Numer. Math. 77: 143–185 (1997)

Numerische Mathematik

Electronic Edition

# A roe-type Riemann solver for hyperbolic systems with relaxation based on time-dependent wave decomposition

**François Bereux**[1], **Lionel Sainsaulieu**[2]

[1] Centre de Mathématiques Appliquées, Ecole Polytechnique, F-91128 Palaiseau Cedex, France
[2] Renault, DR, 9-11, avenue du 18 juin 1940, F-92500 Rueil-Malmaison, France

Received October 7, 1994 / Revised version received September 27, 1995

**Summary.** The paper is devoted to the construction of a higher order Roe-type numerical scheme for the solution of hyperbolic systems with relaxation source terms. It is important for applications that the numerical scheme handles both stiff and non stiff source terms with the same accuracy and computational cost and that the relaxation variables are computed accurately in the stiff case. The method is based on the solution of a Riemann problem for a linear system with constant coefficients: a study of the behavior of the solutions of both the nonlinear and linearized problems as the relaxation time tends to zero enables to choose a convenient linearization such that the numerical scheme is consistent with both the hyperbolic system when the source terms are absent and the correct relaxation system when the relaxation time tends to zero. The method is applied to the study of the propagation of sound waves in a two-phase medium. The comparison between our numerical scheme, usual fractional step methods, and numerical simulation of the relaxation system shows the necessity of using the solutions of a fully coupled hyperbolic system with relaxation terms as the basis of a numerical scheme to obtain accurate solutions regardless of the stiffness.

*Mathematics Subject Classification (1991):* 35L99, 65C20, 76T05

## 1. Introduction

Hyperbolic systems with relaxation are used in the modeling of a variety of physical phenomena of great practical importance such as thermally non equilibrium fluid flows, non reacting two-phase fluid flows composed of solid particles suspended in gas, viscoelasticity, ... The relaxation terms are source terms whose

*Correspondence to*: L. Sainsaulieu

effect is the relaxation to zero of some algebraic quantity, namely the relaxation variables. For instance, in the case of two-phase fluid flows composed of solid particles in gas the relaxation terms model the drag force whose effect is the relaxation to zero of the relative velocity between the two phases. In the case of thermally non equilibrium fluid flows the relaxation term depend upon the different temperatures involved in the modeling: here the thermal equilibrium of the flow is characterized by a single temperature. A relaxation time $\tau$ may be introduced to characterize the stiffness of the relaxation. In the case of the two-phase fluid flows considered above the relaxation time is the drag time which is proportionnal to inverse of the the square of the radius of the particles. In the case of thermally non equilibrium flows the relaxation time depends on the heat exchanges.

When the relaxation time $\tau$ tends to zero the model may be simplified. We expect indeed that the for any position $\mathbf{x}$ and any time $t$ the state vector $\mathbf{u}_\tau(\mathbf{x}, t)$ where $\mathbf{u}_\tau$ is the solution of the model tends to an equilibrium, *i.e.,* a state such that the relaxation variables are zero. It is then possible to use less variables to describe the system. For instance, in the case of the two-phase fluid flows considered above, the velocities of both phases should be equal in the limit of small drag times and it suffices then to use a single velocity for the two phases instead of the two velocities involved in the initial model. In the same manner only one temperature may be used to describe a fluid flow in the limit of thermal equilibrium. Following Chen-Levermore-Liu (1994) the relaxation system that models the flow in the limit of small relaxation times may be deduced from the original model with relaxation by using a Chapman-Enskog expansion. Following Whitam (1974) and Liu (1987) the source terms are relaxation terms if Liu's subcharacteristic condition is satisfied: this condition requires that the characteristic velocities of the relaxation system are interlaced with those of the convection system extracted from the original hyperbolic system with relaxation. The examples given at the beginning of this section are hyperbolic systems with relaxation. Chen-Levermore and Liu (1994) give sufficient conditions that ensure that a system with source terms is an hyperbolic system with relaxation. In this paper we shall only consider systems with relaxation.

Although very efficient and accurate methods have been developped for both hyperbolic systems and systems of ordinary differential equations, many numerical schemes for hyperbolic systems with relaxation are unsatisfactory and the main difficulty arises from the need to handle very different relaxation times with the same scheme. For instance solid particles are usually added in rocket engines in order to damp the combustion instabilities. The particles burn inside the rocket so that the stiffness of the drag terms range from nonstiff to very stiff. On the other hand the computation of an initial value problem for an hyperbolic system with relaxation also involves a wide range of stiffness of the source terms: if the initial data is away from equilibrium, there is a boundary layer in time of order $\tau$ after which the solution is close to equilibrium. During a time interval of order $\tau$ the relaxation terms are thus stiff while they become nonstiff after a time of order $\tau$. The chalenge is thus to construct a numerical scheme that

may handle any stiffness and whose computational cost is of the same order as the cost of usual methods such as the Strang splitting for instance: in order that the computational cost of the method be of the same order as the cost of usual methods for hyperbolic systems of conservation laws, we want to chose the time step only on the CFL condition relative to the convection terms and the source terms should be underresolved in the stiff case.

The construction of numerical schemes for hyperbolic systems with relaxation has attracted a lot of attention in recent years. See for instance Pember (1993) *a*, Pember (1993) *b*, Jin (1995), Jin-Levermore (1995), Bereux (1995), Calfisch-Jin-Russo (1995). Following Pember (1993) *b* and Calfisch-Jin-Russo (1995) we want to construct a numerical scheme that satisfies the following conditions:

(a) The scheme is second order accurate in both space and time in the nonstiff regime ($\tau = O(1)$).

(b) The limit scheme obtained when $\tau$ tends to zero is a second order accurate upwind scheme for the relaxation system deduced from the original model when $\tau \to 0$.

Condition (b) ensures that no spurious solutions are observed: when $\tau$ tends to zero the numerical solution tends to a solution of the relaxation system. Next, conditions (a) and (b) should ensure that the scheme gives second order accurate approximations of the quantities involved in the relaxation system, independently of the relaxation time. But the original model involves both the quantities in the relaxation model and the relaxation variables. However nothing is told in conditions (a) and (b) about the relaxation variables when the relaxation time is not infinitely stiff. In fact, we expect that the relaxation variables are of order $\tau$ for small $\tau$. Furthermore the accurate computation of the relaxation variables is of primary importance for many practical applications so that we add to the two conditions (a) and (b) the following requirement:

(c) the relaxation variables should be accurately computed.

The organization of the paper is the following: we review in Sect. 2 several schemes that were introduced recently for the numerical simulation of hyperbolic systems with relaxation. A discussion of the three conditions above is included. Next we introduce in Sect. 3 a new numerical scheme and Sect. 4 is devoted to its extension at second order. This is a stagered numerical scheme which coincides with the Lax-Friedrichs scheme at first order and with the Nesshayu-Tadmor scheme at second order when the source terms are omitted. The basic step of the construction of this scheme is the solution of a Riemann problem for a linearization of the initial hyperbolic system with relaxation. The constant convection matrix is chosen so that when the source terms are omitted, the shock waves solution of the nonlinear problem are still solutions of the linearized problem. Next the constant matrix used for the linearization of the relaxation terms is chosen so that the limit scheme obtained when $\tau$ tends to zero is an upwind scheme for the relaxation system deduced from the original model. We test the accuracy of our scheme on a practical example in Sects. 5 and 6: Sect. 5 is devoted to the statement of the problem and we compare in Sect. 6 our scheme with two other different schemes, known for their good behavior in the stiff case

and the nonstiff case respectively. It turns that our scheme enables us to compute very accurately the relaxation variables both in the stiff and the nonstiff case (and this is the difficult quantity to compute).

## 2. Review of several schemes for systems with relaxation

The purpose of this section is a brief discussion of the conditions (a), (b) and (c) of Sect. 1. We consider an hyperbolic system with relaxation in the form

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = \frac{1}{\tau}\mathbf{R}(\mathbf{u}) \tag{2.1}$$

where the state vector $\mathbf{u}$ belongs to some given subset $\Omega$ of $\mathbb{R}^p$ and where the flux function $\mathbf{f}$ is such that for any $\mathbf{u} \in \Omega$, the matrix $\mathbf{f}'(\mathbf{u})$ can be diagonalized on $\mathbb{R}$: the first order system

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = 0 \tag{2.2}$$

is an hyperbolic system of conservation laws.

In the sense of Whitham (1977) and Liu (1992) the source terms are relaxation terms if there exists a constant $r \times p$ matrix $\mathbf{Q}$ with rank $r < p$ such that

$$\mathbf{QR}(\mathbf{u}) = \mathbf{0}, \quad \forall \mathbf{u} \in \Omega \tag{2.3}$$

and if for any given $\mathbf{u}^0 \in \Omega$, the differential equation

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{R}(\mathbf{u}(t)), \quad \mathbf{u}(0) = \mathbf{u}^0$$

defines a function $\mathbf{u} : \mathbb{R}_+ \to \Omega$ such that $\lim_{t\to+\infty} \mathbf{u}(t) = \mathbf{Q}\mathbf{u}^0$. On the other hand we assume that there exists a vector valued function $\mathscr{E} : \omega = \mathbf{Q}\Omega \subset \mathbb{R}^r \to \Omega$ such that

$$\mathbf{R}(\mathscr{E}(\mathbf{v})) = \mathbf{0}, \quad \mathbf{Q}\mathscr{E}(\mathbf{v}) = \mathbf{v}, \qquad \forall \mathbf{v} \in \omega. \tag{2.4}$$

The image $\mathscr{E}(\omega)$ of $\mathscr{E}$ is thus the equilibrium manifold $\mathscr{M}$ or the manifold of local equilibria for $\mathbf{R}$:

$$\mathscr{M} = \big\{\mathbf{u} \in \Omega,\ \mathbf{R}(\mathbf{u}) = 0\big\}. \tag{2.5}$$

We assume further that $\mathbf{Q} : \mathscr{M} \to \omega$ defines a bijection. Then, if $\mathbf{u} \in \Omega$ is such that $\mathbf{R}(\mathbf{u}) = \mathbf{0}$, we have $\mathbf{u} = \mathscr{E}\big(\mathbf{Q}\mathbf{u}\big)$. (Indeed, $\mathbf{Q}\big(\mathbf{u} - \mathscr{E}(\mathbf{Q}\mathbf{u})\big) = \mathbf{0}$ and both $\mathbf{u}$ and $\mathscr{E}(\mathbf{Q}\mathbf{u})$ belong to $\mathscr{M}$ by assumption.)

In Chen-Levermore-Liu (1994) the authors study the behavior of a solution of (2.1) when $\tau$ is small: assume that $\mathbf{u}$ is a solution of (2.1). Then, since $\mathbf{Q}$ is a constant matrix and $\mathbf{QR}(\mathbf{u}) = 0$, $\forall \mathbf{u} \in \Omega$, we obtain:

$$\frac{\partial \mathbf{Q}\mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{Q}\mathbf{f}(\mathbf{u}) = 0.$$

When $\tau$ is small we expect that the solution $\mathbf{u}$ is close to the euiqlibrium manifold $\mathscr{M}$ and more precisely,

$$\mathbf{u} = \mathscr{E}(\mathbf{v}) + O(\tau)$$

where we have set $\mathbf{v} = \mathbf{Qu}$. At zero-th order in $\tau$ we deduce that $\mathbf{v}$ is a solution of the following system:

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial}{\partial x}\mathbf{g}(\mathbf{v}) = 0 \tag{2.6}$$

where the flux function $\mathbf{g}$ is given by

$$\mathbf{g}(\mathbf{v}) = \mathbf{Qf}\big(\mathscr{E}(\mathbf{v})\big). \tag{2.7}$$

Following Chen-Levermore-Liu (1994) one can proceed further to the first order expansion: the function $\mathbf{v}$ is then formally proved to be a solution of the following second order system:

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial}{\partial x}\mathbf{g}(\mathbf{v}) - \tau \frac{\partial}{\partial x}\left(\mathbf{D}(\mathbf{v})\frac{\partial \mathbf{v}}{\partial x}\right) = 0 \tag{2.8}$$

where the matrix valued function $\mathbf{D}$ is independent of $\tau$ and may be computed in function of $\mathbf{f}$, $\mathbf{R}$, $\mathscr{E}$ and $\mathbf{Q}$. Furthermore the relaxation variables $\mathbf{R}(\mathbf{u})$ may be evaluated in function of the solution $\mathbf{v}$ of (2.8).

It is proved in Chen-Levermore-Liu that if the continuous solutions of (2.1) satisfy an additional conservation in the form

$$\frac{\partial S}{\partial t} + \frac{\partial F}{\partial x} \leq 0$$

where $S : \Omega \to \mathbb{R}$ is a strictly convex function, the system (2.7) is hyperbolic and the eigenvalues of the matrix $\mathbf{g}'(\mathbf{v})$ are interlaced with those of $\mathbf{f}'(\mathscr{E}(\mathbf{v}))$, $\forall \mathbf{v} \in \omega$. Furthermore the system (2.8) is well posed.

The Chapman-Enskog expansion as described in Chen-Levermore-Liu (1994) is thus a very powerfull tool for the investigation of the solutions of system (2.1) in the limit of small relaxation times. What we expect from a numerical scheme is that it gives a good prediction of both the conserved quantities $\mathbf{Qu}$ and the relaxation variables $\mathbf{R}(\mathbf{u})$ when $\tau$ is small. To this extent a simple way to obtain a numerical scheme for system (2.1) in the limit of small relaxation times is the discretization of system (2.8) by using a fractionnal step method: we introduce the spatial grid points $x_j$, $j \in \mathbb{Z}$ with uniform mesh spacing $\Delta x = x_{j+1} - x_j$ for all $j$. The time levels $t_n$, $n = 0, 1, \ldots$ are also spaced uniformly with time step $\Delta t = t_{n+1} - t_n$. System (2.8) does not contain stiff terms any longer and Strang's fractionnal step method is convenient:

$$\begin{aligned} \mathbf{v}_j^{n+1/3} = &\mathbf{v}_j^n + \frac{\tau \Delta t}{2\Delta x^2}\Big(\mathbf{D}(\mathbf{v}_{j+1/2}^n)(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) \\ &- \mathbf{D}(\mathbf{v}_{j-1/2}^n)(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n)\Big) \end{aligned} \tag{2.9.$i$}$$

$$\mathbf{v}_j^{n+2/3} = \mathbf{v}_j^{n+1/3} - \frac{\Delta t}{\Delta x}\left(\phi_{j+1/2}\left(\mathbf{v}^{n+1/3}\right) - \phi_{j-1/2}\left(\mathbf{v}^{n+1/3}\right)\right) \tag{2.9.ii}$$

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^{n+2/3} + \frac{\tau\Delta t}{2\Delta x^2}\left(\mathbf{D}(\mathbf{v}_{j+1/2}^{n+2/3})(\mathbf{v}_{j+1}^{n+2/3} - \mathbf{v}_j^{n+2/3}) - \mathbf{D}(\mathbf{v}_{j-1/2}^{n+2/3})(\mathbf{v}_j^{n+2/3} - \mathbf{v}_{j-1}^{n+2/3})\right) \tag{2.9.iii}$$

where we have set

$$\mathbf{v}_{j+1/2} = \frac{\mathbf{v}_j + \mathbf{v}_{j+1}}{2}.$$

In the step (2.9.ii) the quantity $\phi_{j+1/2}$ is a numerical flux function taken in $j+1/2$, second order accurate in space and time, and consistant with (2.7). The sequence $(\mathbf{v}_j^{n+1/3})_{j\in\mathbb{Z}}$ is written $\mathbf{v}^{n+1/3}$ for shortness. The scheme (2.9) gives a second order accurate approximation of the solutions of (2.8). The implementation of the scheme (2.9) is very simple and gives excellent results in the stiff case: Bereux (1995) compares different methods for the computation of accoustic waves in a two-phase medium composed of solid particles suspended in gas. It is possible to derive an analytic expression of the sound velocity in the two-phase medium and of the attenuation coefficient of the sound waves (see Culick (1981)). He shows that when the problem becomes stiff, *i.e.*, when the time period of the sound wave is large in comparison with the relaxation time $\tau$, the relative velocity between the two phases, which is here the relaxation variable, is very accurately computed from the numerical solution $\mathbf{v}$ of (2.9) by applying the formulae in Chen-Levermore-Liu (1994). Nevertheless the matrices $\mathbf{f}'(\mathscr{E}(\mathbf{v}))$ and $\mathbf{g}'(\mathbf{v})$ have indeed different eigenvalues and one should not replace system (2.1) by (2.8) in the nonstiff case. Obviously the numerical scheme (2.9) is not uniformly accurate but this scheme will provide us with a reference solution when the relaxation time $\tau$ is very small.

The most commonly used method for the numerical solution of system (2.1) is probably Strang's splitting:

$$\mathbf{u}_j^{n+1/3} = \mathbf{u}_j^n + \frac{\Delta t}{2\tau}\mathbf{R}\left(\mathbf{u}_j^{n+1/3}\right) \tag{2.10.i}$$

$$\mathbf{u}_j^{n+2/3} = \mathbf{u}_j^{n+1/3} - \frac{\Delta t}{\Delta x}\left(\psi_{j+1/2}\left(\mathbf{u}^{n+1/3}\right) - \psi_{j-1/2}\left(\mathbf{u}^{n+1/3}\right)\right) \tag{2.10.ii}$$

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^{n+2/3} + \frac{\Delta t}{2\tau}\mathbf{R}\left(\mathbf{u}_j^{n+1}\right). \tag{2.10.iii}$$

Here $\psi_{j+1/2}$ is a numerical flux function, second order accurate in space and time, and consistant with (2.2). This gives a second order accurate scheme in the nonstiff case: see for instance Langseth-Tveito-Winther (1993). The first and third stages of the scheme are implicit in order to achieve stability independently of the relaxation time $\tau$.

Bereux (1995) compares the numerical solution given by (2.10) with the analytical solution proposed in Culick (1981) for the computation of the propagation of sound waves in a two-phase medium: in the nonstiff case Strang's splitting

(2.10) gives excellent results at a low computational cost. But this scheme proves unable to compute accurately the solutions of system (2.1) in the stiff case: numerical experiments prove that the attenuation coefficient of the sound wave given by the numerical method does not converge to zero as the relaxation time tends to zero, as it should do according to analytical calculations, but tends to a finite value. This positive value decreases extremely slowly when the time step and the space step tend to zero. Furthermore it is proved in Pember (1993) *a* that the scheme (2.10) is only first order accurate in the stiff case, even if a second order accurate numerical flux is used for the solution of (2.2): for $\Delta x$ fixed, if one diminishes the time step $\Delta t$ in order to better resolve the stiffness of system (2.1), the CFL based on the eigenvalues of $\mathbf{f}'(\mathbf{u})$ becomes small and the numerical diffusion brought by the step (2.10.ii) polutes the numerical solution.

