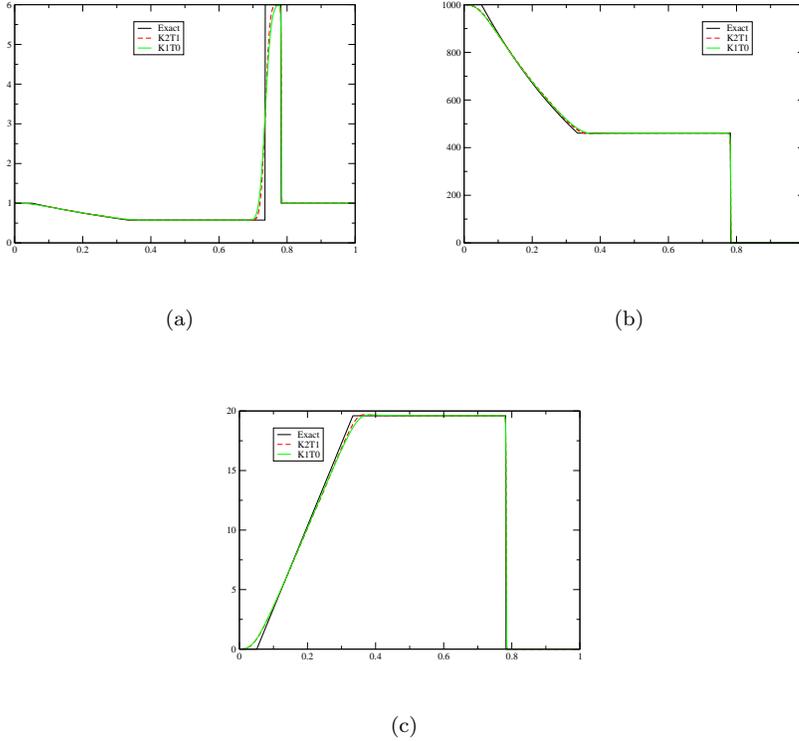


FIG. 5. *Solution of the strong shock problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.012$  with  $CFL = 0.4$ . In each figure, the exact solution and the solutions obtained with  $K2T1$  and  $K1T0$  are depicted.*

1000 cells and the chosen Courant number coefficient is 0.1. Boundary conditions are transmissive. The results for the density, velocity and pressure compared to the exact solution are shown in Figure 8 .

**4.2.6. 2D test case.** The scheme is a straightforward extension of the one dimensional one. In this test case we look at a 2D shock propagation in the domain  $[0, 1] \times [0, 1]$ . The initial data is given by:

$$(31) \quad (\rho_0, u_0, v_0, p_0) = \begin{cases} (1, 0, 0, 1), & \text{if } \|x\| < 0.25, \\ (0.125, 0, 0, 0.1), & \text{if } \|x\| \geq 0.25. \end{cases}$$

The solutions for the density, velocity and pressure computed with the K1T0 first order-scheme as well as the used triangular mesh can be found in Figure 9. It shows that the staggered scheme works in a stable way also in higher dimensions. By comparing the staggered scheme with the solution obtained by a conservative scheme in Figure 10 we demonstrate that the shock front is correctly resolved.

We have also displayed in Figure 11 the relative error of conservation, i.e

$$\int_{\Omega} \mathbf{m}_x^n d\mathbf{x}, \quad \int_{\Omega} \mathbf{m}_y d\mathbf{x}, \quad \text{and} \quad \frac{\int_{\Omega} E^n d\mathbf{x}}{\int_{\Omega} E^0 d\mathbf{x}} - 1.$$

