A relaxation scheme for two-phase multi-component flows

Michaël Baudin, Frédéric Coquel, Quang Huy Tran

Unpublished paper
Article non-publié

A RELAXATION SCHEME FOR TWO-PHASE MULTI-COMPONENT FLOWS

Michaël Baudin[1,3], Frédéric Coquel[2] and Quang-Huy Tran[3]

Abstract. Pursuing the program launched our previous works [*Numer. Math.* **99** (2005), 411–440] and [*SIAM J. Sci. Comput.* **27** (2005), 914–936], we propose a relaxation scheme for the numerical simulation of one-dimensional two-phase multi-component flows governed by a drift-flux model, the main features of which are a large number of components and a high degree of nonlinearity in the closure laws. In the explicit setting, the relaxation approach allows to ensure positivity for the densities and the mass fractions. The relaxation method is worked out further so as to fit into a hybrid explicit-implicit setting, where fast acoustic waves are treated implicitly to save computational time while slow kinematic waves are treated explicitly in order to maintain accuracy on the transportation of materials.

1991 Mathematics Subject Classification. 76T10, 76N15, 35L65, 65M06.

August 7, 2008.

1. Introduction

As a sequel to [2,3], which were primarily concerned with the numerical simulation of a simplistic two-phase flow drift-flux model [25,28] via a relaxation method, this contribution is devoted to a more realistic model, where several *species* —or *components*— may appear in each phase. Before elaborating on how the materials of [1–3] can be adapted and enriched by new ingredients, we would like to set out the scope which this project comes within, as well as the problems that await us.

In the transportation of oil and gas mixtures along a pipeline, it is capital to correctly model the mass exchange between the two phases. This cannot be achieved with the *gas-liquid* model considered in [2,3], which is also referred to as the *non-compositional* model and where mass balances are written for each phase. Instead, we need a more physically relevant version, where mass balances are written for each specie and which is called *multi-component* or *compositional* model.

Due heed must be paid to the number of components involved, since it now equates the number of continuity equations in the model. Traditional techniques such as VFRoe schemes [17,24] turn out not to be efficient in terms computational cost, insofar as these require the numerical evaluation of the eigenvalues and the eigenvectors of a large-sized matrix at each edge and at each time-step. The problem is worsened by the high degree of nonlinearity exhibited by the algebraic closure laws. It is well-known [31] that the thermodynamics of a multi-component mixture is far more complex than that of a non-compositional one. At a practical level, the

Keywords and phrases: Two-phase flows, multi-component fluids, finite volume schemes, relaxation methods, Whitham condition

[1] presently at: ALTRAN Aeronautics-Space & Defence, 2 rue Paul Vaillant Couturier, 92300 Levallois-Perret, France

[2] UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France ; CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

[3] Département Mathématiques Appliquées, Institut Français du Pétrole, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France, Q-Huy.TRAN@ifp.fr, +33 1 47 52 74 22

pressure, the partial fractions and the slip velocity are usually the outputs of a lengthy and expensive process consisting of nonlinear inversions and/or of look-up into tables.

In addition to the foregoing aspects, the novelty of which lies in the multi-component nature of the model, we must keep in mind two issues that we already encountered in [2, 3]. The first issue is *positivity*. In [2], we put forward an explicit relaxation scheme that enables one to guarantee that the total density and the gas mass fraction remain positive from one time-step to another under an adequate tuning of the relaxation parameters. This physically sound requirement meets the need for robustness of the method. For the multi-component model, positivity means that the total density and the various component mass fractions should remain positive.

The second issue we have to face is the design of a *hybrid explicit-implicit* time integration, the motivation of which comes from the following observations. On one hand, we want to use large time-steps in order to bring down the CPU cost to an acceptable level. This advocates the use of an implicit strategy. On the other hand, in the context of the flow regimes considered, there co-exist two kinds of waves that are clearly separated by the magnitude of their characteristic speeds: fast acoustic waves and slow kinematic waves. From the petroleum engineer's standpoint, however, only the kinematic ones are of interest since they represent mass transportation. In order to maintain accuracy on the slow waves, we want these slow waves to be treated as if the scheme were explicit. This seemingly paradoxical constraint gives rise to a hybrid time integration that is implicit with respect to fast waves while remaining explicit with respect to slow waves. In [3], we paved the way toward such a *selectively implicit* time integration, based on a reinterpretation of the explicit relaxation scheme as a Roe method [29] and inspired from a paradigm by Faille and Heintzé [16]. Another attempt via the Arbitrary Lagrange-Euler formalism was recently made in [13].

The present work aims at taking up the challenge of the multi-component case by extending the strategies of [2, 3]. The path we suggest differs from the one sketched out in [4]. Nevertheless, it is powerful enough to be valid for any slip law, unlike the latter. This new extension is made possible and tractable thanks to the versatility of the relaxation philosophy [7, 10, 21, 22], from which we inherit valuable properties regarding stability, accuracy and simplicity on the basis of a virtual freedom from the closure laws. The feasibility of the extension also relies on a deepened and refreshed understanding of the methods formerly proposed for the gas-liquid model.

This paper is outlined as follows. We start, in §2, by introducing the two-phase multi-component model. The explicit framework of the relaxation scheme is presented in §3, where we focus on the ability to guarantee positivity with the help of only two parameters. The lower-bounds for these parameters are derived in an improved fashion, which is more incisive than in [2]. A thorough discussion on the relaxation parameters is provided in the Appendix. In §4, we turn to the hybrid explicit-implicit scheme and shed new light on the "correct" way to perform relaxation in an implicit context, as proposed by Chalons [9] and ourselves [3]. We also revisit Roe's form of the relaxation scheme via the Osher-Solomon interpretation [27]. Numerical results are shown in §5 before we conclude.

2. Two-phase multi-component drift-flux model

Let $K \geq 2$ be the number of components or species, indexed by $k \in \{1, 2, \ldots, K\}$, each of which may appear in the two phases gas g and liquid ℓ. Let ξ_k (resp. η_k) be mass fraction of component k in the gas (resp. liquid) phase. We also refer to $(\xi_k, \eta_k) \in [0,1]^2$ as *partial fractions*. Naturally, the partial fractions satisfy the consistency relations

$$\xi_1 + \ldots + \xi_K = \eta_1 + \ldots + \eta_K = 1. \tag{1}$$

Furthermore, it is assumed that each component k must be present in at least one phase, that is,

$$(\xi_k, \eta_k) \neq (0, 0). \tag{2}$$

2.1. **Physical formulation**

The isothermal model we consider consists of the $K+1$ partial differential equations

$$\partial_t(\rho_g R_g \xi_k + \rho_\ell R_\ell \eta_k) + \partial_x(\rho_g R_g \xi_k v_g + \rho_\ell R_\ell \eta_k v_\ell) = 0 \tag{3a}$$

$$\partial_t(\rho_g R_g v_g + \rho_\ell R_\ell v_\ell) + \partial_x(\rho_g R_g v_g^2 + \rho_\ell R_\ell v_\ell^2 + p) = 0, \tag{3b}$$

where the first K equations (3a) express the mass balances of the K components, and the last equation (3b) express the total momentum balance of the mixture. The notation R_g (resp. R_ℓ) stands for the volumetric fraction of the gas (resp. liquid) phase, and we have

$$R_g + R_\ell = 1. \tag{4}$$

The symbol ρ_g (resp. ρ_ℓ) denotes the gas (resp. liquid) density, assumed to be given functions of the pressure p. It is customary and convenient to define the apparent density of the kth component as

$$\rho_k = \rho_g R_g \xi_k + \rho_\ell R_\ell \eta_k. \tag{5}$$

The PDE system (3) must be accompanied by two sets of algebraic closure laws. The first set is concerned with thermodynamics and returns

$$p = p(\{\rho_k\}_{1\le k\le K}), \quad \xi_h = \xi_h(\{\rho_k\}_{1\le k\le K}), \quad \eta_h = \eta_h(\{\rho_k\}_{1\le k\le K}), \tag{6}$$

for all $h \in \{1,2,\dots,K\}$. The second set comprises a single function returning the *slip* —or *drift*— velocity

$$v_g - v_\ell = \phi(\{\rho_k\}_{1\le k\le K},\, \rho_g R_g v_g + \rho_\ell R_\ell v_\ell), \tag{7}$$

that is, the difference between the gas velocity and the liquid velocity. Note that, for convenience, we have changed the sign convention from $\phi = v_\ell - v_g$ in [2,3] to $\phi = v_g - v_\ell$ in this paper. Incidentally, the name *drift-flux model* originates from the choice of a single momentum balance (3b) and from the subsequent necessity to resort to the additional drift law (7) in order to close the system. The slip law encapsulates the knowledge of physicists about flow regimes based on experimental data.

We shall not dwell on this first statement of the model, because it is in reality more helpful to work with an alternative formulation that bears more resemblance to gas dynamics.

2.2. **Mathematical formulation**

We define the total density ρ, the gas mass fraction Y, the mass fraction c_k of the kth-component and the mass averaged velocity v by the equalities

$$\begin{array}{ll} \text{(a)} \quad \rho = \rho_g R_g + \rho_\ell R_\ell, & \text{(b)} \quad \rho Y = \rho_g R_g, \\ \text{(c)} \quad \rho c_k = \rho_k, & \text{(d)} \quad \rho v = \rho_g R_g v_g + \rho_\ell R_\ell v_\ell. \end{array} \tag{8}$$

With some abuse of notation, the closure laws (6)–(7) read

$$p = p(\rho, \{\rho c_k\}_{1\le k\le K-1}), \quad \xi_h = \xi_h(\rho, \{\rho c_k\}_{1\le k\le K-1}), \quad \eta_h = \eta_h(\rho, \{\rho c_k\}_{1\le k\le K-1}), \tag{9}$$

which accounts for the thermodynamics, with $h \in \{1,2,\dots,K\}$, and

$$\phi = \phi(\rho, \rho v, \{\rho c_k\}_{1\le k\le K-1}), \tag{10}$$

which accounts for the hydrodynamics. Consider the natural phase space

$$\Omega_{\mathbf{u}} = \{\mathbf{u} = (\rho, \rho v, \{\rho c_k\}_{1\le k\le K-1}) \in \mathbb{R}^{K+1} \\ | \ \rho > 0,\ v \in \mathbb{R},\ c_k \ge 0,\ c_1 + \ldots + c_{K-1} \le 1\}. \tag{11}$$

We are now in a position to derive an equivalent system for (3).

Lemma 2.1. *Weak solutions of (3) are equivalent to weak solution of the system*

$$\partial_t(\rho) \quad + \partial_x(\rho v) \qquad\qquad = 0 \tag{12a}$$
$$\partial_t(\rho v) \ + \partial_x(\rho v^2 \ + \ P(\mathbf{u})) = 0 \tag{12b}$$
$$\partial_t(\rho c_k) + \partial_x(\rho c_k v + \sigma_k(\mathbf{u})) = 0, \tag{12c}$$

defined for $\mathbf{u} \in \Omega_{\mathbf{u}}$*, where the subscript* k *runs from 1 to* $K-1$ *in (12c), and*

$$(a)\ P(\mathbf{u}) = p + \rho Y(1-Y)\phi^2, \quad (b)\ \sigma_k(\mathbf{u}) = (\xi_k - \eta_k)\sigma, \quad (c)\ \sigma = \rho Y(1-Y)\phi. \tag{13}$$

Proof. For smooth solutions, the first equation (12a) results from adding the K continuity equations (3a) together and from the equality

$$\rho = \rho_1 + \ldots + \rho_K, \tag{14}$$

as well as (1) and (8d). By inverting the 2×2 linear system

$$Yv_g + (1-Y)v_\ell = v, \qquad v_g - v_\ell = \phi, \tag{15}$$

we find

$$v_g = v + (1-Y)\phi, \qquad v_\ell = v - Y\phi, \tag{16}$$

and substituting in (3a)–(3b) yields (12b)–(12b).

The converse follows the same steps. For weak solutions, equivalence still holds because the only combination we made is a mere sum or subtraction. □

From now on, the new system (12) is credited the abstract but concise form

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}. \tag{17}$$

At this juncture, two difficulties arise. First, as pointed out in the Introduction, the closure laws (9)–(10) are highly nonlinear and very costly. Put another way, the flux $\mathbf{f}(\mathbf{u})$ in (17) often does not have any closed-form expression and is expensive. Second, because of nonlinearities, the hyperbolicity property cannot be ascertained for system (17), i.e., we do not know beforehand whether the Jacobian matrix $\nabla_{\mathbf{u}}\mathbf{f}(\mathbf{u})$ has real eigenvalues and is $\mathbb{R}$-diagonalizable for $\mathbf{u} \in \Omega_{\mathbf{u}}$, except for trivial instances such as when $\phi \equiv 0$.

Numerical computations reveal, however, that hyperbolicity usually holds within the scope of our application. Moreover, for the type of simulations under consideration, the increasingly ordered eigenvalues

$$\mu_1(\mathbf{u}) < \mu_2(\mathbf{u}) \le \ldots \le \mu_K(\mathbf{u}) < \mu_{K+1}(\mathbf{u}) \tag{18}$$

always satisfy $\mu_1(\mathbf{u})\mu_{K+1}(\mathbf{u}) < 0$ and

$$|\mu_1(\mathbf{u})| \approx |\mu_{K+1}(\mathbf{u}))| \gg |\mu_2(\mathbf{u})| \approx \ldots \approx |\mu_K(\mathbf{u})|. \tag{19}$$

We label this phenomenon as the *separation of velocity scales*: the 2 extreme characteristic speeds $\mu_1(\mathbf{u})$ and $\mu_{K+1}(\mathbf{u})$ correspond to fast acoustic waves, while the $K-1$ remaining ones are associated with slow kinematic waves representing the actual transportation of the mixture.

3. Explicit relaxation scheme

3.1. **Relaxation model**

As explained in [2], instead of relaxing every nonlinearity like Jin and Xin [21], we carry out a partial relaxation procedure in the same vein as Jin and Slemrod [20] or Coquel and Perthame [14]. The advantage of this approach is to minimize dissipation by concentrating on the "true" nonlinearities of the system.