Even fourth order splittings are unsatisfactory and Pember proves that one should use an unsplit approach in order to achieve second order accuracy (see Pember (1993) *a*). We would like to emphazise the fact that conditions (a) and (b) in Sect. 1 can not ensure that a numerical scheme gives satisfactory results. In fact condition (b) only deals with the conserved quantities $\mathbf{v}$ but not with the relaxation variables. This is made clear by the following

**Proposition 2.1.** *Let $\psi$ denote a numerical flux, second order accurate and consistant with the hyperbolic system (2.2). The Strang splitting (2.10) is second order accurate in the nonstiff regime and the limit scheme obtained when $\tau$ tends to zero is a second order accurate scheme for the relaxation system (2.7).*

*Proof of Proposition 2.1.* Consider the numerical scheme (2.10): in the limit $\tau \to 0$ the first and third steps write:

$$\mathbf{u}_j^{n+1/3} = \mathscr{E}\left(\mathbf{Q}\mathbf{u}_j^n\right), \quad \mathbf{u}_j^{n+1} = \mathscr{E}\left(\mathbf{Q}\mathbf{u}_j^{n+2/3}\right).$$

Setting $\mathbf{v}_j^n = \mathbf{Q}\mathbf{u}_j^n$, $j \in \mathbb{Z}$, $n \in \mathbb{N}$, we deduce that in the limit $\tau \to 0$,

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\Delta t}{\Delta x}\mathbf{Q}\left(\psi_{j+1/2}\left(\mathscr{E}(\mathbf{v}^n)\right) - \psi_{j-1/2}\left(\mathscr{E}(\mathbf{v}^n)\right)\right) \tag{2.11}$$

where the sequence $(\mathscr{E}(\mathbf{v}_j^n))_{j\in\mathbb{Z}}$ is written $\mathscr{E}(\mathbf{v}^n)$ for shortness.

Next let $\mathbf{v}$ denote a smooth solution of (2.6) and set $\mathbf{v}_j^n = \mathbf{v}(x_j, t_n)$, $j \in \mathbb{Z}$, $n \in \mathbb{N}$. Set next $\mathbf{u} = \mathscr{E}(\mathbf{v})$. The function $\mathbf{u}$ does not satisfy (2.1) but is a solution of the following system:

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = \left(\mathbf{1} - \mathscr{E}'(\mathbf{v})\mathbf{Q}\right)\mathbf{f}'\left(\mathscr{E}(\mathbf{v})\right)\frac{\partial \mathbf{u}}{\partial x}$$

and, since $\mathbf{Q}\mathscr{E}(\mathbf{v}) = \mathbf{v}$, $\forall \mathbf{v} \in \omega$ we deduce that

$$\mathbf{Q}\left(\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u})\right) = \mathbf{0}. \tag{2.12}$$

On the other hand, since $\psi$ is second order accurate we have the following estimate:

$$\frac{\mathbf{u}_j^{n+1}-\mathbf{u}_j^n}{\Delta t}+\frac{\psi_{j+1/2}(\mathbf{u}^n)-\psi_{j-1/2}(\mathbf{u}^n)}{\Delta x}=\frac{\partial \mathbf{u}}{\partial t}+\frac{\partial}{\partial x}\mathbf{f}(\mathbf{u})+O(\Delta t^2+\Delta x^2)$$

where $\mathbf{u}_j^n=\mathbf{u}(x_j,t_n)$, $j\in\mathbb{Z}$, $n\in\mathbb{N}$. Applying next on the left the projector $\mathbf{Q}$ to the last identity gives

$$\mathbf{Q}\left(\frac{\mathbf{u}_j^{n+1}-\mathbf{u}_j^n}{\Delta t}+\frac{\psi_{j+1/2}(\mathbf{u}^n)-\psi_{j-1/2}(\mathbf{u}^n)}{\Delta x}\right)$$

$$=\mathbf{Q}\left(\frac{\partial \mathbf{u}}{\partial t}+\frac{\partial}{\partial x}\mathbf{f}(\mathbf{u})\right)+O(\Delta t^2+\Delta x^2). \tag{2.13}$$

But $\mathbf{v}=\mathbf{Qu}$ by assumption and we deduce from (2.12)-(2.13) that

$$\frac{\mathbf{v}_j^{n+1}-\mathbf{v}_j^n}{\Delta t}+\mathbf{Q}\left(\frac{\psi_{j+1/2}\big(\mathscr{E}(\mathbf{v}^n)\big)-\psi_{j-1/2}\big(\mathscr{E}(\mathbf{v}^n)\big)}{\Delta x}\right)=O(\Delta t^2+\Delta x^2).$$

But in the limit of small relaxation times Strang's splitting takes the form (2.11) and the proof of Proposition 2.1 is complete. □

In fact, when Liu's subcharacteristic condition is satisfied, spurious solutions are not observed with a splitting method, unlike what happens in the case of the ZND detonation model for instance: this result is proved by Pember (Pember (1993) *a*) and Jin (Jin (1994)). But, as proved in Pember (1993) *a*, splitting methods are at most first order in the stiff case and the relaxation variables are miscomputed. Following Pember (1993) *b* one should thus use an unsplit approach: to our knowledge this author was the first one to introduce an unsplit approach for the solution of an hyperbolic system with relaxation that is uniformly accurate independently of the stiffness of the system. However, as Pember notices it, his method does not reduce to an upwind method for the relaxation system as the model system becomes increasingly stiff: small oscillations are observed in the numerical profiles. The oscillations arise from the use of the characteristic velocities of (2.2) in both the nonstiff and the stiff case while in the stiff case, the correct characteristic velocities are those of system (2.6). Pember's scheme writes

$$\mathbf{u}_j^{n+1}=\mathbf{u}_j^n-\frac{\Delta t}{\Delta x}\big(\psi_{j+1/2}^{n+1/2}-\psi_{j-1/2}^{n+1/2}\big)+\frac{\Delta t}{\tau}\mathbf{R}(\mathbf{u}_j^{n+1}) \tag{2.14}$$

where

$$\psi_{j+1/2}^{n+1/2}=\mathbf{f}\big(\mathbf{u}_{j+1/2}^{n+1/2}\big)$$

and where the state $\mathbf{u}_{j+1/2}^{n+1/2}$ is obtained by solving the Riemann problem for (2.2) between two states $\mathbf{u}_{j+1/2,l}^{n+1/2}$ and $\mathbf{u}_{j+1/2,r}^{n+1/2}$ whose computation is described in Pember (1993) *b*. The latter two states depend upon the relaxation $\mathbf{R}$ and the characteristic velocities of system (2.2). But the characteristic velocities of system (2.6) rather than (2.2) should be used in the stiff case in the Riemann solver step. Pember believes that this is the reason of the presence of the small oscillations in his numerical profiles.

## 3. A first order numerical scheme based on a linearized Riemann problem

Let us consider an hyperbolic system with relaxation in the form (2.1) and let $\mathbf{u}$ denote a solution of (2.1). We introduce the spatial grid points $x_j = j\Delta x$, $j \in \mathbb{Z}$ and the time levels $t_n = n\Delta t$, $n \in \mathbb{N}$. Our aim is to derive a numerical scheme that enables to compute an approximation $\mathbf{u}_j^n$ of the quantity $\frac{1}{\Delta x}\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} \mathbf{u}(x,t_n)\,\mathrm{d}x$. Following Pember (1993) *b* we wish to use an unsplit approach and, following the ideas of Godounov we wish to compute a generalized Riemann solver:

$$\frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{w}) = \frac{\mathbf{R}(\mathbf{w})}{\tau}, \quad \mathbf{w}(x,0) = \begin{cases} \mathbf{w}^L, & x < 0 \\ \mathbf{w}^R, & x > 0. \end{cases} \tag{3.1}$$

The solution of the Riemann problem (3.1) is denoted by $W(x,t,\mathbf{w}^L,\mathbf{w}^R)$ (note that this solution is not self-similar as in the case of hyperbolic systems with no source terms) and a stagerred consistent first order numerical scheme is given by the following expression:

$$\mathbf{u}_{j+1/2}^{n+1} = \frac{1}{\Delta x}\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} W\left(x,\Delta t,\mathbf{u}_j^n,\mathbf{u}_{j+1}^n\right)\mathrm{d}x. \tag{3.2}$$

This is obviously an unsplit approach and we expect that under an appropriate CFL like condition this scheme is stable. But the explicit solution of the Riemann problem (3.1) seems out of reach and, following Roe (1984) we replace the nonlinear problem (3.1) by a linear problem with constant coefficients:

$$\frac{\partial \mathbf{w}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{w}}{\partial x} = \frac{\mathbf{B}\mathbf{w}}{\tau}, \quad \mathbf{w}(x,0) = \begin{cases} \mathbf{w}^L, & x < 0 \\ \mathbf{w}^R, & x > 0. \end{cases} \tag{3.3}$$

The choice of the matrices $\mathbf{A}$ and $\mathbf{B}$ should of course depend upon the left and right states $\mathbf{w}^L$ and $\mathbf{w}^R$ in order that the linearization (3.3) is consistent with (3.1).

Before we precise the choice of the matrices $\mathbf{A}$ and $\mathbf{B}$ let us first write explicitly our numerical scheme. The scheme is initialized by setting

$$\mathbf{u}_j^0 = \frac{1}{\Delta x}\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} \mathbf{u}^0(x)\,\mathrm{d}x, \quad j \in \mathbb{Z} \tag{3.4}$$

where $\mathbf{u}^0$ is the initial data. We consider next the following familly of linear generalized Riemann problems with constant coefficients:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}_{j-1/2}^n\frac{\partial \mathbf{u}}{\partial x} = \frac{\mathbf{B}_{j-1/2}^n}{\tau}\mathbf{u} \tag{3.5}$$

and with initial data:

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_{j-1}^n & \text{if } x < (j-1/2)\Delta x \\ \mathbf{u}_j^n & \text{if } x > (j-1/2)\Delta x \end{cases} \tag{3.6}$$

A staggered approximate numerical scheme is then obtained by setting

$$\mathbf{u}_{j-1/2}^{n+1/2} = \frac{1}{\Delta x}\int_{(j-1)\Delta x}^{j\Delta x} \mathbf{u}\big(x, \Delta t/2\big)\,\mathrm{d}x \tag{3.7}$$

where $\mathbf{u}$ is the solution of (3.5)-(3.6). The explicit computation of the formula (3.7) relies on the following

**Lemma 3.1.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two given matrices. We assume that there exists a positive definite symmetric matrix* $\mathbf{S}$ *such that the matrix* $\mathbf{SA}$ *is symmetric. The matrix* $\mathbf{A}$ *may then be diagonalized on* $\mathbb{R}$ *and we denote by* $\lambda_k$, $1 \le k \le p$ *its eigenvalues. Let* $\mathbf{u}$ *denote the solution of the generalized Riemann problem*

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = \frac{\mathbf{B}}{\epsilon}\mathbf{u} \tag{3.8.$i$}$$

*with initial data*

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}^L & \text{if } x < 0 \\ \mathbf{u}^R & \text{if } x > 0. \end{cases} \tag{3.8.$ii$}$$

*Then, under the CFL like condition*

$$\max_{1\le k\le p} \frac{\Delta t|\lambda_k|}{\Delta_x} \le \frac{1}{2}, \tag{3.9}$$

*the function*

$$\mathbf{H}(t) = \frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} \mathbf{u}(x,t)\,dx \tag{3.10}$$

*is obtained for* $0 \le t \le \Delta t$ *by solving the following system of ordinary differential equations:*

$$\frac{d\mathbf{H}}{dt}(t) = \frac{1}{\epsilon}\mathbf{B}\mathbf{H}(t) - \frac{1}{\Delta x}\mathbf{A}\big(\mathbf{u}^R(t) - \mathbf{u}^L(t)\big), \quad \mathbf{H}(0) = \frac{\mathbf{u}^L + \mathbf{u}^R}{2} \tag{3.11.$i$}$$

*where the functions* $\mathbf{u}^R(t)$ *and* $\mathbf{u}^L(t)$ *are given by*

$$\begin{aligned} \frac{d\mathbf{u}^R}{dt}(t) &= \frac{\mathbf{B}}{\epsilon}\mathbf{u}^R(t), \quad \mathbf{u}^R(0) = \mathbf{u}^R \\ \frac{d\mathbf{u}^L}{dt}(t) &= \frac{\mathbf{B}}{\epsilon}\mathbf{u}^L(t), \quad \mathbf{u}^L(0) = \mathbf{u}^L. \end{aligned} \tag{3.11.$ii$}$$

*Proof.* Before proving Lemma 3.1 we need to prove some properties of the solution of (3.8): we assume that the matrix $\mathbf{A}$ satisfies the same properties as in Lemma 3.1. Then,

**Proposition 3.2.** *Let be given two states* $\mathbf{u}^L$ *and* $\mathbf{u}^R$. *For any given positive time* $T$, *there exists a unique solution* $\mathbf{u} \in C^1\big([0,T], BV(\mathbb{R})\big)$ *of system (3.8). Here* $BV(\mathbb{R})$ *denotes the set of functions with bounded variations.*

Next proposition precises the propagation of the discontinuities in the solution $\mathbf{u}$ of (3.8): decompose the jump of the initial data on the right eigenbasis $\{\mathbf{r}_k\}_{1\le k\le p}$ of matrix $\mathbf{A}$:

$$\mathbf{u}^R - \mathbf{u}^L = \sum_{k=1}^{p} a_k^0 \mathbf{r}_k . \tag{3.12}$$

**Proposition 3.3.** *Let* $\mathbf{u}$ *denote the unique solution of (3.8). The function* $\mathbf{z}$ *defined by*

$$\mathbf{z}(x,t) = \mathbf{u}(x,t) - \sum_{k=1}^{p} a_k(t)\mathbf{r}_k H(x-\lambda_k t) - \exp(\mathbf{B}t)\,\mathbf{u}^L \tag{3.13}$$

*where* $H$ *denotes Heavyside's function and where the functions* $a_k$, $1 \le k \le p$ *are given by*

$$a_k(t) = \exp(\mathbf{l}_k \cdot \mathbf{B}\mathbf{r}_k\, t)\, a_k^0 \tag{3.14}$$

*is continuous. Furthermore its derivative* $\frac{\partial \mathbf{z}}{\partial x}$ *belongs to the set* $L^\infty\big([0,T], BV(\mathbb{R})\big)$ *for any positive time* $T$.

Finally the solution of (3.8) remains constant outside a bounded interval:

**Proposition 3.4.** *Let* $\mathbf{u}$ *denote the unique solution of (3.8). Then*

$$\mathbf{u}(x,t) = \exp(\mathbf{B}t)\,\mathbf{u}^L, \quad \text{if } x < \lambda_1 t$$
$$\mathbf{u}(x,t) = \exp(\mathbf{B}t)\,\mathbf{u}^R, \quad \text{if } x > \lambda_p t .$$

The proofs of Propositions 3.2, 3.3 and, 3.4 are given in Appendix A. We can now proceed to the proof of Lemma 3.1: according to Proposition 3.3 the solution $\mathbf{u}$ of system (3.8) is composed with $p$ shock waves with respective velocities $\lambda_k$, $1 \le k \le p$ separated by functions whose derivatives have bounded variation. We may thus compute for $1 \le k \le p$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\lambda_k t}^{\lambda_{k+1}t} \mathbf{u}(x,t)\,\mathrm{d}x = \int_{\lambda_k t}^{\lambda_{k+1}t} \frac{\partial \mathbf{u}}{\partial t}\,\mathrm{d}x + \lambda_{k+1}\mathbf{u}\big((\lambda_{k+1}t)^-, t\big) - \lambda_k \mathbf{u}\big((\lambda_k t)^+, t\big)$$

where $\mathbf{u}\big((\lambda_k t)^\pm, t\big)$ denotes the right (resp. left) value of function $\mathbf{u}$ on the discontinuity line $x = \lambda_k t$. But $\mathbf{u}$ satisfies (3.8.i) and we deduce:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\lambda_k t}^{\lambda_{k+1}t} \mathbf{u}(x,t)\,\mathrm{d}x = \frac{1}{\epsilon}\mathbf{B}\int_{\lambda_k t}^{\lambda_{k+1}t} \mathbf{u}(x,t)\,\mathrm{d}x - \big(\mathbf{A} - \lambda_{k+1}\mathbf{1}\big)\mathbf{u}\big((\lambda_{k+1}t)^-, t\big) + \big(\mathbf{A} - \lambda_k \mathbf{1}\big)\mathbf{u}\big((\lambda_k t)^+, t\big).$$

In the same manner,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\Delta x/2}^{\lambda_1 t}\mathbf{u}(x,t)\,\mathrm{d}x = \frac{1}{\epsilon}\mathbf{B}\int_{-\Delta x/2}^{\lambda_1 t}\mathbf{u}(x,t)\,\mathrm{d}x - \big(\mathbf{A}-\lambda_1\mathbf{1}\big)\mathbf{u}\big((\lambda_1 t)^-,t\big) + \mathbf{A}\mathbf{u}\big(-\Delta x/2,t\big)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\lambda_p t}^{\Delta x/2}\mathbf{u}(x,t)\,\mathrm{d}x = \frac{1}{\epsilon}\mathbf{B}\int_{\lambda_p t}^{\Delta x/2}\mathbf{u}(x,t)\,\mathrm{d}x - \mathbf{A}\mathbf{u}\big(\Delta x/2,t\big) + \big(\mathbf{A}-\lambda_p\mathbf{1}\big)\mathbf{u}\big((\lambda_p t)^+,t\big).$$

Summing the $p+2$ above differential equations gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\Delta x/2}^{\Delta x/2}\mathbf{u}(x,t)\,\mathrm{d}x = \frac{1}{\epsilon}\mathbf{B}\int_{-\Delta x/2}^{\Delta x/2}\mathbf{u}(x,t)\,\mathrm{d}x - \mathbf{A}\big(\mathbf{u}(\Delta x/2,t)-\mathbf{u}(-\Delta x/2),t)\big) + \sum_{k=1}^{p}\big(\mathbf{A}-\lambda_k\mathbf{1}\big)\big(\mathbf{u}\big((\lambda_k t)^+,t\big)-\mathbf{u}\big((\lambda_k t)^-,t\big)\big). \tag{3.15}$$

But we have by Proposition 3.3 that for $1\le k\le p$,

$$\big(\mathbf{A}-\lambda_k\mathbf{1}\big)\big(\mathbf{u}\big((\lambda_k t)^+,t\big)-\mathbf{u}\big((\lambda_k t)^-,t\big)\big)=\mathbf{0}.$$

