Contact Discontinuity Capturing Schemes for Linear Advection and Compressible Gas Dynamics

Bruno Després* Frédéric Lagoutière*

June 20, 2001

Abstract We present a non-diffusive and contact discontinuity capturing scheme for linear advection and compressible Euler system. In the case of advection, this scheme is equivalent to the Ultra-Bee limiter of [20], [24]. We prove for the Ultra-Bee scheme a property of exact advection for a large set of piecewise constant functions. We prove that the numerical error is uniformly bounded in time for such prepared (i.e. piecewise constant) initial data, and state a conjecture of non-diffusion at infinite time based on some local over-compressivity of the scheme for general initial data. We generalize the scheme to compressible gas dynamics and present some numerical results.

Keywords TVD limiter, contact discontinuity capturing scheme, linear advection, Euler system.

1 Introduction

This work deals with non-dissipative schemes for hyperbolic equations. For sake of simplicity we restrict theoretical developments to one dimensional hyperbolic problems, both in the linear and non linear ranges. Linear advection is the basic model that will serve to present the analysis.

Linear advection : *let us consider in one dimension ($x \in \mathbb{R}$, $t \in \mathbb{R}^+$) a constant velocity $a > 0$. The unknown is $u(x,t)$, solution of the equation*

$$\partial_t u + a\partial_x u = 0, \tag{1}$$

with a given initial value

$$u(x,0) = u^0(x) \quad \forall x \in \mathbb{R}. \tag{2}$$

Recall that the solution of this problem is

$$u(x,t) = u^0(x - at). \tag{3}$$

*Laboratoire d'analyse numérique, Université P. et M. Curie, 4, place Jussieu, 75005 Paris ; and Commissariat à l'Énergie Atomique, CEA/DIF, BP12, 91680 Bruyères-le-Châtel, FRANCE. tel : 01 69 26 52 44. e-mail : bruno.despres@cea.fr

⋆Commissariat à l'Énergie Atomique, CEA/DIF, BP12, 91680 Bruyères-le-Châtel, FRANCE. e-mail : lagoutie@ann.jussieu.fr

The question of non-dissipative schemes in conjunction with high order limited schemes has been extensively studied in the past ([20], [24], [25], [21]). Nowadays many researchers still investigate that question (cf. [2], [10], [15], [1],[19], [5], [12] and references therein); see also in the context of Discontinuous Galerkin Methods [4], [8], [3] and references therein.

This problem is a basic one, with no satisfactory general solution. It is an understatement to claim that the question of non-dissipative schemes has the greatest importance in numerical simulation and scientific computing. We take as a general rule that a numerical scheme should ideally respect two points which may be viewed as incompatible : a) a numerical method must have enough dissipation in order to be stable and to capture discontinuous solutions when applied to non-linear hyperbolic problems; b) on the other hand it is important to use a numerical method with as low numerical dissipation as possible, at least one order of magnitude below the real physical dissipation. To our opinion, a consequence of our work is that it is possible to design one order stable schemes, stable enough to capture discontinuous solutions with in some sense zero diffusion. In fact we will prove that for linear advection a particular scheme is exact for a "dense" set of discrete profiles. These discrete profiles consist essentially on piecewise constant functions, step functions (the minimum size of the step is three mesh points). Moreover the scheme is a one order scheme: the flux is chosen as close as possible to the **downwind** value of the numerical unknown. In the case of linear advection, this scheme is equivalent to the so called Ultra-Bee limiter (see [20] and [24]). Nevertheless we propose a constructive way for the derivation of the scheme: this derivation is different from the classical one (presented for instance in ([24])).

The paper is organized as follows. In section 2 we present a stability and consistency analysis for linear advection with constant velocity. We prove that it is equivalent to the limiter analysis and that the particular scheme we choose is the Ultra-Bee limiter. In section 3 we prove for this scheme a property of exact advection of a "dense" in L^1 set of functions, and propose a conjecture of convergence at infinite time. Numerical examples sustain the theoretical results. We generalize the scheme to compressible Euler system, in section 4, and explain why the scheme exactly captures contact discontinuities, what is visible on some numerical results. Finally we propose an extension of this analysis to dimension 2 in section 5 and conclude in section 6.