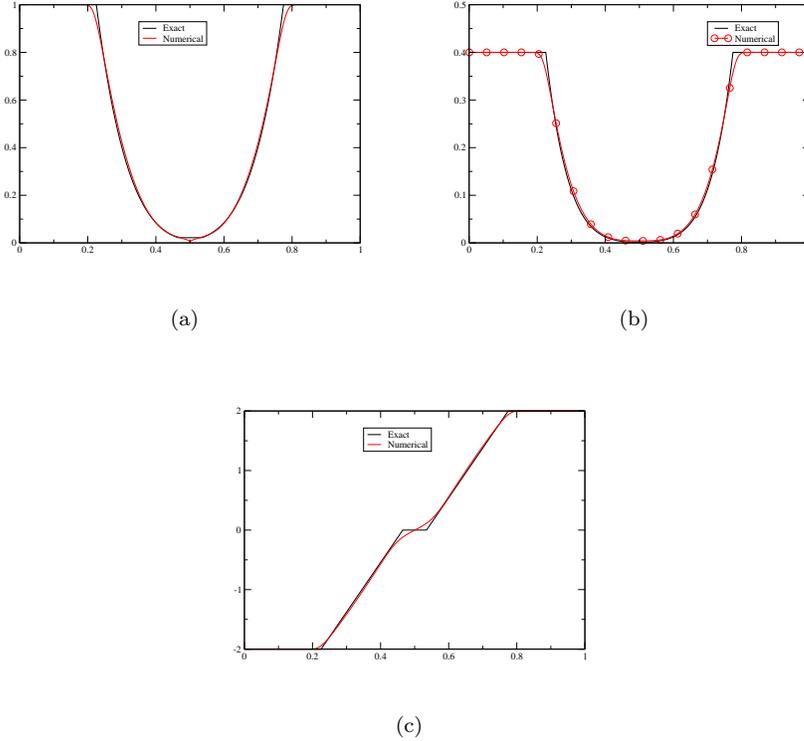


FIG. 6. *Exact and numerical solutions of the 123-problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.15$  with  $CFL = 0.4$ .*

There is no question on the density, since we use a dG scheme for that variable. We see that the errors are negligible. They also are independent of the choice of the quadrature formula (not shown) used to compute the parameters needed in (4.1) and (26).

**4.3. Some remarks concerning the polynomial orders of the velocity and the thermodynamics parameters.** In all the simulations, we have made the choice of polynomials of degree  $p$  for the thermodynamic parameter, and  $p + 1$  for the velocity one, but there is no justification except a rule of the thumb inspired of what is done for incompressible flows. We have encountered stability problems in the case of equal degree polynomials. We discuss this in one dimension, the two dimensional case has not been explored. More precisely, when computing the quantities  $p^*$  for the velocity update, and the mass flux, several choices can be made in addition to the polynomial orders:

- We can take a centered flux (arithmetic average between the left and right states), and  $p^*$  is also the arithmetic average. This will be the "centered" choice.
- We can compute the flux using the exact Godunov solver, as well as for  $p^*$ . This will be the "exact" choice,
- We can compute these quantities using the HLLC methodology. This will be the "HLLC" choice.

