WELL-BALANCED TIME IMPLICIT FORMULATION OF RELAXATION SCHEMES FOR THE EULER EQUATIONS

CHRISTOPHE CHALONS *, FRÉDÉRIC COQUEL †, AND CLAUDE MARMIGNON ‡

Abstract. We show how to derive time implicit formulations of relaxation schemes for the Euler equations for real materials in several space dimensions. In the fully time explicit setting, the relaxation approach has been proved to provide efficient and robust methods. It thus turns interesting to answer the open question of the time implicit extension of the procedure. A first natural extension of the classical time explicit strategy is shown to fail in producing discrete solutions which converge in time to a steady state. We prove that this first approach does not permit a proper balance between the stiff relaxation terms and the flux gradients. We then show how to achieve a well-balanced time implicit method which yields approximate solutions at a perfect steady state.

Key words. time implicit schemes, relaxation methods, well-balanced schemes

AMS subject classifications.

1. Introduction. This work is devoted to the study of time implicit formulations of relaxation schemes for the Euler equations with general pressure laws.

Over the past decade, relaxation schemes have received a considerable attention. Such schemes are primarily intended to approximate the solutions of highly nonlinear hyperbolic systems. The design principle of relaxation schemes consists in approximating the solutions (say the Riemann solutions) of a given highly nonlinear system by the solutions of a larger but weakly nonlinear system with singular perturbations. These perturbations take the form of stiff relaxation source terms which restore the algebraic nonlinearities of the original PDE model in the regime of an infinite relaxation parameter. Here the key issue is that these source terms must of course facilitate the derivation of the approximation procedure together with its nonlinear stability analysis as well.

At the theoretical level, relevant relaxation methods have been proved to obey in their time explicit formulation several important stability properties ranging from L^1 stability (the phase space is preserved) to nonlinear stability like entropy inequalities, see [3], [6], [5], [15], [9] and the references therein. Moreover, some of these methods ([3], [6]) also enjoy accuracy properties like the exact capture of stationary contact discontinuities. From a numerical point of view, the simplicity in the time explicit formulation of these methods guarantees a very low computational effort. In addition, the property that their upwind mechanism stays virtually free from the exact pressure law makes them very useful in practice. At last, the relaxation strategy allows for a fruitful reinterpretation of some of the prominent approximate Riemann solvers as underlined by Bouchut [3] and also Leveque and Pelanti [17]. Such a reintrepretation has made tractable the extension of Riemann solvers to various and difficult settings for complex compressible materials (see [1], [2], [4] for recent contributions), but always in a fully time explicit framework.

It therefore seems useful to extend the relaxation schemes to a time implicit setting

* *Université Paris 7 & Laboratoire Jacques-Louis Lions, U.M.R. 7598, Boîte courrier 187, 75252 Paris Cedex 05, France.* chalons@math.jussieu.fr

† *CNRS & Laboratoire Jacques-Louis Lions, U.M.R. 7598, Boîte courrier 187, 75252 Paris Cedex 05, France.* coquel@ann.jussieu.fr

‡ *Office National d'Etudes et de Recherches Aérospatiales, BP 72, 92322 Châtillon Cédex, France.* Claude.Marmignon@onera.fr

in order to inherit from their valuable properties of stability, accuracy and simplicity in the calculation of steady state solutions via the very classical time marching technique.

Here the solutions of the Euler equations for real gases in several space dimensions are approximated by the solutions of relaxation PDE model due to [24]. From the mathematical standpoint, Chalons and Coulombel [7] have recently proposed a comprehensive study of the convergence properties of the solutions of the relaxation system to the solutions of the original 3×3 Euler system. At the discrete level, the stability properties of the resulting scheme have been studied by [3], [5], [6] in a fully time explicit framework.

In this paper, we prove that the natural extension of the time explicit formulation to a time implicit one fails generally speaking in producing perfectly stationary solutions. We show how to correct the natural extension so as to end up with a robust time implicit relaxation scheme producing discrete solution at a perfect steady state. Indeed, let us briefly report on such an origin. In the commonly used time explicit approach, the relaxation approximation procedure makes use of a fractional step method alternating between solving the homogeneous relaxation conservation laws with the ODE equations associated with the (infinitely) stiff relaxation source terms. It is known after Leveque and Yee [18] that the usual fractional step method may grossly fail in the capture of unsteady solutions of some PDE systems with stiff relaxation (see [13] for the capture of detonation wave and a cure). But we stress that the fractional step method actually fairly succeeds within the fully time explicit framework for relaxation schemes to reproduce the dynamics in unsteady solutions of the equilibrium system. This claim, grounded on numerous numerical evidences (see [3], [5], [6], [15] and [17]), has been given recently a mathematical fundation by Gosse [11] in the context of a 2×2 model for chromatography, proving the convergence of two time splitting techniques. But when dealing with solutions nearly in steady state, splitting techniques are known, already in the setting of finite relaxation rate, to suffer from inaccuracy in balancing source terms and flux gradients. The reader is referred to LeVeque [16], Greenberg and Leroux [12], and to the recent book by Bouchut [3]. As put forward in the present work devoted to large time step methods for capturing stationary solutions, extension to the case of an infinite relaxation rate yields a closely related difficulty: the fractional step method brings in a robust manner solutions of the equilibrium system to be nearly in steady state but not at perfect steady state. In this work, we prove that perfectly stationary solutions cannot be reached in general, unless the relaxation source terms and the flux gradients in the augmented system are kept in a proper balance in the regime of an infinite relaxation rate. In the light of the analysis we propose, we design a well-balanced time implicit formulation of the relaxation approximation procedure. Numerical evidences assess the relevance of the method: perfectly stationary solutions are achieved.

The format of this paper is as follow. Section 2 addresses the relaxation PDE model for approximating the solutions of the Euler equations. The stability of this approximation procedure requires the satisfaction of a so-called subcharacteristic condition. This one is briefly re-investigated on the ground of a Chapman-Enskog expansion. This rather classical issue is addressed here since it turns very useful to understand why the well-balanced time implicit relaxation scheme is stable under the subcharacteristic condition. Section 3 is devoted to the numerical issue. The time explicit relaxation method is shortly revisited to provide all the formulas required in the time implicit formulations. We then prove that the relaxation scheme under consideration can be understood equivalently as a Roe-type method for the relaxation system but not for the original Euler equations. We refer the reader to LeVeque and Pelanti [17] for a reinterpretation of the relaxation scheme of Jin and Xin [15] in term of a Roe method. In the present setting, this equivalence with a Roe linearization stays at the very basis of the time implicit formulations we derive. The natural time implicit extension of the relaxation scheme is then described and analyzed to understand the roots of the reported failure of convergence in time. We propose a correction procedure dictated by the property that the flux gradients in the relaxation system must properly balance the relaxation source term in the regime of an infinite relaxation parameter. The last paragraph gives numerical evidences assessing the correct design of the time implicit formulation we promote in the relaxation framework.

2. Statement of the problem. The present work treats the numerical approximation of the steady state solutions of the Euler equations for real gases :

$$\left\{\begin{array}{l}\partial_t\rho + \nabla.\ \rho\mathbf{w} = 0, \qquad t>0, \quad \mathbf{x}\in\mathcal{D},\\ \partial_t\rho\mathbf{w} + \nabla.\ (\rho\mathbf{w}\otimes\mathbf{w} + p(\mathbf{U})\ \mathbf{I_d}) = 0,\\ \partial_t\rho E + \nabla.\ (\rho E + p(\mathbf{U}))\mathbf{w} = 0,\end{array}\right. \tag{2.1}$$

where $\mathcal{D}$ is a bounded domain of $\mathbb{R}^d$ with $d \geq 1$. The pressure law $p(\mathbf{U})$ is a given smooth function of the unknown $\mathbf{U} = (\rho, \rho\mathbf{w}, \rho E)$ in the form:

$$p(\mathbf{U}) = p(\rho, \rho e) \quad \text{with} \quad \rho e = \rho E - \frac{||\rho\mathbf{w}||^2}{2\rho}, \tag{2.2}$$

with the property that the first order system (2.1) is hyperbolic, namely

$$c^2(\mathbf{U}) = c^2(\rho,\rho e) = \frac{\partial p}{\partial\rho}(\rho,\rho e) + \frac{1}{\rho}(p+\rho e)\frac{\partial p}{\partial\rho e}(\rho,\rho e)\ >\ 0, \tag{2.3}$$

for all state $\mathbf{U}$ in the natural phase space:

$$\Omega_{\mathbf{U}} = \Big\{\mathbf{U} = (\rho,\rho\mathbf{w},\rho E)\in\mathbb{R}^{d+2}/\rho>0, \rho\mathbf{w}\in\mathbb{R}^d, \rho E - \frac{||\rho\mathbf{w}||^2}{2\rho} > 0\Big\}. \tag{2.4}$$

System (2.1) is given the following condensed form:

$$\partial_t\mathbf{U}\ +\ \nabla.\ \mathbf{F}(\mathbf{U}) = 0, \tag{2.5}$$

with clear definition for the vector-valued flux function $\mathbf{F}(\mathbf{U}) = (\mathbf{F}_{x_j}(\mathbf{U}))_{1\leq j\leq d}$.

In the present work, we deserve a central attention to numerical methods for the approximation of the solutions to the Euler equations whose formulation stays free, as far as possible, of the exact form of the required pressure law. In addition, the methods we seek must achieve robustness in the setting of closure laws coming from the physics of compressible real material. The reader is referred to Menikoff and Plorh [21] for a discussion and several examples. It is well-known that the severe nonlinearities involved in these pressure laws make particularly delicate the capture of the nonlinear phenomena induced by the acoustic waves. To tackle these nonlinearities and to allow for an efficient and robust numerical procedure, we propose to adopt a relaxation approach: the weak solutions of the system (2.1) are approximated by the solutions of a larger but simpler PDE model with (infinitely) stiff relaxation source terms. By simpler, it is understood that the underlying nonlinearities are easier to handle. Motivated by the works by Bouchut [3], Chalons and Coquel [6], Coquel *et al.*

[9] and Siliciu [24], simplicity is achieved when no longer understanding the pressure $p(\rho, \rho e)$ as a nonlinear function but as a new unknown we denote by Π, equipped with its own partial differential equation. This new unknown Π is subject to a relaxation procedure which purpose is to restore the original pressure law p in the regime of an infinite relaxation rate. More precisely, the relaxation PDE model developed in [9], [3], [6], [24] reads:

$$\left\{\begin{array}{l}\partial_t \rho^\lambda + \nabla.\, (\rho \mathbf{w})^\lambda = 0, \qquad t > 0, \quad \mathbf{x} \in \mathcal{D},\\ \partial_t (\rho \mathbf{w})^\lambda + \nabla.\, (\rho \mathbf{w} \otimes \mathbf{w} + \Pi\, \mathbf{I_d})^\lambda = 0,\\ \partial_t (\rho E)^\lambda + \nabla.\, ((\rho E + \Pi)\mathbf{w})^\lambda = 0,\\ \partial_t (\rho \Pi)^\lambda + \nabla.\, ((\rho \Pi + a^2)\mathbf{w})^\lambda = \lambda \rho^\lambda (p(\rho^\lambda, (\rho e)^\lambda) - \Pi^\lambda),\end{array}\right. \tag{2.6}$$

where the parameter $\lambda > 0$ stands for the relaxation coefficient rate. Here, the parameter a is a given positive real number and the PDE model (2.6) can be seen to be invariant by rotation. To simplify the notations, the relaxation model (2.6) is given the following condensed form:

$$\partial_t \mathbf{V}^\lambda + \nabla.\, \mathcal{G}(\mathbf{V}^\lambda) = \lambda \mathcal{R}(\mathbf{V}^\lambda), \tag{2.7}$$

with $\mathbf{V} = (\rho, \rho\mathbf{w}, \rho E, \rho\Pi)$, $\mathcal{R}(\mathbf{V}) = (0, 0_{\mathbb{R}^d}, 0, \rho(p(\rho, \rho e) - \Pi))$ and clear definitions for the vector-valued function $\mathcal{G}(\mathbf{V}) = (\mathcal{G}_{x_j}(\mathbf{V}))_{1 \le j \le d}$. The precise role played by the parameter a is explained just hereafter but first, it is worth to stress the reason why (2.6) is easier to handle than (2.1). To that purpose, observe that setting $\lambda = 0$ in (2.6) decouples the total energy equation from the others. In other words, the total energy only enters the algebraic relaxation source term via the definition (2.2) of the original pressure law. This weak coupling is responsible for the following attractive result (see [3] for instance):