2 Linear advection

In order to present the main ingredient of our study let us consider a general scheme for the discretization of (1). We use the standard finite-volume-like discretization

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{\Delta x}(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n). \tag{4}$$

Here Δt and Δx are the time step and cell size. The value of the numerical solution at step n (time $n\Delta t$, $n \in \mathbb{N}$) and in cell j (abscissa $j\Delta x$, $j \in \mathbb{Z}$) is denoted as u_j^n and the updated value in the same cell is denoted as u_j^{n+1}. We assume that the initial condition is given by the constant mass approximation

$$u_j^0 = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u^0(x)dx$$

(we assume that $u^0 \in L^1_{loc}(\mathbb{R})$).

Using simplified notations, we rewrite (4) as

$$\bar{u}_j = u_j - a\frac{\Delta t}{\Delta x}(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}). \tag{5}$$

2.1 L^∞ and TVD stability requirements

We look at some values of the fluxes such that the following L^∞ stability condition and TVD (Total Variation Diminishing) condition is fulfilled (recall that $a > 0$)

$$m_j = \min(u_j, u_{j-1}) \leq \bar{u}_j \leq M_j = \max(u_j, u_{j-1}). \tag{6}$$

It is straightforward to check that (6) may be rewritten as

$$\bar{u}_j = \alpha_j u_j + (1-\alpha_j)u_{j-1} = u_j - D_{j+\frac{1}{2}}(u_j - u_{j-1})$$

where $0 \leq \alpha_j \leq 1$ and $0 \leq D_{j+\frac{1}{2}} = 1 - \alpha_j \leq 1$. Using this formulation we recover the Harten criterion for L^∞ stability and TVD (see [13]). It proves the

Lemma 1 A scheme as (5) which satisfies (6) is L^∞ stable and TVD.

However the updated value $\bar{u}_j$ depends on the fluxes $u_{j+\frac{1}{2}}$ and $u_{j-\frac{1}{2}}$: so we need to study some conditions on the fluxes $u_{j+\frac{1}{2}}$ such that (6) is true. In our opinion the main difficulty is to obtain the values of the fluxes in a somewhat constructive way. Next we present what we think to be such a constructive way.

An equivalent condition to (6) is

$$m_j \leq u_j - a\frac{\Delta t}{\Delta x}(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}) \leq M_j,$$

that is

$$u_{j-\frac{1}{2}} + \frac{\Delta x}{a\Delta t}(u_j - M_j) \leq u_{j+\frac{1}{2}} \leq u_{j-\frac{1}{2}} + \frac{\Delta x}{a\Delta t}(u_j - m_j), \quad \forall j. \tag{7}$$

This inequality is of no use for practical computation since all fluxes are coupled with these inequalities for $j-1$, j, $j+1$, ... The idea is then to derive some sufficient conditions (also written in inequality form) such that (7) is true. We propose to base these sufficient conditions on (8) and (9)

$$m_{j+1} \leq u_{j+\frac{1}{2}} \leq M_{j+1}, \quad \forall j \in \mathbb{Z}, \tag{8}$$

$$M_j + \frac{\Delta x}{a\Delta t}(u_j - M_j) \leq u_{j+\frac{1}{2}} \leq m_j + \frac{\Delta x}{a\Delta t}(u_j - m_j), \quad \forall j \in \mathbb{Z}. \tag{9}$$

The first inequality, (8), is simply the flux counterpart of (6) (a consistency condition), while we have just eliminated $u_{j-\frac{1}{2}}$ in the second. This simple manipulation is a constructive way for the derivation the fluxes. It is straightforward to check that (8) and (9) imply (7). The question now is "are (8) and (9) compatible ?" In other terms, "is this possible to take fluxes verifying the two inequalities (8) and (9) ?". The answer, positive, is given in the following theorem.

Theorem 1 The following two properties hold.

a) Assume the CFL condition

$$0 < a\frac{\Delta t}{\Delta x} \leq 1; \tag{10}$$

then

$$u_j \in [m_{j+1}, M_{j+1}] \cap \left[M_j + \frac{\Delta x}{a\Delta t}(u_j - M_j), m_j + \frac{\Delta x}{a\Delta t}(u_j - m_j) \right] \neq \emptyset.$$

Therefore it is possible to find a value of the flux $u_{j+\frac{1}{2}}$ such that (8) and (9) are true.

b) Inequalities (8) and (9) imply inequality (7) and (8).

Proof: it is obvious that $u_j \in [m_{j+1}, M_{j+1}]$ (from the definition of m_{j+1} and M_{j+1}). Thus we have only to prove that

$$u_j \in \left[M_j + \frac{\Delta x}{a\Delta t}(u_j - M_j), m_j + \frac{\Delta x}{a\Delta t}(u_j - m_j) \right].$$

We here prove that $u_j \leq m_j + \frac{\Delta x}{a\Delta t}(u_j - m_j)$, the other inequality $u_j \geq M_j + \frac{\Delta x}{a\Delta t}(u_j - M_j)$ can be proved with the same arguments.