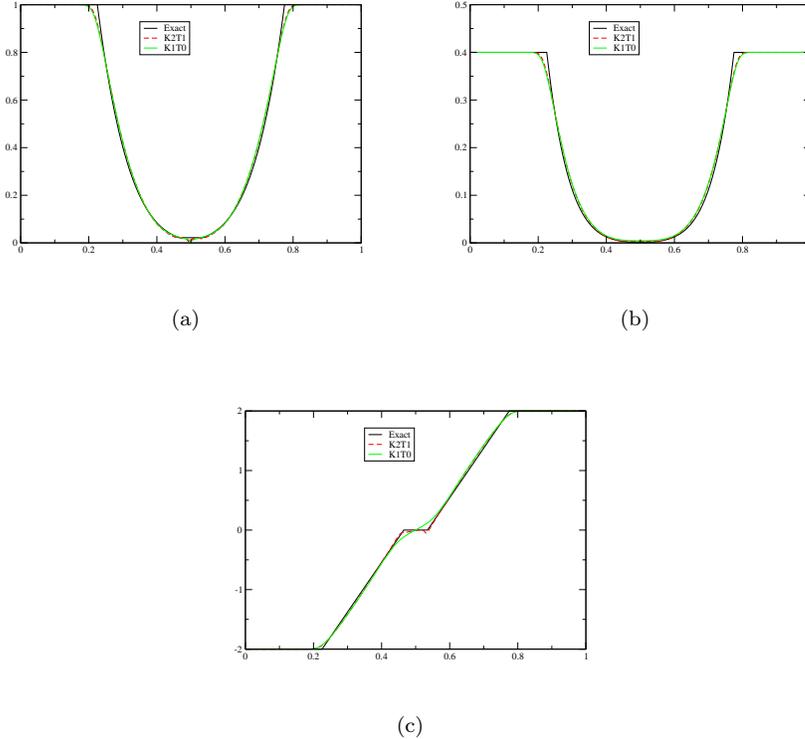


FIG. 7. Solution of the 123-problem for (a) density, (b) velocity and (c) pressure at time  $T = 0.15$  with  $CFL = 0.4$ . In each figure, the exact solution and the solutions obtained with K2T1 and K1T0 are depicted.

In the figure 12 we report the results on the density for the smooth case of section 4.2.1 and time  $t = 0.025$ . We see that all the combinations in polynomial orders with the exact, HLLC choices are stable. The combination velocity with degree 1 and thermodynamics with degree 0 is also stable, while with linear velocity and thermodynamics the scheme is not stable. This is why we have chosen  $t = 0.0025$  because soon after the code blows up. Only the polynomial degrees and the flux have been changed, all of the other elements of the schemes remain the same. No limiting has been used, and the time accuracy is only first order. We have no mathematical explanation, only heuristic ones. We conjecture that for a centered approximation, which is the closest with what is done for incompressible flows, we suffer of a kind of LBB stability problem. This stability problem is cured because of some "upwinding" mechanism with the exact and HLLC solver.

In all the other simulations, the velocity has been one degree more than the thermodynamics, and we take the exact or the HLLC solver, for security.

**5. Conclusions.** In this paper, we have proposed a method to construct staggered high order schemes for compressible flows. The method has been illustrated on a staggered higher-order Residual Distribution (RD) scheme for compressible flow, but as explained in the text, this is not restricted to this particular class of schemes. The key elements are (i) a reformulation of the local conservation properties at the level