On the other hand, under the CFL condition (3.9), Proposition 3.4 gives

$$\mathbf{u}(\Delta x/2,t)=\mathbf{u}^R(t),\qquad \mathbf{u}(-\Delta x/2,t)=\mathbf{u}^L(t)$$

where the functions $t\to\mathbf{u}^L(t),\ \mathbf{u}^R(t)$ are given by (3.11.ii). Then (3.11.i) follows from (3.12). □

We precise now the choice of the matrices $\mathbf{A}^n_{j-1/2}$ and $\mathbf{B}^n_{j-1/2}$ in (3.5). This choice is such that the linearized problem (3.3) is consistent with the non linear problem (3.1). First, when the source terms are absent ($\mathbf{R}=\mathbf{0}$), system (2.1) is an hyperbolic system of conservation laws and, following Roe (1984), the matrix $\mathbf{A}$ in (3.3) is taken as a Roe linearization of the flux function $\mathbf{f}$ between the two states $\mathbf{u}^L$ and $\mathbf{u}^R$: we assume that there exists a matrix valued function $\mathbf{A}:(\mathbf{u}_1,\mathbf{u}_2)\in\Omega^2\to\mathbf{A}(\mathbf{u}_1,\mathbf{u}_2)$ with the following properties:

$$\mathbf{A}(\mathbf{u}_1,\mathbf{u}_2)\big(\mathbf{u}_2-\mathbf{u}_1\big)=\mathbf{f}(\mathbf{u}_2)-\mathbf{f}(\mathbf{u}_1),\quad \forall\mathbf{u}_1,\mathbf{u}_2\in\Omega$$
$$\mathbf{A}(\mathbf{u},\mathbf{u})=\mathbf{f}'(\mathbf{u}),\quad \forall\mathbf{u}\in\Omega$$
$$\mathbf{A}(\mathbf{u}_1,\mathbf{u}_2)\text{ can be diagonalized on }\mathbb{R}\ ,\ \forall\mathbf{u}_1,\mathbf{u}_2\in\Omega.$$

The matrix $\mathbf{A}^n_{j-1/2}$ in (3.5) is then

$$\mathbf{A}^n_{j-1/2}=\mathbf{A}\big(\mathbf{u}^n_{j-1},\mathbf{u}^n_j\big). \tag{3.16}$$

When $\tau$ tends to zero the system (2.1) formally tends to the relaxation system (2.6). On the other hand, we expect that the linear system with constant coefficients

$$\frac{\partial\mathbf{u}}{\partial t}+\mathbf{A}\frac{\partial\mathbf{u}}{\partial x}=\frac{\mathbf{B}\mathbf{u}}{\tau} \tag{3.17}$$

converges when $\tau$ tends to zero to a system in the form

$$\frac{\partial \mathbf{v}}{\partial t} + \widetilde{\mathbf{A}}\frac{\partial \mathbf{v}}{\partial x} = 0 \tag{3.18}$$

under some appropriate assumptions on the matrices **A** and **B**. In order to ensure that the linearized problem (3.3) is consistent with (3.1) we require that the matrix $\widetilde{\mathbf{A}}$ is a Roe linearization of the flux function **g** in (2.6): we have then the following commutative diagram:

$$\begin{array}{ccc} \dfrac{\partial \mathbf{u}}{\partial t} + \dfrac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = \dfrac{\mathbf{R}(\mathbf{u})}{\tau} & \xrightarrow{\tau\to 0} & \dfrac{\partial \mathbf{v}}{\partial t} + \dfrac{\partial}{\partial x}\mathbf{g}(\mathbf{v}) = 0 \\ \text{linearization}\downarrow & & \text{linearization}\downarrow \\ \dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\dfrac{\partial \mathbf{u}}{\partial x} = \dfrac{\mathbf{B}\mathbf{u}}{\tau} & \xrightarrow{\tau\to 0} & \dfrac{\partial \mathbf{v}}{\partial t} + \widetilde{\mathbf{A}}\dfrac{\partial \mathbf{v}}{\partial x} = 0 \end{array} \tag{3.19}$$

Assume that

[H1] There exists a $r \times p$ matrix **Q** and a $p \times r$ matrix **E** such that **QB** = **0**, **BE** = **0** and **QE** = $\mathbf{1}_r$. (The matrices **Q** and **E** are respectively left and right pseudo-inverses of matrix $\mathbf{1}_p - \mathbf{B}$.)

[H2] The $r \times r$ matrix $\widetilde{\mathbf{A}}$ = **QAE** has $r$ distinct eigenvalues $\widetilde{\lambda}_k$, $1 \le k \le r$. We denote by $\widetilde{\mathbf{r}}_k$, $1 \le k \le r$ (resp. $\widetilde{\mathbf{l}}_k$, $1 \le k \le r$) the associated right (resp. left) eigenvectors.

[H3] The matrix **B** has $p-r$ negative eigenvalues $\eta_k$, $r+1 \le k \le p$. The associated eigenvectors are denoted by $\mathbf{q}_k$, $r+1 \le k \le p$.

Then, the behavior of the solutions of the linear system with constant coefficients (3.3) is given by the following

**Proposition 3.5.** *Assume that assumptions [H1] to [H3] hold true. Let* **u** *be a solution of system (3.3) and let* $\widetilde{\mathbf{u}}$ *be the solution of the following system*

$$\frac{\partial \widetilde{\mathbf{u}}}{\partial t} + \widetilde{\mathbf{A}}\frac{\partial \widetilde{\mathbf{u}}}{\partial x} = 0 \tag{3.20}$$

*with the following initial data:*

$$\widetilde{\mathbf{u}}(x,0) = \begin{cases} \mathbf{Q}\mathbf{u}^L, & \text{if } x < 0 \\ \mathbf{Q}\mathbf{u}^R, & \text{if } x > 0. \end{cases} \tag{3.21}$$

*Then, for $t > 0$ given, we can find a positive number $\beta_0$ small enough such that*

$$\left|\mathbf{Q}\widehat{\mathbf{u}}(\xi,t) - \widehat{\widetilde{\mathbf{u}}}(\xi,t)\right| \le C(\beta_0)\big(\epsilon|\xi| + \epsilon\xi^2\big), \quad \text{for } \epsilon|\xi| \le \beta_0 \tag{3.22}$$

*where* $\widehat{\mathbf{u}}$ *and* $\widehat{\widetilde{\mathbf{u}}}$ *denote the Fourier transform of the functions* **u** *and* $\widetilde{\mathbf{u}}$ *respectively.*

The proof of Proposition 3.5 is given in appendix B.

In order that we obtain the commutative diagram (3.19) we need thus that the matrix $\widetilde{\mathbf{A}}$ is a Roe linearization of the flux function **g**:

**Proposition 3.6.** *Assume that there exists a constant matrix* $\mathbf{Q}$ *with rank* $p$ *and a vector valued function* $\mathscr{E} : \omega \to \Omega$ *that satisfy (2.4). Next assume that there exists a Roe linearization* $\mathbf{A} : (\mathbf{u}^L, \mathbf{u}^R) \in \Omega^2 \to \mathbf{A}(\mathbf{u}^L, \mathbf{u}^R)$ *of the flux function* $\mathbf{f}$ *and a matrix valued function* $\mathbf{E}$ *with*

$$\mathbf{E}(\mathbf{v}^L, \mathbf{v}^R)(\mathbf{v}^R - \mathbf{v}^L) = \mathscr{E}(\mathbf{v}^R) - \mathscr{E}(\mathbf{v}^L), \quad \forall \mathbf{v}^L, \mathbf{v}^R \in \omega. \tag{3.23}$$

*Assume further that there exists a parameter vector* $\varphi(\mathbf{v}^L, \mathbf{v}^R)$ *such that*

$$\mathbf{E}(\mathbf{v}^L, \mathbf{v}^R) = \mathscr{E}'\big(\varphi(\mathbf{v}^L, \mathbf{v}^R)\big), \quad \forall \mathbf{v}^L, \mathbf{v}^R \in \omega. \tag{3.24}$$

*Set finally*

$$\mathbf{B}(\mathbf{u}^L, \mathbf{u}^R) = \mathbf{R}'\Big(\mathscr{E}\Big(\varphi\big(\mathbf{Q}\mathbf{u}^L, \mathbf{Q}\mathbf{u}^R\big)\Big)\Big). \tag{3.25}$$

*Then, the matrices* $\mathbf{Q}$, $\mathbf{B}$ *and* $\mathbf{E}$ *satisfy the assumption [H1] and the matrix valued function* $\widetilde{\mathbf{A}}$ *defined by*

$$\widetilde{\mathbf{A}}(\mathbf{v}^L, \mathbf{v}^R) = \mathbf{Q}\mathbf{A}\big(\mathscr{E}(\mathbf{v}^L), \mathscr{E}(\mathbf{v}^R)\big)\mathbf{E}\big(\mathbf{v}^L, \mathbf{v}^R\big) \tag{3.26}$$

*is a linearization of the flux function* $\mathbf{g}$*:*

$$\mathbf{g}(\mathbf{v}^R) - \mathbf{g}(\mathbf{v}^L) = \widetilde{\mathbf{A}}(\mathbf{v}^L, \mathbf{v}^R)(\mathbf{v}^R - \mathbf{v}^L).$$

*Proof of Proposition 3.6.* Since $\mathbf{QR}(\mathbf{u}) = \mathbf{0}$, $\forall \mathbf{u} \in \Omega$, we have by definition of $\mathbf{B}$ that $\mathbf{QB} = \mathbf{0}$. Next, $\mathbf{BE} = \mathbf{0}$ follows by derivation from the identity $\mathbf{R}\big(\mathscr{E}(\mathbf{v})\big) = \mathbf{0}$, $\forall \mathbf{v} \in \omega$. We obtain in the same manner $\mathbf{QE} = \mathbf{1}_r$ from the identity $\mathbf{Q}\mathscr{E}'(\mathbf{v}) = \mathbf{1}_r$, $\forall \mathbf{v} \in \omega$, and the assumption [H1] is satisfied.

Define next the matrix valued function $\widetilde{\mathbf{A}}$ by (3.26): by the definition (2.7) of the flux function $\mathbf{g}$ we compute:

$$\begin{aligned}\mathbf{g}(\mathbf{v}^R) - \mathbf{g}(\mathbf{v}^L) &= \mathbf{Q}\big(\mathbf{f}\big(\mathscr{E}(\mathbf{v}^R)\big) - \mathbf{f}\big(\mathscr{E}(\mathbf{v}^L)\big)\big)\\ &= \mathbf{Q}\mathbf{A}\big(\mathscr{E}(\mathbf{v}^L), \mathscr{E}(\mathbf{v}^R)\big)\big(\mathscr{E}(\mathbf{v}^R) - \mathscr{E}(\mathbf{v}^L)\big)\\ &= \mathbf{Q}\mathbf{A}\big(\mathscr{E}(\mathbf{v}^L), \mathscr{E}(\mathbf{v}^R)\big)\mathbf{E}\big(\mathbf{v}^L, \mathbf{v}^R\big)\big(\mathbf{v}^R - \mathbf{v}^L\big)\end{aligned}$$

since $\mathbf{A}$ and $\mathbf{E}$ are respectively Roe linearizations of functions $\mathbf{f}$ and $\mathscr{E}$. The proof of Proposition 3.6 is complete. □

We can finally write our numerical scheme: when a picewise constant function $\mathbf{u}^n$ is given, we first solve for $j \in \mathbb{Z}$ the following systems of ODEs:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{u}^L(t)}{\mathrm{d}t} &= \frac{1}{\epsilon}\mathbf{B}^n_{j-1/2}\mathbf{u}^L(t), \quad \mathbf{u}^L(0) = \mathbf{u}^n_j\\ \frac{\mathrm{d}\mathbf{u}^R(t)}{\mathrm{d}t} &= \frac{1}{\epsilon}\mathbf{B}^n_{j-1/2}\mathbf{u}^R(t), \quad \mathbf{u}^R(0) = \mathbf{u}^n_{j-1}\\ \frac{\mathrm{d}\mathbf{H}^n_{j-1/2}}{\mathrm{d}t}(t) &= \mathbf{B}^n_{j-/2}\mathbf{H}^n_{j-1/2}(t) - \mathbf{A}^n_{j-1/2}\Big(\frac{\mathbf{u}^L(t) - \mathbf{u}^R(t)}{\Delta x}\Big),\\ \mathbf{H}^n_{j-1/2}(0) &= \frac{\mathbf{u}^n_{j-1} + \mathbf{u}^n_j}{2}\end{aligned} \tag{3.27}$$

where the matrices $\mathbf{A}^n_{j-1/2}$ and $\mathbf{B}^n_{j-1/2}$ are respectively

$$\mathbf{A}^n_{j-1/2} = \mathbf{A}\big(\mathbf{u}^n_{j-1}, \mathbf{u}^n_j\big), \quad \mathbf{B}^n_{j-1/2} = \mathbf{B}\big(\mathbf{u}^n_{j-1}, \mathbf{u}^n_j\big)$$

for **A**, a Roe linearization of **f** and **B** given by (3.25). We set next

$$\mathbf{u}^{n+1/2}_{j-1/2} = \mathbf{H}^n_{j-1/2}\big(\Delta t/2\big). \tag{3.28}$$

Applying again the above staggered scheme gives an approximation of the solution **u** of (2.1) at time $t_{n+1}$. We expect that this scheme is stable under the CFL like condition

$$|\overline{\lambda}|\frac{\Delta x}{\Delta t} \leq 1 \tag{3.29}$$

where $\overline{\lambda}$ denotes the largest propagation velocity involved in the solution of the different linear Riemann problems with constant coefficients.

## 4. A higher order numerical scheme

Following Van Leer (1977), in order to obtain a second order version of our numerical scheme, we approximate a solution **u** of (2.1) by a piecewise linear function instead of a piecewise constant function:

$$\mathbf{u}^n(x) = \mathbf{u}^n_j + \mathbf{p}^n_j\big(x - j\Delta x\big), \quad \text{for } \left(j - \frac{1}{2}\right)\Delta x \leq x \leq \left(j + \frac{1}{2}\right)\Delta x. \tag{4.1}$$

The numerical scheme runs as follows: given a piecewise constant approximation of a solution **u** of system (2.1) we first compute the following slopes:

$$\mathbf{s}^n_{j+1/2} = \frac{\mathbf{u}^n_{j+1} - \mathbf{u}^n_j}{\Delta x} \tag{4.2}$$

which are next corrected with the *min-mod* limiter:

$$\mathbf{p}^n_j = \textit{min-mod}\big(\mathbf{s}^n_{j-1/2}, \mathbf{s}^n_{j+1/2}\big). \tag{4.3}$$

We thus obtain a piecewise linear approximation of function **u** whose total variation is bounded by that of **u**. Setting next

$$\mathbf{u}^n_{j-1/2,+} = \mathbf{u}^n_j - \frac{\Delta x}{2}\mathbf{p}^n_j, \quad \mathbf{u}^n_{j+1/2,-} = \mathbf{u}^n_j + \frac{\Delta x}{2}\mathbf{p}^n_j \tag{4.4}$$

and

$$\mathbf{A}^n_{j-1/2} = \mathbf{A}\big(\mathbf{u}^n_{j-1/2,-}, \mathbf{u}^n_{j-1/2,+}\big), \quad \mathbf{B}^n_{j-1/2} = \mathbf{B}\big(\mathbf{u}^n_{j-1/2,-}, \mathbf{u}^n_{j-1/2,+}\big) \tag{4.5}$$

where **A** is a Roe linearization of the flux function **f** and **B** is chosen as in Sect. 4, we replace the solution of the Riemann problem (3.5)-(3.6) by the solution of the following generalized Riemann problem as in Van Leer (1977):

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}^n_{j-1/2}\frac{\partial \mathbf{u}}{\partial x} = \frac{\mathbf{B}^n_{j-1/2}}{\epsilon}\mathbf{u} \tag{4.6}$$

with initial data:

(4.7)

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_{j-1} + \mathbf{p}_{j-1}\big(x - (j-1)\Delta x\big) = \\ \qquad = \mathbf{u}^n_{j-1/2,-} + \mathbf{p}_{j-1}\left(x - \left(j - \dfrac{1}{2}\right)\right) \quad \text{if } x < (j-1/2)\Delta x \\ \mathbf{u}_j + \mathbf{p}_j\big(x - (j-1)\Delta x\big) = \\ \qquad = \mathbf{u}^n_{j-1/2,+} + \mathbf{p}_j\left(x - \left(j - \dfrac{1}{2}\right)\right) \quad \text{if } x > (j-1/2)\Delta x. \end{cases}$$

We are thus left with the computation of function $\mathbf{H}$ defined by

$$\mathbf{H}(t) = \frac{1}{\Delta x}\int_{\frac{-\Delta x}{2}}^{\frac{\Delta x}{2}} \mathbf{u}(x,t)\,\mathrm{d}x,$$

where $\mathbf{u}$ is a solution of (2.1) with the following initial data:

$$\mathbf{u}^0(x) = \begin{cases} \mathbf{u}^L + \mathbf{p}^L x & \text{if } x < 0 \\ \mathbf{u}^R + \mathbf{p}^R x & \text{if } x > 0. \end{cases}$$

We obtain after a straightforward computation similar to the proof of Lemma 3.1 that

$$\frac{\mathrm{d}\mathbf{H}}{\mathrm{d}t}(t) = \mathbf{B}\mathbf{H}(t) - \frac{\mathbf{A}}{\Delta x}\big(\mathbf{u}(\Delta x/2,t) - \mathbf{u}(-\Delta x/2,t)\big).$$

Unfortunately the explicit value of functions $t \to \mathbf{u}\big(-\Delta x/2,t\big)$ is here much more difficult to compute than in Sect. 3 where it remained constant in space under the CFL condition (3.9). Noticing that the function $\mathbf{v}(x,t) = \mathbf{v}(t) + \mathbf{q}(t)x$ is a solution of (2.1) if

$$\mathbf{q}'(t) = \mathbf{B}\mathbf{q}(t), \quad \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\mathbf{A}\mathbf{q}(t) + \mathbf{B}\mathbf{v}(t), \tag{4.8}$$

we may compute explicitly $\mathbf{u}\big(-\Delta x/2,t\big)$ by solving the coupled differential system (4.8) with initial data $\mathbf{u}^L + \mathbf{p}^L x$ provided that the wave pattern produced by the initial discontinuity in $x = 0$ has not reached yet the position $-\Delta x/2$ at time $t$. (This is true, at least at the first order in time: see Ben Artzi (1989)).

This numerical scheme requires the solution of several differential systems which may lead to complicated computations in practical applications. Following Godlewski-Raviart (1994), we introduce below a simplification of Van Leer's method that leads to a low cost second order numerical scheme. This scheme coincides with Nessyahu-Tadmor non-oscillatory central differencing (see Nessyahu-Tadmor (1990)) when the relaxation terms are omitted.