We know that $u_j \geq m_j = \min(u_{j-1}, u_j)$. Under the CFL condition $a\frac{\Delta t}{\Delta x} \leq 1$, we have $a\frac{\Delta t}{\Delta x} - 1 \leq 0$, so that

$$u_j(a\frac{\Delta t}{\Delta x} - 1) \leq m_j(a\frac{\Delta t}{\Delta x} - 1).$$

This leads to

$$u_j a\frac{\Delta t}{\Delta x} \leq m_j a\frac{\Delta t}{\Delta x} + u_j - m_j,$$

and, provided that $a\frac{\Delta t}{\Delta x} > 0$, to

$$u_j \leq m_j + \frac{u_j - m_j}{a\Delta t/\Delta x},$$

which is the expected upper bound. The second point, b), is straightforward.

The consistency inequality (8) implies that the fluxes define what is called an essentially three points scheme (see [11]): an essentially three points schemes is such that if $u_j = u_{j+1}$ then $u_{j+\frac{1}{2}} = u_j = u_{j+1}$. As a consequence of theorem 1, lemma 1 and a standart convergence result for L^∞ stable and TVD schemes (see [11]), we have

Corollary 1 Under the CFL condition (10), a scheme satisfying (8) and (9) is essentially three points, L^∞ stable and TVD. Assume moreover that the fluxes are locally Lipschitz-continuous functions of their arguments $(u_j)_{j\in\mathbb{Z}}$. Then the numerical solution converges in L^1 to the unique solution of (1).

It turns out that it is possible to base the derivation of the fluxes for practical computations on formulae (8) and (9) since the inequalities for the fluxes are now decoupled.

2.2 Choice of the fluxes

We are interested mainly in non-dissipative schemes and we are not satisfied with the classical upwind scheme. In order to understand what this requirement implies let us look at the very simple situation where the initial solution u^0 is a Heavyside function

$$u_l^0 = 1, \quad \forall l \leq j, \text{ and } u_l^0 = 0, \quad \forall l > j. \tag{11}$$

Let us consider the situation where the time step is not the maximal time step. For example

$$a\frac{\Delta t}{\Delta x} = \frac{1}{3} < 1. \tag{12}$$

The initial condition is plotted on figure 1.

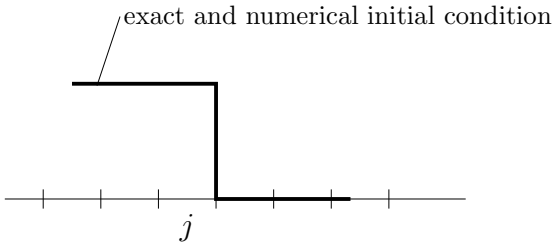

Figure 1: Initial condition.

For the initial condition (11), the exact solution at the first time step is $u^0(x - \frac{\Delta x}{3})$. After projection on the grid, it is

$$u_l^1 = \frac{1}{\Delta x}\int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} u^0(x - \frac{\Delta x}{3})dx,$$

that is

$$\begin{cases} u_l^1 = 1, & \forall l \leq j, \\ u_{j+1}^1 = \frac{1}{3}, \\ u_l^1 = 0, & \forall l > j+1, \end{cases} \tag{13}$$

represented on figure 2.

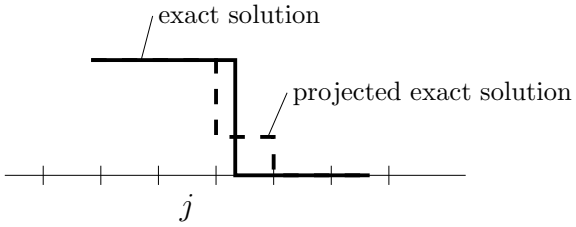

Figure 2: First time step.

At the second time step, the exact solution is $u^0(x-\frac{2\Delta x}{3})$ while the projected exact solution is

$$u_l^2 = \frac{1}{\Delta x}\int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} u^0(x - \frac{2\Delta x}{3})dx,$$

that is

$$\begin{cases} u_l^2 = 1, & \forall l \leq j, \\ u_{j+1}^2 = \frac{2}{3}, \\ u_l^2 = 0, & \forall l > j+1, \end{cases} \tag{14}$$

which is represented in figure 3.