In the same spirit as in [2], we define the relaxation state space $\Omega_{\mathbb{U}}$ as

$$\begin{aligned}\Omega_{\mathbb{U}} = \{\mathbb{U} = (\rho,\, \rho v,\, \rho\Pi,\, \{\rho c_k\}_{1\le k\le K-1},\, \{\rho\Sigma_k\}_{1\le k\le K-1}) \in \mathbb{R}^{2K+1} \\ \mid \rho > 0,\ (v,\Pi)\in\mathbb{R}^2,\ \Sigma_k\in\mathbb{R},\ c_k\ge 0,\ c_1+\ldots+c_{K-1}\le 1\}.\end{aligned} \tag{20}$$

For later use, let us introduce a few more notations. To every

$$\mathbb{G} = (G_1, G_2, \underline{G_3}, \{G\}_{4\to K+2}, \underline{\{G\}_{K+3\to 2K+1}}) \in \mathbb{R}^{2K+1}, \tag{21}$$

we associate its projected value

$$\blacktriangledown\mathbb{G} = (G_1, G_2, \{G\}_{4\to K+2}) \in \mathbb{R}^{K+1}. \tag{22}$$

Since the projection operator $\blacktriangledown$ merely deletes the components pertaining to the relaxation variables, it can act on any vector of $\mathbb{R}^{2K+1}$, including those which do not belong to $\Omega_{\mathbb{U}}$. The subset

$$\mathscr{E} = \{\mathbb{U}\in\Omega_{\mathbb{U}} \mid \Pi = P(\blacktriangledown\mathbb{U}) \quad\text{and}\quad \Sigma_k = \sigma_k(\blacktriangledown\mathbb{U}),\ 1\le k\le K-1\} \tag{23}$$

is called *equilibrium manifold*, and its elements are said to be *at equilibrium*.

In the opposite direction, we define

$$\overline{\overline{\wedge}}\mathbf{u} = (\rho,\, \rho v,\, \rho P(\mathbf{u}),\, \{\rho c_k\}_{1\le k\le K-1},\, \{\rho\sigma_k(\mathbf{u})\}_{1\le k\le K-1}) \in \mathscr{E} \tag{24}$$

to be the *equilibrium extension* of $\mathbf{u}\in\Omega_{\mathbf{u}}$. The extension operator $\overline{\overline{\wedge}}$ will be of great help at the end of §3 and the beginning of §4.

Definition 3.1. *For $\lambda\ge 0$ and $a>b>0$, the system*

$$\partial_t(\rho)^\lambda + \partial_x(\rho v)^\lambda = 0 \tag{25a}$$

$$\partial_t(\rho v)^\lambda + \partial_x(\rho v^2 + \Pi)^\lambda = 0 \tag{25b}$$

$$\partial_t(\rho\Pi)^\lambda + \partial_x(\rho\Pi v + a^2 v)^\lambda = \lambda\rho[P(\blacktriangledown\mathbb{U}^\lambda) - \Pi^\lambda] \tag{25c}$$

$$\partial_t(\rho c_k)^\lambda + \partial_x(\rho c_k v + \Sigma_k)^\lambda = 0 \tag{25d}$$

$$\partial_t(\rho\Sigma_k)^\lambda + \partial_x(\rho\Sigma_k v + b^2 c_k)^\lambda = \lambda\rho[\sigma_k(\blacktriangledown\mathbb{U}^\lambda) - \Sigma_k^\lambda], \tag{25e}$$

defined for $\mathbb{U}^\lambda\in\Omega_{\mathbb{U}}$, is said to be the relaxation model *for the original model (12).*

A few remarks are in order. To begin with, this relaxation system is obtained through the same procedure as in [2]. We first switch to Lagrangian coordinates in order to highlight $(P,\{\sigma_k\})$ as the nonlinear cores to "get rid of." A relaxation model is then proposed, replacing the nonlinear functions $(P,\{\sigma_k\})$ by the new variables $(\Pi,\{\Sigma_k\})$ to which evolution equations are imposed. Reverting back to Eulerian coordinates, we wind up with (25). Noteworthy is the fact that the latter is split into an acoustic block (25a)–(25c) acting as a "nucleus" around which revolve $K-1$ kinematic blocks (25d)–(25e). For $\lambda=0$, the blocks are totally unrelated, at least in Lagrangian coordinates, where they have been studied by several authors [6, 15]. For $\lambda>0$, the blocks are coupled through the stiff relaxation terms appearing in right-hand side.

It is of paramount importance to mention that all the $K-1$ kinematic blocks are assigned the same relaxation parameter $b>0$. This fundamental choice is justified not only by our desire for simplicity, but also by our will to be able to ensure positivity for the mass fractions in a predictive manner. Had we opted for a set of component-dependent parameters $\{b_k\}$, we would have been overwhelmed with troubles. On the side of "simplicity," the solution of the Riemann problem (see §3.3) would have been more painful to write down, because the ordering of the eigenvalues $v \pm b_k \rho^{-1}$ is not known in advance. The scheme would have been more expensive and would not have lent itself to a selectively implicit version. On the side of "positivity," no analysis leading to an adequate lower-bound for $\{b_k\}$ such as Theorem 3.3 would have been possible, because a preliminary consistency property stated in Lemma 3.1 would have been violated.

The relaxation system (25) can be given the convenient abstract form

$$\partial_t \mathbb{U}^\lambda + \partial_x \mathbb{F}(\mathbb{U}^\lambda) = \lambda \mathbb{S}(\mathbb{U}^\lambda), \tag{26}$$

where $\mathbb{F}$ and $\mathbb{S}$ receive clear definitions. Let us recapitulate the classical properties that we will be using. From now on, we set

$$\tau = \rho^{-1} \tag{27}$$

to be the specific volume of the mixture.

Proposition 3.1. *The first-order system (25) is hyperbolic over $\Omega_{\mathbb{U}}$, i.e., for all $\mathbb{U} \in \Omega_{\mathbb{U}}$, the Jacobian matrix $\nabla \mathbb{F}(\mathbb{U})$ has real eigenvalues and is $\mathbb{R}$-diagonalizable. The increasingly arranged eigenvalues*

$$v - a\tau < v - b\tau < v < v + b\tau < v + a\tau, \tag{28}$$

where $v \pm b\tau$ is of multiplicity $K-1$, all correspond to linearly degenerate fields and are associated with the strong Riemann invariants

$$\Pi - av,\ \{\Sigma_k - bc_k\}_{1\le k\le K-1},\ \Pi + a^2\tau,\ \{\Sigma_k + bc_k\}_{1\le k\le K-1},\ \Pi + av. \tag{29}$$

The right eigenvectors for the eigenvalues (28) is given by the matrix

$$\mathscr{P}(\mathbb{U}) = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 \\ v-a\tau & 0 & \dots & 0 & v & 0 & \dots & 0 & v+a\tau \\ \Pi + a^2\tau & 0 & \dots & 0 & \Pi & 0 & \dots & 0 & \Pi + a^2\tau \\ c_1 & 1 & \dots & 0 & c_1 & 1 & \dots & 0 & c_1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ c_{K-1} & 0 & \dots & 1 & c_{K-1} & 0 & \dots & 1 & c_{K-1} \\ \Sigma_1 & -b & \dots & 0 & \Sigma_1 & b & \dots & 0 & \Sigma_1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \Sigma_{K-1} & 0 & \dots & -b & \Sigma_{K-1} & 0 & \dots & b & \Sigma_{K-1} \end{bmatrix}. \tag{30}$$

Proof. The calculations are easily adapted from [2, Lemma 3]. □

The two extreme fields $v \pm a\tau$ represent fast acoustic waves, while the $2K-1$ remaining ones represent slow kinematic waves. The separation of velocity scales is now reflected by $a \gg b$, which implies

$$|v - a\tau| \approx |v + a\tau| \gg |v - b\tau| \approx |v + b\tau|. \tag{31}$$

3.2. **Linear asymptotic stability via a Whitham condition**

Since no entropy pair is known for the system under consideration, our approach cannot enter the general relaxation theory developed by Liu [22] and Chen et al. [12]. The only way for us to investigate the asymptotic

stability of the relaxation system is the framework proposed by Whitham [32] which relies on the Chapman-Enskog expansion in the limit $\lambda \to +\infty$.

Let

$$\mathbf{v} = (\tau, v, \{c_k\}_{1\le k\le K-1}) \in \mathbb{R}_+^* \times \mathbb{R} \times [0,1]^{K-1} \tag{32}$$

be the Lagrangian variables associated to $\mathbf{u}$. Abusing notations once again, we write $P = P(\mathbf{v})$ and $\sigma_k = \sigma_k(\mathbf{v})$ whenever needed. Within this new set of unknowns, we define the partial derivatives P_τ, P_v, $(\sigma_k)_{c_h}$... the argument of which may be viewed as either $\mathbf{v}$ or $\mathbf{u}$ depending on the context.

Theorem 3.1. *At the first order approximation in λ^{-1}, the projected value $\mathbf{u}^\lambda = {}_\blacktriangledown\mathbb{U}^\lambda$ of the solution $\mathbb{U}^\lambda$ to the relaxation system (25) satisfies the* equivalent equation

$$\partial_t \mathbf{u}^\lambda + \partial_x \mathbf{f}(\mathbf{u}^\lambda) = \lambda^{-1}\partial_x\{\mathbf{D}_{a,b}(\mathbf{u}^\lambda)\partial_x \mathbf{u}^\lambda\} \tag{33}$$

where the $(K+1)\times(K+1)$ matrix $\mathbf{D}(\mathbf{u})$ has the form

$$\mathbf{D}_{a,b}(\mathbf{u}) = \frac{1}{\rho^2}\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ \times & a^2 - A(\mathbf{v}) & \times & \dots & \times \\ \times & \times & b^2 - B_1(\mathbf{v}) & \dots & \times \\ \dots & \dots & \dots & \dots & \dots \\ \times & \times & \times & \dots & b^2 - B_{K-1}(\mathbf{v}) \end{bmatrix} \tag{34}$$

in which

$$A(\mathbf{v}) = \boxed{-P_\tau + P_v^2} + P_{c_1}(\sigma_1)_v + \ldots + P_{c_{K-1}}(\sigma_{K-1})_v \tag{35a}$$

$$B_k(\mathbf{v}) = P_{c_k}(\sigma_k)_v + (\sigma_k)_{c_1}(\sigma_1)_{c_k} + \ldots + \boxed{[(\sigma_k)_{c_k}]^2} + \ldots + (\sigma_k)_{c_{K-1}}(\sigma_{K-1})_{c_k} \tag{35b}$$

and the possibly non-zero entries denoted by $\times$ do not depend on (a,b).

Proof. The idea is to plug the Chapman-Enskog expansions

$$\Pi^\lambda = P(\mathbf{u}^\lambda) + \lambda^{-1}P_1^\lambda + O(\lambda^{-2}) \tag{36a}$$

$$\Sigma^\lambda = \sigma(\mathbf{u}^\lambda) + \lambda^{-1}\Sigma_1^\lambda + O(\lambda^{-2}) \tag{36b}$$

into the relaxation system and to keep the first-order terms. Barring from the opposite sign convention for σ and from the heaviness of the multi-component case, the calculations are similar to those of [2, Proposition 2]. □

Formally, as $\lambda \to +\infty$, the original model (12) is recovered. As for linear stability, the widely and commonly used practice consists in requiring that the matrix $\mathbf{D}_{a,b}(\mathbf{u})$ be diffusive, i.e., its eigenvalues have positive real parts. This is what we did in [2, Proposition 1, Lemma 4] to determine the admissible region for (a,b). Here, such a task is out of reach, in view of the size of the matrix.

Instead of the eigenvalues, we are going to impose positivity to the diagonal entries of $\mathbf{D}_{a,b}(\mathbf{u})$. This heuristic simplification is strongly supported by our experience with the gas-liquid model, as summarized in [3, Appendix]. It is neither more nor less "rigorous" than the requirement that the eigenvalues of $\mathbf{D}_{a,b}(\mathbf{u})$ have positive real parts, which is in itself an approximation. We recall that the exact condition for L^2 stability, obtained by linearizing (33) in the neighborhood of any steady state solution $\mathbf{u}$, is that all eigenvalues of the matrix

$$\imath\kappa\nabla\mathbf{f}(\mathbf{u}) - \lambda^{-1}\kappa^2\mathbf{D}_{a,b}(\mathbf{u}), \qquad \kappa \in \mathbb{R},\ \imath^2 = -1, \tag{37}$$

should have negative real parts [23, 30].

Definition 3.2. *The pair* $(a,b) \in \mathbb{R}_+^2$*, with* $a > b > 0$*, is said to satisfy the* diagonal Whitham condition *if*

$$a^2 > -P_\tau(\mathbf{u}) + P_v^2(\mathbf{u}) \quad \textit{and} \quad b > \max_{1\le k\le K-1} |(\sigma_k)_{c_k}|(\mathbf{u}) \tag{38}$$

for all $\mathbf{u} \in \Omega_{\mathbf{u}}$ *under consideration.*

Obviously, the double inequality (38) is somewhat different from the positivity of the diagonal entries of $\mathbf{D}_{a,b}(\mathbf{u})$. It amounts to merely keeping the framed terms in (35). This choice is corroborated by three observations. First, a close inspection of the magnitudes of the summands in the right-hand side of (35) testifies to the fact that —at least for the numerical data we are working with— the leading terms coincide indeed with the framed ones.

Secondly, from a more theoretical point of view, the inequalities (38) could have been derived by the traditional procedure (requiring positivity for the eigenvalues of the allegedly diffusive matrix) with the additional assumptions that P does not depend on the c_k's and the σ_k's do not depend on (τ, v). The calculations are similar to those of [3, Theorem 2.3, Theorem 2.4]. This corresponds to the situation in which the acoustic block and the kinematic block are uncoupled from each other, whence the name of the condition. The last observation pleading in favor of (38) is that the numerical simulations using (38) work fine, behave stably while giving sharper results than those using the whole right-hand side of (35).