Given an approximation of a solution $\mathbf{u}$ of system (2.1) in the form of a piecewise constant function, we first determine a piecewise linear approximation of $\mathbf{u}$ in the form (4.1) using (4.2) and (4.3). We compute next

$$\mathbf{u}^n_{j-1/4} = \mathbf{u}^n_j - \frac{\Delta x}{4}\mathbf{p}^n_j, \quad \mathbf{u}^n_{j+1/4} = \mathbf{u}^n_j + \frac{\Delta x}{4}\mathbf{p}^n_j. \tag{4.9}$$

A prediction of solution $\mathbf{u}$ in $x = \left(j \pm \frac{1}{4}\right)\Delta x$ at time $\left(n + 1/2\right)\Delta t$ is then:

$$\mathbf{u}^{n+1/2}_{j\pm1/4} = \mathbf{u}^n_{j\pm1/4} - \frac{\Delta t}{\Delta x}\left(\mathbf{f}\left(\mathbf{u}^n_{j+1/4}\right) - \mathbf{f}\left(\mathbf{u}^n_{j-1/4}\right)\right) + \frac{\Delta t}{2\epsilon}\mathbf{R}\left(\mathbf{u}^{n+1/2}_{j\pm1/4}\right). \tag{4.10}$$

Then $\mathbf{u}^{n+1}_{j+1/2}$ is obtained by applying the first order numerical scheme of Sect. 4 for time $\Delta t/2$ between the two states $\mathbf{u}^{n+1/2}_{j+1/4}$ and $\mathbf{u}^{n+1/2}_{j+3/4}$, distant from $\Delta x/2$: we set $\mathbf{u}^{n+1/2}_{j+1/2} = \mathbf{H}\left(\Delta t/2\right)$ where function $\mathbf{H}$ is the solution of the following ordinary differential equations:

$$\begin{aligned}
&\frac{\mathrm{d}\mathbf{u}^R}{\mathrm{d}t}(t) = \frac{1}{\epsilon}\mathbf{B}\left(\mathbf{u}^{n+1/2}_{j+1/4}, \mathbf{u}^{n+1/2}_{j+3/4}\right)\mathbf{u}^R(t), \quad \mathbf{u}^R(0) = \mathbf{u}^{n+1/2}_{j+3/4}\\
&\frac{\mathrm{d}\mathbf{u}^L}{\mathrm{d}t}(t) = \frac{1}{\epsilon}\mathbf{B}\left(\mathbf{u}^{n+1/2}_{j+1/4}, \mathbf{u}^{n+1/2}_{j+3/4}\right)\mathbf{u}^L(t), \quad \mathbf{u}^L(0) = \mathbf{u}^{n+1/2}_{j+1/4}\\
&\frac{\mathrm{d}\mathbf{H}}{\mathrm{d}t}(t) = \frac{1}{\epsilon}\mathbf{B}\left(\mathbf{u}^{n+1/2}_{j+1/4}, \mathbf{u}^{n+1/2}_{j+3/4}\right)\mathbf{H}(t) - \frac{2}{\Delta x}\mathbf{A}\left(\mathbf{u}^{n+1/2}_{j+1/4}, \mathbf{u}^{n+1/2}_{j+3/4}\right)\left(\mathbf{u}^R(t) - \mathbf{u}^L(t)\right),\\
&\qquad\qquad \mathbf{H}(0) = \frac{\mathbf{u}^{n+1/2}_{j+1/4} + \mathbf{u}^{n+1/2}_{j+3/4}}{2}.
\end{aligned} \tag{4.11}$$

We prove the

**Lemma 4.1.** *Applied to the linear hyperbolic problem with constant coefficients* $\partial_t\mathbf{u} + \mathbf{A}\partial_x\mathbf{u} = \mathbf{0}$*, our numerical scheme coincides with the Nessyahu-Tadmor numerical scheme:*

$$\mathbf{u}^{n+1}_{j+1/2} = \frac{\mathbf{u}^n_j + \mathbf{u}^n_{j+1}}{2} - \frac{\Delta x}{8}\left(\mathbf{p}^n_{j+1} - \mathbf{p}^n_j\right) - \frac{\Delta t}{\Delta x}\mathbf{A}\left(\mathbf{u}^n_{j+1} - \mathbf{u}^n_j\right) + \frac{\Delta t^2}{2\Delta x}\mathbf{A}^2\left(\mathbf{p}^n_{j+1} - \mathbf{p}^n_j\right). \tag{4.12}$$

**Lemma 4.2.** *When the initial data* $\mathbf{u}^0$ *is a constant function, system (2.1) reduces to the following ordinary differential equation*

$$\frac{d\mathbf{u}}{dt} = \frac{1}{\epsilon}\mathbf{R}\left(\mathbf{u}(t)\right) \tag{4.13}$$

*and the numerical scheme (4.10)-(4.11) reduces to the second order following ODE solver consistent with (4.13):*

$$\mathbf{u}^{n+1/2} = \mathbf{u}^n + \frac{\Delta t}{2\epsilon}\mathbf{R}\left(\mathbf{u}^{n+1/2}\right), \quad \mathbf{u}^{n+1} = \Phi\left(\frac{\Delta t}{2}, \mathbf{u}^{n+1/2}\right)$$

*where* $\Phi$ *is the integral flow of the ODE (4.13).*

*Proof of Lemma 4.1.* We compute indeed:

$$\mathbf{u}^{n+1/2}_{j\pm1/4} = \mathbf{u}^n_j \pm \frac{\Delta x}{4}\mathbf{p}^n_j - \frac{\Delta t}{2}\mathbf{A}\mathbf{p}^n_j$$

so that

(4.14)
$$\mathbf{H}(\Delta t/2) = \mathbf{H}(0) - \frac{\Delta t}{\Delta x}\mathbf{A}\left(\mathbf{u}_{j+1}^n - \mathbf{u}_j^n - \frac{\Delta x}{4}\left(\mathbf{p}_{j+1} + \mathbf{p}_j\right) - \frac{\Delta t\mathbf{A}}{2}\left(\mathbf{p}_{j+1}^n - \mathbf{p}_j^n\right)\right).$$

But

$$\mathbf{H}(0) = \frac{\mathbf{u}_{j+1/4}^{n+1/2} + \mathbf{u}_{j+3/4}^{n+1/2}}{2} = \frac{\mathbf{u}_{j+1}^n + \mathbf{u}_j^n}{2} - \frac{\Delta x}{8}\left(\mathbf{p}_{j+1}^n - \mathbf{p}_j^n\right) - \frac{\Delta t\mathbf{A}}{4}\left(\mathbf{p}_{j+1}^n + \mathbf{p}_j^n\right)$$

and we insert the latter identity in (4.14) to conclude the proof of Lemma 4.1. □

The *proof* of Lemma 4.2 is straightforward. □

## 5. A mathematical model of two-phase fluid flows

Let us consider the flow of a spray of solid particles in a gas in a one-dimensional slab geometry: the mathematical model of this flow is a pertinent example of an hyperbolic system with relaxation terms whose numerical solution involves the approximation of both stiff and nonstiff problems.

We denote by $\alpha$ the volume fraction of the gas phase so that $1-\alpha$ is the volume fraction of the dispersed phase. The gas is characterized by its mass density $\rho_g$ and by its velocity $u_g$. Similarly the dispersed phase is characterized by quantities $\rho_p$ and $u_p$. We assume that $\rho_p$ is a positive constant. Omitting the specific internal energies of the gas and the liquid phases, we only retain in our model the mass and impulsion conservation equations which write as follows:

(5.1.*i*) $$\frac{\partial}{\partial t}(\rho_g) + \frac{\partial}{\partial x}(\rho_g u_g) = 0$$

(5.1.*ii*) $$\frac{\partial}{\partial t}(\rho_g u_g) + \frac{\partial}{\partial x}(\rho_g u_g^2) + \frac{\partial p_g}{\partial x} = \frac{G}{\tau}$$

(5.1.*iii*) $$\frac{\partial}{\partial t}\left((1-\alpha)\rho_p\right) + \frac{\partial}{\partial x}\left((1-\alpha)\rho_p u_p\right) = 0$$

(5.1.*iv*) $$\frac{\partial}{\partial t}\left((1-\alpha)\rho_p u_p\right) + \frac{\partial}{\partial x}\left((1-\alpha)\rho_p u_p^2\right) + \frac{\partial\theta}{\partial x} = -\frac{G}{\tau}$$

where $p_g$ and $\theta$ are pressure functions of the form

(5.2.*i*) $$p_g = K\left(\rho_g\right)^\gamma$$

(5.2.*ii*) $$\theta = \theta_0(1-\alpha).$$

Here $p_g$ is the isotropic gas pressure law and $K$ is an appropriate positive constant while $\theta$ is a function of the dispersed phase volume fraction and $\theta_0$ is representative of the gas rest pressure on the particle. In fact

$$p_{\text{eff}} = p_g + \theta$$

represents the effective pressure of the two-phase flow. Usually $\theta$ is small in comparison with the gas pressure $p_g$ and makes a small contribution to the effective pressure of the two-phase flow. (We refer the reader to Raviart-Sainsaulieu (1994) for more detail about the model (5.1)-(5.2))

Taking as the drag force acting on a single particles the Stokes force, the impulsion exchanges between the two phases read $G/\tau$ where the function $G$ and the relaxation time $\tau$ are in the following form (see Williams (1985))

$$G = (1-\alpha)\rho_p(u_p - u_g) \tag{5.3.$i$}$$

$$\tau = \frac{16r^2\rho_p}{81\mu_g} \tag{5.3.$ii$}$$

Here $r$ denotes the radius of the particles and $\mu_g$ is the viscosity of the gas: they are taken as positive constants.

We can write system (5.1) in the following condensed form:

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\big(\mathbf{f}(\mathbf{u})\big) = \frac{\mathbf{R}(\mathbf{u})}{\tau} \tag{5.4.$i$}$$

or formally in the following non conservation form

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = \frac{\mathbf{B}(\mathbf{u})}{\tau}\cdot \mathbf{u} \tag{5.4.$ii$}$$

where the state vector $\mathbf{u}$ is

$$\mathbf{u} = \begin{pmatrix} \rho_g \\ \rho_g u_g \\ (1-\alpha)\rho_p \\ (1-\alpha)\rho_p u_p \end{pmatrix} \tag{5.5.$i$}$$

and where the matrix valued functions $\mathbf{A}$ and $\mathbf{B}$ are respectively:

$$\mathbf{A}(\mathbf{u}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ c_g^2 - u_g^2 & 2u_g & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & c_p^2 - u_p^2 & 2u_p \end{pmatrix},$$

$$\mathbf{B}(\mathbf{u}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\frac{(1-\alpha)\rho_p}{\rho_g} & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{(1-\alpha)\rho_p}{\rho_g} & 0 & -1 \end{pmatrix}. \tag{5.5.$ii$}$$

Then a straightforward computation proves the following (we refer the reader to Godlewski-Raviart (1991) or Smoller (1982) for any notion about hyperbolic systems):

**Proposition 5.1.** *When the function $G$ is set to zero, system (5.1) consists of two decoupled hyperbolic systems of conservation laws whose characteristic velocities are respectively*

(5.6.*i*) $$\lambda_1 = u_g - c_g,\ \lambda_2 = u_g + c_g$$

*where* $c_g$ *is the gas sound speed:* $c_g = \sqrt{\frac{\gamma p_g}{\rho_g}}$ *and*

(5.6.*ii*) $$\lambda_3 = u_p - c_p,\ \lambda_4 = u_p + c_p$$

*where* $c_p$ *is the following positive number:* $c_p = \sqrt{\theta_0/\rho_p}$. *Furthermore the four characteristic fields of the hyperbolic system extracted from (5.1) are genuinely non linear.*

Next, we can derive an entropy associated with system (5.1):

**Proposition 5.2.** *Let* $S$ *and* $F$ *denote the following two functions:*

(5.7) $$S = \frac{1}{2}\left(\rho_g u_g^2 + (1-\alpha)\rho_p u_p^2\right) + \rho_g P_g(\rho_g) + \theta_0(1-\alpha)\log\left(1-\alpha\right)$$

*where the function* $P_g$ *is given by* $P_g(\rho_g) = \frac{K}{\gamma-1}(\rho_g)^{\gamma-1}$, *(i.e.,* $P_g' = p_g/\rho_g^2$*) and*

$$\begin{aligned} F =& \frac{1}{2}\left(\rho_g u_g^3 + (1-\alpha)\rho_p u_p^3\right) + \rho_g u_g P_g(\rho_g) + p_g\left(u_g + (1-\alpha)u_p\right) \\ &+ \theta_0(1-\alpha)u_p \log\left(1-\alpha\right). \end{aligned} \tag{5.8}$$

*Then, the function* $\mathbf{u} = \left(\rho_g, \rho_g u_g, (1-\alpha), (1-\alpha)u_p\right)^T \to S(\mathbf{u})$ *is strictly convex and if* $\mathbf{u}$ *is a continuous solution of system (5.1), it satisfies the following additional equation:*

$$\frac{\partial}{\partial t}\left(S(\mathbf{u})\right) + \frac{\partial}{\partial x}\left(F(\mathbf{u})\right) = \frac{(u_g - u_p)G}{\tau} = -\frac{(1-\alpha)\rho_p\left(u_g - u_p\right)^2}{\tau}. \tag{5.9}$$

(See Bereux (1994) for the proof of Proposition 5.2.) We deduce from Proposition 2.3 that $S$ is an entropy function for system (5.1) and that the relaxation terms are compatible with the entropy in the sense that the entropy production in (5.9) is always non positive. Furthermore the convection terms extracted from (5.1) can be symmetrized: for any $\mathbf{u} \in \Omega$, the matrix $S''(\mathbf{u})$ is symmetric and positive definite and the matrix $S''(\mathbf{u})\mathbf{A}(\mathbf{u})$ is symmmetric: see Harten-Lax-Van Leer (1981) for the proof.

We consider next the limit $\tau$ tends to zero. According to Propositions 5.1 and 5.2 our system satisfies the assumptions in Chen-Levermore-Liu (1994) and we may thus apply their formalism to write the relaxation system limit of (5.1) when $\tau$ tends to zero. It is however more instructive to perform explicitly the computations.

The relaxation variable is here the relative velocity between the two phases and, in order to derive the relaxation system, we introduce the set of dependent variables $\mathbf{v}$ defined by

(5.10.*i*) $$\mathbf{v}^T = \left(\rho_g, (1-\alpha)\rho_p, \rho_g u_g + (1-\alpha)\rho_p u_p, u_g - u_p\right).$$

For simplicity we denote respectively by $\rho$ and $u$ the mean mass density and mean velocity of the two-phase flow:

$$\rho = \rho_g + (1-\alpha)\rho_p, \qquad u = \frac{\rho_g u_g + (1-\alpha)\rho_p u_p}{\rho_g + (1-\alpha)\rho_p} \tag{5.11.ii}$$

and by $v$ the relative velocity between the two phases:

$$v = u_g - u_p. \tag{5.11.iii}$$

We first have the following

**Lemma 5.3.** *For continuous solutions, system (5.1) takes the equivalent form:*

$$\frac{\partial}{\partial t}\left(\rho_g\right) + \frac{\partial}{\partial x}\left(\rho_g u + \frac{\rho_g(1-\alpha)\rho_p}{\rho} v\right) = 0 \tag{5.12.i}$$

$$\frac{\partial}{\partial t}\left((1-\alpha)\rho_p\right) + \frac{\partial}{\partial x}\left((1-\alpha)\rho_p u - \frac{(1-\alpha)\rho_p\rho_g}{\rho} v\right) = 0 \tag{5.12.ii}$$

$$\frac{\partial}{\partial t}\left(\rho u\right) + \frac{\partial}{\partial x}\left(\rho u^2 + \frac{\rho_g(1-\alpha)\rho_p}{\rho} v^2\right) + \frac{\partial}{\partial x}\left(p_g + \theta\right) = 0 \tag{5.12.iii}$$

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x}\left(\left(\frac{u_g + u_p}{2}\right) v\right) + \\ + \frac{\partial}{\partial x}\left(\frac{K\gamma}{\gamma-1}(\rho_g)^{\gamma-1} - \frac{\theta_0}{\rho_p}\log\left(1-\alpha\right)\right) = -\frac{\rho}{\rho_g \tau} v. \tag{5.12.iv}$$

*Proof.* By definition of the total mass density and of the mean velocity of the two-phase flow, we have

$$u_g = u + \frac{(1-\alpha)\rho_p v}{\rho}$$
$$u_p = u - \frac{\rho_g v}{\rho}.$$

Let **u** be a continuous solution of system (5.1). Then (5.12.i) and (5.12.ii) are obtained respectively from (5.1.i) and (5.1.ii). Next the total impulsion conservation equation (5.12.iii) is obtained by summing equations (5.1.ii) and (5.1.iv). To obtain equation (5.12.iv), insert the gas mass conservation (5.1.i) in the gas impulsion conservation equation (5.1.ii):

$$\frac{\partial u_g}{\partial t} + u_g \frac{\partial u_g}{\partial x} + \frac{1}{\rho_g}\frac{\partial p_g}{\partial x} = \frac{G}{\tau\rho_g}. \tag{5.13.i}$$

Similarly, equations (5.1.iii) and (5.1.iv) yield

$$\frac{\partial u_p}{\partial t} + u_p \frac{\partial u_p}{\partial x} + \frac{1}{(1-\alpha)\rho_p}\frac{\partial \theta}{\partial x} = -\frac{G}{\tau(1-\alpha)\rho_p}. \tag{5.13.ii}$$

Subtracting (5.13.ii) from (5.13.i) gives

$$\frac{\partial v}{\partial t} + u_g \frac{\partial u_g}{\partial x} - u_p \frac{\partial u_p}{\partial x} + \frac{1}{\rho_g} \frac{\partial p_g}{\partial x} - \frac{1}{(1-\alpha)\rho_p} \frac{\partial \theta}{\partial x} = \frac{\rho G}{\tau \rho_g (1-\alpha)\rho_p}.$$

But **u** is continuous. Inserting the definition (5.2) of the pressure functions $p_g$ and $\theta$ in the latter equation gives:

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x}\left(\left(\frac{u_g + u_p}{2}\right) v\right) + \frac{\partial}{\partial x}\left(\frac{K\gamma}{\gamma - 1}(\rho_g)^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right) = \frac{\rho G}{\tau \rho_g (1-\alpha)\rho_p}.$$

Equation (5.12.iv) is finally obtained by inserting the expression (5.3) of $G$ in the latter equation. □