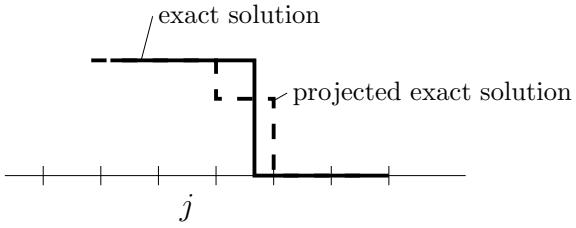

Figure 3: Second time step.

After a third time step, the exact solution is $u^0(x - \Delta x)$. Its projection on the grid is

$$u_l^3 = 1, \quad \forall l \leq j+1, \text{ and } u_l^3 = 0, \quad \forall l > j+1, \tag{15}$$

and is again equal to the exact solution 4.

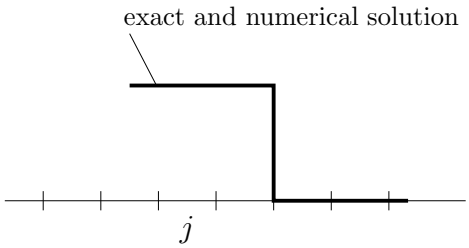

Figure 4: Third time step.

Now we forget that (11), (13), (14) and (15) are some cell-averages of the exact solution, and consider that these numerical profiles are given by a scheme as (4). If we try to define some numerical fluxes such that the scheme (4) applied to the initial condition (11) (resp. (13) or (14)) gives (13) (resp. (14) or (15)), a solution for the cell j is

$$u_{j-\frac{1}{2}}^{0,1,2} = 1 \text{ and } u_{j+\frac{1}{2}}^{0,1,2} = 0.$$

Indeed it implies

$$\begin{array}{ll} u_j^1 = \frac{1}{3}, & u_{j+1}^1 = 0, \\ u_j^2 = \frac{2}{3}, & u_{j+1}^2 = 0, \\ u_j^3 = \frac{2}{3}, & u_{j+1}^3 = 0. \end{array}$$

So we arrive at the conclusion that in this situation the "exact numerical flux", between cell j and cell $j+1$, is equal to the down-winded value of the exact solution, that is $u_{j+\frac{1}{2}}^{1,2} = u_{j+1}^{1,2} = 0$.

Now we raise this simple fact is a general principle for the choice of the numerical flux. The numerical flux will be chosen as closed as possible to the **down-winded** value of the numerical solution.

Gathering previous discussions of stability and consistency requirements we arrive at the following definition of the flux : $u_{j+\frac{1}{2}}$ is the closest value to the downwind value u_{j+1}, subjected to the upwind stability constraints (8) and (9):

$$\left\{ \begin{array}{l} m_{j+1} \leq u_{j+\frac{1}{2}} \leq M_{j+1}, \quad \forall j \in \mathbb{Z}, \\ M_j + \dfrac{\Delta x}{a\Delta t}(u_j - M_j) \leq u_{j+\frac{1}{2}} \leq m_j + \dfrac{\Delta x}{a\Delta t}(u_j - m_j), \quad \forall j \in \mathbb{Z}, \\ |u_{j+1} - u_{j+\frac{1}{2}}| \text{ is minimum}, \quad \forall j \in \mathbb{Z}. \end{array} \right. \tag{16}$$

Of course this definition is of practical use since the choice of the value of the flux is now decoupled from one interface to the other. Moreover, the minimization problem reduces to the explicit formula (17). Note

$$\left\{ \begin{array}{l} b_j = \max(m_{j+1}, M_j + \frac{\Delta x}{a\Delta t}(u_j - M_j)), \\ B_j = \min(M_{j+1}, m_j + \frac{\Delta x}{a\Delta t}(u_j - m_j)). \end{array} \right.$$

$$u_{j+\frac{1}{2}} = \left\{ \begin{array}{ccc} b_j & \text{if} & u_{j+1} < b_j, \\ u_{j+1} & \text{if} & b_j \leq u_{j+1} \leq B_j, \\ B_j & \text{if} & B_j < u_{j+1}. \end{array} \right. \tag{17}$$

The scheme defined with this formula and all the schemes derived from it will be called **limited downwind schemes**, or, more simply, **downwind schemes**.

The scheme defined in (17) is not monotone neither linear. Nevertheless we can very easily state a first result, lemma 2.

Lemma 2 Let $(u_j)_{j\in\mathbb{Z}}$ and $(v_j)_{j\in\mathbb{Z}}$ be two discrete functions and let us apply the limited downwind scheme (4), (17) to these functions.