LEMMA 2.1. *Let be given $a > 0$ in (2.6). Then the first order system in (2.6) is hyperbolic over the following phase space*

$$\Omega_{\mathbf{V}} = \Big\{ \mathbf{V} = (\rho, \rho\mathbf{w}, \rho E, \rho\Pi) \in \mathbb{R}^{d+3} / \rho > 0, \rho\mathbf{w} \in \mathbb{R}^d, \rho E - \frac{||\rho\mathbf{w}||^2}{2\rho} > 0, \rho\Pi \in \mathbb{R} \Big\}. \tag{2.8}$$

Namely, for any given unit vector $\mathbf{n} = (n_i)_{1 \le i \le d} \in \mathbb{R}^d$, the matrix

$$\mathbf{A}(\mathbf{V}, \mathbf{n}) = \sum_{i=1}^{d} n_i \nabla_{\mathbf{V}} \mathcal{G}_{n_i}(\mathbf{V})$$

is $\mathbb{R}$-diagonalizable for all $\mathbf{V} \in \Omega_{\mathbf{V}}$, with the following increasingly ordered eigenvalues:

$$\lambda_1(\mathbf{V}, \mathbf{n}) = \mathbf{w}.\mathbf{n} - \frac{a}{\rho} < \lambda_2(\mathbf{V}, \mathbf{n}) = \mathbf{w}.\mathbf{n} < \lambda_3(\mathbf{V}, \mathbf{n}) = \mathbf{w}.\mathbf{n} + \frac{a}{\rho}, \tag{2.9}$$

where the intermediate eigenvalue $\lambda_2(\mathbf{V}, \mathbf{n})$ has $d{+}1$ order of multiplicity. In addition, all the fields are linearly degenerate: all the propagating waves behave as linear waves.

Observe from the definition (2.9) that the free real parameter a entering (2.6) has the dimension of $\rho c(\rho, \rho e)$, i.e. the dimension of a Lagrangian sound speed. In any given direction $\mathbf{n} \in \mathbb{R}^d$, the extreme waves associated with the eigenvalues $\lambda_1(\mathbf{V}, \mathbf{n})$ and $\lambda_3(\mathbf{V}, \mathbf{n})$ may be thus understood as an approximation of the acoustic waves in the original equations (2.1). Hence the reported linear degeneracy is in clear contrast with the strong nonlinearities involved in the original PDE model and stays at the basis of the efficient numerical method to be discussed in the next sections.

We now come to highlight the importance of a correct definition of the parameter a in the procedure of approximation of the solutions of (2.1) by those of (2.6). To that purpose, let us rewrite the last governing equation in (2.6) as follows:

$$(\rho\Pi)^\lambda = (\rho p(\rho,\rho e))^\lambda - \frac{1}{\lambda}\{\partial_t(\rho\Pi)^\lambda + \nabla.\ ((\rho\Pi + a^2)\mathbf{w})^\lambda\}. \tag{2.10}$$

Clearly in the limit of an infinite relaxation rate $\lambda \to +\infty$, Π^λ formally coincides with the original pressure law $p(\rho,\rho e)$:

$$\lim_{\lambda\to+\infty} \Pi^\lambda = p(\rho,\rho e). \tag{2.11}$$

After the works by Liu [19] and Chen, Levermore and Liu [8], it is known (see also the pioneering work by Whitham [26]) that to prevent a general relaxation procedure from instabilities in the regime of a large parameter $\lambda >> 1$, the so-called subcharacteristic condition, or Whitham condition [26], must be met: the eigenvalues (2.9) of the relaxation PDE model and those of the original system must be properly interlaced. In the present relaxation setting (2.6) for approximating the solutions of (2.1), these stability conditions are satisfied provided that the free coefficient $a > 0$ in (2.9) upper-bounds the exact Lagrangian sound speed $\rho c(\rho,\rho e)$; namely:

$$a > \rho \times c(\mathbf{U}) = \rho c(\rho,\rho e) \tag{2.12}$$

must be valid for all the state $\mathbf{U}$ under consideration. The reader is referred to the quoted works [3], [6], [9] for a detailed discussion of (2.12) and its relationship with the validity of entropy inequalities for the relaxation PDE model (2.6) that are closely related to those of the original equations.

To assess the relevance of some of our forthcoming conclusions, it is worth to briefly shade light on the Whitham condition (2.12) on the simpler ground of a Chapman-Enskog expansion (see Whitham [26] , Chen, Levermore and Liu [8]). According to this approach, the unknown $(\rho\Pi)^\lambda$ is given the following expansion for large but finite values of $\lambda > 0$:

$$(\rho\Pi)^\lambda = (\rho p(\rho,\rho e))^\lambda + \frac{1}{\lambda}(\rho\Pi)_1^\lambda + \mathcal{O}(\frac{1}{\lambda^2}). \tag{2.13}$$

The first order corrector $(\rho\Pi)_1^\lambda$ is found when plugging (2.13) in the PDE (2.10) to obtain:

$$(\rho\Pi)^\lambda = (\rho p(\rho,\rho e))^\lambda - \frac{1}{\lambda}\{\partial_t(\rho p(\rho,\rho e))^\lambda + \nabla.\ ((\rho p(\rho,\rho e) + a^2)\mathbf{w})^\lambda\} + \mathcal{O}(\frac{1}{\lambda^2}), \tag{2.14}$$

so that the first order corrector reads:

$$\begin{aligned} \Pi_1^\lambda &= -\tfrac{1}{\rho^\lambda}\{\partial_t(\rho p(\rho,\rho e))^\lambda + \nabla.\ ((\rho p(\rho,\rho e) + a^2)\mathbf{w})^\lambda\} \\ &= -\{\partial_t(p(\rho,\rho e))^\lambda + \mathbf{w}^\lambda.\nabla(p(\rho,\rho e))^\lambda\} - \tfrac{a^2}{\rho^\lambda}\nabla.\ \mathbf{w}^\lambda. \end{aligned} \tag{2.15}$$

To go further, we notice that the first $d+2$ equations in (2.6) reads as follows at the first order in $\frac{1}{\lambda}$ and in view of the near-equilibrium identity $\Pi^\lambda = (p(\rho,\rho e))^\lambda + \mathcal{O}(\frac{1}{\lambda})$:

$$\left\{ \begin{array}{l} \partial_t\rho^\lambda + \nabla.\ (\rho\mathbf{w})^\lambda = 0, \qquad t > 0, \quad \mathbf{x} \in \mathcal{D}, \\ \partial_t(\rho\mathbf{w})^\lambda + \nabla.\ (\rho\mathbf{w}\otimes\mathbf{w} + p(\rho,\rho e)\ \mathbf{I_d})^\lambda = \mathcal{O}(\frac{1}{\lambda}), \\ \partial_t(\rho E)^\lambda + \nabla.\ ((\rho E + p(\rho,\rho e))\mathbf{w})^\lambda = \mathcal{O}(\frac{1}{\lambda}), \end{array} \right. \tag{2.16}$$

so that classical manipulations prove that smooth solutions of (2.6) obey at the first order in $\frac{1}{\lambda}$ the next equation for the original pressure law $p(\rho, \rho e)$:

$$\partial_t p(\rho, \rho e)^\lambda + \mathbf{w}^\lambda.\nabla p(\rho, \rho e)^\lambda + \rho^\lambda c^2(\rho, \rho e)^\lambda \nabla.\ \mathbf{w}^\lambda = \mathcal{O}(\frac{1}{\lambda}). \tag{2.17}$$

As a consequence, the first order corrector Π_1^λ in (2.15) writes equivalently:

$$\Pi_1^\lambda = -\frac{1}{\rho^\lambda}(a^2 - (\rho^\lambda c(\mathbf{U})^\lambda)^2)\nabla.\ \mathbf{w}^\lambda. \tag{2.18}$$

Invoking the expansion (2.13) with the formula (2.18), the first order asymptotic system governing the solutions of the relaxation system (2.6) for large values $\lambda >> 1$ then reduces to:

$$\begin{aligned} &\partial_t \rho^\lambda + \nabla.\ (\rho \mathbf{w})^\lambda = 0, \\ &\partial_t (\rho \mathbf{w})^\lambda + \nabla.\ (\rho \mathbf{w} \otimes \mathbf{w} + p(\rho, \rho e)\ \mathbf{I_d})^\lambda = -\tfrac{1}{\lambda}\nabla.\ (\Pi_1^\lambda\ \mathbf{I_d}), \\ &\qquad\qquad = \tfrac{1}{\lambda}\nabla.\ (\tfrac{1}{\rho^\lambda}(a^2 - (\rho c)^2)^\lambda \nabla.\ \mathbf{w}^\lambda\ \mathbf{I_d}), \\ &\partial_t (\rho E)^\lambda + \nabla.\ ((\rho E + p(\rho, \rho e))\mathbf{w})^\lambda = \tfrac{1}{\lambda}\nabla.\ (\tfrac{1}{\rho^\lambda}(a^2 - (\rho c)^2)^\lambda \mathbf{w}^\lambda \nabla.\ \mathbf{w}^\lambda). \end{aligned} \tag{2.19}$$

Observe that (2.19) takes the form of the original Euler equations (2.1) but in the presence of a viscous perturbation with viscosity like coefficient $\frac{1}{\lambda \rho^\lambda}(a^2 - (\rho^\lambda c(\mathbf{U})^\lambda)^2)$. As it is well-known (see Majda and Pego [20] for instance), this viscosity coefficient must be positive for the solutions of the near-equilibrium system (2.19) to be stable: this requirement is nothing but the Whitham condition expressed in (2.12).

3. The numerical procedure. We show how to take advantage of the relaxation system (2.6) in the derivation of an efficient time implicit method for the approximation of the solutions to the original Euler equations (2.1). We first briefly revisit the time explicit relaxation framework for the sake of completeness in the required formulas. We then propose a seemingly natural extension of this framework to a time implicit setting (see [15], [9] for instance). We prove that such a natural extension fails to produce perfectly steady state approximate solutions: residues stop decreasing after a few order of magnitude to then reach a plateau. We then show how to correct this first extension so as to end up with a robust time implicit method that yields converged in time discrete solutions corresponding to a ten order of magnitude decrease for the residues.

For the sake of simplicity in the notations, we only address the case of bidimensionnal problems when focusing on cartesian grids with constant space step $\Delta x > 0$ and $\Delta y > 0$, the extension to curvilinear grids being a classical matter. The time variable is discretized using a constant time step $\Delta t > 0$. The approximate solution $\mathbf{U}_h(x, y, t)$, $h = \max(\Delta x, \Delta y)$, is sought under the form of a piecewise constant function at each time level $t^n = n\Delta t$, $n \geq 0$. An initial data $\mathbf{U}_0(x, y)$ being prescribed for (2.1), we classically define:

$$\mathbf{U}_h(x, y, 0) = \mathbf{U}_{i,j}^0 = \frac{1}{\Delta x \Delta y}\int_{\mathcal{C}_{ij}} \mathbf{U}_0(x, y)dxdy, \qquad i, j \in \mathbb{Z}, \tag{3.1}$$

with $(x, y) \in \mathcal{C}_{ij} = ((i - \frac{1}{2})\Delta x, (i + \frac{1}{2})\Delta x) \times ((j - \frac{1}{2})\Delta y, (j + \frac{1}{2})\Delta y)$.
The boundary conditions considered in the present work are extremely classical in the setting of the Euler equations: namely far field boundary conditions and wall conditions. Their treatment is a classical matter described for instance in Hirsch [14]. Besides, MUSCL second order in space enhancement is used in the forthcoming numerical evidences. We again refer the reader to [14] for the required material.