When $\tau$ is very small, we expect the relative velocity between the two phases to be of order $\tau$. The Chapman-Enskog expansion developed in Chen-Levermore-Liu (1994) consists in doing a formal expansion of the solutions of system (5.12) in form of a power serie in $\tau$. At zero-th order in $\tau$, we obtain readily the following hyperbolic system of conservation laws in the unknown vector valued function $(\rho_g, (1-\alpha)\rho_p, \rho u)$:

$$\frac{\partial}{\partial t}(\rho_g) + \frac{\partial}{\partial x}(\rho_g u) = 0 \tag{5.14.i}$$

$$\frac{\partial}{\partial t}((1-\alpha)\rho_p) + \frac{\partial}{\partial x}((1-\alpha)\rho_p u) = 0 \tag{5.14.ii}$$

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2) + \frac{\partial}{\partial x}(p_g + \theta) = 0. \tag{5.14.iii}$$

Next, at first order in $\tau$, function $v$ is obtained from equation (5.12.iv) in the form

$$v = -\frac{\tau \rho_g}{\rho} \frac{\partial}{\partial x}\left(\frac{K\gamma}{\gamma - 1}(\rho_g)^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right) \tag{5.15}$$

and inserting expression (5.15) in equations (5.12.i), (5.12.ii) and (5.12.iii) gives the following convection-diffusion system:

$$\frac{\partial}{\partial t}(\rho_g) + \frac{\partial}{\partial x}(\rho_g u) - \tau \frac{\partial}{\partial x}\left(\frac{\rho_g}{\rho}\frac{\partial}{\partial x}\left(\frac{K\gamma}{\gamma-1}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right)\right) = 0 \tag{5.16.i}$$

$$\frac{\partial}{\partial t}((1-\alpha)\rho_p) + \frac{\partial}{\partial x}((1-\alpha)\rho_p u) + \tau \frac{\partial}{\partial x}\left(\frac{\rho_g}{\rho}\frac{\partial}{\partial x}\left(\frac{K\gamma}{\gamma-1}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right)\right) = 0 \tag{5.16.ii}$$

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p_g + \theta) = 0. \tag{5.16.iii}$$

It follows from Chen-Levermore-Liu (1994) that system (5.15) is hyperbolic, and that the diffusion matrix in (5.16) is compatible with the entropy $T = S\left(\rho_g, (1-\alpha)\rho_p, \rho_g u, (1-\alpha)\rho_p u\right)$. In fact we prove the

**Proposition 5.4.** *System (5.15) is an hyperbolic system of conservation laws whose characteristic velocities are*

$$\widehat{\lambda}_1 = u - c_{\text{diph}},\ \widehat{\lambda}_2 = u,\ \widehat{\lambda}_3 = u + c_{\text{diph}} \tag{5.17}$$

*where the sound speed in the two-phase medium is* $c_{\text{diph}} = \sqrt{\frac{\gamma p_g + \theta}{\rho}}$. *Furthermore the pair* $\left(T, E\right)$ *defined by*

$$\begin{aligned} T &= \frac{1}{2}\rho u^2 + \rho_g P(\rho_g) + \theta_0(1-\alpha)\log\left(1-\alpha\right) \\ E &= \frac{1}{2}\rho u^3 + u\left(\rho_g P(\rho_g) + \theta_0(1-\alpha)\log\left(1-\alpha\right)\right) + u p_g(2-\alpha) \end{aligned}$$

*is an entropy-flux pair for system (5.15), compatible with the diffusion terms in (5.16): any continuous solution of (5.16) satisfies the following additional conservation:*

$$\begin{aligned} &\frac{\partial T}{\partial t} + \frac{\partial E}{\partial x} - \tau\frac{\partial}{\partial x}\left[\frac{\rho_g}{\rho}\left(\frac{\gamma K}{\gamma-1}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\left(1+\log\left(1-\alpha\right)\right)\right)\times\right. \\ &\qquad\left.\times\frac{\partial}{\partial x}\left(\frac{\gamma K}{(\gamma-1)}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right)\right] \\ &\qquad = \tau\frac{\rho_g}{\rho}\left[\frac{\partial}{\partial x}\left(\frac{\gamma K}{\gamma-1}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\log\left(1-\alpha\right)\right)\right]^2 \geq 0. \end{aligned} \tag{5.18}$$

*Remark 5.1.* The number $c_{\text{diph}}$ is the sound speed in the two-phase medium in the limit $\tau \to 0$ and is smaller than the sound speed $c_g$ in gas alone: indeed, in the limit $\tau \to 0$, the inertia of the two-phase medium is the sum of the inertia of the gas and dispersed phases while its compressibility is that of the gas alone since the particles are incompressible and this explains that $c_{\text{diph}} < c_g$.

*Proof.* The expression of the characteristic velocities of system (5.15) follow from a straightforward computation. Let us prove the entropy balance (5.18): let $\left(\rho_g, (1-\alpha)\rho_p, \rho u\right)$ be a continuous solution of system (5.16). We first compute:

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\rho u^2\right) + \frac{\partial}{\partial x}\left(\frac{1}{2}\rho u^3\right) = \frac{\rho u}{2}\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x}\right) + \frac{u}{2}\left(\frac{\partial}{\partial t}\left(\rho u\right) + \frac{\partial}{\partial x}\left(\rho u^2\right)\right).$$

But the mass conservation equation of the two phase fluid flow obtained by summing equations (5.16.i) and (5.16.ii) writes

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} = 0$$

and inserting the latter equation in the total impulsion conservation equation (5.16.iii), we obtain:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{1}{\rho}\frac{\partial}{\partial x}\left(p_g + \theta\right) = 0.$$

We deduce that

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\rho u^2\right) + \frac{\partial}{\partial x}\left(\frac{1}{2}\rho u^3\right) = -u\frac{\partial}{\partial x}\left(p_g + \theta\right). \tag{5.19}$$

On the other hand, we have by definition of function $P_g$:

$$\begin{aligned}\frac{\partial}{\partial t}\left(\rho_g P_g + \theta_0\ (1-\alpha)\log(1-\alpha)\right) + \frac{\partial}{\partial x}\left(u\left(\rho_g P_g + \theta_0(1-\alpha)\log(1-\alpha)\right)\right) = \\ = \left(\frac{p_g}{\rho_g} + P_g\right)\left(\frac{\partial \rho_g}{\partial t} + \frac{\partial \rho_g u}{\partial x}\right) - p_g\frac{\partial}{\partial x}u \\ + \frac{\theta_0}{\rho_p}\left(1 + \log(1-\alpha)\right)\left(\frac{\partial}{\partial t}\left((1-\alpha)\rho_p\right) + \frac{\partial}{\partial x}\left((1-\alpha)\rho_p u\right)\right) \\ - \theta_0(1-\alpha)\frac{\partial u}{\partial x}\end{aligned}$$

and inserting the gas and dispersed phases mass conservation equations (5.16.i) and (5.16.ii) in the latter equation, we obtain:

$$\begin{aligned}&\frac{\partial}{\partial t}\left(\rho_g P_g + \theta_0\ (1-\alpha)\log(1-\alpha)\right) + \frac{\partial}{\partial x}\left[u\left(\rho_g P_g + \theta_0(1-\alpha)\log(1-\alpha)\right)\right) \\ &\quad = \tau\left[\frac{p_g}{\rho_g}P_g - \frac{\theta_0}{\rho_p}\left(1 + \log\left(1-\alpha\right)\right)\right] \\ &\frac{\partial}{\partial x}\left(\frac{\rho_g}{\rho}\left(\frac{\gamma K}{\gamma - 1}\rho_g^{\gamma-1} - \frac{\theta_0}{\rho_p}\log(1-\alpha)\right)\right).\end{aligned} \tag{5.20}$$

Then the entropy balance (5.18) follows from (5.19) and (5.20). □

To conclude this section we consider the numbers $a_k(t)$ defined by (3.14) in the particular case of system (5.1). We denote by $\mathbf{r}_k$ (resp. $\mathbf{l}_k(\mathbf{u})$), $1 \le k \le 4$ the right (resp. left) eigenvectors of $\mathbf{A}(\mathbf{u})$. The left eigenvectors are normalized by $\mathbf{l}_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 1$, $1 \le k \le 4$ (we do not need to specify the normalization of the right eigenvectors here). The relaxation terms in system (5.1) are compatible with the convection matrix $\mathbf{A}$ in the following sense:

**Proposition 5.6.** *Assume that the condition*

$$|u_g| < c_g, \quad |u_p| < c_p, \tag{5.21}$$

*is satisfied. Then the left and right eigenvectors of matrix* $\mathbf{A}(\mathbf{u})$ *satisfy the following estimates:*

$$\mathbf{l}_k(\mathbf{u}) \cdot \left(\mathbf{B}(\mathbf{u})\mathbf{r}_k(\mathbf{u})\right) < 0, \quad 1 \le k \le 4. \tag{5.22}$$

**Corollary 5.7.** *Under the condition (5.21) the functions $a_k$ defined by (3.14) tend to zero as $t \to +\infty$ exponentially fast.*

*Proof of Proposition 5.6.* We compute the four right eigenvectors of matrix $\mathbf{A}(\mathbf{u})$:

$$\mathbf{r}_1 = \begin{pmatrix} 1 \\ u_g - c_g \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{r}_2 = \begin{pmatrix} 1 \\ u_g + c_g \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ u_p - c_p \end{pmatrix}, \quad \mathbf{r}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ u_p + c_p \end{pmatrix}$$

while the left eigenvectors are respectively

$$\mathbf{l}_1 = \frac{-1}{2c_g} \begin{pmatrix} -u_g - c_g \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{l}_2 = \frac{1}{2c_g} \begin{pmatrix} -u_g + c_g \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{l}_3 = \frac{-1}{2c_p} \begin{pmatrix} 0 \\ 0 \\ -u_p - c_p \\ 1 \end{pmatrix}, \quad \mathbf{l}_4 = \frac{1}{2c_p} \begin{pmatrix} 0 \\ 0 \\ -u_p + c_p \\ 1 \end{pmatrix}.$$

By definition of matrix $\mathbf{B}$, we compute the following expressions:

$$\mathbf{l}_1(\mathbf{u}) \cdot \mathbf{B}(\mathbf{u})\mathbf{r}_1(\mathbf{u}) = \frac{(u_g - c_g)(1-\alpha)\rho_p}{2\tau c_g \rho_g}$$

$$\mathbf{l}_2(\mathbf{u}) \cdot \mathbf{B}(\mathbf{u})\mathbf{r}_2(\mathbf{u}) = \frac{-(u_g + c_g)(1-\alpha)\rho_p}{2\tau c_g \rho_g}$$

$$\mathbf{l}_3(\mathbf{u}) \cdot \mathbf{B}(\mathbf{u})\mathbf{r}_3(\mathbf{u}) = \frac{u_p - c_p}{2\tau c_p}$$

$$\mathbf{l}_4(\mathbf{u}) \cdot \mathbf{B}(\mathbf{u})\mathbf{r}_4(\mathbf{u}) = -\frac{u_p + c_p}{2\tau c_p}$$

and the proof of Proposition 5.6 is complete. □

## 6. Application of our scheme to the computation of model (5.1)

We apply the numerical schemes written in Sects. 3 and 4 to system (5.1) that models two-phase fluid flows composed of liquid particles in a gas phase. We first note that system (5.1) is an hyperbolic ystem with relaxation. Indeed a convenient pair $(\mathbf{Q}, \mathscr{E})$ is given by

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathscr{E}(\mathbf{v}) = \begin{pmatrix} v_1 \\ \frac{v_1 v_3}{v_1 + v_2} \\ v_2 \\ \frac{v_2 v_3}{v_1 + v_2} \end{pmatrix}. \tag{6.1}$$

Next our numerical scheme is based on a linearization in the form (3.3) of (5.1) which satisfies assumptions [H1] to [H3] of Sect. 3: according to Lemmas 4.2 and 4.3 the assumption [H1] relies on convenient linearizations of the vector valued functions $\mathbf{f}$ and $\mathscr{E}$.

**Lemma 6.1.** *Let be given two given states* $\mathbf{u}^L$ *and* $\mathbf{u}^R$ *in* $\Omega^2$. *Define the intermediate state* $\overline{\mathbf{u}} \in \Omega$ *as*

$$\begin{aligned}
\sqrt{\overline{\rho}_g} &= \left(\sqrt{\rho_g^L} + \sqrt{\rho_g^R}\right)/2 \\
\sqrt{\overline{\rho}_g}\overline{u}_g &= \left(\sqrt{\rho_g^L}u_g^L + \sqrt{\rho_g^R}u_g^R\right)/2 \\
\sqrt{\overline{\rho}_p} &= \left(\sqrt{\rho_p^L} + \sqrt{\rho_p^R}\right)/2 \\
\sqrt{\overline{\rho}_p}\overline{u}_p &= \left(\sqrt{\rho_p^L}u_p^L + \sqrt{\rho_p^R}u_p^R\right)/2
\end{aligned} \tag{6.2}$$

*Then the* $4 \times 4$ *matrix*

$$\mathbf{A}\left(\mathbf{u}^L, \mathbf{u}^R\right) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \overline{c}_g^2 - \overline{u}_g^2 & 2\overline{u}_g & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \overline{c}_p^2 - \overline{u}_p^2 & 2u_p \end{pmatrix} \tag{6.3}$$

*where*

$$\overline{c}_g^2 = \begin{cases} k\dfrac{(\rho_g^R)^\gamma - (\rho_g^L)^\gamma}{\rho_g^R - \rho_g^L} & \text{if } \rho_g^R \neq \rho_g^L \\ k\gamma\rho_g^{\gamma-1} & \text{if } \rho_g^L = \rho_g^R = \rho_g \end{cases} \tag{6.4.$i$}$$

*and*

$$\overline{c}_p^2 = \frac{\theta_0}{\rho_l} \tag{6.4.$ii$}$$

*is a Roe linearization of the flux function* $\mathbf{f}$.

The *proof* of Lemma 6.1 is straightforward. We have next the

**Lemma 6.2.** *Let be given two states* $\mathbf{v}^L$ *and* $\mathbf{v}^R$ *in* $\omega = \mathbf{Q}\Omega$. *There exists a parameter vector* $\varphi(\mathbf{v}^L, \mathbf{v}^R)$ *in* $\omega$ *such that the* $4 \times 4$ *matrix valued function* $\mathbf{E}$ *defined by*

$$\mathbf{E}\left(\mathbf{v}^L, \mathbf{v}^R\right) = \mathscr{E}'\left(\varphi\left(\mathbf{v}^L, \mathbf{v}^R\right)\right) \tag{6.5}$$

*is a linearization of the vector valued function* $\mathscr{E} : \omega \to \Omega$.

*Proof.* It suffices in fact to linearize the real valued function $\chi_1$ defined by

$$\chi_1(\mathbf{v}) = \frac{v_1 v_3}{v_1 + v_2}$$

in the form

$$\chi_1(\mathbf{v}^R) - \chi_1(\mathbf{v}^L) = \chi_1'\big(\varphi(\mathbf{v}^L, \mathbf{v}^R)\big) \cdot \big(\mathbf{v}^R - \mathbf{v}^L\big). \tag{6.6}$$

Assume indeed that (6.6) holds true. Then, setting

$$\chi_2(\mathbf{v}) = \frac{v_2 v_3}{v_1 + v_2},$$

we compute

$$\chi_2(\mathbf{v}) = v_3 - \frac{v_1 v_3}{v_1 + v_2}$$

so that

$$\begin{aligned}\chi_2(\mathbf{v}^R) - \chi_2(\mathbf{v}^L) &= v_3^R - v_3^L - \chi_1(\mathbf{v}^R) - \chi_1(\mathbf{v}^L)\\ &= v_3^R - v_3^L - \chi_1'\big(\varphi(\mathbf{v}^L, \mathbf{v}^R)\big) \cdot \big(\mathbf{v}^R - \mathbf{v}^L\big)\\ &= \chi_2'\big(\varphi(\mathbf{v}^L, \mathbf{v}^R)\big) \cdot \big(\mathbf{v}^R - \mathbf{v}^L\big).\end{aligned}$$

Then by the definition (6.1) of function $\mathscr{E}$:

$$\mathscr{E}(\mathbf{v}^R) - \mathscr{E}(\mathbf{v}^L) = \mathscr{E}'\big(\varphi(\mathbf{v}^L, \mathbf{v}^L)\big) \cdot \big(\mathbf{v}^R - \mathbf{v}^L\big)$$

and the matrix valued function $\big(\mathbf{v}^L, \mathbf{v}^R\big) \to \mathscr{E}'\big(\varphi(\mathbf{v}^L, \mathbf{v}^R)\big)$ is indeed a linearization of $\mathscr{E}$.

Let us linearize $\chi_1$: set $\chi(s) = \chi_1\big(\mathbf{v}^L + s(\mathbf{v}^R - \mathbf{v}^L)\big)$ for $s \in [0,1]$. There exists a real number $s(\mathbf{v}^L, \mathbf{v}^R) \in [0,1]$ such that $\chi(1) - \chi(0) = \chi'(s)$: it suffices to set

$$\varphi\big(\mathbf{v}^L, \mathbf{v}^R\big) = \mathbf{v}^L + s\big(\mathbf{v}^L, \mathbf{v}^R\big)\big(\mathbf{v}^R - \mathbf{v}^L\big)$$

to obtain (6.6). □

Define next the $4 \times 4$ matrix valued function $(\mathbf{u}^L, \mathbf{u}^R) \in \Omega^2 \to \mathbf{B}(\mathbf{u}^L, \mathbf{u}^R)$ by (3.25). Then, according to Lemma 4.3, for $\mathbf{u}^L$ and $\mathbf{u}^R$ given, the following linearization

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\big(\mathbf{u}^L, \mathbf{u}^R\big)\frac{\partial \mathbf{u}}{\partial x} = \frac{1}{\epsilon}\mathbf{B}\big(\mathbf{u}^L, \mathbf{u}^R\big)\mathbf{u} \tag{6.7}$$

satisfies [H1]. Next matrix $\mathbf{B}\big(\mathbf{u}^L, \mathbf{u}^R\big)$ has one negative eigenvalue and the triple eigenvalue zero such that [H3] holds. Finally a straightforward computation shows that for $(\mathbf{u}^L, \mathbf{u}^R) \in \Omega^2$ given the $3\times 3$ matrix $\widetilde{\mathbf{A}} = \mathbf{Q}\mathbf{A}\big(\mathbf{u}^L, \mathbf{u}^R\big)\mathbf{E}\big(\mathbf{Q}\mathbf{u}^L, \mathbf{Q}\mathbf{u}^R\big)$ has the following three distinct eigenvalues at least when $|\overline{u}_g - \overline{u}_p|$ is small enough:

$$\begin{aligned}\lambda_1 =& \frac{\rho_g \overline{u}_g + (1-\alpha)\rho_p \overline{u}_p}{\rho_g + (1-\alpha)\rho_p}\\ &- \sqrt{\frac{\rho_g \overline{c}_g^2 + a\overline{c}_p^2}{\rho} - \frac{\rho_g(1-\alpha)\rho_p}{\rho^2}\big(\overline{u}_g - \overline{u}_p\big)^2}\end{aligned} \tag{6.8.$i$}$$

$$\lambda_2 = u \tag{6.8.$ii$}$$

$$\begin{aligned}\lambda_3 =& \lambda_1 = \frac{\rho_g \overline{u}_g + (1-\alpha)\rho_p \overline{u}_p}{\rho_g + (1-\alpha)\rho_p}\\ &+ \sqrt{\frac{\rho_g \overline{c}_g^2 + a\overline{c}_p^2}{\rho} - \frac{\rho_g(1-\alpha)\rho_p}{\rho^2}\big(\overline{u}_g - \overline{u}_p\big)^2}\end{aligned} \tag{6.8.$iii$}$$

where

$$u = \frac{v_3}{v_1 + v_2}, \quad \mathbf{v} = \mathbf{Q}\varphi\big(\mathbf{Q}\mathbf{u}^L, \mathbf{Q}\mathbf{u}^R\big).$$

Hence our linearization satisfies [H1] to [H3] and is compatible with the equilibrium system (2.15) in the sense that for $\mathbf{v}^L$ and $\mathbf{v}^R$ given in $\omega$ the $3 \times 3$ matrix

$$\widetilde{\mathbf{A}}\big(\mathbf{v}^L, \mathbf{v}^R\big) = \mathbf{Q}\mathbf{A}\big(\mathscr{E}(\mathbf{v}^L), \mathscr{E}(\mathbf{v}^R)\big)\mathbf{E}(\mathbf{v}^L, \mathbf{v}^R)$$

is a Roe linearization of the flux function in (5.14). We may thus apply the numerical schemes described in Sects. 3 and 4.