Assume $v_j = cu_j + d \ \forall j \in \mathbb{Z}$ (with $c \in \mathbb{R}$, $d \in \mathbb{R}$); then $v_{j+\frac{1}{2}} = cu_{j+\frac{1}{2}} + d$ $\forall j \in \mathbb{Z}$, and so $\bar{v}_j = c\bar{u}_j + d \ \forall j \in \mathbb{Z}$.

The proof is obvious, the relations between the bounds for $u_{j+\frac{1}{2}}$ and $v_{j+\frac{1}{2}}$ being straightforward.

2.3 Link with limiters

It is interesting to make the link with the classical theory of limiters (in [13], [18], [20], [23], see also [24] and [11]).

Let us recall some basic facts about numerical schemes with limiters for the discretization of (1). In the limiter framework a general scheme is rewritten as

$$\bar{u_j} = u_j - \nu(u_j - u_{j-1}) - \frac{\nu(1-\nu)}{2}(\varphi_{j+\frac{1}{2}}(u_{j+1} - u_j) - \varphi_{j-\frac{1}{2}}(u_j - u_{j-1})), \quad (18)$$

where ν is the Courant number, $\nu = a\frac{\Delta t}{\Delta x}$. In the classical theory (cf. [23]) the coefficient $\varphi_{j+\frac{1}{2}}$ is a function of the rate of increase $r_{j+\frac{1}{2}}$ and the Courant number ν

$$\begin{array}{l} r_{j+\frac{1}{2}} = (u_j - u_{j-1})/(u_{j+1} - u_j), \\ \varphi_{j+\frac{1}{2}} = \varphi(r_{j+\frac{1}{2}}, \nu). \end{array} \quad (19)$$

Function φ is either subjected to the following limitations (if we allow the Courant number to enter the definition of φ)

$$\left\{ \begin{array}{l} \varphi(r,\nu) = 0 \quad \text{for } r \leq 0, \\ 0 \leq \varphi(r,\nu) \leq \min(\frac{2}{1-\nu}, \frac{2r}{\nu}), \end{array} \right. \quad (20)$$

or subjected to the following restrictions (if we do not want that the Courant number enters the definition of φ)

$$\left\{ \begin{array}{l} \varphi(r,\nu) = \varphi(r), \\ \varphi(r) = 0 \quad \text{for } r \leq 0, \\ 0 \leq \varphi(r) \leq \min(2, 2r). \end{array} \right.$$

A standard result of Sweby proves the following stability result (see [23]).

Theorem 2 The scheme (18), (19), (20) is L^∞ stable and TVD under the condition: $0 < \nu \leq 1$.

Let us recall briefly that the proof is based on a Harten reformulation

$$\bar{u}_j = u_j - D_{j+\frac{1}{2}}(u_j - u_{j-1})$$

with $0 \leq D_{j+\frac{1}{2}} \leq 1$.

Simple manipulations of our inequalities prove that our scheme is equivalent to the upper bound of (20)

$$\varphi(r,\nu) = \text{minmod}(\frac{2r}{\nu}, \frac{2}{1-\nu}) = \max(0, \min(\frac{2r}{\nu}, \frac{2}{1-\nu})). \tag{21}$$

Lemma 3 Let us consider linear advection at constant velocity $a > 0$. The scheme (5) defined by (16) is the same scheme than the Ultra-Bee scheme (18), (19), (21).

Proof: We here do not report the whole proof (see in [16]). It is only a matter of obvious calculus, which may be split in two steps.

- The Ultra-Bee scheme is a finite volume scheme in the form (4) with flux

$$u_{j+\frac{1}{2}} = u_j + (u_{j+1} - u_j)\max(0, \min((\frac{1}{a\Delta t/\Delta x} - 1)\frac{u_j - u_{j-1}}{u_{j+1} - u_j}), 1),$$

  so that it verifies

$$\frac{u_{j+\frac{1}{2}} - u_j}{u_{j+1} - u_j} = \max(0, \min((\frac{1}{a\Delta t/\Delta x} - 1)\frac{u_j - u_{j-1}}{u_{j+1} - u_j}), 1).$$

- The limited downwind fluxes verifies too

$$\frac{u_{j+\frac{1}{2}} - u_j}{u_{j+1} - u_j} = \max(0, \min((\frac{1}{a\Delta t/\Delta x} - 1)\frac{u_j - u_{j-1}}{u_{j+1} - u_j}), 1).$$

Theorem 2 is equivalent to our corollary 1.