3.3. **Riemann problem and positivity of intermediate states**

As a prerequisite to the full numerical scheme in §3.4, the Riemann problem corresponding to the homogeneous ($\lambda = 0$) relaxation system needs to be solved. Besides the structure of the Riemann solution, we lay emphasis on the sufficient conditions for various quantities of the intermediate states to be positive. This will, in turn, be useful for the positivity of the updated variables.

Let $\mathbb{U}_L \in \Omega_{\mathbb{U}}$ and $\mathbb{U}_R \in \Omega_{\mathbb{U}}$ be the left and right states defining the initial data

$$\mathbb{U}(t=0,x) = \mathbb{U}_L\, \mathbf{1}_{\{x<0\}} + \mathbb{U}_R\, \mathbf{1}_{\{x>0\}}, \tag{39}$$

where $\mathbf{1}_{\{.\}}$ is the characteristic function. We introduce

$$\begin{aligned}
&\text{(a)}\quad v^* = \frac{v_R + v_L}{2} - \frac{\Pi_R - \Pi_L}{2a}\\
&\text{(b)}\quad \Pi^* = \frac{\Pi_R + \Pi_L}{2} - a\,\frac{v_R - v_L}{2}\\
&\text{(c)}\quad \tau_L^* = \frac{v_R - v_L}{2a} - \frac{\Pi_R - \Pi_L}{2a^2} + \tau_L\\
&\text{(d)}\quad \tau_R^* = \frac{v_R - v_L}{2a} + \frac{\Pi_R - \Pi_L}{2a^2} + \tau_R
\end{aligned} \tag{40}$$

and

$$\begin{aligned}
&\text{(a)}\quad c_k^* = \frac{(c_k)_R + (c_k)_L}{2} - \frac{(\Sigma_k)_R - (\Sigma_k)_L}{2b}\\
&\text{(b)}\quad \Sigma_k^* = \frac{(\Sigma_k)_R + (\Sigma_k)_L}{2} - b\,\frac{(c_k)_R - (c_k)_L}{2}
\end{aligned} \tag{41}$$

The above quantities are intended to appear in the formulae for the solution, as shown in the upcoming Proposition. Note that, in (41), the subscript k runs from 1 to $K-1$.

Proposition 3.2. *If* $a > b > 0$ *and if* a *is large enough, e.g., in accordance with (47) of Theorem 3.2, then the solution to the homogeneous relaxation system (25) with the initial data (39) is the self-similar function*

$\mathbb{U}(t,x) = \mathscr{U}(x/t;\mathbb{U}_L,\mathbb{U}_R)$ *made up of 6 constant states, given by*

$$\begin{aligned}\mathscr{U}(x/t;\mathbb{U}_L,\mathbb{U}_R) = \mathbb{U}_L\,\mathbf{1}_{\{\frac{x}{t}<\mu_L\}} \quad &+ \mathbb{U}_L^{\dashv}\,\mathbf{1}_{\{\mu_L<\frac{x}{t}<\mu_L^*\}} + \mathbb{U}_L^{\dashv\shortmid}\,\mathbf{1}_{\{\mu_L^*<\frac{x}{t}<\mu^*\}} \\ &+ \mathbb{U}_R^{\shortmid\vdash}\,\mathbf{1}_{\{\mu^*<\frac{x}{t}<\mu_R^*\}} + \mathbb{U}_R^{\vdash}\,\mathbf{1}_{\{\mu_R^*<\frac{x}{t}<\mu_R\}} + \mathbb{U}_R\,\mathbf{1}_{\{\frac{x}{t}>\mu_R\}}\end{aligned} \tag{42}$$

and defined for $(t,x)\in\mathbb{R}_+^*\times\mathbb{R}$*, where the characteristic speeds are*

$$\begin{array}{ccccccccc} v_L - a\tau_L & < & v^* - b\tau_L^* & < & v^* & < & v^* + b\tau_R^* & < & v_R + a\tau_R \\ \| & & \| & & \| & & \| & & \| \\ \mu_L & & \mu_L^* & & \mu^* & & \mu_R^* & & \mu_R \end{array} \tag{43}$$

and the consecutively connected intermediate states are

$$\begin{bmatrix} \rho_L^* \\ \rho_L^* v^* \\ \rho_L^* \Pi^* \\ \rho_L^* \{c_k\}_L \\ \rho_L^* \{\Sigma_k\}_L \end{bmatrix} \rightsquigarrow \begin{bmatrix} \rho_L^* \\ \rho_L^* v^* \\ \rho_L^* \Pi^* \\ \rho_L^* \{c_k^*\} \\ \rho_L^* \{\Sigma_k^*\} \end{bmatrix} \rightsquigarrow \begin{bmatrix} \rho_R^* \\ \rho_R^* v^* \\ \rho_R^* \Pi^* \\ \rho_R^* \{c_k^*\} \\ \rho_R^* \{\Sigma_k^*\} \end{bmatrix} \rightsquigarrow \begin{bmatrix} \rho_R^* \\ \rho_R^* v^* \\ \rho_R^* \Pi^* \\ \rho_R^* \{c_k\}_R \\ \rho_R^* \{\Sigma_k\}_R \end{bmatrix} \tag{44}$$
$$\begin{array}{ccccc} \| & & \| & & \| \\ \mathbb{U}_L^{\dashv} & & \mathbb{U}_L^{\dashv\shortmid} & & \mathbb{U}_R^{\shortmid\vdash} \quad \mathbb{U}_R^{\vdash} \end{array}$$

Proof. The proof makes use of the Riemann invariants (29) of Proposition 3.1 and follows along the same lines as in [2, Proposition 3]. □

The Riemann problem can be solved in the most general case including configurations of eigenvalues other than (43). But in view of the separation of velocity scales (31), only the ordering (43) is of interest to us. However, in contrast with (28), the assumption $a > b > 0$ alone is not sufficient to secure (43).

Theorem 3.2. *For* $a > b > 0$*, we have the ordering*

$$v_L - a\tau_L < v^* - b\tau_L^* < v^* < v^* + b\tau_R^* < v_R + a\tau_R \tag{45}$$

if and only if the specific volumes of the intermediate states are positive, i.e.,

$$\tau_L^* > 0 \qquad \textit{and} \qquad \tau_R^* > 0. \tag{46}$$

A sufficient condition for the equivalent conditions (45) and (46) to hold is

$$a > \frac{-(v_R - v_L) + \sqrt{(v_R - v_L)^2 + 8\min(\tau_L,\tau_R)|\Pi_R - \Pi_L|}}{4\min(\tau_L,\tau_R)}. \tag{47}$$

Proof. Since the $v \pm a\tau$ waves are linearly degenerate, we have

$$v_L - a\tau_L = v^* - a\tau_L^* \qquad \text{and} \qquad v_R + a\tau_R = v^* + a\tau_R^*. \tag{48}$$

This can also be directly checked from (40). Therefore, the double inequality

$$v_L - a\tau_L < v^* < v_R + a\tau_R \tag{49}$$

is equivalent to the positivity property (46).

Inserting the value of v^* from (40a) into (49) yields two second-degree polynomial inequalities in a, namely,

$$\wp_L(a) = 2\tau_L a^2 + (v_R - v_L)a - (\Pi_R - \Pi_L) > 0, \tag{50a}$$

$$\wp_R(a) = 2\tau_R a^2 + (v_R - v_L)a + (\Pi_R - \Pi_L) > 0. \tag{50b}$$

For $a > 0$, we have $\wp_L(a) \geq \wp(a)$ and $\wp_R(a) \geq \wp(a)$ with the lower-bound

$$\wp(a) = 2\min(\tau_L, \tau_R)a^2 + (v_R - v_L)a - |\Pi_R - \Pi_L|. \tag{51}$$

Hence, it suffices to ask for $\wp(a) > 0$. As a second-degree polynomial, $\wp$ is convex and has two roots of opposite signs. The positive one is given by (47). □

In practice, and as will be seen in §3.4, the left and right states usually belong to the equilibrium manifold $\mathscr{E}$, so that $\Pi_L = P(\blacktriangledown\mathbb{U}_L)$ and $\Pi_R = P(\blacktriangledown\mathbb{U}_R)$. For short, we write P_L and P_R for the equilibrium values of Π_L and Π_R. Imposing the lower-bound

$$a > a^\sharp(\mathbf{u}_L, \mathbf{u}_R) = \frac{-(v_R - v_L) + \sqrt{(v_R - v_L)^2 + 8\min(\tau_L, \tau_R)|P_R - P_L|}}{4\min(\tau_L, \tau_R)}, \tag{52}$$

we may legitimately wonder about a possible logical connection between this lower-bound and the first part of the diagonal Whitham condition (38), i.e.,

$$a > \sqrt{-P_\tau + P_v^2}. \tag{53}$$

In other words, is (52) or (49) implied by some discrete version of (53)? As shown in the Appendix, the answer is in the affirmative for a pure acoustic system, in which P does not depend on $(v, \{c_k\})$.

We now shift attention to the intermediate fractions c_k^*. As said earlier, the formulae (41a) are valid only for $1 \leq k \leq K-1$. It is tempting to define the intermediate mass fraction for the Kth-component as

$$c_K^* = 1 - (c_1^* + \ldots + c_{K-1}^*). \tag{54}$$

The good news is that this definition is consistent with formula (41a) for the Riemann solution.

Lemma 3.1. *If $\mathbb{U}_L \in \mathscr{E}$ and $\mathbb{U}_R \in \mathscr{E}$, then the intermediate fraction c_K^* defined by (54) is equal to the value obtained by formally putting $k = K$ in (41), that is,*

$$c_K^\star = \frac{(c_K)_R + (c_K)_L}{2} - \frac{(\sigma_K)_R - (\sigma_K)_L}{2b}, \tag{55}$$

where $(\sigma_K)_L = \sigma_K(\blacktriangledown\mathbb{U}_L) = (\Sigma_K)_L$ and $(\sigma_K)_R = \sigma_K(\blacktriangledown\mathbb{U}_R) = (\Sigma_K)_R$.

Anticipating the proof, we underline that this miracle occurs solely when the same relaxation parameter b is used for every component. As will be clarified in Theorem 3.3, the very possibility of guaranteeing $c_k^* \in [0,1]$ for $1 \leq k \leq K$ relies on this essential property. This is a compelling argument in favor of a unique b for all components.

Proof. Since the left and right states are at equilibrium, we have $\Sigma_k = \sigma_k$ for $1 \leq k \leq K-1$. If we sum the equalities (41a) over $1 \leq k \leq K-1$ and subtract the result to 1, we get

$$c_K^\star = \frac{(c_K)_R + (c_K)_L}{2} - \frac{\varsigma_R - \varsigma_L}{2b}, \tag{56}$$

where

$$\varsigma = -(\sigma_1 + \ldots + \sigma_{K-1}). \tag{57}$$

Now, from definition (13), it is readily verified that

$$\sigma_1 + \ldots + \sigma_{K-1} + \sigma_K = 0 \tag{58}$$

because of the consistency relations (1). As a consequence,

$$\varsigma = \sigma_K \tag{59}$$

at the left and right states, which proves the claim. □

We are now able to state a sufficient condition for the positivity of the intermediate mass fractions.

Theorem 3.3. *If $\mathbb{U}_L \in \mathscr{C}$ and $\mathbb{U}_R \in \mathscr{C}$, then a sufficient condition to secure*

$$c_k^* \in [0,1], \qquad 1 \le k \le K, \tag{60}$$

is that

$$b \ge b^\sharp(\mathbf{u}_L, \mathbf{u}_R) = \max\left\{|\rho\phi|_L, |\rho\phi|_R\right\}. \tag{61}$$

The reader's attention is drawn to the fact that the right-hand side of (61) is exactly equal to that of [2, Eq. (64)], notwithstanding a more complex model. As far as the proof is concerned, the version below is shorter than in [2].

Proof. Since $c_1^* + \ldots + c_{K-1}^* + c_K^* = 1$, we are going to seek a sufficient condition to have

$$c_k^* \ge 0, \qquad \text{for } 1 \le k \le K. \tag{62}$$

From (41a) and (55), we infer that $c_k^* \ge 0$ is equivalent to

$$b \ge \frac{(\sigma_k)_R - (\sigma_k)_L}{(c_k)_R + (c_k)_L} \tag{63}$$

if $(c_k)_L > 0$ or $(c_k)_R > 0$. The case $(c_k)_L = (c_k)_R = 0$ will be settled later on. By the triangle inequality, we have

$$\frac{(\sigma_k)_R - (\sigma_k)_L}{(c_k)_R + (c_k)_L} \le \frac{|\sigma_k|_R + |\sigma_k|_L}{(c_k)_R + (c_k)_L} \tag{64a}$$

$$\le \frac{|Y(1-Y)\theta_k|_R + |Y(1-Y)\theta_k|_L}{(c_k)_R + (c_k)_L} \times \max\left\{|\rho\phi|_L, |\rho\phi|_R\right\} \tag{64b}$$

with

$$\theta_k = \xi_k - \eta_k. \tag{65}$$

Combining (8c), (5) and (8b), we obtain

$$c_k = Y\xi_k + (1-Y)\eta_k. \tag{66}$$

This equality can be cast under three different forms, namely,

$$c_k = \eta_k + Y\theta_k = \xi_k - (1-Y)\theta_k = \gamma_k + Z_k|\theta_k| \tag{67}$$

where the pair

$$(\gamma_k, Z_k) = (\eta_k, Y)\,\mathbf{1}_{\{\theta_k \ge 0\}} + (\xi_k, 1-Y)\,\mathbf{1}_{\{\theta_k < 0\}} \tag{68}$$

still belongs to $[0,1]^2$. Therefore, dropping out the subscript k momentarily in c_k, γ_k, Z_k, θ_k so as to alleviate the notations, we have

$$\frac{|Y(1-Y)\theta|_R + |Y(1-Y)\theta|_L}{c_R + c_L} = \frac{Z_R(1-Z_R)|\theta|_R + Z_L(1-Z_L)|\theta|_L}{\gamma_R + Z_R|\theta|_R + \gamma_L + Z_L|\theta|_L} \tag{69a}$$

$$\leq \frac{Z_R|\theta|_R + Z_L|\theta|_L}{\gamma_R + Z_R|\theta|_R + \gamma_L + Z_L|\theta|_L} \tag{69b}$$

$$\leq 1. \tag{69c}$$

Resuming the bounding process (64), we end up with

$$\frac{(\sigma_k)_R - (\sigma_k)_L}{(c_k)_R + (c_k)_L} \leq \max\{|\rho\phi|_L, |\rho\phi|_R\}, \tag{70}$$

from which $b > b^\sharp(\mathbf{u}_L, \mathbf{u}_R)$ clearly appears to be a sufficient condition for $c_k^* \geq 0$.