From now on we work with the second order version (4.9)-(4.10)-(4.11) of our scheme. It is important to note that the system of differential equations (4.11) may be solved explicitly: these are the formulae that are actually encoded.

In order to test the quality of our scheme, we compute the solution of some Riemann problems and the propagation of sound waves in a two-phase medium and we compare our numerical results with those obtained with two other methods. The study of the propagation of sound waves at different frequencies and for different particles radii is very interesting since we dispose of analytical solutions and we may easily cover the whole range of stiffness for system (5.1). Govern indeed the time step used for the computation of the propagation of a sound wave with frequency $f$ and celerity $c$ by the CFL condition relative to the convection terms alone:

$$\Delta t \simeq \Delta x / c$$

where $c$ is the velocity of the sound wave. A good resolution of the wave requires that the space step is several times smaller than the wavelength $\Lambda$: say

$$\Delta x = \frac{\Lambda}{N}$$

where $N \simeq 50$. Then, the ratio between the time step $\Delta t$ and the relaxation time $\epsilon$ writes:

$$\frac{\Delta t}{\epsilon} = \frac{\Delta x}{c\epsilon} = \frac{\Lambda}{Nc\epsilon} = \frac{1}{fN\epsilon}. \tag{6.9}$$

System (5.1) is stiff if ratio (6.9) is large, *i.e.*, if the frequency $f$ of the sound wave is small or if the relaxation time $\epsilon$ is small. The quality of the numerical scheme may thus be tested by computing several sound waves for different dimensionless numbers

$$B = \frac{1}{f\epsilon} \tag{6.10}$$

and by comparing the numerical results with the following analytical solution (see Bereux (1994) and Culick (1981)):

$$\mathbf{u} = \mathbf{u}^0 + \mathbf{u}' \exp(-at) \exp\left(i 2\pi f \left(t - \frac{x}{c}\right)\right) \tag{6.11.$i$}$$

where the attenuation coefficient $a$ is given in function of the frequency $f$ by

(6.11.*ii*) $$a = \frac{\epsilon(2\pi f)^2}{2} \frac{1}{1 + \left(\frac{2\pi f \epsilon \alpha^0 \rho_g^0}{\rho^0}\right)^2} \frac{\rho_p(1-\alpha^0)}{\rho^0}.$$

The second order numerical scheme of Sect. 5 is referred as Method I in the sequel. Let us describe the two other methods used in the comparison: Method II is a fractional step method which runs as follows: let $\phi$ denote a numerical flux at second order consistent with the convection system extracted from (5.1). (For instance we choose a second order scheme based on the Roe linearization (6.3) of the flux function **f**.) Then a second order fractional step method is the following:

$$\begin{aligned}
\mathbf{u}_j^{n+1/3} &= \mathbf{u}_j^n + \frac{\Delta t}{2} \frac{\mathbf{R}\left(\mathbf{u}_j^{n+1/3}\right)}{\epsilon} \\
\mathbf{u}_j^{n+2/3} &= \mathbf{u}_j^{n+1/3} - \frac{\Delta t}{\Delta x}\left(\phi_{j+1/2}^{n+1/3} - \phi_{j-1/2}^{n+1/3}\right) \\
\mathbf{u}_j^{n+1} &= \mathbf{u}_j^{n+2/3} + \frac{\Delta t}{2} \frac{\mathbf{R}\left(\mathbf{u}_j^{n+2/3}\right)}{\epsilon}.
\end{aligned}$$

Finally Method III is a second order fractional step applied to the relaxation system (5.16): let $\psi$ denote a second order numerical flux that approximates the flux in system (5.14). Then Method III allows to compute an approximation of the solution **v** of (5.16) and runs as follows:

$$\begin{aligned}
\mathbf{v}_j^{n+1/3} &= \mathbf{v}_j^n - \frac{\Delta t}{\Delta x}\left(\psi_{j+1/2}^n - \psi_{j-1/2}^n\right) \\
\mathbf{v}_j^{n+2/3} &= \mathbf{v}_j^{n+1/3} - \frac{1}{\Delta x^2}\Big(\mathbf{D}\left(\mathbf{v}_{j+1/2}^{n+1/3}\right) \cdot \left(\mathbf{v}_{j+1}^{n+1/3} - \mathbf{v}_j^{n+1/3}\right) - \\
&\qquad - \mathbf{D}\left(\mathbf{v}_{j-1/2}^{n+1/3}\right) \cdot \left(\mathbf{v}_j^{n+1/3} - \mathbf{v}_{j-1}^{n+1/3}\right)\Big) \\
\mathbf{v}_j^{n+1} &= \mathbf{v}_j^{n+2/3} - \frac{\Delta t}{\Delta x}\left(\psi_{j+1/2}^{n+2/3} - \psi_{j-1/2}^{n+2/3}\right)
\end{aligned}$$

where $\mathbf{v}_{j+1/2}^{n+1/3} = \left(\mathbf{v}_j^{n+1/3} + \mathbf{v}_{j+1}^{n+1/3}\right)/2$ and where **D** is the diffusion matrix such that the diffusion terms in system (5.16) write $\frac{\partial}{\partial x}\left(\mathbf{D}(\mathbf{v})\frac{\partial \mathbf{v}}{\partial x}\right)$. Function **u** is then deduced from the numerical solution **v** thanks to (5.15). This method is described in more details in Bereux (1994).

We compute the propagation of sound waves in a two-phase medium for different numbers $B$ with the three methods: the computation is initialized with a given rest state $\mathbf{u}^0$. We use Neumann boundary conditions at the right boundary of the computational domain and Dirichlet boundary conditions at the left boundary. The left state is the rest state $\mathbf{u}^0$ whose gas mass density is modified to impose a gas pressure oscillation with a frequency $f$ and an amplitude $P_{\text{osc}}$. The rest state $\mathbf{u}^0$ in (6.11.i) is defined by

$$\alpha^0 = .9991,\ \rho_g^0 = 3.78\,\text{kg.m}^{-3},\ u_g = u_p = 0\,\text{m.s}^{-1}.$$

On the other hand, the constant mass density $\rho_p$ and the gas viscosity $\mu_g$ are respectively

$$\rho_p = 1766\,\mathrm{kg.m}^{-3},\ \mu_g = 8.8510^{-5}\ \mathrm{kg.m}^{-1}.\mathrm{s}^{-1}.$$

The coefficients in the pressure laws (2.2) are:

$$K = 974216\,\mathrm{Pa},\quad \gamma = 1.23,\quad \theta_0 = 10000\mathrm{Pa}.$$

Finally the amplitude and the frequency of the pressure oscillation are

$$P_{\mathrm{osc}} = 50000\,\mathrm{Pa},\quad f = 70\,\mathrm{Hz}.$$

Figure 1 gives the attenuation coefficient of the three numerical methods considered here compared with the theoretical attenuation coefficient given by (6.11.ii) in function of the radius of the particles: by definition of number $\epsilon$ number $B$ writes

$$B = \frac{81\mu_g}{16r^2 f}$$

and is large for small particle radius. Hence the relaxation terms in (5.1) are stiff for small particle radius and non stiff for large radius. We used 400 grid points for the computations and the time step is governed by the CFL condition alone: $CFL = 0.8$. The profiles given in fig. 1 lack smoothness due to the postprocessing of the numerical solutions.

Method II gives good results for large particle radius but is unable to deal with stiff systems. Indeed it is proved in Langseth-Tveito-Whinter (1993) that the solutions computed with the fractional step method converge to the solutions of system (5.1) when both the time step $\Delta t$ and the ratio $\Delta t/\epsilon$ tend to zero. Here the second condition is not fulfilled since we want to govern the time step with the CFL condition alone and when the particle radius is small the attenuation given by Method II is large and does not tend to zero with the particle radius as it should according to the theoretical formula (6.11.ii). Hence Method II can not be used even when low precision is required. On the contrary Method III gives excellent results for very small particle radius: here the attenuation coefficient given by method III indeed tends to zero with $r$. However this method is based on system (5.16) which was derived assuming that the relative velocity between the two phases is very small and this is definitely false when number $B$ becomes of order unity. The attenuation predicted by Method III is an increasing function of $B$, in contradiction with the theoretical formula (6.11.ii). Furthermore, in some applications, particles are included in some gas in order to reduce combustion instabilities and one tries to maximize the damping produced by the particles, *i.e.*, to have $B = 1$. Method III is thus of no help in the simulation of such devices. Method I gives excellent results in the whole range of numbers $B$. (See in particular the zoom in the small radius zone.)

The numerical solutions of a representative Riemann problem are computed for different particle radius using the three methods: the left and right values of the initial state are listed in table below:

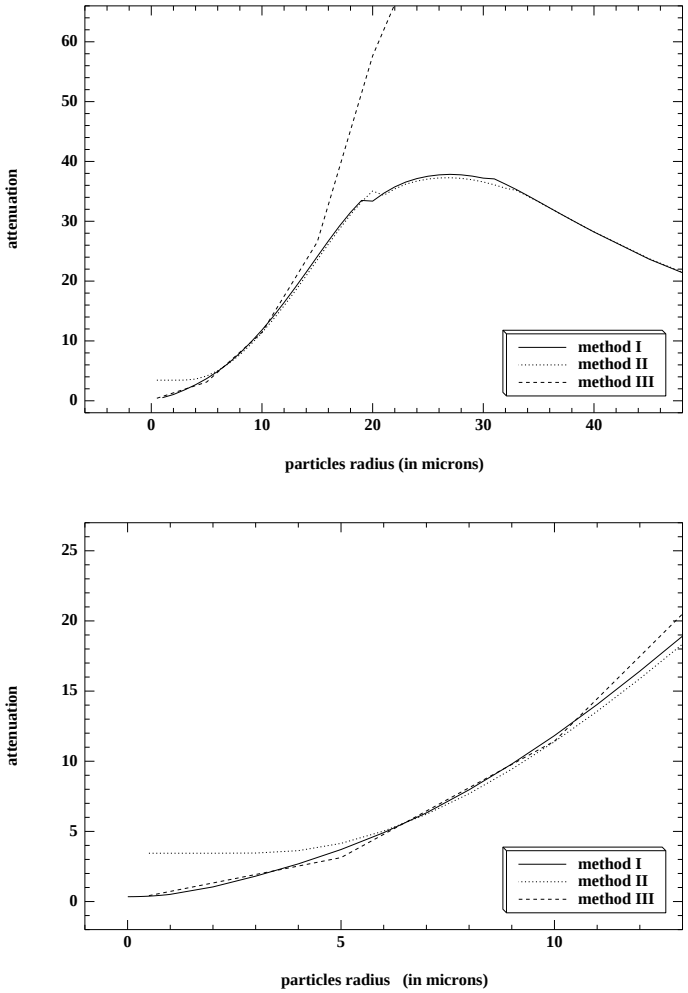


**Fig. 1.**

**Table 1.**

| | $\rho_g(\mathrm{kgm}^{-3})$ | $\alpha$ | $u_g\ (\mathrm{ms}^{-1})$ | $u_p\ (\mathrm{ms}^{-1})$ |
|---|---|---|---|---|
| $\mathbf{u}^L$ | 1. | .9991 | 5. | 10. |
| $\mathbf{u}^R$ | 2. | .991 | 10. | 5. |

The different physical constants used in the computation are:

$$\rho_p = 1766\,\mathrm{kg.m}^{-3},\ \mu_g = 8.8510^{-5}\ \mathrm{kg.m}^{-1}.\mathrm{s}^{-1},$$

$$K = 100000\,\mathrm{Pa},\quad \gamma = 1.4,\quad \theta_0 = 10000\mathrm{Pa}.$$

The particle radius are respectively

$$r_1 = 0.1\,10^{-6}\,\mathrm{m},\quad r_2 = 1.\,10^{-6}\,\mathrm{m},\quad r_3 = 10.\,10^{-6}\,\mathrm{m}.$$

We use 400 grid points and the CFL is $CFL = 0.6$.

The profile of the relative velocity between the two phases at time $t = 5.68\,10^{-4}$ s are given in Figs. 2–7: Figs. 2, 3 and 4 compare methods 1 and 3 for the three radii $r_1$, $r_2$ and $r_3$ respectively. Next Figs. 5, 6 and 7 compare methods 1 and 2 for the three radii $r_1$, $r_2$ and $r_3$ respectively. We observe that for large radius method 3 induces too much numerical attenuation of the waves while methods 1 and 2 give results of the same quality for $r_1$ and $r_2$. On the contrary method 2 induces too much attenuation for small radii (see Figs. 5 and 6) but gives almost the same profiles as method 1 for $r_3$. These results confirm the behavior observed in the case of the propagation of sound waves in a two-phase medium: method 2 induces a too important attenuation of the waves in the range of small $B$ numbers and gives correct profiles for larger $B$. On the contrary method 3 is very accurate for in the region of small $B$ numbers but definitely inaccurate when number $B$ is of order unity. Method 1 is very satisfactory in the whole range of numbers $B$ and is the best among the three methods tested in this paper.

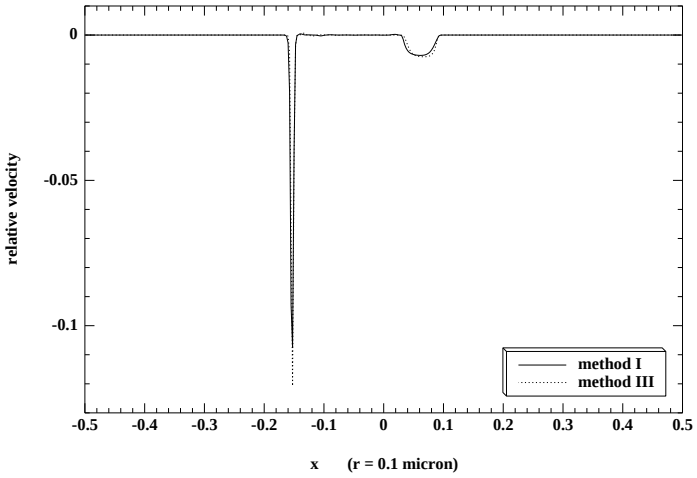


**Fig. 2.**

Finally Method I turns to be an excellent tool for the numerical computation of sound waves in a two-phase medium and gives the expected results in the solution of Riemann problems. The relaxation variables are accurately computed, no matter the stiffess of the source terms. This proves the necessity of using solutions of a fully coupled linearization of system (5.1) as the basis of a numerical scheme in order to obtain accurate solutions regardless of the stiffness.