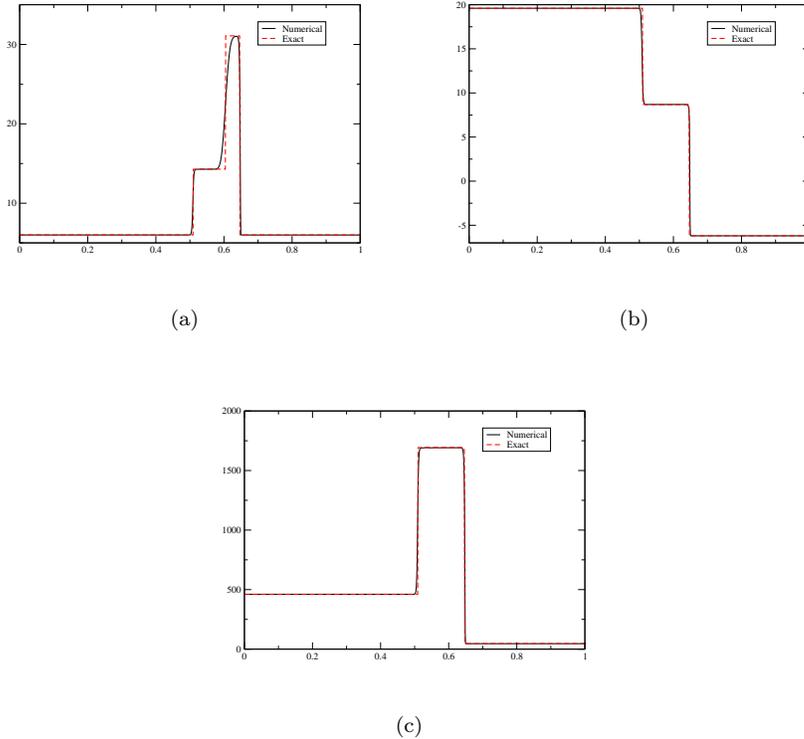


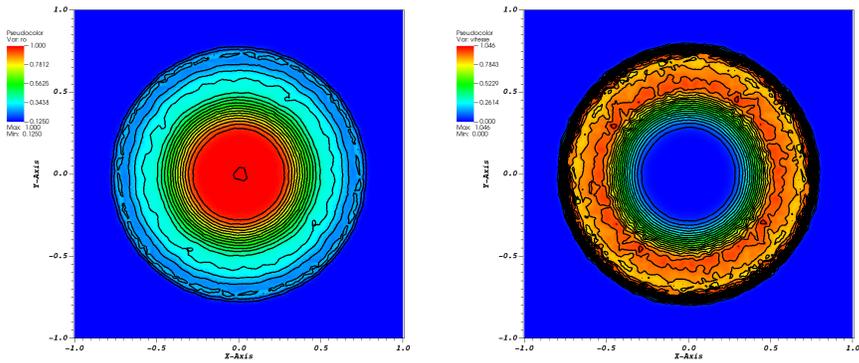
FIG. 8. *Exact and numerical solutions of the Collela and Woodward test case for (a) density, (b) velocity and (c) pressure at time  $T = 0.012$  with  $CFL = 0.1$  computed on a mesh with 1000 points.*

of elements, and not faces as it is classically done, (ii) that the type stepping method is obtained by a combination of Euler forward steps, but this is more general than, for example SSP Runge Kutta: Defect correction methods can also be used.

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

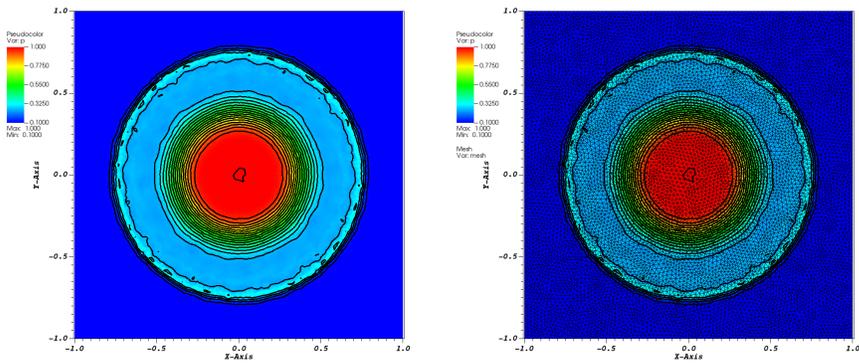
Let us write a series of remarks to conclude this paper.

1. Though the numerical examples are mostly one-dimensional (because in this case one can compute the exact solution for comparison), the description of the correction introduced in Section 3.2, as well as the conditions introduced in Proposition 1 are formulated for general elements.
2. The Residual Distribution formalism introduced here is not restrictive. In [13], 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,



(a)

(b)



(c)

(d)

FIG. 9. Numerical solutions of the 2D shock test case for (a) density, (b) vertical velocity and (c) pressure at time  $T = 0.16$  computed on a mesh consisting of triangles which is depicted for the pressure in (d).

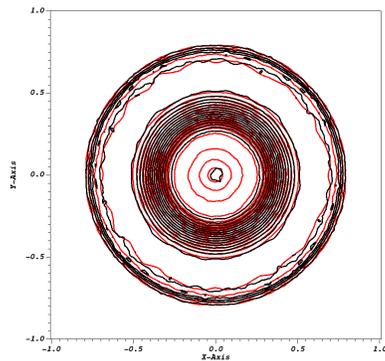


FIG. 10. Comparison of the solutions of the 2D shock test case for the pressure obtained by a conservative scheme (red) and the staggered one (black) at time  $T = 0.16$ .

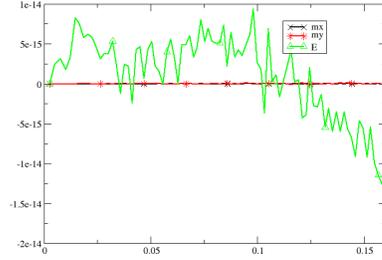


FIG. 11. Conservation errors, in time.