3 An optimal property of the downwind scheme

It may appear as striking that we propose to take the Ultra-Bee scheme as a "good" scheme, since it is well known that the Ultra-Bee scheme is a first order scheme subjected to the over-compressive pathology, as it is the case for the Super-Bee scheme reported for example in [24].

Being first order means in the limiter framework that $\varphi(1,\nu) \neq 1$, while being over-compressive means that if we take, say, a Gaussian as initial solution, then after a "long" time (that is if we look at the numerical solution for large n) then the Gaussian is squared off. A square numerical profile has taken place of the initial Gaussian function.

In the following we would like to present a somewhat different discussion of this point. The conclusions that we will draw from the analysis will be completely different. We claim that on the contrary the downwind scheme is a very good scheme.

3.1 Exact advection of "piecewise constant" functions

The following result (new to our knowledge) states that the downwind scheme is an exact scheme for a "dense" in L^1 set of functions.

Theorem 3 Let us assume that the discrete function $(u_j)_{j\in\mathbb{Z}}$ is such that $\exists\alpha \in [0,1[$ such that $\forall j \in \mathbb{Z}$,

$$\left\{ \begin{array}{l} u_{3j+1} = u_{3j} \\ u_{3j+2} = \alpha u_{3j+1} + (1-\alpha)u_{3j+3}. \end{array} \right.$$

Then

- either $0 \leq \alpha + \nu < 1$ and for all j

$$\left\{ \begin{array}{l} \bar{u}_{3j+1} = \bar{u}_{3j} = u_{3j} \\ \bar{u}_{3j+2} = (\bar{\alpha})\bar{u}_{3j+1} + (1-\bar{\alpha})\bar{u}_{3j+3} \end{array} \right.$$

  with $0 \leq \bar{\alpha} = \alpha + \nu \leq 1$;

- or $1 \leq \alpha + \nu < 2$ and for all j

$$\left\{ \begin{array}{l} \bar{u}_{3j+2} = \bar{u}_{3j+1} = u_{3j+1} \\ \bar{u}_{3j+3} = (\bar{\alpha})\bar{u}_{3j+2} + (1-\bar{\alpha})\bar{u}_{3j+4} \end{array} \right.$$

  with $0 \leq \bar{\alpha} = \alpha + \nu - 1 \leq 1$.

The set of functions verifying hypothesis of theorem (3) is a set of step functions.

Proof: For sake of simplicity, we only prove the result for a one-step function:

$$\left\{ \begin{array}{ll} u_l = 1 & \forall l \leq j, \\ u_{j+1} = \alpha & \text{with } 0 \leq \alpha \leq 1, \\ u_l = 0 & \forall l \geq j+2. \end{array} \right.$$

We need to prove that

- either $0 \leq \alpha + \nu < 1$, then

$$\left\{ \begin{array}{ll} \bar{u}_l = 1 & \forall l \leq j, \\ \bar{u}_{j+1} = \bar{\alpha}, & \\ \bar{u}_l = 0 & \forall l \geq j+2. \end{array} \right.$$

  with $0 \leq \bar{\alpha} = \alpha + \nu \leq 1$;

- or $1 \leq \alpha + \nu < 2$, then

$$\left\{ \begin{array}{ll} \bar{u}_l = 1 & \forall l \leq j+1, \\ \bar{u}_{j+2} = \bar{\alpha}, & \\ \bar{u}_l = 0 & \forall l \geq j+3. \end{array} \right.$$

  with $0 \leq \bar{\alpha} = \alpha + \nu - 1 \leq 1$.

Firstly remark that the stability constraint impose that $\bar{u}_l = 1 \; \forall l \leq j$ and $\bar{u}_l = 0$ $\forall l \geq j+2$.

Let us compute the flux $u_{j+\frac{1}{2}}$. We have

$$\left\{ \begin{array}{l} m_j = 1, \\ M_j = 1 \end{array} \right.$$

so that

$$\left\{ \begin{array}{l} m_j + \frac{u_j - m_j}{\nu} = 1, \\ M_j + \frac{u_j - M_j}{\nu} = 1 \end{array} \right.$$

and necessarily

$$u_{j+\frac{1}{2}} = 1.$$

Now we compute the flux on the interface $j + 3/2$. We have

$$\left\{ \begin{array}{l} m_{j+1} = \alpha, \\ M_{j+1} = 1 \end{array} \right.$$

and

$$\left\{ \begin{array}{l} m_{j+1} + \frac{u_{j+1} - m_{j+1}}{\nu} = \alpha, \\ M_{j+1} + \frac{u_{j+1} - M_{j+1}}{\nu} = 1 + \frac{\alpha - 1}{\nu}, \end{array} \right.$$

and

$$\left\{ \begin{array}{l} m_{j+2} = 0, \\ M_{j+2} = \alpha. \end{array} \right.$$