3.1. Revisiting the time explicit procedure. In this section, we briefly revisit the usual approach for deriving time explicit scheme in a relaxation framework. The discrete solution $\mathbf{U}_h(x,y,t^n)$ being known at time $t^n = n\Delta t$, $n \geq 1$, this one is evolved to the next time level $t^{n+1} = t^n + \Delta t$ thanks to a time explicit finite volume scheme:

$$\mathbf{U}_h(x,y,t^{n+1}) \equiv \mathbf{U}_{i,j}^{n+1} = \mathbf{U}_{i,j}^{n} - \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+\frac{1}{2},j}^{n} - \mathbf{F}_{i-\frac{1}{2},j}^{n}) - \frac{\Delta t}{\Delta y}(\mathbf{F}_{i,j+\frac{1}{2}}^{n} - \mathbf{F}_{i,j-\frac{1}{2}}^{n}),\ (x,y) \in \mathcal{C}_{ij}, \tag{3.2}$$

where $\mathbf{F}_{i+\frac{1}{2},j}^{n}$ and $\mathbf{F}_{i,j+\frac{1}{2}}^{n}$ must be defined at time t^n from two numerical flux functions that are respectively consistent with the exact flux functions $\mathbf{F}_x$ and $\mathbf{F}_y$ in the x and y direction. Their required definition follows from the next classical two steps relaxation procedure (see [15] for instance). This approach can be understood as a splitting technique for the relaxation system (2.6) when setting λ to 0 in a first step and then letting the relaxation parameter λ go to infinity in a second step.

First step: Evolution in time $(t^n \to t^{n+1-})$
Starting from the discrete solution $\mathbf{U}_h(x,y,t^n)$, we consider a relaxation approximate solution at time t^n setting:

$$\mathbf{V}_h(x,y,t^n) = (\mathbf{U}_h(x,y,t^n), (\rho\Pi)_h(x,y,t^n)), \tag{3.3}$$

where the relaxation pressure is defined at equilibrium

$$(\rho\Pi)_h(x,y,t^n) = \rho_h p(\rho_h, (\rho e)_h)(x,y,t^n). \tag{3.4}$$

We then solve for times $t \in [0, \Delta t[$, Δt small enough, the following Cauchy problem for the frozen relaxation system (2.7) with $\lambda = 0$:

$$\begin{cases} \partial_t \mathbf{V} + \nabla.\ \mathcal{G}(\mathbf{V}) = 0, & t > 0, \quad \mathbf{x} \in \mathcal{D}, \\ \mathbf{V}(x,y,0) = \mathbf{V}_h(x,y,t^n). \end{cases} \tag{3.5}$$

Within the finite volume framework, this amounts to update the relaxation approximate solution at time t^{n+1-} setting in each cell $\mathcal{C}_{ij}$:

$$\mathbf{V}_{i,j}^{n+1,-} = \mathbf{V}_{i,j}^{n} - \frac{\Delta t}{\Delta x}(\mathcal{G}_{i+\frac{1}{2},j}^{n} - \mathcal{G}_{i-\frac{1}{2},j}^{n}) - \frac{\Delta t}{\Delta y}(\mathcal{G}_{i,j+\frac{1}{2}}^{n} - \mathcal{G}_{i,j-\frac{1}{2}}^{n}), \quad (x,y) \in \mathcal{C}_{ij}, \tag{3.6}$$

where the definitions of $\mathcal{G}_{i+\frac{1}{2},j}^{n}$ and $\mathcal{G}_{i,j+\frac{1}{2}}^{n}$ will be given hereafter.

Second step: relaxation $(t^{n+1-} \to t^{n+1})$
In each cell $\mathcal{C}_{ij}$, we solve the following EDO problem in the limit $\lambda \to \infty$:

$$\begin{cases} \partial_t \rho^\lambda = 0, \\ \partial_t (\rho\mathbf{w})^\lambda = 0, \\ \partial_t (\rho E)^\lambda = 0, \\ \partial_t (\rho\Pi)^\lambda = \lambda \rho^\lambda (p(\rho, \rho e)^\lambda - \Pi^\lambda), \end{cases} \tag{3.7}$$

with as initial data $\mathbf{V}(x,y,\Delta t^-)$ the solution of the Cauchy problem (3.5) at time Δt. In other words, the approximate relaxation solution $\mathbf{V}_h(x,y,t^{n+1})$ is set at equilibrium at time t^{n+1} in each cell when keeping unchanged ρ, $\mathbf{w}$ and E:

$$\rho_{i,j}^{n+1} = \rho_{i,j}^{n+1-},\ (\rho\mathbf{w})_{i,j}^{n+1} = (\rho\mathbf{w})_{i,j}^{n+1-},\ (\rho E)_{i,j}^{n+1} = (\rho E)_{i,j}^{n+1-}, \tag{3.8}$$

but when redefining the relaxation pressure at t^{n+1} so as to enforce equilibrium in agreement with (3.4) :

$$(\rho\Pi)_{i,j}^{n+1} = (\rho p)(\rho_{i,j}^{n+1}, (\rho e)_{i,j}^{n+1}).$$

The Euler approximate solution $\mathbf{U}_h(x, y, t^{n+1})$ is then defined at time t^{n+1} in each cell $\mathcal{C}_{ij}$ using (3.8)

$$\mathbf{U}_{i,j}^{n+1} = \Big(\rho_{i,j}^{n+1},\ (\rho\mathbf{w})_{i,j}^{n+1},\ (\rho E)_{i,j}^{n+1}\Big). \tag{3.9}$$

This concludes the method.

Let us now give a detailed description of the required numerical fluxes $\mathcal{G}^n_{i+\frac{1}{2},j}$, $\mathcal{G}^n_{i,j+\frac{1}{2}}$ in (3.6) in order to eventually infer the definition of the required fluxes $\mathbf{F}^n_{i+\frac{1}{2},j}$ and $\mathbf{F}^n_{i,j+\frac{1}{2}}$ in (3.2). We classically take advantage of the invariance by rotation of the relaxation PDE model (2.6) to focus solely on the definitions of $\mathcal{G}^n_{i+\frac{1}{2},j}$ and thus of $\mathbf{F}^n_{i+\frac{1}{2},j}$. The fluxes $\mathcal{G}^n_{i,j+\frac{1}{2}}$ and $\mathbf{F}^n_{i,j+\frac{1}{2}}$ are given symmetrical definitions. Here, the numerical flux function $\mathcal{G}^n_{i+\frac{1}{2},j}$ is built from the Godunov approach (see Godlewski and Raviart [10]) when solving a Riemann problem for (2.6) and with $\lambda = 0$ in the x direction. Indeed, denoting $\mathcal{G}_x$ the exact flux function in (2.6) in the x direction, we consider for two given states $\mathbf{V}_L$ and $\mathbf{V}_R$ in $\Omega_\mathbf{V}$, $\mathcal{W}(.;\mathbf{V}_L,\mathbf{V}_R)$ the self-similar solution of

$$\begin{cases} \partial_t\mathbf{V} + \partial_x\mathcal{G}_x(\mathbf{V}) = 0, \\ \mathbf{V}(x,0) = \begin{cases} \mathbf{V}_L & \text{if} \quad x<0, \\ \mathbf{V}_R & \text{if} \quad x>0, \end{cases} \end{cases} \tag{3.10}$$

to then define the required flux $\mathcal{G}^n_{i+\frac{1}{2},j}$ as follows:

$$\begin{aligned} \mathcal{G}^n_{i+\frac{1}{2},j} &= \mathcal{G}_x(\mathcal{W}(0^+;\mathbf{V}^n_{i,j},\mathbf{V}^n_{i+1,j})) \\ &\equiv ((\mathcal{G}^\rho_x)^n_{i+\frac{1}{2},j}, (\mathcal{G}^{\rho u}_x)^n_{i+\frac{1}{2},j}, (\mathcal{G}^{\rho v}_x)^n_{i+\frac{1}{2},j}, (\mathcal{G}^{\rho E}_x)^n_{i+\frac{1}{2},j}, (\mathcal{G}^{\rho\Pi}_x)^n_{i+\frac{1}{2},j}). \end{aligned} \tag{3.11}$$

Here, u (respectively v) denotes the velocity component in the x (respectively y) direction. Let us precise that the states $\mathbf{V}_L \equiv \mathbf{V}^n_{i,j}$ and $\mathbf{V}_R \equiv \mathbf{V}^n_{i+1,j}$ in (3.11) necessarily read in view of (3.3), (3.4):

$$\mathbf{V}^n_{i,j} = \Big(\mathbf{U}^n_{i,j},\ \rho^n_{i,j}p^n_{i,j}\Big), \quad \mathbf{V}^n_{i+1,j} = \Big(\mathbf{U}^n_{i+1,j},\ \rho^n_{i+1,j}p^n_{i+1,j}\Big). \tag{3.12}$$

We are now in a position to define the required numerical flux function $\mathbf{F}^n_{i+\frac{1}{2},j} = ((\mathbf{F}^\rho_x)^n_{i+\frac{1}{2},j}, (\mathbf{F}^{\rho u}_x)^n_{i+\frac{1}{2},j}, (\mathbf{F}^{\rho v}_x)^n_{i+\frac{1}{2},j}, (\mathbf{F}^{\rho E}_x)^n_{i+\frac{1}{2},j})$ from the formula (3.11) and (3.8):

$$\begin{aligned} &(\mathbf{F}^\rho_x)^n_{i+\frac{1}{2},j} = (\mathcal{G}^\rho_x)^n_{i+\frac{1}{2},j}, \quad (\mathbf{F}^{\rho u}_x)^n_{i+\frac{1}{2},j} = (\mathcal{G}^{\rho u}_x)^n_{i+\frac{1}{2},j}, \\ &(\mathbf{F}^{\rho v}_x)^n_{i+\frac{1}{2},j} = (\mathcal{G}^{\rho v}_x)^n_{i+\frac{1}{2},j}, \quad (\mathbf{F}^{\rho E}_x)^n_{i+\frac{1}{2},j} = (\mathcal{G}^{\rho E}_x)^n_{i+\frac{1}{2},j}. \end{aligned}$$

Observe from (3.12) that the proposed numerical flux function $\mathbf{F}^n_{i+\frac{1}{2},j}$ is consistent with the exact flux function $\mathbf{F}_x$.
Being given two states $\mathbf{V}_L$ and $\mathbf{V}_R$ in $\Omega_\mathbf{V}$, we now define for the sake of completeness

the self-similar solution $\mathcal{W}(.;\mathbf{V}_L,\mathbf{V}_R)$ of system (3.10) which expanded form reads:

$$\left\{\begin{array}{l}\partial_t\rho+\partial_x(\rho u)=0,\\ \partial_t(\rho u)+\partial_x(\rho u^2+\Pi)=0,\\ \partial_t(\rho v)+\partial_x(\rho uv)=0,\\ \partial_t(\rho E)+\partial_x((\rho E+\Pi)u)=0,\\ \partial_t(\rho\Pi)+\partial_x((\rho\Pi+a^2(\mathbf{V}_L,\mathbf{V}_R)u)=0,\end{array}\right. \tag{3.13}$$

with initial data $\mathbf{V}_0(x)=\mathbf{V}_L$, $x<0$; $\mathbf{V}_R$ otherwise. In (3.13), we propose to define the coefficient $a(\mathbf{V}_L,\mathbf{V}_R)$ in order to simultaneously satisfy the following simplified Whitham condition

$$a(\mathbf{V}_L,\mathbf{V}_R)=\max(\ (\rho c)(\mathbf{U}_L),(\rho c)(\mathbf{U}_R)\ ), \tag{3.14}$$

together with

$$\sigma_1(\mathbf{V}_L,\mathbf{V}_R)=u_L-\frac{a(\mathbf{V}_L,\mathbf{V}_R)}{\rho_L}<\sigma_2(\mathbf{V}_L,\mathbf{V}_R)=u^\star(\mathbf{V}_L,\mathbf{V}_R)<\sigma_3(\mathbf{V}_L,\mathbf{V}_R)=u_R+\frac{a(\mathbf{V}_L,\mathbf{V}_R)}{\rho_R}, \tag{3.15}$$

where

$$u^\star(\mathbf{V}_L,\mathbf{V}_R)=\frac{1}{2}(u_R+u_L)-\frac{1}{2a(\mathbf{V}_L,\mathbf{V}_R)}(\Pi_R-\Pi_L). \tag{3.16}$$

We refer the reader to Bouchut [3], Chalons and Coquel [6], for a nonlinear version of the Whitham condition (3.14) which is actually required for the validity of entropy inequalities. In the present work, we promote (3.14) for practical reasons but this simplified form must be supplemented with the requirement (3.15) which is easy to fulfil in practice. The additional condition (3.15) asks for a natural ordering of the waves in the Riemann solution : it thus sounds natural and it will be seen in the proof of the next statement to be equivalent to the property that the intermediate states in the Riemann solution are kept within the physical phase space.