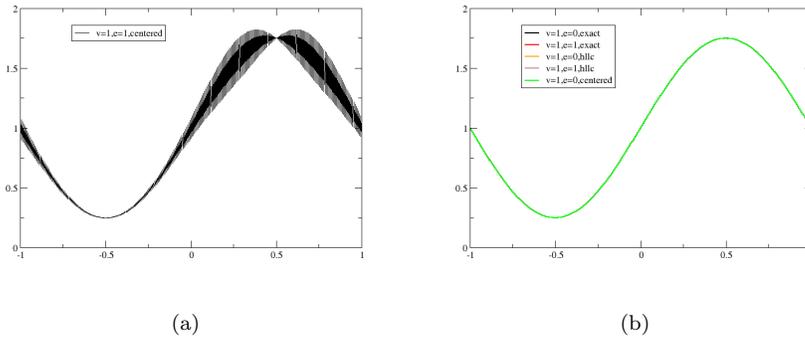


FIG. 12. (a) linear velocity, piecewise constant thermodynamics, (b) all the other cases.

it is certainly possible to write the contribution at element level, as here, and then to rewrite the scheme in the semi-discrete form (9) (if first-order accuracy in time is chosen), 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 elements themselves: this is what is behind the proof of Proposition 1, thus corrections of the form (19) and (20) 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 [24, 23, 29], and the conclusions are the same. This is however not a proof.

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

**Acknowledgements.** I thank Dr. Bettina Wieber for her very constructive comments. I am also grateful to Dr. Ksenya Ivanova during her stay at I-Math for our discussions on this problem. I am also in debt with the two unknown reviewers that have led to drastic improvement with respect to the original submission.

**Appendix A. Short introduction to the Deferred Correction (DeC) approach.** We consider again a hyperbolic system in the form

$$(32) \quad \frac{\partial U}{\partial t} + L(U) = 0$$

which we want to approximate with a high-order accurate scheme in time. 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 which will be denoted by  $\mathcal{L}^1$ . The DeC algorithm is providing an iterative procedure that approximates the solution of the  $\mathcal{L}^2$  scheme  $U^*$  in the following way:

$$(33) \quad \mathcal{L}^1(U^{n+1}) = 0,$$

$$(34) \quad \mathcal{L}^1(U^{(k)}) = \mathcal{L}^1(U^{(k-1)}) - \mathcal{L}^2(U^{(k-1)}), \quad \text{with } k = 2, \dots, K,$$

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 [30]:

PROPOSITION A.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  $\Delta$ , such that for any  $U, V$  we have that

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

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

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

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

Following the proceeding in Section 3, we get for a second-order DeC scheme:

$$\Phi_{\sigma\nu, K}^u = \frac{1}{2}(\Phi_{\sigma\nu, K}^u(U^{(k)}) + \Phi_{\sigma\nu, K}^u(U^n)), \quad \Phi_{\sigma\varepsilon, K}^\rho = \frac{1}{2}(\Phi_{\sigma\varepsilon, K}^\rho(U^{(k)}) + \Phi_{\sigma\varepsilon, K}^\rho(U^n)).$$

The calculations can also be immediately extended to higher accuracy in time by modifying the above half sums.

**Appendix B. Proof of Proposition 1.** We first show some estimates for scalar functions (the system case is identical), and then we use them to show Proposition 1. We start with 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 diameters 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 below. In  $\mathbb{R}^d$ , we have a functional description of the density, the velocity and the energy: we call them  $\rho_h$ ,  $\mathbf{u}_h$  and  $e_h$  to refer they are defined from  $\mathbb{R}^d = \cup K$ .

Let  $T > 0$  and let  $0 < t_1 < \dots < t_n < \dots < t_N \leq T$  be a time discretisation 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  $\{u_h^p\}_{p=0 \dots N}$ , where  $u_h^p$  belongs to  $V^h$  or  $W^h$  (see Section 2.1). They are defined from degrees of freedom that are again denoted by  $\sigma$ . We can define a function  $u_\Delta$  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  $X_\Delta$  and is equipped with the  $L^\infty$  and  $L^2$  norms.