To settle the case $(c_k)_L = (c_k)_R = 0$, we invoke the identity (66) to have

$$c_k = 0 \Rightarrow Y\xi_k = (1-Y)\eta_k = 0. \tag{71}$$

But by virtue of assumption (2), ξ_k and η_k cannot vanish simultaneously. Thus, $Y = 0$ or $1 - Y = 0$, and in any case $Y(1-Y) = 0$. This implies

$$(\sigma_k)_L = (\sigma_k)_R = 0, \tag{72}$$

and by applying (41a) or (55), we arrive at the equality $c_k^* = 0$. The proof is now completed. □

Again, the question naturally arises as to whether or not there is a link between condition (61) or the positivity of c_k^* and some discrete version of the second part of the diagonal Whitham condition (38), i.e.,

$$b \geq |(\sigma_k)_{c_k}|. \tag{73}$$

Again, as shown in the Appendix, the answer is in the affirmative for the pure scalar case, where σ_k depends only on c_k.

3.4. **Numerical scheme and positivity of updated variables**

The Godunov flux associated with the relaxation system is classically defined from the Riemann solution as

$$\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) = \mathbb{F}(\mathscr{U}(x/t = 0^+; \mathbb{U}_L, \mathbb{U}_R)). \tag{74}$$

From now on, we make heavy use of the operators $\blacktriangledown$ and $\overline{\overline{\wedge}}$ previously defined in (22) and (24).

Definition 3.3. *Let $\mathbf{u}_L \in \Omega_{\mathbf{u}}$ and $\mathbf{u}_R \in \Omega_{\mathbf{u}}$ be two states, and $a > b > 0$ be the relaxation parameters subject to*

$$a > a^\sharp(\mathbf{u}_L, \mathbf{u}_R) \qquad \text{and} \qquad b \geq b^\sharp(\mathbf{u}_L, \mathbf{u}_R) \tag{75}$$

as in (52) and (61). The vector

$$\mathbf{h}(\mathbf{u}_L, \mathbf{u}_R) = \blacktriangledown\mathbb{H}(\overline{\overline{\wedge}}\mathbf{u}_L, \overline{\overline{\wedge}}\mathbf{u}_R) \tag{76}$$

is said to be the relaxation numerical flux *for the original problem (12).*

The domain is divided into cells of size Δx. We look for an approximation $\mathbf{u}_i^n$ of $\mathbf{u}(x_i, t^n)$ at the center x_i of cell i and at time t^n. Let $\Delta t = t^{n+1} - t^n$ be the time step. The first-order explicit relaxation numerical scheme reads

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x}[\mathbf{h}(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n) - \mathbf{h}(\mathbf{u}_{i-1}^n, \mathbf{u}_i^n)], \tag{77}$$

where $\mathbf{h}(.,.)$ is given by Definition 3.3. Note that (a,b) can be chosen locally: at each edge $i+1/2$, there are two relaxation parameters $(a_{i+1/2}, b_{i+1/2})$ exerting an influence on the numerical flux $\mathbf{h}(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n)$.

The overall scheme (76)–(77) can be thought of as the outcome of the following splitting procedure, by means of which we initially described the relaxation method in [2].

(1) *Set to equilibrium.* Put $\lambda = +\infty$ and build the equilibrium manifold $\Pi = P(\mathbf{u})$ and $\Sigma = \Sigma(\mathbf{u})$ by applying the extension operator $\mathbb{U}_i^n = \overline{\overline{\wedge}}\mathbf{u}_i^n$;

(2) *Relax evolution.* Put $\lambda = 0$ and solve the relaxation system $\partial_t \mathbb{U} + \partial_x \mathbb{F}(\mathbb{U}) = \mathbb{O}$ by the first-order explicit Godunov scheme

$$\mathbb{U}_i^{n+1} = \mathbb{U}_i^n - \frac{\Delta t}{\Delta x}\big[\mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) - \mathbb{H}(\mathbb{U}_{i-1}^n, \mathbb{U}_i^n)\big] \tag{78}$$

from which deduce $\mathbf{u}_i^{n+1} = \blacktriangledown \mathbb{U}_i^{n+1}$ by projection.

The benefit of the splitting interpretation is that it is valid at the continuous level. However, the conciseness of the new formulation (76)–(77) turns out to be a judicious starting point for the design of the selectively implicit scheme in §4. Before getting to this, let us examine the explicit scheme (77).

Theorem 3.4. *Assume $\mathbf{u}_i^n \in \Omega_{\mathbf{u}}$ for all $i \in \mathbb{Z}$. If*

- *the parameters $a_{i+1/2}^n > b_{i+1/2}^n > 0$ are in agreement with (52) and (61), i.e.,*

$$a_{i+1/2}^n > a^\sharp(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n) \qquad \text{and} \qquad b_{i+1/2}^n > b^\sharp(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n); \tag{79}$$

- *the time-step complies with the CFL condition*

$$\frac{\Delta t}{\Delta x} \max_i \max\{\, |v^* - a\tau_L^*|_{i+1/2}^n,\, |v^* + a\tau_R^*|_{i+1/2}^n \,\} < \frac{1}{2}, \tag{80}$$

then the first-order explicit relaxation scheme (77) satisfies

$$\rho_i^{n+1} > 0 \qquad \text{and} \qquad (c_k)_i^{n+1} \in [0,1], \;\; 1 \le k \le K. \tag{81}$$

Proof. By Theorem 3.2 and Theorem 3.3, we are sure that the intermediate states appearing in the Riemann solution at each edge $i+1/2$ satisfy

$$(\tau_L^*)_{i+1/2} > 0, \quad (\tau_R^*)_{i+1/2} > 0, \quad \text{and} \quad (c_k^*)_{i+1/2} \in [0,1], \quad 1 \le k \le K. \tag{82}$$

The proof from this to (81) is similar to that of [2, Corollary 1]. In a nutshell, the positivity of ρ_i^{n+1} stems from the fact that, as a result of the Godunov scheme (78) for the relaxation system, it is equal to the average of a piecewise-constant function at time $n+1$, and each of the pieces returning a strictly positive value. As for $(c_k)_i^{n+1}$, it is also equal to an average —albeit in the sense of a positive measure induced by the total density— of a piecewise-constant function bounded by $[0,1]$. □

As explained in [3, §2.3.2], the relaxation parameters are in practice computed as

$$a_{i+1/2}^n > \max\{\, a^\flat(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n),\, a^\sharp(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n) \,\} \tag{83a}$$

$$b_{i+1/2}^n \ge \max\{\, b^\flat(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n),\, b^\sharp(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n) \,\} \tag{83b}$$

where the additional bounds

$$a^\flat(\mathbf{u}_L, \mathbf{u}_R) = \max\Big\{\sqrt{-P_\tau(\mathbf{u}_L) + P_v^2(\mathbf{u}_L)},\, \sqrt{-P_\tau(\mathbf{u}_R) + P_v^2(\mathbf{u}_R)}\Big\} \tag{84a}$$

$$b^\flat(\mathbf{u}_L, \mathbf{u}_R) = \max \max_{1\le k\le K-1} \{|(\sigma_k)_{c_k}(\mathbf{u}_L)|, |(\sigma_k)_{c_k}(\mathbf{u}_R)|\} \tag{84b}$$

are intended to mimic the diagonal Whitham condition (38) at the discrete level. Numerical experiments corroborate to the fact that the reinforced choice (83) is a good one. A discussion about possible connections between the $\flat$ and $\sharp$ bounds is supplied in the Appendix.

4. Hybrid explicit-implicit relaxation scheme

The design of a hybrid explicit-implicit time-integration in the context of relaxation methods is a delicate matter which requires several ingredients. We start by making right away some general but crucial observations about the fully and linearly implicit time integrations in such a context. Afterwards, we revisit the relaxation numerical flux in order to give it a "differentiable" form via the Roe interpretation. Finally, we propose a strategy for the selectively implicit time-integration.

4.1. **Interplay between relaxation and implicit time integrations**

Let us rewrite the first-order explicit scheme (76)–(77) as

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x}[\blacktriangledown\mathbb{H}(\overline{\overline{\wedge}}\mathbf{u}_i^n, \overline{\overline{\wedge}}\mathbf{u}_{i+1}^n) - \blacktriangledown\mathbb{H}(\overline{\overline{\wedge}}\mathbf{u}_{i-1}^n, \overline{\overline{\wedge}}\mathbf{u}_i^n)]. \tag{85}$$

Most naturally, the fully implicit counterpart of (85) reads

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x}[\blacktriangledown\mathbb{H}(\overline{\overline{\wedge}}\mathbf{u}_i^{n+1}, \overline{\overline{\wedge}}\mathbf{u}_{i+1}^{n+1}) - \blacktriangledown\mathbb{H}(\overline{\overline{\wedge}}\mathbf{u}_{i-1}^{n+1}, \overline{\overline{\wedge}}\mathbf{u}_i^{n+1})]. \tag{86}$$

Introducing the enlarged variables

$$\mathbb{U}_i^{n+1} = \overline{\overline{\wedge}}\mathbf{u}_i^{n+1}, \qquad \mathbb{U}_i^n = \overline{\overline{\wedge}}\mathbf{u}_i^n, \tag{87}$$

which are at equilibrium by construction of $\overline{\overline{\wedge}}$, we can easily check that the implicit formulation (86) is equivalent to

$$\blacktriangledown\mathbb{U}_i^{n+1} = \blacktriangledown\mathbb{U}_i^n - \frac{\Delta t}{\Delta x}[\blacktriangledown\mathbb{H}(\mathbb{U}_i^{n+1}, \mathbb{U}_{i+1}^{n+1}) - \blacktriangledown\mathbb{H}(\mathbb{U}_{i-1}^{n+1}, \mathbb{U}_i^{n+1})] \tag{88a}$$

$$\mathbb{U}_i^{n+1} \in \mathbb{E}. \tag{88b}$$

The equilibrium condition (88b) compensates for the "loss" of information created by the projection operator $\blacktriangledown$. In order to express it algebraically, let

$$\perp\mathbb{U} = (\rho\Pi, \{\rho\Sigma_k\}_{1\le k\le K-1}) \in \mathbb{R}\times\mathbb{R}^{K-1}. \tag{89}$$

The operator $\perp$ is best seen as the map sending any vector of $\mathbb{R}^{2K+1}$ to the components pertaining to the relaxation variables deleted by the operator $\blacktriangledown$. The Cartesian equations for the equilibrium manifold $\mathbb{E}$ are abstractly condensed as

$$\perp\mathbb{U} = \mathbb{L}(\blacktriangledown\mathbb{U}). \tag{90}$$

Proposition 4.1. *The fully implicit scheme (86) in the original variables* $\mathbf{u}$ *is equivalent to the system*

$$\blacktriangledown\mathbb{U}_i^{n+1} = \blacktriangledown\left\{\mathbb{U}_i^n - \frac{\Delta t}{\Delta x}[\mathbb{H}(\mathbb{U}_i^{n+1}, \mathbb{U}_{i+1}^{n+1}) - \mathbb{H}(\mathbb{U}_{i-1}^{n+1}, \mathbb{U}_i^{n+1})]\right\} \tag{91a}$$

$$\perp\mathbb{U}_i^{n+1} = \mathcal{L}(\blacktriangledown\mathbb{U}_i^{n+1}) \tag{91b}$$

in the enlarged variables $\mathbb{U}$.

Proof. The first line (91a) results from (88a) and the linearity of the operator $\blacktriangledown$. The second line (91b) is none other than (88b). □

This is a most fundamental result, insofar as that it tells us the correct form for the implicit equations in the relaxation variables $\mathbb{U}$. The naive idea

$$\mathbb{U}_i^{n+1} = \mathbb{U}_i^n - \frac{\Delta t}{\Delta x}[\mathbb{H}(\mathbb{U}_i^{n+1}, \mathbb{U}_{i+1}^{n+1}) - \mathbb{H}(\mathbb{U}_{i-1}^{n+1}, \mathbb{U}_i^{n+1})], \tag{92}$$

even followed by some projection step, leads to an erroneous discretization and disastrous results, as evidenced by Chalons [9,11] for steady-state solutions and by Baudin et al. [1,3] for transient solutions. A partial version of the correct form (91) was suggested by the same authors under the name of *implicit projection* [3, §3.2.2] and justified by convincing but sophisticated arguments based on the splitting procedure *relax evolution–set to equilibrium* (see §3.4). To our knowledge, however, equation (86) has never been given a prominent role in the derivation the proposed modifications.