Proposition 3.1. *Let be given two states $\mathbf{V}_L$ and $\mathbf{V}_R$ in $\Omega_\mathbf{V}$. Choose the coefficient $a(\mathbf{V}_L,\mathbf{V}_R)$ so as to verify (3.14) and (3.15). Then, the self-similar solution $\mathcal{W}(.,\mathbf{V}_L,\mathbf{V}_R)$ of the Cauchy problem (3.13) with initial data:*

$$\mathbf{V}_0(x)=\left\{\begin{array}{lll}\mathbf{V}_L & \text{if} & x<0,\\ \mathbf{V}_R & \text{if} & x>0,\end{array}\right. \tag{3.17}$$

is made of four constant states $\mathbf{V}_L,\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R),\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R),\mathbf{V}_R$ separated by contact discontinuities propagating with speed $\sigma_i(\mathbf{V}_L,\mathbf{V}_R)$, $i=1,2,3$:

$$\mathcal{W}(x/t;\mathbf{V}_L,\mathbf{V}_R)=\left\{\begin{array}{rll}\mathbf{V}_L & \text{if} & \frac{x}{t}<\sigma_1(\mathbf{V}_L,\mathbf{V}_R),\\ \mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R) & \text{if} & \sigma_1(\mathbf{V}_L,\mathbf{V}_R)<\frac{x}{t}<\sigma_2(\mathbf{V}_L,\mathbf{V}_R),\\ \mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R) & \text{if} & \sigma_2(\mathbf{V}_L,\mathbf{V}_R)<\frac{x}{t}<\sigma_3(\mathbf{V}_L,\mathbf{V}_R),\\ \mathbf{V}_R & \text{if} & \sigma_3(\mathbf{V}_L,\mathbf{V}_R)<\frac{x}{t}.\end{array}\right. \tag{3.18}$$

The intermediate states $\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)$ and $\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)$ belong to the phase space $\Omega_\mathbf{V}$ (i.e. $\rho_1(\mathbf{V}_L,\mathbf{V}_R)>0$, $\rho_2(\mathbf{V}_L,\mathbf{V}_R)>0$) and are recovered from the next formulas *with* $a = a(\mathbf{V}_L, \mathbf{V}_R)$*:*

$$
(3.19)\qquad
\begin{aligned}
&\Pi^\star = \Pi_1 = \Pi_2 = \tfrac{1}{2}(\Pi_L + \Pi_R) - \tfrac{a}{2}(u_R - u_L),\\
&u^\star = u_1 = u_2 = \tfrac{1}{2}(u_L + u_R) - \tfrac{1}{2a}(\Pi_R - \Pi_L),\\
&\tfrac{1}{\rho_1} = \tfrac{1}{\rho_L} - \tfrac{1}{a}(u_L - u^\star),\\
&\tfrac{1}{\rho_2} = \tfrac{1}{\rho_R} - \tfrac{1}{a}(u^\star - u_R),\\
&v_1 = v_L, \quad v_2 = v_R,\\
&E_1 = E_L - \tfrac{1}{a}(\Pi^\star u^\star - \Pi_L u_L),\\
&E_2 = E_R + \tfrac{1}{a}(\Pi^\star u^\star - \Pi_R u_R),
\end{aligned}
$$

where $u^\star = u^\star(\mathbf{V}_L, \mathbf{V}_R)$ *is given in (3.16).*

For our forthcoming purpose, it is useful to briefly sketch the proof of this statement (see [3] and [6] for a detailed proof and additional results).

Proof. The linear degeneracy of all the fields in the hyperbolic system (3.10) implies that the Riemann solution is systematically made of three contact discontinuities separating four constant states $\mathbf{V}_L$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_R$, respectively propagating with speed $\sigma_1 = \lambda_1(\mathbf{V}_L) = \lambda_1(\mathbf{V}_1)$, $\sigma_2 = \lambda_2(\mathbf{V}_1) = \lambda_2(\mathbf{V}_2)$ and $\sigma_3 = \lambda_3(\mathbf{V}_2) = \lambda_3(\mathbf{V}_R)$. The intermediate states $\mathbf{V}_1$ and $\mathbf{V}_2$ are then determined from the Rankine Hugoniot conditions that must be satisfied across each discontinuity in the solution. To conclude, observe from (3.19) that

$$
(3.20)\qquad \frac{1}{\rho_1} = \frac{1}{a}(\sigma_2 - \sigma_1), \quad \frac{1}{\rho_2} = \frac{1}{a}(\sigma_3 - \sigma_2),
$$

so that the condition (3.15) to be satisfied by the coefficient a ensures $0 < \rho_1 < +\infty$ and $0 < \rho_2 < +\infty$. ◻

3.2. Towards time implicit formulations. The relaxation approximation procedure we have discussed naturally makes use of the Godunov numerical flux function. As it is well-known, this flux function is only Lipschitz continuous and not differentiable. This lack of smoothness makes rather delicate the derivation of efficient time implicit formulations mostly based on a linearized form of the flux function. The usual way to circumvent this difficulty is to perform a partial linearization of the flux function obtained when freezing in the time expansion the unsmooth parts. This approach is very well exemplified by the Roe numerical flux function which is also only Lipschitz continuous. We refer the reader to the pioneering work by Mülder and van Leer [22]. By contrast to the Roe method, the form of the Godunov flux function does not allow for a simple localization of the non-differentiable parts so as to end up with a robust partial linearization of it. Due to the linear degeneracy property of all the fields of the relaxation system (3.13), we prove in this section that the Godunov flux function (3.11) is algebraically equivalent to a Roe flux function. The existence of an equivalent Roe-type linearization will permit to investigate the time implicit formulations of the relaxation procedure. We refer the reader to LeVeque and Pelanti [17] where the equivalence with a Roe linearization is also valid for the Jin and Xin relaxation procedure [15]. The equivalence property we claim relies on a series of statements given hereafter. The required proofs are somehow lengthy and are thus postponed to Appendix A.

PROPOSITION 3.2. *For any given pair of states $(\mathbf{V}_L, \mathbf{V}_R) \in \Omega^2_{\mathbf{V}}$, with the notations of Proposition 3.1, let us define the following five vectors of $\mathbb{R}^5$:*

$$\mathbf{r}_1(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} 1 \\ u_L - a/\rho_L \\ v_L \\ E_L + \Pi_L/\rho_L - au^\star/\rho_L \\ \Pi_L + a^2/\rho_L \end{pmatrix},$$

$$\mathbf{r}_2(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{r}_3(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{r}_4(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} 1 \\ u^\star \\ 0 \\ 0 \\ \Pi^\star \end{pmatrix},$$

$$\mathbf{r}_5(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} 1 \\ u_R + a/\rho_R \\ v_R \\ E_R + \Pi_R/\rho_R + au^\star/\rho_R \\ \Pi_R + a^2/\rho_R \end{pmatrix},$$

where $u^\star$ and $\Pi^\star$ are given in (3.19). Then, the family $(\ \mathbf{r}_i(\mathbf{V}_L, \mathbf{V}_R)\ )_{1\le i\le 5}$ spans $\mathbb{R}^5$, i.e. the matrix

$$\mathbf{R}(\mathbf{V}_L, \mathbf{V}_R) = \Big(\mathbf{r}_1(\mathbf{V}_L, \mathbf{V}_R), \mathbf{r}_2(\mathbf{V}_L, \mathbf{V}_R), \mathbf{r}_3(\mathbf{V}_L, \mathbf{V}_R), \mathbf{r}_4(\mathbf{V}_L, \mathbf{V}_R), \mathbf{r}_5(\mathbf{V}_L, \mathbf{V}_R)\Big) \tag{3.21}$$

is invertible. In addition, we have:

$$\begin{aligned} \mathbf{V}_1(\mathbf{V}_L, \mathbf{V}_R) - \mathbf{V}_L &= (\rho_1 - \rho_L)\ \mathbf{r}_1(\mathbf{V}_L, \mathbf{V}_R), \\ \mathbf{V}_2(\mathbf{V}_L, \mathbf{V}_R) - \mathbf{V}_1(\mathbf{V}_L, \mathbf{V}_R) &= (\rho_2 v_R - \rho_1 v_L)\ \mathbf{r}_2(\mathbf{V}_L, \mathbf{V}_R) \\ &\quad +(\rho_2 E_2 - \rho_1 E_1)\ \mathbf{r}_3(\mathbf{V}_L, \mathbf{V}_R) \\ &\quad +(\rho_2 - \rho_1)\ \mathbf{r}_4(\mathbf{V}_L, \mathbf{V}_R), \\ \mathbf{V}_R - \mathbf{V}_2(\mathbf{V}_L, \mathbf{V}_R) &= (\rho_R - \rho_2)\ \mathbf{r}_5(\mathbf{V}_L, \mathbf{V}_R). \end{aligned} \tag{3.22}$$

PROPOSITION 3.3. *For any given pair of states $(\mathbf{V}_L, \mathbf{V}_R) \in \Omega^2_{\mathbf{V}}$, let us consider the well-defined matrix $\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R) \in \mathcal{M}at(\mathbb{R}^5)$ given by:*

$$\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R) = \mathbf{R}(\mathbf{V}_L, \mathbf{V}_R)\mathbf{D}(\mathbf{V}_L, \mathbf{V}_R)\mathbf{R}^{-1}(\mathbf{V}_L, \mathbf{V}_R), \tag{3.23}$$

where $\mathbf{R}(\mathbf{V}_L, \mathbf{V}_R)$ is the invertible matrix introduced in (3.21) and $\mathbf{D}(\mathbf{V}_L, \mathbf{V}_R)$ the diagonal matrix defined by:

$$\mathbf{D}(\mathbf{V}_L, \mathbf{V}_R) = diag(\sigma_1(\mathbf{V}_L, \mathbf{V}_R), \sigma_2(\mathbf{V}_L, \mathbf{V}_R), \sigma_2(\mathbf{V}_L, \mathbf{V}_R), \sigma_2(\mathbf{V}_L, \mathbf{V}_R), \sigma_3(\mathbf{V}_L, \mathbf{V}_R)). \tag{3.24}$$