We then have the following lemma:

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

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

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

$$(35) \quad \lim_{h \rightarrow 0, \Delta t \rightarrow 0} \left( \sum_{n=0}^N \Delta t_n \sum_{K \subset \mathcal{Q}} |K| \sum_{\sigma \in K} \|(u_h)_\sigma - \overline{(u_h^n)}_K\| \right) = 0.$$

*Proof.* The proof is inspired from [31] and can be found in [18]. □

Now we have all the prerequisites for proving Proposition 1. We will perform the proof for the momentum since the proof for the energy is similar and can be done in a straightforward manner. We proceed the proof with several lemmas.

LEMMA 3. *Under the conditions of Proposition 1, for any  $\varphi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+)$  we have*

$$\begin{aligned} \lim_{\Delta t \rightarrow 0, \Delta \rightarrow 0} \sum_{n=0}^{\infty} \int_{\mathbb{R}^d} \varphi_h(\rho_h^{n+1} \mathbf{u}_h^{n+1} - \rho_h^n \mathbf{u}_h^n) d\mathbf{x} \\ = - \int_{\mathbb{R} \times \mathbb{R}^+} \frac{\partial \varphi}{\partial t} u dxdt + \int_{\mathbb{R}} \varphi(x, 0) u_0 dxdt, \end{aligned}$$

where

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

*Proof.* This is the classical lemma. □

*Proof of Proposition 1.* We start from (13a)

$$\begin{aligned} & \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} \\ & + \Delta t_n \sum_K \psi_K^n \left[ \sum_{\sigma_\nu \in K} \omega_{\sigma_\nu}^{\rho, n+1} \Phi_{\sigma_\nu, K}^{\mathbf{u}} + \sum_{\sigma_\varepsilon \in K} \omega_{\sigma_\nu}^{\mathbf{u}, n, K} \Phi_{\sigma_\varepsilon, K}^\rho \right] \\ & + \Delta t_n \sum_K \left( F_K(\mathbf{u}^n) + \sum_{\sigma_\nu \in K} D_{\sigma_\nu}(\mathbf{u}^n) \right) = 0, \end{aligned}$$

and by using the assumptions of Proposition 1 we obtain

$$\begin{aligned} & \int_{\mathbb{R}^d} \psi(\mathbf{x}, t) (\rho^{n+1} \mathbf{u}^{n+1} - \rho^n \mathbf{u}^n) d\mathbf{x} + \sum_K \psi_K^n \left[ \Delta t_n \int_{\partial K} \mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n} d\gamma \right] \\ & + \Delta t_n \sum_K \left( F_K(\mathbf{u}^n) + \sum_{\sigma_\nu \in K} D_{\sigma_\nu}(\mathbf{u}^n) \right) = 0. \end{aligned}$$

From (13b), we see that

$$\begin{aligned} F_K(\mathbf{u}^n) &= \sum_{\sigma_\nu \in K} (\psi_{\sigma_\nu}^n - \psi_K^n) \omega_{\sigma_\nu}^{\rho, n+1, K} \Phi_{\sigma_\nu, K}^{\mathbf{u}}, \\ D_{\sigma_\nu}(\mathbf{u}^n) &= \sum_{K', \sigma_\nu \in K'} \left[ \sum_{K, \sigma_\nu \in K \cap K'} \omega_{\sigma_\nu}^{\rho, n+1, K} (\psi_K^n - \psi_{\sigma_\nu}^n) \Phi_{\sigma_\nu, K'}^{\mathbf{u}} \right], \end{aligned} \quad \square$$

so that, since  $\psi_K^n - \psi_{\sigma_\nu}^n = O(h)$ , using the estimates of Lemma 2, we have

$$\lim_{\Delta t_n, h \rightarrow 0} \Delta t_n \sum_K F_K(\mathbf{u}^n) = 0 \quad \text{and} \quad \lim_{\Delta t_n, h \rightarrow 0} \Delta t_n \sum_{\sigma_\nu \in K} D_{\sigma_\nu}(\mathbf{u}^n) = 0$$

because the mesh is shape regular and  $\Delta t_n/h$  is bounded. Last, using the same technique as in [18], and due again to the Lemma 2, we see that

$$\lim_{\Delta t_n, h \rightarrow 0} \sum_K \psi_K^n \Delta t_n \int_{\partial K} \mathbf{f}^{\mathbf{m}}(U^n) \cdot \mathbf{n} d\gamma = \int_{\mathbb{R}^+} \int_{\mathbb{R}^d} \nabla_{\mathbf{x}} \psi(\mathbf{x}, t) \mathbf{f}^{\mathbf{m}}(U) d\mathbf{x}.$$

The convergence result for the energy is done with exactly the same method which then finishes the proof of Proposition 1.