In industrial applications, a handy and cheaper substitute for the full implicit scheme is the so-called *linearly implicit* strategy. It consists in performing just one Newton iteration on the full implicit system. This amounts to using the Taylor expansion

$$\mathbb{H}(\mathbb{U}_i^{n+1}, \mathbb{U}_{i+1}^{n+1}) \approx \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) + \nabla_L \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) \cdot \delta\mathbb{U}_i + \nabla_R \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) \cdot \delta\mathbb{U}_{i+1} \tag{93}$$

with the increments

$$\delta\mathbb{U} = \mathbb{U}^{n+1} - \mathbb{U}^n, \tag{94}$$

provided of course that $\mathbb{H}$ is differentiable. In our case, the Godunov flux $\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R)$ is notoriously not differentiable, but we will find a way to circumvent this difficulty in §4.2. Therefore, assuming differentiability, we denote by $\nabla_L \mathbb{H}$ and $\nabla_R \mathbb{H}$ the $(K+1) \times (K+1)$ matrices representing the partial derivatives of $\mathbb{H}$ with respect to $\mathbb{U}_L$ and $\mathbb{U}_R$. We introduce

$$\mathtt{B}_i = -\frac{\Delta t}{\Delta x} \nabla_L \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n), \qquad \mathtt{C}_{i+1} = \frac{\Delta t}{\Delta x} \nabla_R \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) \tag{95}$$

and

$$\mathtt{A}_i = \mathtt{I}_{2K+1} - \mathtt{B}_i - \mathtt{C}_i. \tag{96}$$

We also define the Jacobian matrix

$$\Lambda_i = \nabla \mathbf{\pounds}(\mathbf{u}_i^n), \tag{97}$$

the "explicit" update

$$\mathbb{U}_i^{n\curlyvee} = \mathbb{U}_i^n - \frac{\Delta t}{\Delta x}[\mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) - \mathbb{H}(\mathbb{U}_{i-1}^n, \mathbb{U}_i^n)], \tag{98}$$

which is out of equilibrium, and $\mathbf{u}_i^{n\curlyvee} = \blacktriangledown \mathbb{U}_i^{n\curlyvee}$ its projected value. The quotes are simply due to the fact that the time-step Δt at hand generally does not comply with the usual CFL condition for a "true" explicit scheme.

For any matrix $\mathtt{M}$ having $2K+1$ rows, we obtain $\blacktriangledown\mathtt{M}$ from $\mathtt{M}$ by extracting the rows corresponding to the indices kept by the projector $\blacktriangledown$. For any matrix $\mathtt{M}$ having $2K+1$ columns, we obtain $\mathtt{M}^{\blacktriangledown}$ (resp. $\mathtt{M}^{\perp}$) by selecting the columns corresponding to the indices preserved by the projector $\blacktriangledown$ (resp. $\perp$). The following Proposition gives the correct form for the linearly implicit scheme.

Proposition 4.2. *The linearly implicit scheme in the context of relaxation takes the form*

$$\begin{bmatrix} \blacktriangledown\mathtt{B}_{i-1}^{\blacktriangledown} & \blacktriangledown\mathtt{B}_{i-1}^{\perp} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \blacktriangledown\mathtt{A}_{i}^{\blacktriangledown} & \blacktriangledown\mathtt{A}_{i}^{\perp} \\ -\Lambda_i & \mathtt{I}_K \end{bmatrix} \begin{bmatrix} \blacktriangledown\mathtt{C}_{i+1}^{\blacktriangledown} & \blacktriangledown\mathtt{C}_{i+1}^{\perp} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \blacktriangledown\delta\mathbb{U}_{i-1} \\ \perp\delta\mathbb{U}_{i-1} \\ \blacktriangledown\delta\mathbb{U}_{i} \\ \perp\delta\mathbb{U}_{i} \\ \blacktriangledown\delta\mathbb{U}_{i+1} \\ \perp\delta\mathbb{U}_{i+1} \end{bmatrix} = \begin{bmatrix} \blacktriangledown(\mathbb{U}_i^{n\times} - \mathbb{U}_i^n) \\ 0 \end{bmatrix} \tag{99}$$

in the enlarged increments $\delta\mathbb{U}$. It is algebraically equivalent to the form

$$\begin{bmatrix} \blacktriangledown\mathtt{B}_{i-1}^{\blacktriangledown} + \blacktriangledown\mathtt{B}_{i-1}^{\perp}\Lambda_i & \blacktriangledown\mathtt{A}_{i}^{\blacktriangledown} + \blacktriangledown\mathtt{A}_{i}^{\perp}\Lambda_i & \blacktriangledown\mathtt{C}_{i+1}^{\blacktriangledown} + \blacktriangledown\mathtt{C}_{i+1}^{\perp}\Lambda_i \end{bmatrix} \begin{bmatrix} \delta\mathbf{u}_{i-1} \\ \delta\mathbf{u}_{i} \\ \delta\mathbf{u}_{i+1} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_i^{n\times} - \mathbf{u}_i^n \end{bmatrix} \tag{100}$$

in the original increments $\delta\mathbf{u}$. The latter can also be obtained by linearizing (86) directly.

Proof. Plug the approximate value $\mathbb{H}(\mathbb{U}_i^{n+1}, \mathbb{U}_i^{n+1})$ of (93), as well as the linear approximation

$$\perp\delta\mathbb{U}_i = \Lambda_i \cdot \blacktriangledown\delta\mathbb{U}_i \tag{101}$$

into the generic form (91). The calculations are similar to those in [3]. Eliminating the $\perp\delta\mathbb{U}$'s in (99) results in (100). Finally, it it not difficult to verify that a direct linearization of (86) yields (100). □

In both formulations, we have to cope with a block-tridiagonal linear system. On the grounds of computational efficiency, it is recommended to implement (100), the size of which is almost twice as small —$K+1$ versus $2K+1$— as (99). Nevertheless, the matrices in the enlarged formulation (99) are necessary for the design of a hybrid explicit-implicit time integration. Indeed, the matrices $\mathtt{A}_i$, $\mathtt{B}_i$ and $\mathtt{C}_i$ will eventually be altered in §4.3.

4.2. **Roe's form of the relaxation flux**

The Godunov flux $\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R)$ defined in (74) is only Lipschitz continuous and not differentiable. Moreover, the way it is expressed so far does not help us spot the non-smooth parts. Note that a Roe- or VFRoe-type flux, which is of the form

$$\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) = \tfrac{1}{2}[\mathbb{F}(\mathbb{U}_L) + \mathbb{F}(\mathbb{U}_R)] - \tfrac{1}{2}|\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)|(\mathbb{U}_R - \mathbb{U}_L), \tag{102}$$

would be easier to handle, because non-differentiability is concentrated in the absolute value of the matrix $\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)$. In such a situation, the cure could be to resort to a partial linearization by freezing the non-smooth parts, as advocated by Mulder and van Leer [26]. Concretely, we would decide that

$$\nabla_L\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) := \tfrac{1}{2}[\,\nabla\mathbb{F}(\mathbb{U}_L) + |\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)|\,], \tag{103a}$$

$$\nabla_R\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) := \tfrac{1}{2}[\,\nabla\mathbb{F}(\mathbb{U}_R) - |\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)|\,]. \tag{103b}$$

This motivates our search for a Roe-type interpretation (102), the existence of which is ultimately a consequence of the linear degeneracy of all fields in the relaxation model. The success of such an undertaking relies on the following technical Lemma, which was taken for granted in [3], but which needs to be brought to light for the sake of rigor.

Lemma 4.1. *If the parameter a fulfills condition (47), i.e.,*

$$a > \frac{-(v_R - v_L) + \sqrt{(v_R - v_L)^2 + 8\min(\tau_L, \tau_R)|\Pi_R - \Pi_L|}}{4\min(\tau_L, \tau_R)}, \tag{104}$$

then the $(2K+1)\times(2K+1)$ matrix

$$\mathscr{R}(\mathbb{U}_L, \mathbb{U}_R) = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 \\ v_L - a\tau_L & 0 & \dots & 0 & v^* & 0 & \dots & 0 & v_R + a\tau_R \\ \Pi_L + a^2\tau_L & 0 & \dots & 0 & \Pi^* & 0 & \dots & 0 & \Pi_R + a^2\tau_R \\ (c_1)_L & 1 & \dots & 0 & c_1^* & 1 & \dots & 0 & (c_1)_R \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ (c_{K-1})_L & 0 & \dots & 1 & c_{K-1}^* & 0 & \dots & 1 & (c_{K-1})_R \\ (\Sigma_1)_L & -b & \dots & 0 & \Sigma_1^* & b & \dots & 0 & (\Sigma_1)_R \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ (\Sigma_{K-1})_L & 0 & \dots & -b & \Sigma_{K-1}^* & 0 & \dots & b & (\Sigma_{K-1})_R \end{bmatrix}, \tag{105}$$

using the notations (40)–(41), is invertible.

Proof. Let $\{R_q\}$, $1 \le q \le 2K+1$, be the column vectors of $\mathscr{R}$, and assume that there is a linear combination such that $\sum_{1\le q\le 2K+1} \alpha_q R_q = \mathbf{0}$, with $\alpha_q \in \mathbb{R}$. We claim that

$$(\alpha_1, \alpha_{K+1}, \alpha_{2K+1}) = (0,0,0), \tag{106}$$

since otherwise, the 3×3 matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ v_L - a\tau_L & v^* & v_R + a\tau_R \\ \Pi_L + a^2\tau_L & \Pi^* & \Pi_R + a^2\tau_R \end{bmatrix}, \tag{107}$$

extracted from the first three rows of $\mathscr{R}$, would have to be singular. By subtracting the middle column to the other columns and by expanding with respect to the new first row, we find for the —supposedly equal to zero— determinant of this 3×3 matrix

$$-2a[v^* - (v_L - a\tau_L)][v^* - (v_R + a\tau_R)]. \tag{108}$$

By virtue of Theorem 3.2, however, this quantity is strictly positive if the parameter a meets condition (47). In turn, the equality (106) implies $\alpha_2 = \alpha_{K+2} = 0$ by examining rows 4 and $K+4$, columns 2 and $K+2$. We proceed likewise until $\alpha_K = \alpha_{2K} = 0$. □

The invertibility of $\mathscr{R}$ allows us to state the following Theorem, the most remarkable feature of which is, from a practical point of view, the fact that the matrix $\mathscr{A}$ can be computed by a closed formula.

Theorem 4.1. *If $a > b > 0$ and if condition (47) is satisfied for all $(\mathbb{U}_L, \mathbb{U}_R)$ under consideration, then there exists a $(2K+1)\times(2K+1)$ matrix $\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)$ such that the Godunov flux (74) can be expressed as*

$$\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) = \tfrac{1}{2}[\mathbb{F}(\mathbb{U}_L) + \mathbb{F}(\mathbb{U}_R)] - \tfrac{1}{2}|\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R)|(\mathbb{U}_R - \mathbb{U}_L). \tag{109}$$

This matrix can be given, for instance, by the formula

$$\mathscr{A}(\mathbb{U}_L, \mathbb{U}_R) = \mathscr{R}(\mathbb{U}_L, \mathbb{U}_R)\mathscr{D}(\mathbb{U}_L, \mathbb{U}_R)\mathscr{R}^{-1}(\mathbb{U}_L, \mathbb{U}_R), \tag{110}$$

in which the invertible matrix $\mathscr{R}(\mathbb{U}_L, \mathbb{U}_R)$ is defined by (105) of Lemma 4.1, and

$$\mathscr{D}(\mathbb{U}_L, \mathbb{U}_R) = \mathrm{Diag}(v_L - a\tau_L, v^* - b\tau_L^*, \dots, v^* - b\tau_L^*, v^*, v^* + b\tau_R^*, \dots, v^* + b\tau_R^*, v_R + a\tau_R) \tag{111}$$

using the notations (40)–(41). Formula (110) defines a Roe-type linearization for the relaxation system (25), in the sense that

1. $\mathscr{A}(\mathbb{U},\mathbb{U}) = \nabla\mathbb{F}(\mathbb{U})$*;*
2. $\mathscr{A}(\mathbb{U}_L,\mathbb{U}_R)$ *is* $\mathbb{R}$*-diagonalizable;*
3. $\mathscr{A}(\mathbb{U}_L,\mathbb{U}_R)(\mathbb{U}_R-\mathbb{U}_L) = \mathbb{F}(\mathbb{U}_R)-\mathbb{F}(\mathbb{U}_L)$.

Proof. The proof is similar to [3, Proposition 3.2] regarding the claim that $\mathscr{A}$ is a Roe matrix. For the readers interested in the details, the starting point of this part is the set of equalities below, which are the counterparts of the jump relations in [3, Lemma 3.1]:

$$\mathbb{U}_L^{\dashv} - \mathbb{U}_L = [\rho_L^* - \rho_L]R_1 \tag{112a}$$
$$\mathbb{U}_L^{\dashv\!\shortmid} - \mathbb{U}_L^{\dashv} = [\rho_L^* c_1^* - \rho_L^*(c_1)_L]R_2 + \ldots + [\rho_L^* c_{K-1}^* - \rho_L^*(c_{K-1})_L]R_K \tag{112b}$$
$$\mathbb{U}_R^{\shortmid\vdash} - \mathbb{U}_L^{\dashv\!\shortmid} = [\rho_R^* - \rho_L^*]R_{K+1} \tag{112c}$$
$$\mathbb{U}_R^{\vdash} - \mathbb{U}_R^{\shortmid\vdash} = [\rho_R^*(c_1)_R - \rho_R^* c_1^*]R_{K+2} + \ldots + [\rho_R^*(c_{K-1})_R - \rho_R^* c_{K-1}^*]R_{2K} \tag{112d}$$
$$\mathbb{U}_R - \mathbb{U}_R^{\vdash} = [\rho_R^* - \rho_R]R_{2K+1}. \tag{112e}$$

The vectors $\{R_q\}$, $1 \le q \le 2K+1$, are the columns of $\mathscr{R}$. The intermediate states $\mathbb{U}_L^{\dashv}$, $\mathbb{U}_L^{\dashv\!\shortmid}$, $\mathbb{U}_L^{\shortmid\vdash}$, $\mathbb{U}_L^{\vdash}$ are those of the Riemann solution in Proposition 3.2.

As for checking that this matrix gives rise to the flux equality (109), the steps are identical to those in [3, Theorem 3.3]: the formalism by Harten et al. [19] is invoked in to express the flux at issue in two different ways, the half-sum of which finally yields the desired result. In comparison with [3], the only difference lies in the multiplicity of the eigenvalues $v^* - b\tau_L^*$ and $v^* + b\tau_R^*$, but no serious complication arises in any part. □

For informative purpose, we wish to briefly report a second way to establish Theorem 4.1 via the Osher-Solomon interpretation [27]. This theoretical tool not only makes the checking of (109) simpler, but also brings new insight into the relaxation model.