Then, $\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R)$ is a Roe-type linearization for the quasi-1D relaxation system (3.13); namely:

$$\begin{aligned} &(i)\ \ \mathbf{A}_x(\mathbf{V}, \mathbf{V}) = \nabla_{\mathbf{V}} \mathcal{G}_x(\mathbf{V}), \\ &(ii)\ \ \mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R)\ (\mathbf{V}_R - \mathbf{V}_L) = \mathcal{G}_x(\mathbf{V}_R) - \mathcal{G}_x(\mathbf{V}_L), \\ &(iii)\ \mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R) \quad \text{is } \mathbb{R}\text{-diagonalizable.} \end{aligned} \tag{3.25}$$

Equipped with these notations and results, the main statement of this section is:

THEOREM 3.4. *For any given pair of states $(\mathbf{V}_L, \mathbf{V}_R) \in \Omega^2_{\mathbf{V}}$, the Godunov numerical flux function for the quasi-1D relaxation system (3.13) is algebraically equivalent to the following Roe numerical flux function:*

$$\mathcal{G}_x(\mathcal{W}(0^+; \mathbf{V}_L, \mathbf{V}_R)) = \frac{1}{2}\Big(\mathcal{G}_x(\mathbf{V}_L) + \mathcal{G}_x(\mathbf{V}_R) - |\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R)|(\mathbf{V}_R - \mathbf{V}_L)\Big), \tag{3.26}$$

where $\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R)$ denotes the Roe linearization (3.23).
Let us emphasize that the matrix $\mathbf{A}_x(\mathbf{V}_L, \mathbf{V}_R) \in \mathcal{M}at(\mathbb{R}^5)$ is a Roe-type linearization for the 5×5 quasi-1D relaxation system but not for the 4×4 quasi-1D original Euler equations. Let us conclude when underlying that the numerical flux function in the y direction, $\mathcal{G}^n_{i,j+\frac{1}{2}}$, can be also equivalently reexpressed as a Roe method following symmetrical steps.

The proof of the main result of this section, Theorem 3.4, relies on the following technical lemma.

LEMMA 3.5. *For any given pair of states $(\mathbf{V}_L, \mathbf{V}_R) \in \Omega^2_{\mathbf{V}}$, the Godunov numerical flux function (3.26) equivalently reads*

$$\begin{aligned} \mathcal{G}_x(\mathcal{W}(0^+; \mathbf{V}_L, \mathbf{V}_R)) \;=\; & \tfrac{1}{2}\Big(\mathcal{G}_x(\mathbf{V}_L) + \mathcal{G}_x(\mathbf{V}_R) - (\ |\sigma_1(\mathbf{V}_L, \mathbf{V}_R)|(\mathbf{V}_1 - \mathbf{V}_L) \\ & + |\sigma_2(\mathbf{V}_L, \mathbf{V}_R)|(\mathbf{V}_2 - \mathbf{V}_1) + |\sigma_3(\mathbf{V}_L, \mathbf{V}_R)|(\mathbf{V}_R - \mathbf{V}_2)\ )\Big), \end{aligned} \tag{3.27}$$

where we have used the notations introduced in Proposition 3.1.

Recall that the reader is referred to Appendix A for the proofs of the above claims.

3.3. A first time implicit method. The two step relaxation procedure we have described in a time explicit framework can be given a straightforward time implicit formulation. For the sake of efficiency, this time implicit method is classically linearized thanks to the existence of a Roe linearization for equivalently re-expressing the numerical fluxes. We refer the reader to the work by Mülder and van Leer [22] and to the book by Hirsch [14]. More precisely, the two steps of the previous section now read:

First step: evolution in time $(t^n \to t^{n+1-})$
Solve the following linearized time implicit scheme

$$\mathbf{V}^{n+1-}_{i,j} = \mathbf{V}^n_{i,j} - \frac{\Delta t}{\Delta x}(\mathcal{G}^{n+1-}_{i+\frac{1}{2},j} - \mathcal{G}^{n+1-}_{i-\frac{1}{2},j}) - \frac{\Delta t}{\Delta y}(\mathcal{G}^{n+1-}_{i,j+\frac{1}{2}} - \mathcal{G}^{n+1-}_{i,j-\frac{1}{2}}), \quad (i,j) \in \mathbb{Z}^2, \tag{3.28}$$

where thanks to the equivalent form (3.26) we have classically set (see [22] and [14]):

$$\begin{aligned} \mathcal{G}^{n+1-}_{i+\frac{1}{2},j} = \mathcal{G}^n_{i+\frac{1}{2},j} \quad & + \quad \tfrac{1}{2}(\nabla_{\mathbf{V}}\ \mathcal{G}_x(\mathbf{V}^n_{i,j}) + |\mathbf{A}_x(\mathbf{V}^n_{i,j}, \mathbf{V}^n_{i+1,j})|)\ \delta(\mathbf{V}^n_{i,j}) \\ & + \quad \tfrac{1}{2}(\nabla_{\mathbf{V}}\ \mathcal{G}_x(\mathbf{V}^n_{i+1,j}) - |\mathbf{A}_x(\mathbf{V}^n_{i,j}, \mathbf{V}^n_{i+1,j})|)\ \delta(\mathbf{V}^n_{i+1,j}), \end{aligned} \tag{3.29}$$

where the time increments are defined by:

$$\delta(\mathbf{V}^n_{i,j}) = \mathbf{V}^{n+1-}_{i,j} - \mathbf{V}^n_{i,j}. \tag{3.30}$$

A symmetrical definition applies to the numerical flux $\mathcal{G}^{n+1-}_{i,j+\frac{1}{2}}$ in the y direction.

Solving (3.28) then classically amounts to solve a linear system in the unknown $\delta(\mathbf{V}_{i,j}^n)_{i,j\in\mathbb{Z}^2}$ with a pentadiagonal 5×5 block matrix. The non-zero entries of a block-line of the corresponding matrix read

$$(\mathbf{L}_x^n)_{i,j},\ (\mathbf{L}_y^n)_{i,j},\ (\mathbf{D}^n)_{i,j},\ (\mathbf{R}_y^n)_{i,j},\ (\mathbf{R}_x^n)_{i,j}, \tag{3.31}$$

where the diagonal 5×5 matrix writes

$$\begin{aligned}\mathbf{D}_{i,j}^n = \mathbf{Id} + \tfrac{\Delta t}{2\Delta x}\Big(&|\mathbf{A}_x(\mathbf{V}_{i,j}^n,\mathbf{V}_{i+1,j}^n)| + |\mathbf{A}_x(\mathbf{V}_{i-1,j}^n,\mathbf{V}_{i,j}^n)|\Big)\\ +\tfrac{\Delta t}{2\Delta y}\Big(&|\mathbf{A}_y(\mathbf{V}_{i,j}^n,\mathbf{V}_{i,j+1}^n)| + |\mathbf{A}_y(\mathbf{V}_{i,j-1}^n,\mathbf{V}_{i,j}^n)|\Big),\end{aligned} \tag{3.32}$$

while the extradiagonal 5×5 matrices are defined by

$$\begin{aligned}(\mathbf{L}_x^n)_{i,j} &= -\tfrac{\Delta t}{2\Delta x}\Big(\nabla_{\mathbf{V}}\mathcal{G}_x(\mathbf{V}_{i-1,j}^n) + |\mathbf{A}_x(\mathbf{V}_{i-1,j}^n,\mathbf{V}_{i,j}^n)|\Big),\\ (\mathbf{R}_x^n)_{i,j} &= +\tfrac{\Delta t}{2\Delta x}\Big(\nabla_{\mathbf{V}}\mathcal{G}_x(\mathbf{V}_{i+1,j}^n) - |\mathbf{A}_x(\mathbf{V}_{i,j}^n,\mathbf{V}_{i+1,j}^n)|\Big),\end{aligned} \tag{3.33}$$

with symmetrical definitions for $(\mathbf{L}_y^n)_{i,j}$ and $(\mathbf{R}_y^n)_{i,j}$.

Second step: Relaxation $(t^{n+1-}\to t^{n+1})$
From the solution $\mathbf{V}_{i,j}^{n+1-}$ of the above linear problem (3.28), we keep unchanged:

$$\rho_{i,j}^{n+1} = \rho_{i,j}^{n+1-},\ (\rho\mathbf{w})_{i,j}^{n+1} = (\rho\mathbf{w})_{i,j}^{n+1-},\ (\rho E)_{i,j}^{n+1} = (\rho E)_{i,j}^{n+1-},$$

while we define so as to enforce equilibrium:

$$(\rho\Pi)_{i,j}^{n+1} = (\rho p)(\rho_{i,j}^{n+1},(\rho e)_{i,j}^{n+1}).$$

In other words, the second step is kept unchanged. This concludes the presentation of the first time implicit formulation of the relaxation approximation procedure.

Despite being natural and robust, this first extension is numerically shown hereafter to fail in producing converged in time discrete solutions. The origin of the failure may be understood as follows. The steady state solutions of the Euler equations (2.1), namely solutions of

$$\nabla.\ \mathbf{F}(\mathbf{U}) = 0,\quad \mathbf{x}\in\mathcal{D}, \tag{3.34}$$

are intended to be recovered from the steady state solutions of the relaxation system (2.6)

$$\nabla.\ \mathcal{G}(\mathbf{V}^\lambda) - \lambda\mathcal{R}(\mathbf{V}^\lambda) = 0,\quad \mathbf{x}\in\mathcal{D}, \tag{3.35}$$

in the regime of an infinite relaxation rate $\lambda\to\infty$. Observe that one cannot expect stationary solutions $\mathbf{V}^\lambda$ of (3.35) to simultaneously satisfy $\nabla.\mathcal{G}(\mathbf{V}^\lambda)=0$ and $\mathcal{R}(\mathbf{V}^\lambda)=0$ generally speaking in view of the next result:

Proposition 3.6. *Let be given $\mathbf{V}:\mathcal{D}\to\Omega_{\mathbf{V}}$ a smooth function of the space variables such that:*

$$\nabla.\ \mathcal{G}(\mathbf{V}) = 0 \qquad \textit{and} \qquad \mathcal{R}(\mathbf{V}) = 0, \qquad \mathbf{x}\in\mathcal{D}, \tag{3.36}$$

simultaneously hold. Then, the smooth function $\mathbf{V}$ necessarily obeys:

$$(a^2 - \rho^2c^2(\rho,\rho e))\ \nabla.\ \mathbf{w} = 0, \qquad \mathbf{x}\in\mathcal{D}, \tag{3.37}$$

where $c(\rho, \rho e)$ is the sound speed introduced in (2.3). The proof is postponed at the end of the paragraph. Under the mandatory Whitham condition (2.12), a smooth function $\mathbf{V}$ satisfying (3.36) and therefore (3.37) necessarily comes with the property of a divergence free velocity field $\mathbf{w}$, i.e. $\nabla.\ \mathbf{w} = 0$. From $\nabla.\ \rho\mathbf{w} = 0$ stated in (3.36), we would then infer $\mathbf{w}.\nabla\rho = 0$ and thus either ρ necessarily stays constant or the velocity vanishes. Such conditions are far from being general. Therefore, stationary solutions of (3.35) do not obey (3.36) in general. In other words, the singular relaxation source term in (3.35) must come into proper balance with the flux divergence. Here stays the reason of the reported failure in the proposed time marching strategy for capturing steady state solutions of (3.34) via those of (3.35).
Indeed, the discussed time marching method intends to restore solutions of (3.35) on the basis of a splitting strategy in between the flux divergence and the relaxation source term: solve first the frozen relaxation system (2.6) choosing $\lambda = 0$

$$\partial_t \mathbf{V} + \nabla.\ \mathcal{G}(\mathbf{V}) = 0,$$

to then restore the relaxation effects when solving

$$\partial_t \mathbf{V}^\lambda - \lambda \mathcal{R}(\mathbf{V}^\lambda) = 0,$$

in the limit $\lambda \to \infty$. Formally, time convergence in this splitting strategy to some stationary solutions $\mathbf{V}$ would require $\nabla.\ \mathcal{G}(\mathbf{V}) = 0$ together with $\lambda\mathcal{R}(\mathbf{V}^\lambda) \to 0$ in the limit $\lambda \to \infty$, properties which cannot hold for general solutions of (3.35). In other words, splitting the relaxation source term from the flux divergence cannot result in a well-balanced approximation of the solutions of (3.35) and therefore of (3.34).