So

$$\left\{ \begin{array}{l} b_{j+1} = \max(0, 1 + \frac{\alpha - 1}{\nu}), \\ B_{j+1} = \alpha. \end{array} \right.$$

So

- either $0 \leq \alpha + \nu < 1$, then $b_{j+1} = 0$, $u_{j+3/2} = 0$ (recall formula (17)), so that $\bar{u}_{j+1} = u_{j+1} - \nu(0 - 1) = \alpha + \nu = \bar{\alpha}$ and $\bar{u}_{j+2} = 0$;
- or $1 \leq \alpha + \nu < 2$, then $b_{j+1} = 1 + \frac{\alpha - 1}{\nu}$, $u_{j+3/2} = 1 + \frac{\alpha - 1}{\nu}$, so that $\bar{u}_{j+1} = u_{j+1} - \nu((1 + \frac{\alpha - 1}{\nu}) - 1) = \alpha - \nu - \alpha + 1 - \nu = 1$ and $\bar{\alpha_{j+2}} = 0 - \nu(0 - (1 + \frac{\alpha - 1}{\nu})) = \alpha + \nu - 1 = \overline{\alpha}$.

The result is proved for this simple step from 1 to 0. Now, using the lemma 2, and the fact that the scheme is essentially three-point (this property leading to the fact that there are no interactions between steps of two or more cells), we can extrapolate the result of the theorem. A complete and detailed proof can be found in [16].

Remark 1 Iterating this result on many time steps, we see that the steps ($u_{3j+1} = u_{3j}$) are perfectly transported at the right velocity (refer to [16] for a comprehensive proof, or remember the example of the Heavyside function in subsection 2.2).

Remark 2 In theorem 3, α is the same between each three points step. It is possible to replace three points steps ($u_{3j+1} = u_{3j}$, $u_{3j+2} = \alpha u_{3j+1} + (1 - \alpha)u_{3j+2}$), by four (or more) points steps with non-constant α ($u_{3j+2} = u_{3j+1} = u_{3j}$, $u_{3j+3} = \alpha_{3j+2}u_{3j+2} + (1 - \alpha_{3j+2})u_{3j+4}$).

3.2 Uniform bounds on the numerical error

Let u_j^n be a given numerical profile. We define $u^n(x)$ the function equal to u_j^n in each cell, that is $u^n(x) = \sum_{j\in\mathbb{Z}} u_j^n 1_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[}$.

Theorem 4 Consider an initial condition $u^0(x) \in L^1(\mathbb{R}) \bigcap BV(\mathbb{R})$ such that the discrete initial condition verifies

$$\left\{ \begin{array}{l} u_{3j+1}^0 = u_{3j}^0, \\ u_{3j+2}^0 = \alpha u_{3j+1}^0 + (1-\alpha)u_{3j+3}^0, \quad \text{for some } 0 \leq \alpha < 1, \end{array} \right. \tag{22}$$

Then, assuming the CFL condition $a\Delta t/\Delta x \leq 1$, one has

$$\|u^n(\cdot) - u(\cdot, n\Delta t)\|_{L^1(\mathbb{R})} \leq 3\Delta x TV(u^0)$$

where $TV(u^0)$ is the total variation of u^0.

This result means that the numerical error is bounded uniformly in time. It can be viewed as an infinite time error estimate. For all other known finite volume schemes, the bound $3\Delta x TV(u^0)$ is multiplied by an increasing function of the time (or of the number of iterates).

Idea of the proof This non-classical result relies essentially on theorem 3 and remark 1. It consists in remarking that from theorem 3, the step values are conserved for all time. Thus the error estimate can consist only in a shift between the exact and the numerical solution, and an additional term for the error in the intermediate values (linear combinations). This additional term is of course bounded by a constant multiplied by Δx. Then we have to evaluate the error due to the shift. A short analysis reveals that the numerical result does not depend on the time steps, that is, in particular, that the semi-continuous (continuous in time) scheme associated to (17) is equivalent to (17) (for a prepared datum). This leads to the fact that the shift is of less than one cell. The associated error is then bounded by a constant multiplied by Δx. All the details of the proof are reported in [16].

After many numerical experiments and theoretical investigations, we think that the following conjecture, dealing with non-prepared data, is true. For a given Δx we consider both the discrete numerical solution $(u_{\Delta x})_j^n$ and the associate functions $(u_{\Delta x})^n(x) = \sum_{j\in\mathbb{Z}} (u_{\Delta x})_j^n 1_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[}$, for each n (these functions are constant in every cell).