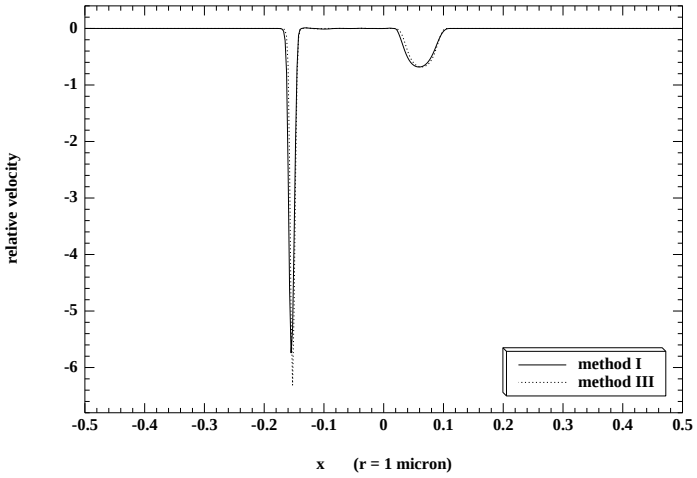


**Fig. 3.**

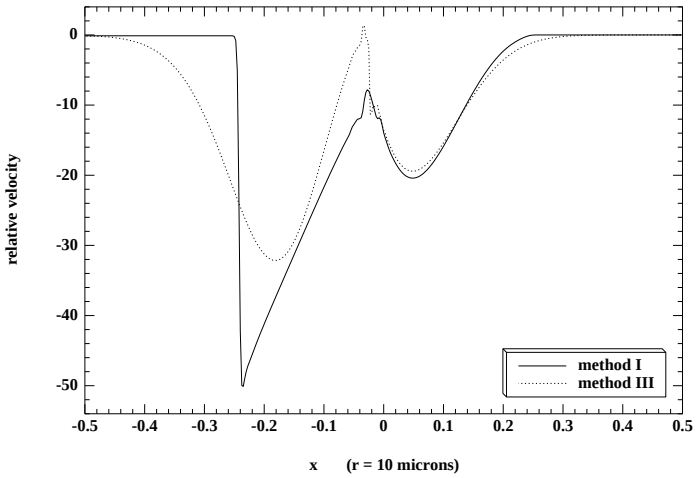


**Fig. 4.**

## Appendix A

This appendix is devoted to the proof of some properties of the solution of the generalized Riemann problem with constant coefficients (3.8): we prove the Propositions 3.2, 3.3 and, 3.4. We first deal with the global existence of a unique solution with bounded variations of (3.8): in Fourier variables system (3.8) takes the form of the following system of ODEs:

$$\frac{\partial \widehat{\mathbf{u}}}{\partial t} = -i\xi \mathbf{A}\widehat{\mathbf{u}} + \mathbf{B}\widehat{\mathbf{u}}$$

whose formal solution writes

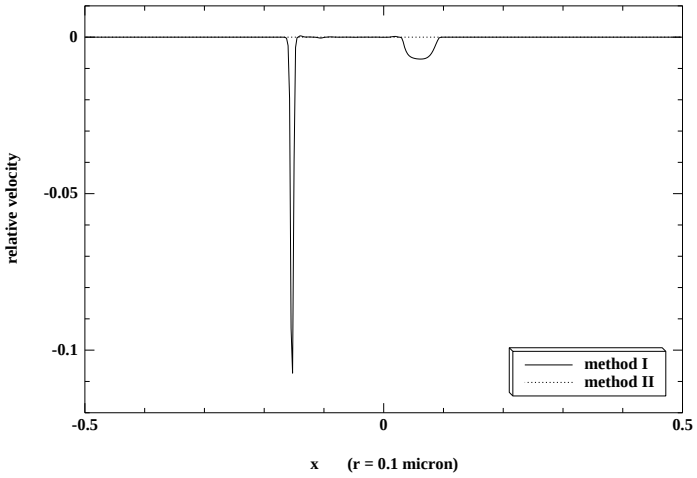


**Fig. 5.**

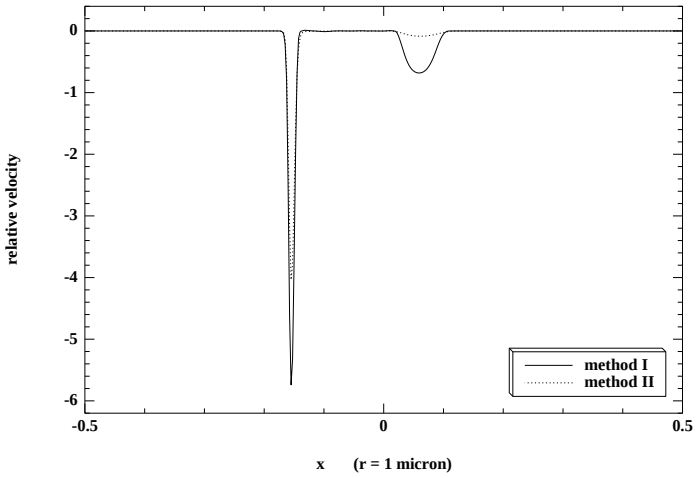


**Fig. 6.**

$$\widehat{\mathbf{u}}(\xi, t) = \exp\big(\big(-i\xi\mathbf{A} + \mathbf{B}\big)t\big)\widehat{\mathbf{u}}(\xi, 0).$$

But the Fourier transform of the initial is given by

$$(A.1) \qquad \widehat{\mathbf{u}}(\xi, 0) = -i\,\frac{\mathbf{u}^R - \mathbf{u}^L}{\xi} + \mathbf{u}^L\delta$$

where $\delta$ denotes Dirac's measure so that

$$(A.2) \qquad \widehat{\mathbf{u}}(\xi, t) = -i\,\exp\big(\big(-i\xi\mathbf{A} + \mathbf{B}\big)t\big)\cdot\frac{\mathbf{u}^R - \mathbf{u}^L}{\xi} + \exp\big(\mathbf{B}t\big)\mathbf{u}^L\delta.$$

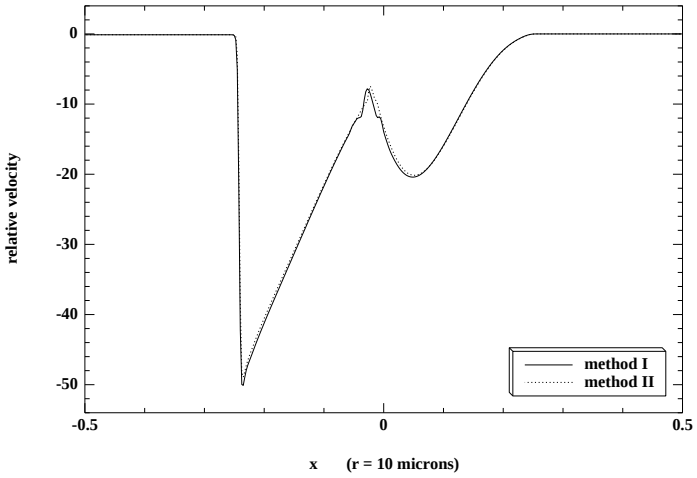


**Fig. 7.**

Recalling that the functions in $BV(\mathbb{R})$ are precisely the functions $x \to \mathbf{z}(x)$ whose Fourier transform $\xi \to \widehat{\mathbf{z}}(\xi)$ satisfies $\xi\widehat{\mathbf{z}}(\xi) \in L^\infty(\mathbb{R})$, we obtain that the function $\mathbf{u}$ defined by (A.2) belongs to the set $C^1\big([0,T], BV(\mathbb{R})\big)$ thanks to the following

**Lemma A.1.** *For any fixed $t$ the function $\xi \to \left|\exp\big(\big(-i\xi\mathbf{A}+\mathbf{B}\big)t\big)\right|$ is bounded over $\mathbb{R}$*

*and the proof of Proposition 3.2 is complete.* □

*Proof of Lemma A.1.* Let be given $\xi$ and let $\mu$ denote an eigenvalue of matrix $-i\xi\mathbf{A}+\mathbf{B}$: there exists a vector $\mathbf{s}$ with $|\mathbf{s}| = 1$ such that

$$\big(-i\xi\mathbf{A}+\mathbf{B}\big)\mathbf{s} = \mu\mathbf{s}.$$

But we have assumed the existence of a symmetrizer $\mathbf{S}$ for matrix $\mathbf{A}$: multiply on the left the above system by the vector $\mathbf{S}\bar{\mathbf{s}}$ where $\bar{\mathbf{s}}$ denotes the complex conjugate of vector $\mathbf{s}$: since $\mathbf{S}$ is symmetric, we get

$$-i\xi\bar{\mathbf{s}}\cdot\mathbf{SAs} + \bar{\mathbf{s}}\cdot\mathbf{SBs} = \mu\bar{\mathbf{s}}\cdot\mathbf{Ss}.$$

Since $\mathbf{S}$ is positive definite $\bar{\mathbf{s}}\cdot\mathbf{Ss}$ is a positive real number. Furthermore matrix $\mathbf{SA}$ is symmetric and $\bar{\mathbf{s}}\cdot\mathbf{SAs}$ is a real number. We deduce that

$$\mathrm{Re}(\mu) = \frac{\mathrm{Re}\,(\bar{\mathbf{s}}\cdot\mathbf{SBs})}{\bar{\mathbf{s}}\cdot\mathbf{Ss}}$$

and the latter expression is bounded independently of $\xi$. This concludes the proof of Lemma A.1. □

We prove next Proposition 3.3: *Proof of Proposition 3.3.* Let be given $\mathbf{u}^L$, $\mathbf{u}^R$ and let $\mathbf{z}$ denote the function defined by (3.13). In Fourier variables, $\mathbf{z}$ writes

$$\widehat{\mathbf{z}}(\xi,t) = -i\exp\left((-i\xi\mathbf{A}+\mathbf{B})\,t\right)\frac{\mathbf{u}^R-\mathbf{u}^L}{\xi} + i\sum_{k=1}^{p} a_k(t)\frac{\exp\big(-i\lambda_k t\big)}{\xi}$$

or, by (3.12),

$$(A.3) \qquad \widehat{\mathbf{z}}(\xi,t) = \frac{-i}{\xi}\sum_{k=1}^{p}\left[\exp\left((-i\xi\mathbf{A}+\mathbf{B})\,t\right)a_k^0 - a_k(t)\exp\left(-i\lambda_k t\right)\right]\mathbf{r}_k.$$

The behavior of the eigenvalues of matrix $-i\xi\mathbf{A}+\mathbf{B}$ is given by the following

**Lemma A.2.** *Let* $\mathbf{A}$ *be a given* $p\times p$ *matrix with* $p$ *distinct eigenvalues* $\lambda_k$, $1\le k\le p$ *and let* $\mathbf{B}$ *be a given matrix. Denote by* $\mathbf{r}_k$, $1\le k\le p$ *(resp.* $\mathbf{l}_l$, $1\le k\le p$*) the right (resp. left) eigenvectors of matrix* $\mathbf{A}$*. When* $\xi$ *is large enough, the matrix* $-i\xi\mathbf{A}+\mathbf{B}$ *has* $p$ *distinct eigenvalues* $\mu_k(\xi)$, $1\le k\le p$ *with the following asymptotic behavior:*

$$(A.4) \qquad \mu_k(\xi) = -i\xi\lambda_k + \mathbf{l}_k\cdot\mathbf{B}\mathbf{r}_k + \left(\frac{1}{\xi}\right).$$

*The associated right eigenvectors* $\mathbf{s}_k(\xi)$ *have the following behavior:*

$$(A.5) \qquad \mathbf{s}_k(\xi) = \mathbf{r}_k + O\left(\frac{1}{\xi}\right).$$

The proof of Lemma A.2 follows from the results in Wilkinson (1978), pp. 66-67.

We deduce from (A.5) that for $0\le t\le T$,

$$\begin{aligned}\exp\left((-i\xi\mathbf{A}+\mathbf{B})\,t\right)\mathbf{r}_k &= \exp\left((-i\xi\mathbf{A}+\mathbf{B})\,t\right)\mathbf{s}_k + \exp\left(\left(-i\xi\mathbf{A}+\mathbf{B}\right)t\right)\left(\mathbf{r}_k-\mathbf{s}_k\right)\\ &= \exp\left(\mu_k(\xi)t\right)\mathbf{s}_k + O\left(1/\xi\right).\end{aligned}$$

We insert next (A.4) in the latter estimate to deduce:

$$\exp\left(\left(-i\xi\mathbf{A}+\mathbf{B}\right)t\right)\mathbf{r}_k = \exp\left(-i\lambda_k t\right)\exp\left(\mathbf{l}_k\cdot\mathbf{B}\mathbf{r}_k t\right)\mathbf{r}_k + O\left(1/\xi\right)$$

or, by definition of the functions $a_k$, $1\le k\le p$:

$$\exp\left(\left(-i\xi\mathbf{A}+\mathbf{B}\right)t\right)\mathbf{r}_k = \exp\left(-i\lambda_k t\right)\frac{a_k(t)}{a_k^0}\mathbf{r}_k + O\left(1/\xi\right).$$

We deduce from (A.3) and the latter estimate that for $0\le t\le T$, the function $\xi\to\xi^2\widehat{\mathbf{z}}(\xi,t)$ is uniformly bounded, *i.e.*, that the function $\frac{\partial\mathbf{z}}{\partial x}$ belongs to the set $L^\infty\left([0,T],BV(\mathbb{R})\right)$. The proof of Proposition 3.3 is complete. □

*Proof of Proposition 3.4.* Let $\mathbf{u}$ denote the unique solution of (3.8) and set

$$\mathbf{z}(x,t) = \mathbf{u}\left(x+\lambda_p t,t\right) - \exp\left(\mathbf{B}t\right)\mathbf{u}^R.$$

According to Proposition 3.3, $\mathbf{z}$ is continuous on $]0,+\infty[$ for $t\ge 0$ and $\mathbf{z}(x,0)=0$, $x\ge 0$. Next let be given $\psi$, a positive decreasing function with $\psi(0)=1$ and $-\psi'(x)\le\psi(x)$, $x\ge 0$. Set

$$I(t) = \frac{1}{2}\int_0^{+\infty}\psi(x)\mathbf{z}(x,t)\cdot\mathbf{S}\mathbf{z}(x,t)\,\mathrm{d}x$$

where **S** is a symmetrizer of **A**. Since **z** is continuous on $]0,+\infty[$ and has bounded variations, we compute:

$$\begin{aligned}\frac{\mathrm{d}I}{\mathrm{d}t}(t) &= \int_0^{+\infty} \psi \mathbf{z}\cdot\mathbf{S}\frac{\partial \mathbf{z}}{\partial t}\,\mathrm{d}x \\ &= -\int_0^{+\infty} \psi \mathbf{z}\cdot\mathbf{S}\big(\mathbf{A}-\lambda_p\mathbf{1}\big)\frac{\partial \mathbf{z}}{\partial x}\,\mathrm{d}x + \int_0^{+\infty}\psi\mathbf{z}\cdot\mathbf{SBz}\,\mathrm{d}x \\ &= \frac{1}{2}\int_0^{+\infty}\frac{\partial\psi}{\partial x}\mathbf{z}\cdot\mathbf{S}\big(\mathbf{A}-\lambda_p\mathbf{1}\big)\mathbf{z}\,\mathrm{d}x + \int_0^{+\infty}\psi\mathbf{z}\cdot\mathbf{SBz}\,\mathrm{d}x \\ &+ \frac{1}{2}\big(\mathbf{z}\cdot\mathbf{S}\big(\mathbf{A}-\lambda_p\mathbf{1}\big)\mathbf{z}\big)(0^+,t).\end{aligned}$$

The matrices **S** and **SA** are indeed symmetric. Decompose next the function **z** on the eigenbasis $\mathbf{r}_k$, $1\le k\le p$ of **A**:

$$\mathbf{z}(x,t)=\sum_{k=1}^{p} z_k(x,t)\mathbf{r}_k.$$

Recalling that the left eigenvectors of matrix **A** write $\mathbf{l}_k = \mathbf{S}\mathbf{r}_k$, $1\le k\le p$, we compute:

$$\mathbf{z}\cdot\mathbf{S}\big(\mathbf{A}-\lambda_p\mathbf{1}\big)\mathbf{z}=\sum_{k=1}^{p}(\lambda_k-\lambda_p)z_k^2$$

and the latter expression is non positive. We deduce that

$$\frac{\mathrm{d}I}{\mathrm{d}t}(t)\le\frac{1}{2}\int_0^{+\infty}\frac{\partial\psi}{\partial x}\mathbf{z}\cdot\mathbf{S}\big(\mathbf{A}-\lambda_p\mathbf{1}\big)\mathbf{z}\,\mathrm{d}x+\int_0^{+\infty}\psi\mathbf{z}\cdot\mathbf{SBz}\,\mathrm{d}x$$

and since $|\psi'|\le\psi$, we get

$$\frac{\mathrm{d}I}{\mathrm{d}t}(t)\le CI(t)$$

for some positive number $C$. But $I(0)=0$ and we deduce that $I(t)\le 0$ for $t\ge 0$. This enables us to conclude that

$$\mathbf{u}(x,t)=\exp\big(\mathbf{B}t\big)\mathbf{u}^R,\quad x>\lambda_p t.$$

The proof of the identity

$$\mathbf{u}(x,t)=\exp\big(\mathbf{B}t\big)\mathbf{u}^L,\quad x<\lambda_0 t$$

is similar and the proof of Proposition 3.4 is complete. □

## Appendix B

We consider in this appendix the proof of Proposition 3.5: we consider the linear Riemann problem with constant coefficients (3.3) and we assume that the three

assumptions [H1], [H2] and [H3] hold true. Before we begin the proof let us notice that when [H1] holds, if a vector $\mathbf{s}$ i such that $\mathbf{Bs} = \mathbf{0}$, $\mathbf{s}$ writes

$$(B.1) \qquad \mathbf{s} = \mathbf{EQs}.$$

We compute indeed $\mathbf{Q}(\mathbf{s} - \mathbf{EQs}) = 0$ and $\mathbf{B}(\mathbf{s} - \mathbf{EQs}) = 0$ so that vector $\mathbf{w} = \mathbf{s} - \mathbf{EQs}$ belongs to the vector space $\ker \mathbf{B} \cap \ker \mathbf{Q}$. But since $\mathbf{BE} = 0$, we have that $\mathrm{range}\,(\mathbf{E}) = \ker(\mathbf{B})$ and we deduce from $\mathbf{QE} = \mathbf{1}_r$ that $\ker \mathbf{B} \cap \ker \mathbf{Q} = \ker \mathbf{Q} \cap \mathrm{range}\,\mathbf{E} = \{\mathbf{0}\}$, which gives (B.1).

On the other hand,

$$(B.2) \qquad \mathbf{Qq}_k = \mathbf{0},\ q+1 \le k \le p.$$

Indeed, apply on the left matrix $\mathbf{Q}$ to the identity $\mathbf{Bq}_k = \eta_k \mathbf{q}_k$: since $\mathbf{QB} = \mathbf{0}$, we get $\eta_k \mathbf{Qq}_k = 0$ and we deduce that $\mathbf{Qq}_k = 0$ since number $\eta_k$ is negative.