Theorem 4.2. *If $a > b > 0$ and if condition (47) is satisfied for all $(\mathbb{U}_L,\mathbb{U}_R)$ under consideration, then there exists a path $\Gamma = \Gamma(\mathbb{U}_L,\mathbb{U}_R) \subset \Omega_{\mathbb{U}}$ such that the Godunov flux (74) can be expressed as*

$$\mathbb{H}(\mathbb{U}_L,\mathbb{U}_R) = \tfrac{1}{2}[\mathbb{F}(\mathbb{U}_L)+\mathbb{F}(\mathbb{U}_R)] - \tfrac{1}{2}\int_{\Gamma(\mathbb{U}_L,\mathbb{U}_R)} |\nabla\mathbb{F}(\mathbb{U})|\,\mathrm{d}\mathbb{U}. \tag{113}$$

This path, which is made up of $2K+1$ straight line portions

$$\Gamma = \Gamma_1 \cup (\Gamma_2 \cup \ldots \cup \Gamma_K) \cup \Gamma_{K+1} \cup (\Gamma_{K+2} \cup \ldots \cup \Gamma_{2K}) \cup \Gamma_{2K+1}, \tag{114}$$

each Γ_q being directed by the eigenvector R_q of the matrix $\mathscr{R}$ defined in (105), connects the intermediate states of the Riemann solution in accordance with the diagram

$$\mathbb{U}_L \xrightarrow{\ \Gamma_1\ } \mathbb{U}_L^{\dashv} \xrightarrow{\ \Gamma_2\cup\ldots\cup\Gamma_K\ } \mathbb{U}_L^{\dashv\!\shortmid} \xrightarrow{\ \Gamma_{K+1}\ } \mathbb{U}_R^{\shortmid\vdash} \xrightarrow{\ \Gamma_{K+2}\cup\ldots\cup\Gamma_{2K}\ } \mathbb{U}_R^{\vdash} \xrightarrow{\ \Gamma_{2K+1}\ } \mathbb{U}_R.$$

Proof. Except for the claim that every Γ_q is a straight line portion, the existence of such a path follows from the linear degeneracy of all fields. The Γ_q's are none other than the integral curves, along which the tangent vector is parallel to the right eigenvector, i.e.,

$$\frac{\mathrm{d}\mathbb{U}}{\mathrm{d}\zeta}(\zeta) = R_q(\mathbb{U}(\zeta)) \tag{115}$$

for some parametrization $\zeta \mapsto \mathbb{U}(\zeta)$ of Γ_q. Again, the multiplicity of the eigenvalues $v \pm b\tau$ is hardly a source of annoyance. This is exemplified by the the jump from $\mathbb{U}_L^{\dashv}$ to $\mathbb{U}_L^{\dashv\!\shortmid}$, which is not a pure contact discontinuity

because $v^* - a\tau_L^*$ is not a "simple" eigenvalue, but which can be decomposed into $K-1$ elementary "simple" jumps.

It comes as an amazing feature of the relaxation model (25) that along each portion Γ_q, the right eigenvector $R_q(\mathbb{U}(\zeta))$ is constant, which implies that Γ_q is a straight line portion! Indeed, inspection of the matrix of eigenvectors (30) reveals that only columns 1, $K+1$ and $2K+1$ may have non-constant entries. Using the strong Riemann invariants (29), it can be verified that the components of R_1 are constant across the 1-contact discontinuity. We remind that a strong Riemann invariant for a given field is a weak Riemann invariants for all other fields. The same line of reasoning applies to the two other eigenvectors. □

For short, let $\{\nu_q\}$, $1 \le q \le 2K+1$ be the diagonal entries of $\mathscr{D}(\mathbb{U}_L, \mathbb{U}_R)$ defined in (111). As before, R_q is the qth column of the matrix $\mathscr{R}(\mathbb{U}_L, \mathbb{U}_R)$ given in (105). On one hand, by invariance of the qth-eigenvalue field and qth-eigenvector field along Γ_q, we have

$$\int_{\Gamma_q} |\nabla \mathbb{H}(\mathbb{U})| \, \mathrm{d}\mathbb{U} = |\nu_q|(\zeta_{q+1} - \zeta_q) R_q, \tag{116}$$

assuming that the segment Γ_q corresponds to the interval $[\zeta_q, \zeta_{q+1}]$ within the parametrization $\zeta \mapsto \mathbb{U}(\zeta)$ for the path Γ. Summing these equalities over q yields

$$\begin{aligned} \int_{\Gamma} |\nabla \mathbb{H}(\mathbb{U})| \, \mathrm{d}\mathbb{U} &= |\nu_1|(\zeta_2 - \zeta_1) R_1 + \ldots + |\nu_{2K+1}|(\zeta_{2K+2} - \zeta_{2K+1}) R_{2K+1} \\ &= \mathscr{R}|\mathscr{D}| \left(\zeta_2 - \zeta_1, \ldots, \zeta_{2K+2} - \zeta_{2K+1}\right)^T \end{aligned} \tag{117}$$

On the other hand, from the plain equality

$$\mathbb{U}_R - \mathbb{U}_L = \int_{\Gamma_1} \mathrm{d}\mathbb{U} + \ldots + \int_{\Gamma_{2K+1}} \mathrm{d}\mathbb{U} = (\zeta_2 - \zeta_1) R_1 + \ldots + (\zeta_{2K+2} - \zeta_{2K+1}) R_{2K+1} \tag{118}$$

and from the invertibility of $\mathscr{R}$ proven in Lemma 4.1, we infer that

$$\left(\zeta_2 - \zeta_1, \ldots, \zeta_{2K+2} - \zeta_{2K+1}\right)^T = \mathscr{R}^{-1}(\mathbb{U}_R - \mathbb{U}_L). \tag{119}$$

Therefore, combining (117) and (119), we end up with

$$\int_{\Gamma} |\nabla \mathbb{H}(\mathbb{U})| \, \mathrm{d}\mathbb{U} = \mathscr{R}|\mathscr{D}|\mathscr{R}^{-1}(\mathbb{U}_R - \mathbb{U}_L) = |\mathscr{R}\mathscr{D}\mathscr{R}^{-1}|(\mathbb{U}_R - \mathbb{U}_L). \tag{120}$$

This alternative proof of (109) relies on the property that the integral curves are straight lines. Had the eigenvector $R_q(\mathbb{U}(\zeta))$ not been invariant along Γ_q, we would have had to replace $\mathscr{R}$ by $\widehat{\mathscr{R}}$, the matrix whose columns are

$$\widehat{R}_q = \frac{1}{\zeta_{q+1} - \zeta_q} \int_{\zeta_q}^{\zeta_{q+1}} R_q(\mathbb{U}(\zeta)) \mathrm{d}\mathbb{U}, \tag{121}$$

in equations (117)–(119). Thus the Roe matrix would have been $\widehat{\mathscr{R}}\mathscr{D}\widehat{\mathscr{R}}^{-1}$. However, the invertibility of $\widehat{\mathscr{R}}$ is unclear.

4.3. **From implicit to selectively implicit**

After solving the non-differentiability problem for $\mathbb{H}(\mathbb{U}_L, \mathbb{U}_R)$, we go back to the Taylor expansions (93). In order to become explicit with respect to kinematic waves, our *modus operandi* is to prevent the first-order terms

$$\nabla_L \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) \cdot \delta\mathbb{U}_i + \nabla_R \mathbb{H}(\mathbb{U}_i^n, \mathbb{U}_{i+1}^n) \cdot \delta\mathbb{U}_{i+1} \tag{122}$$

from capturing any contribution due to slow waves. As initiated by Faille and Heintzé [16] and explained in [3, §3.2], we replace the approximations (103) by

$$\nabla_L \mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) \approx \tfrac{1}{2}[\, \widetilde{\nabla \mathbb{F}}(\mathbb{U}_L) + |\widetilde{\mathscr{A}}(\mathbb{U}_L, \mathbb{U}_R)|\,] \tag{123a}$$

$$\nabla_R \mathbb{H}(\mathbb{U}_L, \mathbb{U}_R) \approx \tfrac{1}{2}[\, \widetilde{\nabla \mathbb{F}}(\mathbb{U}_R) - |\widetilde{\mathscr{A}}(\mathbb{U}_L, \mathbb{U}_R)|\,], \tag{123b}$$

where $\widetilde{\ }$ stands for the *acoustic part* operator, defined as follows. Let M be a $(2K+1)\times(2K+1)$ matrix representing any of the matrices $\nabla\mathbb{F}$ or $\mathscr{A}$, and of course supposed to be $\mathbb{R}$-diagonalizable. From the diagonalized form

$$M = R\,\mathrm{Diag}(\nu_1, \nu_2, \dots, \nu_K, \nu_{K+1}, \nu_{K+2}, \dots \nu_{2K}, \nu_{2K+1})\, R^{-1}, \tag{124}$$

we set

$$\widetilde{M} = R\,\mathrm{Diag}(\nu_1, 0, \dots, 0, 0, 0, \dots 0, \nu_{2K+1})\, R^{-1}. \tag{125}$$

We have to be aware of the dependence of the basis of eigenvectors R on the matrix M. This is why this selectively implicit procedure is best seen as a heuristic, although it is exact for linear hyperbolic systems and works fine in nonlinear cases. Anyhow, it allows us to work with a time-step subject to the CFL condition

$$\frac{\Delta t}{\Delta x} \max_{i\in\mathbb{Z}} \max\{|v^* - b\tau_L^*|_{i+1/2}^n,\, |v^* + b\tau_R^*|_{i+1/2}^n\} < \frac{1}{2} \tag{126}$$

based on the slow characteristic speeds $v \pm b\tau$, thus saving CPU time and maintaining accuracy on kinematic waves. To get a good compromise between a large time-step and an acceptable amount of smearing for the acoustic part of the solution, it is advised to take into account the additional safety limitation

$$\frac{\Delta t}{\Delta x} \max_{i\in\mathbb{Z}} \max\{|v_L - a\tau_L|_{i+1/2}^n,\, |v_R + a\tau_R|_{i+1/2}^n\} < 20 \tag{127}$$

based on fast characteristic speeds $v \pm a\tau$.

5. Numerical results

We carry out two numerical simulations corresponding to $K = 2$ components, endowed with an idealized but physically meaningful pressure law $p = p(\rho, \rho c_1)$. This thermodynamic law assumes a homogenized response of the two phase densities

$$\rho_\ell(p) = \rho_\ell^0 + \frac{p - p^0}{a_\ell^2} \qquad \text{and} \qquad \rho_g(p) = \frac{p}{a_g^2} \tag{128}$$

with respect to the pressure p, but introduces the notion of *dew* and *bubble* curves in the $(\rho, \rho c_1)$-plane. We refer the reader to [8, 31] for details. In the test cases, we set

$$\begin{aligned} &p_0 = 10^5 \text{ Pa}, && \rho_\ell^0 = 997.10 \text{ kg/m}^3, \\ &a_g = 330 \text{ m/s}, && a_\ell = 500 \text{ m/s}. \end{aligned} \tag{129}$$

The simulations are Riemann problems over a sufficiently long domain, discretized by the uniform space meshing $\Delta x = 0.5$ m. The discontinuity in the initial data is located at $x = 50$ m. Actually, we use a second-order versions of the first-order schemes presented in §3 and §4. The enhancement procedures are the same as in [3, §2.4 & §2.5].

5.1. **Fast shock with zero-slip law**

In the first test case, we consider the zero-slip law $\phi \equiv 0$, which implies that the gas and liquid phases move at the same velocity. The reason why we start by taking such a simplified hydrodynamic law is that we first want to see whether or not the above two-phase two-component thermodynamic law, which is far more sophisticated than that used in [3], is well-supported by the schemes. Another reason is that for $\phi \equiv 0$, the original system (12) is known to be hyperbolic, and an exact solution to the Riemann problem associated with (12) is available for comparison.

The initial data

$$\begin{pmatrix} \rho \\ c_1 \\ v \end{pmatrix}_L = \begin{pmatrix} 0.999 \times 10^5 \\ 0.98 \\ 89.569 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} \rho \\ c_1 \\ v \end{pmatrix}_R = \begin{pmatrix} 0.9 \times 10^5 \\ 0.98 \\ 55.076 \end{pmatrix} \tag{130}$$

are tailored so that this exact solution is a pure acoustic shock propagating at 400 m/s. Here and contrarily to the common practice mentioned in the Introduction, we take a close look at the acoustic wave in order to assess the amount of dissipation. Figure 1 compares the pressure computed by the schemes of this paper to their VFRoe counterparts of the TACITE code [28]. In the explicit setting, relaxation is less diffusive than VFRoe. In the hybrid explicit-implicit setting, relaxation is more diffusive than VFRoe. From this and other numerical runs, the general observation is that the hybrid explicit-implicit version of the relaxation scheme captures acoustic waves with a little more dissipation than VFRoe.

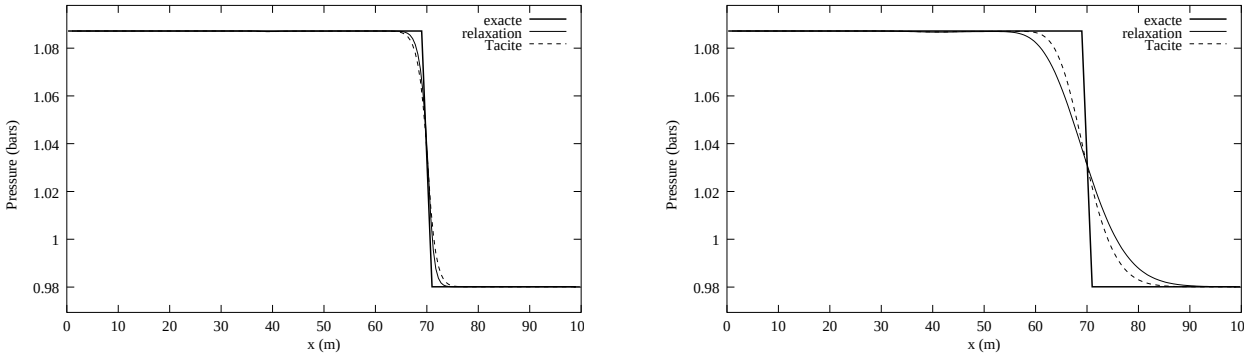

FIGURE 1. Explicit (left) and hybrid explicit-implicit (right) schemes for experiment 1

5.2. **Slow transportation with Zuber-Finlday's law**

We now switch to the more realistic slip law

$$\phi(\mathbf{u}) = -\frac{0.2\,v + 0.35\sqrt{gD}}{1 + 0.2\,\kappa(\mathbf{u})}, \qquad \text{and} \qquad \kappa(\mathbf{u}) = \frac{1 - Y(\mathbf{u})}{\rho/\rho_{\mathrm{L}}(p(\mathbf{u})) - 1}, \tag{131}$$

where $g = 9.81$ m/s^2 is the gravity constant and $D = 0.144$ m is the diameter of the pipeline. The hydrodynamic relation (131) is known as the Zuber-Findlay law [33], and reflects the intermittent flow of medium-sized gas bubbles in a vertical duct.