We end this section when establishing Proposition 3.6.

Proof. By definition, a function $\mathbf{V}$ with the property $\mathcal{R}(\mathbf{V}) = 0$, $x \in \mathcal{D}$ is such that:

$$\Pi(x) = p(\rho(x), \rho e(x)), \quad \mathbf{x} \in \mathcal{D}. \tag{3.38}$$

If in addition it obeys $\nabla.\mathcal{G}(\mathbf{V}) = 0$ according to (3.36), one infers from (3.38) that:

$$\begin{aligned} \nabla.\ \rho\mathbf{w} &= 0, \\ \nabla.\ (\rho\mathbf{w} \otimes \mathbf{w} + p\ \mathbf{I_d}) &= 0, \\ \nabla.\ (\rho E\mathbf{w} + p\mathbf{w}) &= 0, \end{aligned} \tag{3.39}$$

together with

$$\nabla.\ (\rho p\mathbf{w}) + a^2 \nabla.\ \mathbf{w} = 0. \tag{3.40}$$

Decomposing $\mathbf{V} = (\mathbf{U}, \rho\Pi = \rho p)$ gives in view of (3.39) that $\mathbf{U}$ is nothing but a stationary solution of the Euler equation, i.e. obeying $\nabla.\mathbf{F}(\mathbf{U}) = 0$ as stated in (3.34). In view of the balance equation (2.17) for governing the pressure p, such a solution obeys:

$$\rho\ \mathbf{w}.\nabla p + \rho^2 c^2(\mathbf{U})\nabla.\ \mathbf{w} = 0, \quad \mathbf{x} \in \mathcal{D}, \tag{3.41}$$

which recasts as:

$$\nabla.\ (\rho\mathbf{w}p) + \rho c^2(\mathbf{U})\nabla.\ \mathbf{w} = 0, \tag{3.42}$$

invoking $\nabla.\ (\rho\mathbf{w}) = 0$ from (3.39). Substracting (3.42) from (3.40) yields the required identity (3.37). ∎

3.4. Well-balanced time implicit formulation in a relaxation framework. In the light of the previous section, the correct design of a time implicit relaxation procedure requires to handle simultaneously the relaxation source term with the flux divergence during the first evolution step $t^n \to t^{n+1-}$. We propose to adopt the following strategy, still made of two steps for reasons we will explain on due time.

First step: evolution in time $(t^n \to t^{n+1-})$

Instead of the the frozen version (3.5) with $\lambda = 0$, we have to approximate at time Δt the solution of the following Cauchy problem

$$\begin{cases} \partial_t \mathbf{V}^\lambda + \nabla.\, \mathcal{G}(\mathbf{V}^\lambda) = \lambda \mathcal{R}(\mathbf{V}^\lambda), \\ \mathbf{V}^\lambda(x,y,0) = \mathbf{V}_h(x,y,t^n), \end{cases} \tag{3.43}$$

in the regime of an infinite relaxation rate $\lambda \to \infty$. The initial data $\mathbf{V}_h(x,y,t^n)$ is again built at equilibrium from $\mathbf{U}_h(x,y,t^n)$ according to (3.3)-(3.4).
In order to derive the required approximate solution, let us start from the following direct extension of (3.28)

$$\mathbf{V}_{i,j}^{\lambda,\, n+1-} = \mathbf{V}_{i,j}^{\lambda,\, n} - \frac{\Delta t}{\Delta x}(\mathcal{G}_{i+\frac12,j}^{\lambda,\, n+1-} - \mathcal{G}_{i-\frac12,j}^{\lambda,\, n+1-}) - \frac{\Delta t}{\Delta y}(\mathcal{G}_{i,j+\frac12}^{\lambda,\, n+1-} - \mathcal{G}_{i,j-\frac12}^{\lambda,\, n+1-}) + \lambda \Delta t \mathcal{R}(\mathbf{V}_{i,j}^{\lambda,\, n+1-}), \tag{3.44}$$

which we have to deal with in the limit $\lambda \to \infty$. To cope with this limit, let us rewrite the last discrete equation in (3.44) for updating the relaxation pressure, as follows

$$\begin{aligned} \Big(\rho\Pi\Big)_{i,j}^{\lambda,\, n+1-} &= \Big(\rho p(\rho,\rho e)\Big)_{i,j}^{\lambda,\, n+1-} \\ &- \tfrac{1}{\lambda}\Big\{ \tfrac{(\rho\Pi)_{i,j}^{\lambda,\, n+1-} - (\rho p(\rho,\rho e))_{i,j}^n}{\Delta t} + \tfrac{\mathcal{G}_{i+\frac12,j}^{\lambda,\, n+1-} - \mathcal{G}_{i-\frac12,j}^{\lambda,\, n+1-}}{\Delta x} + \tfrac{\mathcal{G}_{i,j+\frac12}^{\lambda,\, n+1-} - \mathcal{G}_{i,j-\frac12}^{\lambda,\, n+1-}}{\Delta y} \Big\}, \end{aligned} \tag{3.45}$$

which is nothing but a time implicit discrete form of (2.10). Under the Whitham condition (3.14) for the sake of stability, we formally let λ go to infinity in (3.45) to consider the implicit formula

$$\Big(\rho\Pi\Big)_{i,j}^{n+1-} = \Big(\rho p(\rho,\rho e)\Big)_{i,j}^{n+1-} \equiv \Big(\rho p(\rho,\rho\mathbf{w},\rho E)\Big)_{i,j}^{n+1-}.$$

To lower the computational effort due to the nonlinear pressure law $p(\mathbf{U})$, we propose to Taylor expand this implicit formula so as to consider the following final definition

$$\begin{aligned} \Big(\rho\Pi\Big)_{i,j}^{n+1-} &= \Big(\rho p(\rho,\rho e)\Big)_{i,j}^{n} + \Big(p(\mathbf{U}) + \rho\tfrac{\partial p}{\partial \rho}(\mathbf{U})\Big)_{i,j}^{n}\Big(\rho_{i,j}^{n+1-} - \rho_{i,j}^n\Big) \\ &+ (\rho\nabla_{\mathbf{w}} p(\mathbf{U}))_{i,j}^n\Big((\rho\mathbf{w})_{i,j}^{n+1-} - (\rho\mathbf{w})_{i,j}^n\Big) + (\rho\tfrac{\partial p}{\partial \rho E}(\mathbf{U}))_{i,j}^n\Big((\rho E)_{i,j}^{n+1-} - (\rho E)_{i,j}^n\Big). \end{aligned} \tag{3.46}$$

It is then convenient to recast (3.46) in terms of the time increments introduced in (3.30) to get the next identity

$$\begin{aligned} &\Big(p(\mathbf{U}) + (\rho\tfrac{\partial p}{\partial \rho}(\mathbf{U})\Big)_{i,j}^{n}\, \delta\Big(\rho_{i,j}^n\Big) + (\rho\nabla_{\mathbf{w}} p(\mathbf{U}))_{i,j}^n\, \delta\Big((\rho\mathbf{w})_{i,j}^n\Big) \\ &+ (\rho\tfrac{\partial p}{\partial \rho E}(\mathbf{U}))_{i,j}^n\, \delta\Big((\rho E)_{i,j}^n\Big) - \delta\Big((\rho\Pi)_{i,j}^n\Big) = 0, \end{aligned} \tag{3.47}$$

since by construction $\Big(\rho\Pi\Big)_{i,j}^n = \Big(\rho p(\rho,\rho e)\Big)_{i,j}^n$ in (3.4).
Equipped with this identity, we are in a position to state the linear problem to be

solved in the unknown $\delta\mathbf{V}^n_{i,j}$. In that aim, we define the pentadiagonal 5×5 block matrix entering this linear problem from the block matrices previously introduced in (3.31), (3.32) and (3.33)

$$(\tilde{\mathbf{L}_x})^n_{i,j},\quad (\tilde{\mathbf{L}_y})^n_{i,j},\quad (\tilde{\mathbf{D}})^n_{i,j},\quad (\tilde{\mathbf{R}_y})^n_{i,j},\quad (\tilde{\mathbf{R}_x})^n_{i,j}, \tag{3.48}$$

where each four first lines of the 5×5 matrices corresponds respectively to the four first lines of:

$$(\mathbf{L}_x)^n_{i,j},\quad (\mathbf{L}_y)^n_{i,j},\quad (\mathbf{D})^n_{i,j},\quad (\mathbf{R}_y)^n_{i,j},\quad (\mathbf{R}_x)^n_{i,j}, \tag{3.49}$$

while solely the last line of each of the 5×5 matrices (3.49) governing the time increments $\delta(\rho\Pi)$ have been modified in the new block line (3.48) to account for the new update formula (3.47). Since this formula written in a given cell $\mathcal{C}_{ij}$ only involves the time increment $\delta\mathbf{V}^n_{i,j}$, the last line of $(\tilde{\mathbf{L}_x})^n_{i,j}$, $(\tilde{\mathbf{L}_y})^n_{i,j}$ and $(\tilde{\mathbf{R}_x})^n_{i,j}$, $(\tilde{\mathbf{R}_y})^n_{i,j}$ are necessarily set identically to the zero line

$$(0,0,0,0,0),$$

while necessarily the last line of the diagonal matrix $\tilde{\mathbf{D}}^n_{ij}$ reads:

$$\Big(p+\Big(\rho\frac{\partial p}{\partial \rho}\Big)^n_{ij},\quad \Big(\rho\frac{\partial p}{\partial \rho u}\Big)^n_{ij},\quad \Big(\rho\frac{\partial p}{\partial \rho v}\Big)^n_{ij},\quad \Big(\rho\frac{\partial p}{\partial \rho E}\Big)^n_{ij},\quad -1\Big). \tag{3.50}$$

At last, the corresponding component of the right side is set to zero so as to restore (3.47) with (3.50).
This completes the description of the first step in the well-balanced time implicit formulation of the relaxation scheme.
The need for a second step stems from the linearized version (3.46) we have introduced at time level t^{n+1-} to reduce the computational effort involved in the initial guess $(\rho\Pi)^{n+1-}_{i,j} = (\rho p(\rho,\rho e))^{n+1-}_{i,j}$. Since equilibrium is not achieved with the linearized form (3.46), a second step is required to be in position to restart the procedure from time t^{n+1}. Since this second step just asks for the identity $(\rho\Pi)^{n+1}_{i,j} = (\rho p(\rho,\rho e))^{n+1}_{i,j}$, this step exactly coincides with the second step described in Section 3.3. This concludes the presentation of the method.

Numerical evidences, discussed in the next paragraph, prove the robustness of the proposed correction together with its efficiency. Let us again underline that the Whitham condition (2.12) plays a major role in the stability of the well-balanced time implicit procedure. Indeed, its design principle based on the formula (3.45) with large values of λ has clear connexion with the Chapman-Enskog expansion detailed in section 2. The proposed procedure can thus be understood as a consistent approximation of the near-equilibrium system (2.19), which is a viscous (i.e. stable) perturbation of the original Euler equation if and only if the Whitham condition is satisfied. To go further, it is useful to observe that the pentadiagonal 5×5 block matrices introduced in (3.48) can be reduced to pentadiagonal 4×4 block matrices when eliminating $\delta\Big((\rho\Pi)^n_{i,j}\Big)$ from the unknowns $\delta\Big(\rho^n_{i,j}\Big)$, $\delta\Big((\rho\mathbf{w})^n_{i,j}\Big)$ and $\delta\Big((\rho E)^n_{i,j}\Big)$ thanks to the formula (3.46). These calculations are straightforward and are left to the reader. The CPU effort in our procedure is thus the expected one for the implicit time discretization of the 2D Euler equations (2.1) or their 4×4 near-equilibrium version (2.19).