#### 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] R. Abgrall, K. Lipnikov, N. Morgan, and Svetlana Tokareva. Multidimensional staggered grid residual distribution scheme for Lagrangian hydrodynamics. *SIAM J. Sci. Comput.*, 42(1):A343–A370, 2020.
- [3] V. A. Dobrev, T. V. Kolev, and R. 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 conservation laws, I*. Ellipse, 1991.
- [5] M.L. Wilkins. *Methods in Computational Physics*, chapter Calculation of elastic-plastic flows, pages 211–263. Advances in Research and Applications. Academic Press, 1964.

- [6] Jan Nikl, Milan Kuchařík, and Stefan Weber. High-order curvilinear finite element magneto-hydrodynamics. I: A conservative Lagrangian scheme. *J. Comput. Phys.*, 464:28, 2022. Id/No 111158.
- [7] R. Abgrall and M. Dumbser. Mhd flows on staggered meshes: local conservation. *in preparation*, 2023.
- [8] Hervé Guillard and Cécile Viozat. On the behaviour of upwind schemes in the low Mach number limit. *Comput. Fluids*, 28(1):63–86, 1999.
- [9] Hervé Guillard. On the behavior of upwind schemes in the low Mach number limit. IV: P0 approximation on triangular and tetrahedral cells. *Comput. Fluids*, 38(10):1969–1972, 2009.
- [10] H. Guillard and B. Nkonga. On the behaviour of upwind schemes in the low Mach number limit: a review. In *Handbook on numerical methods for hyperbolic problems. Applied and modern issues*, pages 203–231. Amsterdam: Elsevier/North Holland, 2017.
- [11] 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.
- [12] Hester Bijl and Pieter Wesseling. A unified method for computing incompressible and compressible flows in boundary-fitted coordinates. *J. Comput. Phys.*, 141(2):153–173, 1998.
- [13] 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.
- [14] L. Gastaldo, R. Herbin, J.-C. Latché, and N. Therme. A MUSCL-type segregated–explicit staggered scheme for the Euler equations. *Comput. & Fluids*, 175:91–110, 2018.
- [15] R. Herbin, J.-C. Latché, and C. Zaza. A cell-centred pressure-correction scheme for the compressible Euler equations. *IMA J. Numer. Anal.*, 40(3):1792–1837, 2020.
- [16] 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.
- [17] 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.
- [18] R. Abgrall and P. L. Roe. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.*, 19(1-3):3–36, 2003.
- [19] R. Abgrall. Essentially non oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys.*, 214(2):773–808, 2006.
- [20] 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.
- [21] R. Abgrall, M. Ricchiuto, and D. de Santis. High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids. *SIAM J. Scientific Computing*, 36(3):A955–A983, 2014.
- [22] 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.
- [23] R. Abgrall, P. Bacigaluppi, and S. Tokareva. A high-order nonconservative approach for hyperbolic equations in fluid dynamics. *Computers and Fluids*, 2018.
- [24] 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.
- [25] Jean-Luc Guermond and Richard Pasquetti. A correction technique for the dispersive effects of mass lumping for transport problems. *Comput. Methods Appl. Mech. Eng.*, 253:186–198, 2013.
- [26] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximation of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1437–1453, 2004.
- [27] J. Cheng and C.-W. Shu. Positivity-preserving Lagrangian scheme for multi-material compressible flow. *J. Comput. Phys.*, 257:143–168, 2014.
- [28] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, third edition, 2009. A practical introduction.
- [29] R. Abgrall, P. Öffner, and H. Ranocha. Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous Galerkin schemes, 2021.
- [30] R. Abgrall and D. Torlo. High order asymptotic preserving deferred correction implicit-explicit schemes for kinetic models. *SIAM J. Sci. Comput.*, 42(3):B816–B845, 2020.
- [31] 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.