This is a difficult experiment, insofar as the initial data

$$\begin{pmatrix} \rho \\ c_1 \\ v \end{pmatrix}_L = \begin{pmatrix} 0.91827 \\ 0.8163265 \\ -10 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} \rho \\ c_1 \\ v \end{pmatrix}_R = \begin{pmatrix} 997.1 \\ 0.0816327 \\ -10 \end{pmatrix} \tag{132}$$

are adjusted so that we cover the whole two-phase domain from $Y_L = 0$ to $Y_R = 1$. Our goal is to put the schemes through the test of robustness at the boundaries of the admissible domain for thermodynamics. Both versions (explicit and selectively implicit) of the VFRoe method in the TACITE code failed this test and stopped after producing a negative mass fraction c_1.

We do not know the exact solution to the Riemann problem equipped with this combination of closure laws. But on the basis of the analysis by Benzoni-Gavage [5] for the Zuber-Findlay law paired with a much simpler pressure law, and considering that $v_L = v_R$ and $p_L \approx p_R \approx 10^5$ Pa, we can legitimately anticipate the "exact" solution as a slow shock propagating at a speed around -10 m/s. This guess is in good a agreement with the curves in Fig. 2, where we display the density and the gas mass fraction.

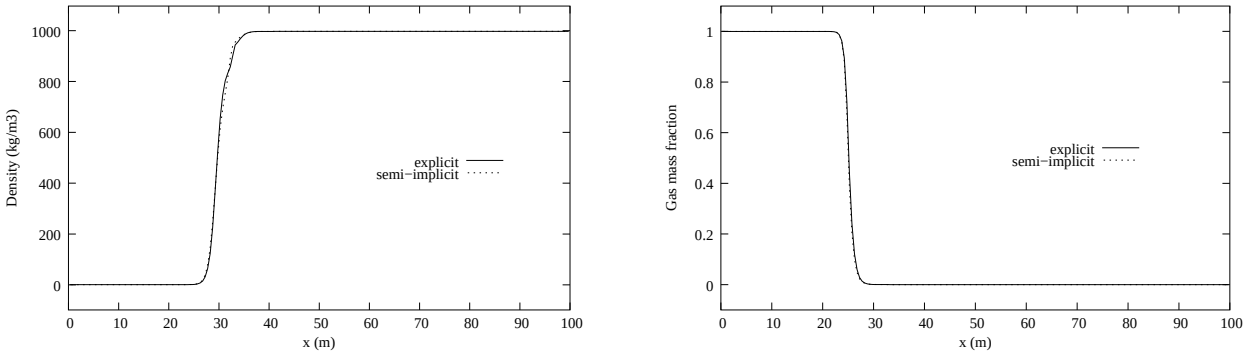

FIGURE 2. Density (left) and the gas mass fraction (right) for experiment 2

6. Conclusion

Once again, relaxation has proved to be a valuable tool in the design of an effective method for the simulation of two-phase flows governed by a drift-flux model. The explicit and the hybrid explicit-implicit versions bring a very satisfactory answer to the difficulties due to the multi-component nature of the mixture, namely: a higher degree of nonlinearity in the closure laws and an impressive increase in the number of unknowns to be updated. In trying to deploy the same strategies as in [2,3], we have gained further insight into the theoretical foundations of the relaxation method. We hope these findings could be useful to other researches in the area.

References

[1] M. Baudin, *Méthodes de relaxation pour la simulation des écoulements polyphasiques dans les conduites pétrolières*, Thèse de doctorat, Université Pierre et Marie Curie, 2003.

[2] M. Baudin, C. Berthon, F. Coquel, R. Masson, and Q. H. Tran, *A relaxation method for two-phase flow models with hydrodynamic closure law*, Numer. Math., 99 (2005), pp. 411–440.

[3] M. Baudin, F. Coquel, and Q. H. Tran, *A semi-implicit relaxation scheme for modeling two-phase flow in a pipeline*, SIAM J. Sci. Comput., 27 (2005), pp. 914–936.

[4] M. Baudin, F. Coquel, and Q. H. Tran, *A relaxation method via the Born-Infeld system*, Math. Models Meth. Appl. Sci., (2008). to appear.

[5] S. Benzoni-Gavage, *Analyse Numérique des Modèles Hydrodynamiques d'Écoulements Diphasiques Instationnaires dans les Réseaux de Production Pétrolière*, thèse de doctorat, École Normale Supérieure de Lyon, 1991.

[6] F. Bouchut, *Entropy satisfying flux vector splittings and kinetic BGK models*, Numer. Math., 94 (2003), pp. 623–672.

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

[8] V. Bouvier, *Algorithmes de résolution compostionnelle dans TACITE*, Technical report 42485, Institut Français du Pétrole, 1995.

[9] C. Chalons, *Bilans d'entropie discrets dans l'approximation numérique des chocs non classiques. Application aux équations de Navier-Stockes multi-pression et à quelques systèmes visco-capillaires*, phd dissertation, École Polytechnique, novembre 2002. in French.
[10] C. Chalons and F. Coquel, *Navier-Stokes equations with several independent pressure laws and explicit predictor-corrector schemes*, Numer. Math., 101 (2005), pp. 451–478.
[11] C. Chalons, F. Coquel, and C. Marmignon, *Well-balanced time implicit formulation of relaxation schemes for the Euler equations*, SIAM J. Sci. Comput., 30 (2008), pp. 349–415.
[12] G. Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.
[13] F. Coquel, Q. L. Nguyen, M. Postel, and Q. H. Tran, *Entropy-satisfying relaxation method with large time-steps for Euler IBVPs*, Math. Comp., (2007). submitted.
[14] F. Coquel and B. Perthame, *Relaxation of energy and approximate Riemann solvers for general pressure laws in fluid dynamics*, SIAM J. Numer. Anal., 35 (1998), pp. 2223–2249.
[15] B. Després, *Lagrangian systems of conservation laws*, Numer. Math., 89 (2001), pp. 99–134.
[16] I. Faille and É. Heintzé, *A rough finite volume scheme for modeling two phase flow in a pipeline*, Computers and Fluids, 28 (1999), pp. 213–241.
[17] T. Gallouët and J.-M. Masella, *A rough Godunov scheme*, Compte-Rendus à l'Académie des Sciences, 323 (1996), p. 77.
[18] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, vol. 118 of Applied Mathematical Sciences, Springer-Verlag, New York, 1996.
[19] A. Harten, P. D. Lax, and B. van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Review, 25 (1983), pp. 35–61.
[20] S. Jin and M. Slemrod, *Regularization of the burnett equations via relaxation*, J. Stat. Phys., 103 (2001), pp. 1009–1033.
[21] S. Jin and Z. P. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimension*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.
[22] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.
[23] A. Majda and R. L. Pego, *Stable viscosity matrices for systems of conservation laws*, J. Diff. Eq., 56 (1985), pp. 229–262.
[24] J. M. Masella, I. Faille, and T. Gallouët, *On an approximate Godunov scheme*, Int. J. Comput. Fluid Dynam., 12 (1999), pp. 133–149.
[25] J.-M. Masella, Q. H. Tran, D. Ferré, and C. Pauchon, *Transient simulation of two-phase flows in pipes*, Int. J. Multiphase Flow, 24 (1998), pp. 739–755.
[26] W. A. Mulder and B. van Leer, *Experiments with implicit upwind methods for the Euler equations*, J. Comput. Phys., 59 (1985), pp. 232–246.
[27] S. Osher and F. Solomon, *Upwind difference schemes for hyperbolic systems of conservation laws*, Math. Comp., 38 (1982), pp. 339–374.
[28] C. Pauchon, H. Dhulesia, G. Binh-Cirlot, and J. Fabre, *TACITE: A transient tool for multiphase pipeline and well simulation*, SPE Annual Technical Conference and Exhibition, New Orleans, September 1994, 1994. SPE Paper 28545.
[29] P. L. Roe, *Approximate Riemann solvers, parameter vectors and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.
[30] D. Serre, *Systèmes de Lois de Conservation I & II*, Collection Fondations, Diderot Éditeur Arts et Sciences, Paris, 1996.
[31] J. Vidal, *Thermodynamique : méthodes appliquées au raffinage et au génie chimique*, Technip, Paris, 1973.
[32] G. B. Whitham, *Linear and Nonlinear Waves*, vol. 258 of Pure and Applied Mathematics, Wiley-Interscience, New York, 1974.
[33] N. Zuber and J. Findlay, *Average volumetric concentration in two-phase flow systems*, J. Heat Transfer, 87 (1965), pp. 453–458.

Appendix A. Whitham-like conditions and positivity principles at the discrete level

This Appendix aims at carefully examining the connection between various requirements on the relaxation parameters and their effects at the discrete level for two basic models. It seems to us that the results gathered here in a unifying view are not as widely known as they deserve.

A.1. **Scalar conservation law**

Let us approximate the scalar hyperbolic equation

$$\partial_t Y + \partial_m \sigma(Y) = 0, \tag{133}$$

with $\sigma \in \mathscr{C}^1([0,1];\mathbb{R})$, by the Jin-Xin relaxation system [21]

$$\partial_t Y^\lambda + \quad \partial_m \Sigma^\lambda = 0 \tag{134a}$$

$$\partial_t \Sigma^\lambda + b^2 \partial_m Y^\lambda = \lambda[\sigma(Y^\lambda) - \Sigma^\lambda] \tag{134b}$$

for $\lambda > 0$ and $b > 0$. By the Chapman-Enskog analysis, it is shown that the asymptotic equilibrium equation for (134) in the limit $\lambda \to +\infty$ is

$$\partial_t Y^\lambda + \partial_m \sigma(Y^\lambda) = \lambda^{-1} \partial_m \{[b^2 - \sigma_Y^2(Y^\lambda)] \partial_m Y^\lambda\}, \tag{135}$$

so that a sufficient condition for linear stability is the Whitham condition

$$b > |\sigma_Y(Y)| \tag{136}$$

at the continuous level.

Let $Y_L \in [0,1]$ and $Y_R \in [0,1]$. The solution to the Riemann problem of the homogeneous system (134) corresponding to the initial data

$$\begin{bmatrix} Y \\ \Sigma \end{bmatrix}(t=0,m) = \begin{bmatrix} Y_L \\ \sigma(Y_L) \end{bmatrix} \mathbf{1}_{\{m<0\}} + \begin{bmatrix} Y_R \\ \sigma(Y_R) \end{bmatrix} \mathbf{1}_{\{m>0\}} \tag{137}$$

is the self-similar function

$$\begin{bmatrix} Y \\ \Sigma \end{bmatrix}\left(\tfrac{m}{t}\right) = \begin{bmatrix} Y_L \\ \sigma_L \end{bmatrix} \mathbf{1}_{\{\frac{m}{t} < -b\}} + \begin{bmatrix} Y^* \\ \Sigma^* \end{bmatrix} \mathbf{1}_{\{-b < \frac{m}{t} < b\}} + \begin{bmatrix} Y_R \\ \sigma_R \end{bmatrix} \mathbf{1}_{\{\frac{m}{t} > b\}}, \tag{138}$$

where

$$\begin{aligned} &\text{(a)} \quad Y^* = \frac{Y_R + Y_L}{2} - \frac{\sigma_R - \sigma_L}{2b} \\ &\text{(b)} \quad \Sigma^* = \frac{\sigma_R + \sigma_L}{2} - b\frac{Y_R - Y_L}{2} \end{aligned} \tag{139}$$

with the shorthand notations $\sigma_L = \sigma(Y_L)$ and $\sigma_R = \sigma(Y_R)$. In imitation of (136), we consider

$$b^\natural(Y_L, Y_R) = \left|\frac{\sigma_R - \sigma_L}{Y_R - Y_L}\right|, \tag{140}$$

which is well-defined for $Y_L = Y_R$.

Theorem A.1. *$Y^* \in \lfloor Y_L, Y_R \rceil$ if and only if $b \geq b^\natural(Y_L, Y_R)$.*

Proof. From (139a), it is straightforward to check that

$$(Y^* - Y_L)(Y^* - Y_R) = \frac{(\sigma_R - \sigma_L)^2}{b^2} - (Y_R - Y_L)^2. \tag{141}$$

Therefore, the left-hand side is negative if and only if $b > b^\natural(Y_L, Y_R)$. □

Hence, for a scalar conservation law, the discrete version $b > b^\natural(Y_L, Y_R)$ of the continuous Whitham condition $b > |\sigma_Y|$ yields a min-max principle on the intermediate state Y^* of the Riemann problem when the left and right data are at equilibrium. This min-max principle on the intermediate state in turn gives rise to the min-max principle on the updated value at the cell-centers. Indeed, if we use the first-order explicit scheme

$$Y_j^{n+1} = Y_j^n - \frac{\Delta t}{\Delta m}[\Sigma^*(Y_j^n, Y_{j+1}^n) - \Sigma^*(Y_{j-1}^n, Y_j^n)], \tag{142}$$

then by an averaging argument, we can show that

$$Y_j^{n+1} \in \lfloor Y_{j-1}^n, Y_j^n, Y_{j+1}^n \rceil \tag{143}$$

as soon as

$$b_{i-1/2} > b^\natural(Y_{i-1}^n, Y_i^n), \quad b_{i+1/2} > b^\natural(Y_i^n, Y_{i+1}^n), \quad \frac{\Delta t}{\Delta m} \max\{b_{i-1/2}, b_{i+1/2}\} < \frac{1}{2}. \tag{144}$$

In practice, the habit is to define

$$b^\flat(Y_L, Y_R) = \max\{|\sigma_Y(Y_L)|, |\sigma_Y(Y_R)|\}. \tag{145}$$

If σ is a convex or concave function over the interval $\lfloor Y_L, Y_R \rceil$, then $b^\flat \geq b^\natural$ and we are justified in choosing b larger than $b^\flat(Y_L, Y_R)$.