4. Numerical illustrations. We investigate the performance of the proposed time implicit formulations in the approximation of the steady state solution of the Euler equations over a blunt body. For simplicity, the pressure law is the one of a polytropic gas with adiabatic coefficient $\gamma = 1.2$. The free-stream conditions follow from a Mach number set to $M_\infty = 10$ and a static pressure $p_\infty = 40 Pa$: they are responsible for the existence a strong bow shock in the steady state solution. The computational domain consists in a curvilinear mesh made of 60×48 cells.
Figure 1 shows the time history of the L^2 norm of the density time derivative obtained using the first time implicit relaxation method. About 15000 time steps have been performed according to the following CFL strategy: the CFL number is set to the constant value 25 during the first 7000 time iterations and decreased down to CFL= 5. Such a strategy makes use of rather small CFL numbers to prove that the two plateaus achieved in the convergence history is characteristic of a time implicit method which fails to produce converged in time discrete solutions. By contrast, the time history of the L^2-norm of the density time derivative obtained thanks to the well-balanced time implicit relaxation method as depicted in Figure 2 proves perfect convergence in time for the discrete solutions. The results of the two runs are compared, respectively using a constant CFL number set to 25 for the sake of comparison and then choosing an increased value CFL= 200 in order to speed up the calculation. At last, Figure 3 displays the density contours in the steady solution obtained with CFL= 200.

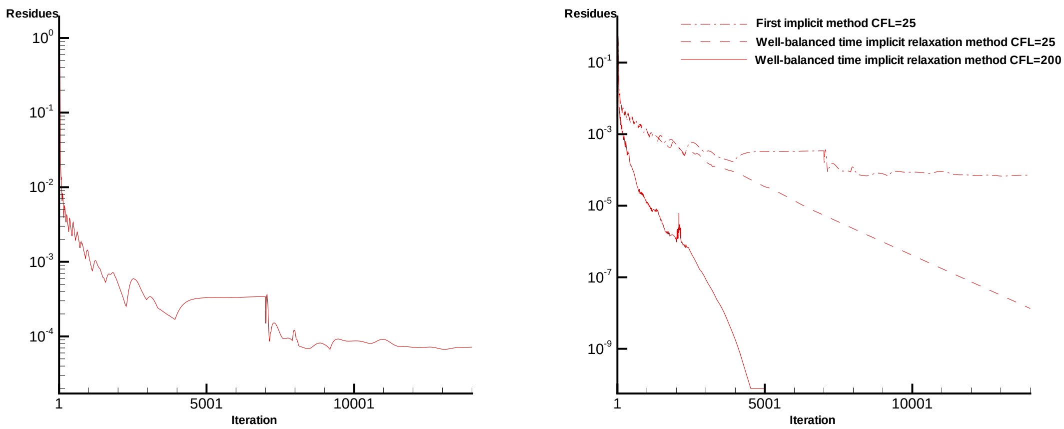

FIG. 4.1. *Time history of the L^2 norm of the density time derivative using the proposed time implicit relaxation schemes*

Acknowledgments. The second author has been partly financially supported by ONERA during the completion of this work.

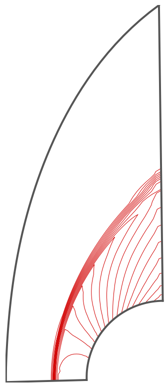

FIG. 4.2. *Density contours with the well-balanced time implicit method at* $CFL = 200$

Appendix A. . Here we give the detailed proofs of the statements proposed in Section 3.2. We first establish Proposition 3.2.

Proof. With the notations of Section 3.2, the five vectors $(\ \mathbf{r}_i(\mathbf{V}_L,\mathbf{V}_R)\ )_{1\le i\le 5}$ are shown in a first step to span $\mathbb{R}^5$ when proving that the identity

$$\sum_{i=1}^{5} \alpha_i\ \mathbf{r}_i(\mathbf{V}_L,\mathbf{V}_R) = 0 \tag{A.1}$$

implies $\alpha_i = 0$ for $i = 1,...,5$. Indeed, under the condition (3.15), namely:

$$\sigma_1(\mathbf{V}_L,\mathbf{V}_R) < \sigma_2(\mathbf{V}_L,\mathbf{V}_R) < \sigma_3(\mathbf{V}_L,\mathbf{V}_R), \tag{A.2}$$

let us check that

$$\begin{vmatrix} 1 & 1 & 1 \\ \sigma_1(\mathbf{V}_L,\mathbf{V}_R) & \sigma_2(\mathbf{V}_L,\mathbf{V}_R) & \sigma_3(\mathbf{V}_L,\mathbf{V}_R) \\ \Pi_L + \frac{a^2}{\rho_L} & \Pi^\star & \Pi_R + \frac{a^2}{\rho_R} \end{vmatrix} \neq 0, \tag{A.3}$$

so that from the first, second and last components of the vectors $\mathbf{r}_1(\mathbf{V}_L,\mathbf{V}_R)$, $\mathbf{r}_4(\mathbf{V}_L,\mathbf{V}_R)$ and $\mathbf{r}_5(\mathbf{V}_L,\mathbf{V}_R)$, the identity (A.1) yields

$$\alpha_1 = \alpha_4 = \alpha_5 = 0.$$

In turn, we successively deduce $\alpha_2 = 0$ and $\alpha_3 = 0$ respectively from the third component of $\mathbf{r}_2(\mathbf{V}_L,\mathbf{V}_R)$ and the fourth one of $\mathbf{r}_3(\mathbf{V}_L,\mathbf{V}_R)$. Hence the required result.

To prove (A.3), observe that the proposed determinant reads equivalently:

$$\begin{vmatrix} 0 & 1 & 0 \\ \sigma_1-\sigma_2 & \sigma_2 & \sigma_3-\sigma_2 \\ \Pi_L + \frac{a^2}{\rho_L} - \Pi^\star & \Pi^\star & \Pi_R + \frac{a^2}{\rho_R} - \Pi^\star \end{vmatrix} = - \begin{vmatrix} \sigma_1-\sigma_2 & \sigma_3-\sigma_2 \\ a(\sigma_2-\sigma_1) & a(\sigma_3-\sigma_2) \end{vmatrix} \tag{A.4}$$

$$= 2a(\sigma_2-\sigma_1)(\sigma_3-\sigma_2) > 0,$$

in view of the condition (A.2). We have used the next identities inferred from the definition $\Pi^\star$ and $u^\star$ in (3.19) of Proposition 3.1:

$$\begin{aligned}\Pi_L + \frac{a^2}{\rho_L} - \Pi^\star &= a\ \Big(\ \frac{1}{2}(u_R+u_L) - \frac{1}{2a}(p_R-p_L)\ \Big) - a(u_L - \frac{a}{\rho_L}),\\ &= a(\sigma_2-\sigma_1),\end{aligned}$$

and

$$\begin{aligned}\Pi_R + \frac{a^2}{\rho_R} - \Pi^\star &= -a\ \Big(\ \frac{1}{2}(u_R+u_L) - \frac{1}{2a}(p_R-p_L)\ \Big) + a(u_R + \frac{a}{\rho_R}),\\ &= a(\sigma_3-\sigma_2).\end{aligned}$$

The identities stated in (3.22) result from direct calculations based on the detailed form of the two states $\mathbf{V}_1$, $\mathbf{V}_2$ given in Proposition 3.1. The details are left to the reader. This concludes the proof. □

We are in a position to give the proof of Proposition 3.3.

Proof. For any given states $\mathbf{V}_L$ and $\mathbf{V}_R$ in $\Omega_\mathbf{V}$, the matrix $\mathbf{R}(\mathbf{V}_L,\mathbf{V}_R)$ is invertible in view of Proposition 3.2 and thus makes well-defined the matrix $\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)$ introduced in (3.23). Let us prove that this matrix is a Roe-type linearization for the quasi-1D relaxation system (3.13). First observe that the definition (3.23) just expresses that the matrix $\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)$ is $\mathbb{R}$-diagonalizable as expected in (3.25) property (*iii*). Next, direct calculations yield that for any given $\mathbf{V}\in\Omega_\mathbf{V}$ the Jacobian matrix $\nabla_\mathbf{V}\mathcal{G}_x(\mathbf{V})$ admits the following eigenvalues

$$\sigma_1(\mathbf{V},\mathbf{V}) < \sigma_2(\mathbf{V},\mathbf{V}) < \sigma_3(\mathbf{V},\mathbf{V}), \tag{A.5}$$

where $\sigma_2(\mathbf{V},\mathbf{V})$ has three order of multiplicity. In addition, $\mathbf{r}_1(\mathbf{V},\mathbf{V})$ (respectively $\mathbf{r}_5(\mathbf{V},\mathbf{V})$) is a right eigenvector associated with $\sigma_1(\mathbf{V},\mathbf{V})$ (respectively $\sigma_3(\mathbf{V},\mathbf{V})$) while the three vectors $(\mathbf{r}_2(\mathbf{V},\mathbf{V}),\mathbf{r}_3(\mathbf{V},\mathbf{V}),\mathbf{r}_4(\mathbf{V},\mathbf{V}))$ yield a basis of the eigenspace associated with the eigenvalue $\sigma_2(\mathbf{V},\mathbf{V})$. In other words, the first consistency property (*i*) stated in (3.25) is valid.

To conclude, we have to check that the property (*ii*) holds true. To that purpose, let us start from the next identity:

$$\mathbf{V}_R-\mathbf{V}_L = (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))+(\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R))+(\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L), \tag{A.6}$$

so that

$$\begin{aligned}\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_L) = {}&\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))\\ &+\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R))\\ &+\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L).\end{aligned}$$

In view of the formula (3.22), we infer that:

$$\begin{aligned}\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_L) \quad = \quad &\sigma_1(\mathbf{V}_L,\mathbf{V}_R)(\rho_R-\rho_2)\mathbf{r}_5(\mathbf{V}_L,\mathbf{V}_R)\\ &+\sigma_2(\mathbf{V}_L,\mathbf{V}_R)\ \big(\ (\rho_2 v_R-\rho_1 v_L)\mathbf{r}_2(\mathbf{V}_L,\mathbf{V}_R)+\\ &(\rho_2E_2-\rho_1E_1)\mathbf{r}_3(\mathbf{V}_L,\mathbf{V}_R)+(\rho_2-\rho_1)\mathbf{r}_4(\mathbf{V}_L,\mathbf{V}_R)\ \big)\\ &+\sigma_3(\mathbf{V}_L,\mathbf{V}_R)(\rho_1-\rho_L)\mathbf{r}_1(\mathbf{V}_L,\mathbf{V}_R),\\ = \quad &\sigma_1(\mathbf{V}_L,\mathbf{V}_R)(\mathbf{V}_R-\mathbf{V}_2)\\ &+\sigma_2(\mathbf{V}_L,\mathbf{V}_R)(\mathbf{V}_2-\mathbf{V}_1)\\ &+\sigma_3(\mathbf{V}_L,\mathbf{V}_R)(\mathbf{V}_1-\mathbf{V}_L).\end{aligned} \tag{A.7}$$

But by definition, $\sigma_1(\mathbf{V}_L,\mathbf{V}_R)$, $\sigma_2(\mathbf{V}_L,\mathbf{V}_R)$ and $\sigma_3(\mathbf{V}_L,\mathbf{V}_R)$ are velocities of the contact discontinuities involved in the Riemann solution (3.18) in Proposition 3.1. Therefore, the Rankine Hugoniot jump relations allow to recast (A.7) according to:

$$\begin{aligned}\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_L) &= (\mathcal{G}_x(\mathbf{V}_R)-\mathcal{G}_x(\mathbf{V}_2))+(\mathcal{G}_x(\mathbf{V}_2)-\mathcal{G}_x(\mathbf{V}_1))\\ &\quad +(\mathcal{G}_x(\mathbf{V}_1)-\mathcal{G}_x(\mathbf{V}_L)),\\ &= \mathcal{G}_x(\mathbf{V}_R)-\mathcal{G}_x(\mathbf{V}_L),\end{aligned}\tag{A.8}$$

which is nothing but the required result. □

We now turn proving Lemma 3.5.