The proof of Proposition 3.5 relies on the study of the behavior of the eigenvalues of the matrices $-i\xi\mathbf{A} + \mathbf{B}/\tau$, at least for $\tau|\xi|$ small enough. When the $r$ negative eigenvalues of $\mathbf{B}$ are distinct, the perturbation Lemma A.1 gives $r$ distinct real negative eigenvalues of matrix $-i\xi\mathbf{A} + \mathbf{B}/\epsilon$, at least for $\epsilon|\xi|$ small enough. These eigenvalues are of order $1/\epsilon$ as $\epsilon$ tends to 0. Next zero is a multiple eigenvalue of $\mathbf{B}$ for which things are more complicated. In fact the splitting of this multiple eigenvalue depends on the matrix $\mathbf{A}$ and is determined with the ingredients used to perform the Chapman-Enskog expansion of system (3.1) (see Chen-Levermore-Liu (1992) or Bereux (1994) for instance). We first prove the

**Lemma B.1.** *We can find a positive $\beta_0$ such that for $\epsilon|\xi| \le \beta_0$, the matrix $-i\xi\mathbf{A} + \frac{1}{\epsilon}\mathbf{B}$ has $p - r$ eigenvalues $\mu_k(\xi,\epsilon)$, $r+1 \le k \le p$ with the following asymptotic behavior:*

$$(B.3.i) \qquad \left|\mu_k(\xi,\epsilon) - \frac{\eta_k}{\epsilon}\right| \le C|\xi|, \quad r+1 \le k \le p.$$

*The associated eigenvectors are denoted by $\mathbf{s}_k(\xi,\epsilon)$ and satisfy*

$$(B.3.ii) \qquad |\mathbf{s}_k(\xi,\epsilon) - \mathbf{q}_k| \le C\epsilon|\xi|, \quad r+1 \le k \le p.$$

The proof of Lemma B.1 follows from Lemma A.1 and assumption [H3]. We consider next the splitting of the multiple eigenvalue 0 of $\mathbf{B}$:

**Lemma B.2.** *We can find a positive number $\beta_0$ such that for $\epsilon|\xi| \le \beta_0$, matrix $-i\xi\mathbf{A} + \frac{1}{\epsilon}\mathbf{B}$ has $r$ distinct eigenvalues $\mu_k(\xi,\epsilon)$, $1 \le k \le r$ with the following behavior:*

$$(B.4.i) \qquad \left|\mu_k(\xi,\epsilon) + i\xi\widetilde{\lambda}_k\right| \le C(\beta_0)\epsilon\xi^2.$$

*The associated eigenvectors satisfy*

$$(B.4.ii) \qquad |\mathbf{s}_k(\xi,\epsilon) - \mathbf{E}\widetilde{\mathbf{r}}_k| \le C(\beta_0)\epsilon|\xi|.$$

*Proof of Lemma B.2.* We are looking for a pair $(\mu, \mathbf{s})$ solution of

$$(B.5) \qquad \Big( -i\xi\mathbf{A} + \frac{1}{\epsilon}\mathbf{B}\Big)\mathbf{s} = \mu\mathbf{s}$$

in the form of the following formal expansion:

$$\mu = \sum_{l\geq 0} \mu^l \epsilon^l, \quad \mathbf{s} = \sum_{l\geq 0} \mathbf{s}^l \epsilon^l.$$

At order $-1$ in $\epsilon$, we obtain that

$$\mathbf{B}\mathbf{s}_0 = 0$$

and (B.1) yields

$$(B.6) \qquad \mathbf{s}^0 = \mathbf{E}\mathbf{Q}\mathbf{s}^0.$$

We obtain next at zero-th order:

$$-i\xi\mathbf{A}\mathbf{s}^0 + \mathbf{B}\mathbf{s}^1 = \mu^0\mathbf{s}^0.$$

Apply matrix $\mathbf{Q}$ on the left to the later equation: recalling that by assumption $\mathbf{QB} = \mathbf{0}$, we deduce that $\mathbf{s}_0$ satisfies

$$-i\xi\mathbf{Q}\mathbf{A}\mathbf{s}^0 = \mu^0\mathbf{Q}\mathbf{s}^0.$$

By using (B.6) we deduce that vector $\widetilde{\mathbf{s}}^0 = \mathbf{Q}\mathbf{s}^0$ satisfies

$$-i\xi\widetilde{\mathbf{A}}\widetilde{\mathbf{s}}^0 = \mu\widetilde{\mathbf{s}}^0,$$

*i.e.,* that vector $\widetilde{\mathbf{s}}^0$ is either zero or an eigenvector of matrix $\widetilde{\mathbf{A}}$. When $\widetilde{\mathbf{s}}^0$ is zero, vector $\mathbf{s}^0 = \mathbf{E}\widetilde{\mathbf{s}}^0$ is itself zero and we are not interested in that case. On the contrary, when $\widetilde{\mathbf{s}}^0$ is non zero, we have

$$(B.7) \qquad \mathbf{s}^0 = \mathbf{E}\widetilde{\mathbf{r}}_k, \quad \mu^0 = -i\xi\widetilde{\lambda}_k$$

for some index $k \in [1, \ldots, r]$.

At zero-th order in $\epsilon$, the pair $\big(\mu^0, \mathbf{s}^0\big)$ is an eigenpair of matrix $\widetilde{\mathbf{A}}$. Let us now look for an eigenpair $\big(\mathbf{s}, \mu\big)$ of matrix $-i\xi\mathbf{A} + \frac{1}{\epsilon}\mathbf{B}$ in the form

$$(B.8) \qquad \mathbf{s} = \mathbf{s}^0 + \epsilon\xi\mathbf{w}, \quad \mu = -i\xi\left(\widetilde{\lambda}_k + \tau\epsilon\xi\right).$$

Equation (B.5) is then replaced by

$$(B.9) \qquad \mathbf{B}\mathbf{w} = i\big(\mathbf{A} - \widetilde{\lambda}_k\mathbf{1}\big)\mathbf{s}^0 + i\delta\big(\mathbf{A} - \widetilde{\lambda}_k\mathbf{1}\big)\mathbf{w} - i\tau\delta\mathbf{s}^0 - i\tau\delta^2\mathbf{w}$$

where we have set $\delta = \epsilon\xi$. Recalling that the range of matrix $\mathbf{B}$ is the kernel of $\mathbf{Q}$, equation (B.9) has a solution provided that

$$(B.10) \qquad \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k\mathbf{1}\big)\mathbf{w} = \tau\mathbf{Q}\mathbf{s}^0 + \delta\tau\mathbf{Q}\mathbf{w}.$$

According to the implicit functions theorem, system (B.9)-(B.10) has a one parameter family of solutions $\delta \to \big(\mathbf{w}(\delta), \tau(\delta)\big)$ provided that there exists a solution $\big(\mathbf{w}(0), \tau(0)\big)$ when $\delta = 0$ of system (B.9)-(B.10) and that system (B.9)-(B.10) determines the derivative $\big(\mathbf{w}'(0), \tau'(0)\big)$ of the one parameter family. Setting $\delta = 0$, system (B.9)-(B.10) takes the form:

$$(B.11) \qquad \mathbf{B}\mathbf{w} = i\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{s}^0, \quad \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{w} = \tau \mathbf{Q}\mathbf{s}^0.$$

Next the formal derivative taken in $\eta = 0$ of system (B.9)-(B.10) writes

$$(B.12) \qquad \mathbf{B}\mathbf{w}'(0) = i\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{w}(0) - i\tau(0)\mathbf{s}^0, \quad \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{w}'(0) = \tau'(0)\mathbf{Q}\mathbf{s}^0 + \tau(0)\mathbf{Q}\mathbf{w}(0).$$

Both systems (B.11) and (B.12) are solved thanks to the following

**Lemma B.3.** *Let be given a vector* $\mathbf{y}$ *in the range of* $\mathbf{B}$ *and two vectors* $\mathbf{z}^0$ *and* $\mathbf{z}^1$ *such that* $\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0 \neq 0$. *Then system*

$$(B.13) \qquad \mathbf{B}\mathbf{w} = \mathbf{y}, \quad \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{w} = \tau \mathbf{Q}\mathbf{z}^0 + \mathbf{Q}\mathbf{z}^1$$

*has a solution* $\big(\mathbf{w}, \tau\big)$. *Furthermore*

$$(B.14.i) \qquad |\tau| \leq C \frac{|\mathbf{y}| + |\mathbf{z}^1|}{|\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0|}$$

$$(B.14.ii) \qquad |\mathbf{w}| \leq C \left(1 + \frac{|\mathbf{z}^0|}{|\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0|}\right) \big(|\mathbf{y}| + |\mathbf{z}^1|\big).$$

Since $\mathbf{s}^0 = \mathbf{E}\widetilde{\mathbf{r}}_k$, we have $\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{s}_0 = \widetilde{\mathbf{l}}_k \cdot \widetilde{\mathbf{r}}_k = 1$. Next, by definition of $\mathbf{s}^0$, $\mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{s}^0 = 0$ so that the vector $i\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{s}^0$ belongs to the vector space range $\mathbf{B}$ and we can apply Lemma B.3 to the solution of system (B.11): we obtain the existence of a pair $\big(\mathbf{w}(0), \tau(0)\big)$ solution of (B.11). We apply again Lemma B.3 to obtain the existence of $\big(\mathbf{w}'(0), \tau'(0)\big)$ solution of (B.12). Indeed by (B.13), the vector $i\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}\big)\mathbf{s}^0$ belongs to range $(\mathbf{B})$. We may thus apply the implicit functions theorem to conclude the existence of a one parameter family $\big(\mathbf{w}(\delta), \tau(\delta)\big)$ of solutions of (B.9), at least for $|\delta|$ small enough or equivalently for $\epsilon|\xi|$ small enough. The corresponding solution of (B.5) then satisfies estimate (B.4). The proof of Lemma B.2 is complete. □

*Proof of Lemma B.3.* Let be given $\mathbf{y}$, $\mathbf{z}^0$ and $\mathbf{z}^1$ as in Lemma B.3. We write vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{w}^1 + \mathbf{w}^2$ where $\mathbf{w}^1 = \mathbf{E}\mathbf{Q}\mathbf{w}$ and $\mathbf{w}^2 = \big(\mathbf{1} - \mathbf{E}\mathbf{Q}\big)\mathbf{w}$. The first equation in (B.13) gives $\mathbf{B}\mathbf{w}^2 = \mathbf{y}$. But $\mathbf{Q}\mathbf{w}^2 = \mathbf{0}$ and because $\ker \mathbf{B} \cap \ker \mathbf{Q} = \{\mathbf{0}\}$, we obtain that $\mathbf{w}^2$ is uniquely determined and satisfies

$$\big|\mathbf{w}^2\big| \leq C\,|\mathbf{y}|.$$

Next, the second equation in (B.13) writes

$$\big(\widetilde{\mathbf{A}} - \widetilde{\lambda}_k \mathbf{1}_r\big)\mathbf{Q}\mathbf{w}^1 = \tau \mathbf{Q}\mathbf{z}^0 + \mathbf{Q}\mathbf{z}^1 - \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}_p\big)\mathbf{w}^2.$$

Since $\widetilde{\lambda}_k$ is an eigenvalue of matrix $\widetilde{\mathbf{A}}$, the latter equation has a solution only if

$$\tau \widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0 = \widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\big(\mathbf{A} - \widetilde{\lambda}_k \mathbf{1}_p\big)\mathbf{w}^2 - \widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^1.$$

But $\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0$ is by assumption non zero so that

$$|\tau| \leq C \frac{|\mathbf{y}| + |\mathbf{z}^1|}{|\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0|}.$$

We deduce that $\mathbf{w}^1$ satisfies:

$$\begin{aligned} |\mathbf{w}^1| &\leq C \left(|\tau|\,|\mathbf{z}^0| + |\mathbf{w}^2| + |\mathbf{z}^1|\right) \\ &\leq C \left(1 + \frac{|\mathbf{z}^0|}{|\widetilde{\mathbf{l}}_k \cdot \mathbf{Q}\mathbf{z}^0|}\right) \left(|\mathbf{y}| + |\mathbf{z}^1|\right). \end{aligned}$$

This concludes the proof of Lemma B.3. □

We can now conclude the proof of Proposition 3.5: let $t$ be fixed. According to Lemmas B.1 and B.2, for $\epsilon|\xi|$ small enough, matrix $-i\xi\mathbf{A} + \frac{1}{\epsilon}\mathbf{B}$ may be diagonalized in basis $\mathbf{s}_k(\xi,\epsilon)$, $1 \leq k \leq p$ of $\mathbb{R}^p$. For $\xi$ and $\epsilon$ given, decompose vector $\widehat{\mathbf{u}}(\xi,0)$ on the latter basis:

$$\widehat{\mathbf{u}}(\xi,0) = \sum_{k=1}^{p} \widehat{u}_k^{\epsilon}(\xi)\mathbf{s}_k(\xi,\epsilon).$$

Then,

$$\widehat{\mathbf{u}}(xit) = \exp\left(-i\xi t\mathbf{A} + \frac{t}{\epsilon}\mathbf{B}\right)\widehat{\mathbf{u}}(\xi,0) = \sum_{k=1}^{p} \exp\big(\mu_k(\xi,\epsilon)t\big)\widehat{u}_k^{\epsilon}(\xi)\mathbf{s}_k(\xi,\epsilon).$$

But, according to Lemma B.1, the eigenvalues $\mu_k(\xi,\epsilon)$, $r+1 \leq k \leq p$ tend to $-\infty$ and more precisely, we deduce from Lemma B.1 and (B.2) that

$$\left|\mathbf{Q}\widehat{\mathbf{u}}(\xi,t) - \sum_{k=1}^{r} \exp\big(\mu_k(\xi,\epsilon)t\big)\widehat{u}_k^{\epsilon}(\xi)\mathbf{Q}\mathbf{s}_k(\xi,\epsilon)\right| \leq C\epsilon|\xi|.$$

Next, according to Lemma B.2, for $1 \leq k \leq r$,

$$\exp\big(\mu_k(\xi,\epsilon)\big) = \exp\left(-i\widetilde{\lambda}_k\xi + O\left(\epsilon\xi^2\right)\right) = \exp\big(-i\widetilde{\lambda}_k\xi\big)\left(1 + O\left(\epsilon\xi^2\right)\right),$$

and we deduce that for $\epsilon|\xi|$ small enough,

$$\left|\mathbf{Q}\widehat{\mathbf{u}}(\xi,t) - \sum_{k=1}^{r} \exp\big(-i\xi\widetilde{\lambda}_k\big)\widehat{u}_k^{\epsilon}(\xi)\mathbf{Q}\mathbf{s}_k(\xi,\epsilon)\right| \leq C\left(\epsilon|\xi| + \epsilon\xi^2\right).$$

Inserting (B.4.ii) in the latter estimate gives

$$\left|\mathbf{Q}\widehat{\mathbf{u}}(\xi,t) - \sum_{k=1}^{r} \exp\big(-i\xi\widetilde{\lambda}_k\big)\widehat{u}_k^{\epsilon}(\xi)\widetilde{\mathbf{r}}_k\right| \leq C\left(\epsilon|\xi| + \epsilon\xi^2\right). \tag{B.15}$$

On the other hand estimates (B.3.ii) and (B.4.ii) give

$$\left| \mathbf{Q}\widehat{\mathbf{u}}(\xi,0) - \sum_{k=1}^{r} \widehat{u}_k^{\epsilon}(\xi)\widetilde{\mathbf{r}}_k \right| \leq C\,\epsilon|\xi|$$

and together with (B.15) we obtain:

$$(B.16) \qquad \left| \mathbf{Q}\widehat{\mathbf{u}}(\xi,t) - \exp\left(-i\xi\widetilde{\mathbf{A}}t\right)\mathbf{Q}\widehat{\mathbf{u}}(\xi,0) \right| \leq C\left(\epsilon|\xi| + \epsilon\xi^2\right).$$

But letting $\widetilde{\mathbf{u}}$ denote the solution of (3.19)-(3.20), we have

$$\widehat{\widetilde{\mathbf{u}}}(\xi,t) = \exp\left(-i\xi t\widetilde{\mathbf{A}}\right)\mathbf{Q}\widehat{\mathbf{u}}(\xi,0)$$

and estimate (3.21) is precisely (B.16). The proof of Proposition 3.5 is complete. □

**Table 2.** Table of figures

| Figure No | Type of calculation | Plotted quantity | Methods used | Radius ($10^{-6}$ m) |
|---|---|---|---|---|
| Figure 1 | Sound wave | Attenuation coefficient | Methods 1, 2, 3 | 0...50 |
| Figure 2 | Riemann problem | Relative velocity | Methods 1, 3 | 0.1 |
| Figure 3 | Riemann problem | Relative velocity | Methods 1, 3 | 1 |
| Figure 4 | Riemann problem | Relative velocity | Methods 1, 3 | 10 |
| Figure 5 | Riemann problem | Relative velocity | Methods 1, 2 | 0.1 |
| Figure 6 | Riemann problem | Relative velocity | Methods 1, 2 | 1 |
| Figure 7 | Riemann problem | Relative velocity | Methods 1, 2 | 10 |

## References

Ben-Artzi, M. (1989): The Generalized Riemann Problem for Reactive Flows. J. Comput. Phys, **81**, 70-101.

Bereux, F. (1994): Zero-Relaxation Limit versus Operator Splitting for Two-Phase Fluid Flows Computations. Submitted for publication in the European J. of mechanics.

Burman, E., Sainsaulieu, L. (1994): Numerical Analysis of Two Operator Splitting Methods for an Hyperbolic System of Conservation Laws with Stiff Relaxation Terms. Submitted for publication in Comp. Methods in Applied Mechanics and Engineering.

Chen, G. Q., Levermore, C. D. and Liu, T. P. (1992): Hyperbolic Conservation Laws with Stiff Relaxation Terms and Entropy. preprint.

Chen, G. C., Liu, T. P. (1993): Zero Relaxation and Dissipation Limits for Hyperbolic Conservation Laws. Comm. in Pure and Applied Math., **46**, 755-781.

Culick, F. (1981): Combustion instability in solid rocket motors. Volume 2: a guide for motor designers. Chemical Propulsion Information Agency publication $N^0$ **290**.

Godlewski, E., Raviart, P.-A. (1991): Hyperbolic Systems of Conservation Laws, Vol 1. Ed. Ellipses.

Godlewski, E., Raviart, P.-A. (1994): Hyperbolic Systems of Conservation Laws. Vol.2. Springer Verlag.

Harten, A., Lax, P., Van Leer B. (1981): On Upstream Differencing and Godunov Type Schemes for Hyperbolic Conservation Laws. SIAM Rev., **25**, 35-61.

Langseth, J. O., Tveito, A., Winther, R. (1993): On the Convergence of Operator Splitting Applied to Conservation Laws with Source Terms. Preprint 1993-1, Universitetet I Oslo, Institutt for Informatikk.

LeVeque, R., Yee, H. (1990): A Study of Numerical Methods for Hyperbolic Conservation Laws with Stiff Source Terms. J. Comput. Phys., **86**, 187-210.
Liu, T. P. (1988): Hyperbolic Conservation Laws with Relaxation. Comm. Math. Phys. **108**, pp. 153-175.
Pember, R. P. (1993) a: Numerical Methods for Hyperbolic Conservation Laws with Stiff Relaxation. I. Spurious Solutions. SIAM J. of applied Math., vol. 53, No. 5, 1293-1330.
Pember, R. P. (1993) b: Numerical Methods for Hyperbolic Conservation Laws with Stiff Relaxation. II. Higher Order Godounov Methods. SIAM J. Sci. Comput., Vol. 14, No. 4, 824-859.
Raviart, P.-A., Sainsaulieu, L. (1994): A Nonconservative Hyperbolic System Modeling Spray Dynamics. Part 1. Solution of the Riemann Problem. Accepted for publication in Mathematical Methods and Models in Applied Sciences.
Roe, P. L. (1981): Approximate Riemann Solvers, Parameter Vectors and Difference Schemes. J. Comput. Phys, **43**, 357-372.
Smoller, J. (1982): Shock waves and Reaction Diffusion Equations. Grundlehren der mathematischen Wissenschaft, No. 258, Springer Verlag, New-York.
Van Leer, B. (1977): Towards the ultimate conservative difference scheme: III. Upstream-centered finite-difference schemes for ideal compressible flows. J. Comp. Phys., **23**, 263-275.
Wilkinson, J. H. (1978): The Algebraic Eigenvalue Problem. Oxford Clarendon Press.
Williams, F. A. (1985): Combustion Theory. Benjamin Cumming.
Nessyahu, H., Tadmor, E. (1990): Non-Oscillatory Central Differencing for Hyperbolic Conservation Laws. J. Comp. Physics, **87**, 408-462.

This article was processed by the author
using the Springer-Verlag TEX QPMZGHB macro package 1991.