Theorem A.2. *For flux functions of the form $\sigma(Y) = Y(1-Y)\phi(Y)$, a sufficient condition to secure $Y^* \in [0,1]$ is that*

$$b \geq b^\sharp(Y_L, Y_R) = \max\{|\phi(Y_L)|, |\phi(Y_R)|\}. \tag{146}$$

Proof. By formally setting $\rho = 1$, the proof follows the same lines as —and is even easier than— Theorem 3.3. □

Plainly, $Y^* \in \lfloor Y_L, Y_R \rceil \Rightarrow Y^* \in [0,1]$. Put another way, for a pure scalar conservation law, we do not have to worry about $b \geq b^\sharp$. The Whitham-like condition $b \geq b^\natural$ is the one and only condition to be taken care of. By the way, it is likely that $b^\natural \geq b^\sharp$.

Things change, however, when instead of $\sigma(Y)$ we have $\sigma(\tau, Y, v)$, as is formally the case of (12c) in Lagrangian coordinates. In such a situation, $b^\natural$ may no longer be finite for $Y_L = Y_R$, and even if it is, there is no longer any clear inequality between $b^\natural$ and $b^\flat$. The advantage of $b^\flat$ over $b^\natural$ is that it remains well-defined, using the partial derivative with respect to Y. This is why we heuristically have to look for $b \geq \max\{b^\sharp, b^\flat\}$, as indicated in (83b).

A.2. **Euler's isentropic model**

Let us approximate the p-system [18]

$$\partial_t \tau - \partial_m v = 0 \tag{147a}$$

$$\partial_t v + \partial_m p(\tau) = 0, \tag{147b}$$

where the pressure law is assumed to satisfy

$$p_\tau < 0 \qquad \text{and} \qquad p_{\tau\tau} > 0, \tag{148}$$

by the relaxation system

$$\partial_t \tau^\lambda - \partial_m v^\lambda = 0 \tag{149a}$$

$$\partial_t v^\lambda + \partial_m \Pi^\lambda = 0 \tag{149b}$$

$$\partial_t \Pi^\lambda + a^2 \partial_m \Pi^\lambda = \lambda[p(\tau^\lambda) - \Pi^\lambda] \tag{149c}$$

for $\lambda > 0$ and $a > 0$. By the Chapman-Enskog analysis, it is shown that the asymptotic equilibrium equation for (149) in the limit $\lambda \to +\infty$ is

$$\partial_t \tau^\lambda - \partial_m v^\lambda = 0 \tag{150a}$$

$$\partial_t v^\lambda + \partial_m p(\tau^\lambda) = \lambda^{-1} \partial_m \{[a^2 + p_\tau(\tau^\lambda)] \partial_m v\}, \tag{150b}$$

so that a sufficient condition for linear stability is the Whitham condition

$$a > \sqrt{-p_\tau(\tau)} \tag{151}$$

at the continuous level.

Introduce $\mathbf{w} = (\tau, v)$ and let $\mathbf{w}_L \in \mathbb{R}_+^* \times \mathbb{R}$ and $\mathbf{w}_R \in \mathbb{R}_+^* \times \mathbb{R}$. The solution to the Riemann problem of the homogeneous system (149) corresponding to the initial data

$$\begin{bmatrix} \mathbf{w} \\ \Pi \end{bmatrix}(t=0,m) = \begin{bmatrix} \mathbf{w}_L \\ p(\tau_L) \end{bmatrix} \mathbf{1}_{\{m<0\}} + \begin{bmatrix} \mathbf{w}_R \\ p(\tau_R) \end{bmatrix} \mathbf{1}_{\{m>0\}} \tag{152}$$

is the self-similar function

$$\begin{bmatrix} \mathbf{w} \\ \Pi \end{bmatrix}\left(\tfrac{m}{t}\right) = \begin{bmatrix} \mathbf{w}_L \\ p_L \end{bmatrix} \mathbf{1}_{\{\frac{m}{t}<-a\}} + \begin{bmatrix} (\tau_L^*, v^*) \\ \Pi^* \end{bmatrix} \mathbf{1}_{\{-a<\frac{m}{t}<0\}} + \begin{bmatrix} (\tau_R^*, v^*) \\ \Pi^* \end{bmatrix} \mathbf{1}_{\{0<\frac{m}{t}<a\}} + \begin{bmatrix} \mathbf{w}_R \\ p_R \end{bmatrix} \mathbf{1}_{\{\frac{m}{t}>a\}}, \tag{153}$$

where

$$\begin{aligned}
&\text{(a)} \quad v^* = \frac{v_R + v_L}{2} - \frac{p_R - p_L}{2a} \\
&\text{(b)} \quad \Pi^* = \frac{p_R + p_L}{2} - a\frac{v_R - v_L}{2} \\
&\text{(c)} \quad \tau_L^* = \frac{v_R - v_L}{2a} - \frac{p_R - p_L}{2a^2} + \tau_L \\
&\text{(d)} \quad \tau_R^* = \frac{v_R - v_L}{2a} + \frac{p_R - p_L}{2a^2} + \tau_R
\end{aligned} \tag{154}$$

with the shorthand notations $p_L = p(\tau_L)$ and $p_R = p(\tau_R)$. In imitation of (151), we consider

$$a^\natural(\tau_L, \tau_R) = \left(-\frac{p_R - p_L}{\tau_R - \tau_L}\right)^{1/2} \tag{155}$$

which is well-defined for $\tau_L = \tau_R$.

Proposition A.1. *$(\tau_R^* - \tau_L^*)(\tau_R - \tau_L) \geq 0$ if and only if $a \geq a^\natural(\tau_L, \tau_R)$.*

Proof. This equivalence is a direct consequence of (154c)–(154c). □

Hence, for the p-system, the discrete version $a > a^\natural(\tau_L, \tau_R)$ of the continuous Whitham condition $a > \sqrt{-p_\tau}$ allows the initial ordering of the specific volumes to be carried over to the intermediate states.

What we want to investigate above all is the positivity of the intermediate volumes τ_L^*, τ_R^* and the extent to which this positivity could be implied by a Whitham-like condition. Let us recall a preliminary result which we already encountered in Theorem 3.2.

Proposition A.2. *The intermediate specific volumes τ_L^*, τ_R^* are positive if and only if we have the ordering*

$$v_L - a\tau_L < v^* < v_R + a\tau_R \tag{156}$$

for the characteristic speeds of the Eulerian version of (149).

Proof. See Theorem 3.2 or directly check using (154). □

The main result we wish to put forward is the following.

Theorem A.3.

(1) *If $v_R \geq v_L$, then for all $a \geq a^\natural(\tau_L, \tau_R)$ we have $\tau_L^* > 0$ and $\tau_R^* > 0$.*

(2) *If* $v_R < v_L$ *but*

$$|v_R - v_L| < a^\natural(\tau_L, \tau_R)(\tau_R + \tau_L), \tag{157}$$

then the choice $a = a^\natural(\tau_L, \tau_R)$ *ensures* $\tau_L^* > 0$ *and* $\tau_R^* > 0$. *Condition (157) holds in particular for a subsonic regime where*

$$|v_L| < a^\natural(\tau_L, \tau_R)\tau_L \qquad \textit{and} \qquad |v_R| < a^\natural(\tau_L, \tau_R)\tau_R. \tag{158}$$

If, in addition to (157), the ratio in the densities does not exceed 3, i.e.,

$$\max\{\tau_L, \tau_R\} \le 3\min\{\tau_L, \tau_R\}, \tag{159}$$

then for all $a \ge a^\natural(\tau_L, \tau_R)$ *we have* $\tau_L^* > 0$ *and* $\tau_R^* > 0$.

Proof. It is convenient to use the notations

$$[\![\Psi]\!] = \Psi_R - \Psi_L \qquad \text{and} \qquad \overline{\Psi} = \frac{\Psi_R + \Psi_L}{2} \tag{160}$$

for any quantity Ψ. Plugging the expression (154a) for v^* into (156) yields two quadratic inequations

$$\wp_L(a) = 2\tau_L a^2 + [\![v]\!]a - [\![p]\!] > 0, \tag{161a}$$

$$\wp_R(a) = 2\tau_R a^2 + [\![v]\!]a + [\![p]\!] > 0. \tag{161b}$$

On putting $a = a^\natural(\tau_L, \tau_R) = (-[\![p]\!]/[\![\tau]\!])^{1/2}$, we obtain

$$\wp_L(a^\natural) = \wp_R(a^\natural) = a^\natural([\![v]\!] + 2a^\natural\overline{\tau}). \tag{162}$$

Case 1. If $[\![v]\!] \ge 0$, then $a = a^\natural$ is a suitable choice because $\wp_L(a^\natural) = \wp_R(a^\natural) > 0$. Furthermore, both functions $\wp_L$ and $\wp_R$ are increasing with respect to $a \in \mathbb{R}_+^*$. Consequently, every $a > a^\natural$ also satisfies (161).

Case 2. If $[\![v]\!] < 0$ but within the smallness constraint (157), we still have $\wp_L(a^\natural) = \wp_R(a^\natural) > 0$, and $a = a^\natural$ remains a suitable choice. To deduce (157) from the subsonic hypothesis (158), we add up the inequalities

$$v_R + a^\natural\tau_R > 0 \qquad \text{and} \qquad -v_L + a^\natural\tau_L > 0. \tag{163}$$

The delicate point with $[\![v]\!] < 0$ is that since $\wp_L$ and $\wp_R$ are not increasing with respect to $a \in \mathbb{R}_+^*$, it may happen that as $a \to +\infty$, the quantities $\wp_L(a)$ and/or $\wp_R(a)$ becomes negative for a while before turning positive again.

To fix ideas, assume $[\![\tau]\!] \le 0$. Then, $[\![p]\!] \ge 0$. The polynomial $\wp_L$ has two real roots of opposite signs. Between the roots, it takes negative values. Therefore, $a^\natural$ must be larger than the positive root. As a consequence, $\wp_L(a) > 0$ for all $a > a^\natural$. As far as $\wp_R$ is concerned, it may have two positive roots or may no real roots, according to the sign of the discriminant $[\![v]\!]^2 - 8\tau_R[\![p]\!]$. If there is no real roots, the conclusion is $\wp_R(a) > 0$ for all $a > a^\natural$. If there are two real roots, their half-sum $-[\![v]\!]/4\tau_L$ is also the critical point of $\wp_R$, where a minimum occurs. Assume that

$$a^\natural < -\frac{[\![v]\!]}{4\tau_L}. \tag{164}$$

This implies

$$4\tau_L a^\natural + [\![v]\!] < 0 < [\![v]\!] + 2\overline{\tau}a^\natural, \tag{165}$$

from which it follows that $4\tau_L < 2\overline{\tau}$, whence $3\tau_R < \tau_L$. This contradicts (159), and establishes that a lies on the right branch

$$a^\natural > -\frac{[\![v]\!]}{4\tau_L} \tag{166}$$

of the parabola. The other case $[\![\tau]\!] \ge 0$ is similar. □

If the ratio condition (159) is violated, counter-examples can be found so as to invalidate the statement about positivity. We see that the Whitham condition $a > a^\natural$ still has strong connections with the positivity of the intermediate densities, even though it is not geared as neatly as in the scalar model. Like the scalar equation, an averaging argument shows that the positivity of the intermediate specific volumes at each edge implies that of the specific volume at the center of the cells, if we use the first-order explicit scheme

$$\tau_j^{n+1} = \tau_j^n - \frac{\Delta t}{\Delta m}\left[v^*(\mathbf{w}_j^n, \mathbf{w}_{j+1}^n) - v^*(\mathbf{w}_{j-1}^n, \mathbf{w}_j^n)\right] \tag{167}$$

under the CFL condition $\frac{\Delta t}{\Delta m}\max\{a_{i-1/2}, a_{i+1/2}\} < \frac{1}{2}$. A direct control of positivity is possible via another lower-bound for a.

Theorem A.4. *A sufficient condition to secure $\tau_L^* > 0$ and $\tau_R^* > 0$ is that*

$$a > a^\sharp(\mathbf{w}_L, \mathbf{w}_R) = \frac{-(v_R - v_L) + \sqrt{(v_R - v_L)^2 + 8\min(\tau_L, \tau_R)|p_R - p_L|}}{4\min(\tau_L, \tau_R)}. \tag{168}$$

Proof. See Theorem 3.2. □

For $[\![v]\!] \geq 0$ or $[\![v]\!] < 0$ subject to (157)–(159), we actually do not to worry about $a > a^\sharp$. The one and only condition to be taken care of is $a > a^\natural$. Since p is convex, we can even work with the stronger requirement $a > a^\flat$, where

$$a^\flat(\tau_L, \tau_R) = \max\{\sqrt{-p_\tau(\tau_L)}, \sqrt{-p_\tau(\tau_R)}\}. \tag{169}$$

Things change, however, when instead of $p(\tau)$ we have $P(\tau, Y, v)$, as is formally the case of (12a)–(12b) in Lagrangian coordinates. In such a situation, $a^\natural$ may no longer be finite for $\tau_L = \tau_R$, and even if it is, there is no longer any clear inequality between $a^\natural$ and $a^\flat$. The advantage of $a^\flat$ over $a^\natural$ is that it remains well-defined, using the partial derivative with respect to τ. This is why we heuristically have to look for $a \geq \max\{a^\sharp, a^\flat\}$, as indicated in (83a).