Proof. Let us consider the self-similar solution $\mathcal{W}(.,\mathbf{V}_L,\mathbf{V}_R)$ (3.18) of the Riemann problem for the quasi-1D relaxation system (3.13) and define the next two half-averages (see Harten, Lax and van Leer [25] for instance):

$$\overline{\mathbf{V}}_L=\frac{2}{\Delta x}\int_{-\frac{\Delta x}{2}}^{0}\mathcal{W}(\frac{x}{\Delta t},\mathbf{V}_L,\mathbf{V}_R)\ dx,\tag{A.9}$$

$$\overline{\mathbf{V}}_R=\frac{2}{\Delta x}\int_{0}^{\frac{\Delta x}{2}}\mathcal{W}(\frac{x}{\Delta t},\mathbf{V}_L,\mathbf{V}_R)\ dx.\tag{A.10}$$

Classical arguments [25] based on the conservation form of the system (3.13) yield:

$$\overline{\mathbf{V}}_L=\mathbf{V}_L-\frac{2\Delta t}{\Delta x}(\ \mathcal{G}_x(\mathcal{W}(0^+;\mathbf{V}_L,\mathbf{V}_R))-\mathcal{G}_x(\mathbf{V}_L)\ ),\tag{A.11}$$

$$\overline{\mathbf{V}}_R=\mathbf{V}_R-\frac{2\Delta t}{\Delta x}(\ \mathcal{G}_x(\mathbf{V}_R)-\mathcal{G}_x(\mathcal{W}(0^+;\mathbf{V}_L,\mathbf{V}_R))\ ).\tag{A.12}$$

But using the property that the self-similar function $\mathcal{W}(.,\mathbf{V}_L,\mathbf{V}_R)$ is piecewise constant, direct calculations allow to reexpress equivalently the averages (A.9), (A.10) as follows:

$$\begin{aligned}\overline{\mathbf{V}}_L &= \tfrac{2}{\Delta x}\Big((\tfrac{\Delta x}{2}+\sigma_1^-(\mathbf{V}_L,\mathbf{V}_R)\Delta t)\mathbf{V}_L+\Delta t\ (\sigma_2^-(\mathbf{V}_L,\mathbf{V}_R)-\sigma_1^-(\mathbf{V}_L,\mathbf{V}_R))\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)\\ &\quad +\Delta t\ (\sigma_3^-(\mathbf{V}_L,\mathbf{V}_R)-\sigma_2^-(\mathbf{V}_L,\mathbf{V}_R))\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\Delta t\ \sigma_3^-(\mathbf{V}_L,\mathbf{V}_R)\mathbf{V}_R\Big),\\ &= \mathbf{V}_L-\tfrac{2\Delta t}{\Delta x}\Big(\sigma_1^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L)\\ &\qquad +\sigma_2^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R))\\ &\qquad +\sigma_3^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))\Big),\end{aligned}\tag{A.13}$$

and

$$\begin{aligned}\overline{\mathbf{V}}_R &= \mathbf{V}_R-\tfrac{2\Delta t}{\Delta x}\Big(\sigma_1^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L)\\ &\qquad +\sigma_2^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R))\\ &\qquad +\sigma_3^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))\Big),\end{aligned}\tag{A.14}$$

with the usual notations $\sigma^-=\min(0,\sigma)$ and $\sigma^+=\max(0,\sigma)$. Hence, substracting (A.13) from (A.11) yields the first formula for the Godunov flux function

$$\begin{aligned}\mathcal{G}_x(\mathcal{W}(0^+;\mathbf{V}_L,\mathbf{V}_R)) &= \mathcal{G}_x(\mathbf{V}_L)+\Big(\sigma_1^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L)\\ &\qquad +\sigma_2^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R))\\ &\qquad +\sigma_3^-(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))\Big),\end{aligned}\tag{A.15}$$

while equalizing (A.12) and (A.14) provides the second formula

$$\begin{aligned}\mathcal{G}_x(\mathcal{W}(0^+;\mathbf{V}_L,\mathbf{V}_R)) &= \mathcal{G}_x(\mathbf{V}_R) - \Big(\sigma_1^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L) \\ &\quad +\sigma_2^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)) \\ &\quad +\sigma_3^+(\mathbf{V}_L,\mathbf{V}_R)\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R))\Big).\end{aligned} \tag{A.16}$$

At last, the arithmetic average of (A.15) and (A.16) gives the required result (3.27) in view of the identity $|\sigma| = \sigma^+ - \sigma^-$. □

We conclude this appendix when establishing Theorem 3.4.

Proof. It suffices to establish the next identity:

$$\begin{aligned}|\mathbf{A}_x(\mathbf{V}_L,\mathbf{V}_R)| &= |\sigma_1(\mathbf{V}_L,\mathbf{V}_R)|\ (\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_L) \\ &\quad +|\sigma_2(\mathbf{V}_L,\mathbf{V}_R)|\ (\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)-\mathbf{V}_1(\mathbf{V}_L,\mathbf{V}_R)) \\ &\quad +|\sigma_3(\mathbf{V}_L,\mathbf{V}_R)|\ (\mathbf{V}_R-\mathbf{V}_2(\mathbf{V}_L,\mathbf{V}_R)),\end{aligned} \tag{A.17}$$

to infer from the flux formula (3.27) stated in Lemma 3.5 the required equivalent form (3.26). Invoking the identities (3.22), we successively get, when omitting the dependency in $\mathbf{V}_L$ and $\mathbf{V}_R$ for simplicity:

$$\begin{aligned}|\sigma_1|\ (\mathbf{V}_1-\mathbf{V}_L) &+ |\sigma_2|\ (\mathbf{V}_2-\mathbf{V}_1) + |\sigma_3|\ (\mathbf{V}_R-\mathbf{V}_2) \\ &= (\rho_1-\rho_L)\ |\sigma_1|\ \mathbf{r}_1 + (\rho_2 v_R - \rho_1 v_L)\ |\sigma_2|\ \mathbf{r}_2 \\ &\quad +(\rho_2 E_2 - \rho_1 E_1)\ |\sigma_2|\ \mathbf{r}_3 + (\rho_2-\rho_1)\ |\sigma_2|\ \mathbf{r}_4 \\ &\quad +(\rho_R-\rho_2)\ |\sigma_3|\ \mathbf{r}_5, \\ &= (\rho_1-\rho_L)\ |\mathbf{A}_x|\ \mathbf{r}_1 + (\rho_2 v_R - \rho_1 v_L)\ |\mathbf{A}_x|\ \mathbf{r}_2 \\ &\quad +(\rho_2 E_2 - \rho_1 E_1)\ |\mathbf{A}_x|\ \mathbf{r}_3 + (\rho_2-\rho_1)\ |\mathbf{A}_x|\ \mathbf{r}_4 \\ &\quad +(\rho_R-\rho_2)\ |\mathbf{A}_x|\ \mathbf{r}_5, \\ &= |\mathbf{A}_x|(\mathbf{V}_R-\mathbf{V}_L).\end{aligned} \tag{A.18}$$

This is the required result. □

REFERENCES

[1] M. Baudin, C. Berthon, F. Coquel and Q.H. Tran, *A relaxation method for two-phase flows models with hydrodynamic closure laws*, Numer. Math., Vol. 99(3), (2005), pp. 411–440.

[2] C. Berthon, *Numerical approximation of the 10-moment Gaussian closure*, Math. Comp., Vol. 75, (2006), pp. 1809–1831.

[3] F. Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for source*, Frontiers in Mathematics Series, Birkhauser, 2004.

[4] F. Bouchut, S. Jin and X. Li, *Numerical approximations of pressureless and isothermal gaz dynamics*, SIAM J. Numer. Anal., Vol. 41, (2003), pp. 135–158.

[5] C. Chalons, *Bilans d'entropie discrets dans l'approximation numérique des chocs non classiques. Application aux équations de Navier-Stokes multi-pression et à quelques systèmes visco-capillaires*, Thèse de Doctorat de l'Ecole Polytechnique, in French, 2002.

[6] C. Chalons and F. Coquel, *Navier-Stokes equations with several independent pressure laws and explicit predictor-corrector schemes*, Numer. Math., Vol. 101(3), (2005), pp. 451–478.

[7] C. Chalons and J.F. Coulombel, *Relaxation approximation of the Euler equations*, submitted.

[8] G.Q. Chen, D. Levermore, and T.P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., Vol. 48(7), (1995), pp. 787–830.

[9] F. Coquel, E. Godlewski, A. In, B. Perthame, and P. Rascle, *Some new Godunov and relaxation methods for two phase flows*, Proceedings of the International Conference on Godunov methods : theory and applications, Kluwer Academic, Plenum Publisher, 2001.

[10] E. Godlewski and P.A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, Springer-Verlag, New York, 1996.

[11] L. Gosse, *Time-splitting schemes and measure source terms for a quasilinear relaxing system*, M3AS, Vol. 13(8), (2003), pp. 1081–1101.
[12] J.M. Greenberg and A.Y. Leroux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., Vol. 33 n.1, (1996), pp. 1–16.
[13] C. Helzel, R.J. Leveque and G. Warnecke, *A modified fractional step method for the accurate approximation of detonation waves*, SIAM J. Sci. Comput., Vol. 22(4), (2000), pp. 1489–1510.
[14] C. Hirsch, *Numerical computation of internal and external flows*, John Wiley and Sons, 1990.
[15] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., Vol. 48, (1995), pp. 235–276.
[16] R.J. LeVeque, *Balancing source terms and flux gradients in high-resolution Godunov methods*, J. Comp. Phys., Vol. 146, (1998), pp. 346–365.
[17] R.J. LeVeque and M. Pelanti, *A class of approximate Riemann solvers and their relation to relaxation schemes*, J. Comp. Phys., Vol. 172, (2001), pp. 572–591.
[18] R.J. LeVeque and H.C. Yee, *A study of numerical methods for hyperbolic conservation laws with source terms*, J. Comp. Phys., Vol. 86(1), (1990), pp. 187–210.
[19] T.P. Liu, *Hyperbolic systems with relaxation*, Comm. Math. Phys., Vol. 57, (1987), pp. 153–175.
[20] A. Majda and R.L. Pego, *Stable viscosity matrices for systems of conservation laws*, J. Diff. Eq., Vol. 56, $n^o 2$, (1985), pp. 229–262.
[21] R. Menikoff and B. Plohr, *The Riemann problem for fluid flow of real materials*, Rev. Mod. Phys., Vol. 61, (1989), pp. 75–130.
[22] W.A. Mülder and B. van Leer, *Experiments with implicit upwind methods for the Euler equations*, J. Comp. Phys., Vol. 59, (1985), pp. 232–246.
[23] P.L. Roe, *Approximate Riemann solvers, parameter vectors and difference scheme*, J. Comp. Phys., Vol. 43, (1981), pp. 357–372.
[24] I. Suliciu,*On the thermodynamics of fluids with relaxation and phase transitions. I-Fluids with relaxation*, Int. J. Eng. Sci., Vol 36, (1998), pp. 921–947.
[25] A. Harten, P.D. Lax, and B. van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., Vol. 25, (1983), pp. 35-61.
[26] G. Whitham, *Linear and nonlinear waves*, Wiley-Interscience, 1974.