Conjecture 1 For all initial condition $u^0(x) \in L^1(\mathbb{R}) \cap BV(\mathbb{R})$, let us consider a sequence of initial approximation $(u_{\Delta x})^0$ with

$$\lim_{\Delta x \to 0} \|(u_{\Delta x})^0 - u^0\|_{L^1(\mathbb{R})} = 0. \tag{23}$$

Then,

$$\lim_{\Delta x \to 0} \left(\limsup_{n \to \infty} \|(u_{\Delta x})^n(\cdot) - u(\cdot, n\Delta t)\|_{L^1(\mathbb{R})} \right) = 0$$

(under CFL condition $0 < a\Delta t/\Delta x \leq 1$, and **adding a supplementary condition** $a\Delta t/\Delta x \neq 1/2$).

Remark 3 Of course if $(u_{\Delta x})^0$ is chosen in a set of numerical profiles satisfying the hypothesis of theorem 3, it is straightforward to prove the result. Moreover, it is always possible to chose a sequence $(u_{\Delta x})^0$ verifying both (22) and (23). This comes from the fact that every L^1 function can be approached by such a prepared discrete function. Step functions are in this meaning a dense set of L^1 (see [16] for a rigorous proof of this density result).

This is an infinite time convergence conjecture. In other words, the error for infinite time tends to 0 with the mesh size Δx, whatever the initial approximation is (even if unprepared). Note that the limit in time is before $\Delta x \to 0$. This result is false for all known classical methods. At present time we do not have a comprehensive proof of that conjecture. However it is possible to understand what is going on combining the results of theorems 3 and 5.

Let us consider for instance a smooth initial condition, that is $r_{j+\frac{1}{2}} \approx 1$. Due to theorem 5 (see in appendix A) the scheme is linearly unstable around all smooth profiles if $\nu \neq 1/2$. Linear instability appears and the smooth initial profile is replaced after a few time steps by a profile such that either $r_{j+\frac{1}{2}} \approx 0$ or $r_{j+\frac{1}{2}} \approx \pm\infty$. We have observed that the number of time steps is approximatively 30 or 40. In fact the numerical solution is then very close to a "piecewise constant" discrete profile described in theorem 3. Then the piecewise constant profile is perfectly transported.

So in some sense the scheme approximates the initial smooth profile by a piecewise constant approximation. It produces an error $C_1\Delta x$ in L^1 norm, where C_1 is function of the number of time steps needed by the scheme to replace the initial smooth profile by the piece-wise constant profile: this number of time steps seems to be more or less independent of Δx. Then the piecewise constant profile is perfectly transported with null error ($C_2\Delta x$ where $C_2 = 0$). So the total error at time T is $C(T)\Delta x$ with $C(T) = C_1 + C_2 = C_1$. The total error at time T is now independent of T !

Nevertheless the reader must be convinced that, even if we think this argument is correct, it is not at all a proof. For example we have observed that the Courant number ν has to be different from one half for the scheme to be able to locally and rapidly project smooth profiles on piece-wise constant profiles. If $\nu = \frac{1}{2}$ the scheme projects globally and slowly smooth profiles on piecewise constant profiles : we recover the squaring effect already reported for the Super-Bee scheme. We do not know the reason of that necessary condition $\nu \neq \frac{1}{2}$.

We may sum up this discussion considering two cases

a) If $a\Delta t/\Delta x \neq \frac{1}{2}$, the scheme is in some sense **locally over-compressive** : the conjecture seems to be true and all smooth smooth initial profiles are approximated by piecewise constant functions.

b) If $a\Delta t/\Delta x = \frac{1}{2}$, many numerical experiments show that the conjecture is not true. **Global over-compressivity** is present (like with the Super-Bee limiter).

3.3 Numerical results

We here present a few numerical results. They have been computed on the interval $[0, 1]$ with periodic boundary conditions in order to be able to observe the result at a very long time. We took for Courant number $\nu = 0.1$. The first one is the result for an indicatrix function. It illustrates the theorem of exact advection, and we see that this initial condition is not at all dissipated (figure 5). The second gives an idea why we conjecture the infinite-time convergence result (figure 6): the initial Gaussian has been transformed into a step Gaussian in a few time steps, and after has been exactly computed. Here, the result is given after 1000 periods, it is only indicative, the result being the same after any number of periods. It can be here compared to the result given by the Super-Bee scheme, showing its global over-compressive property.

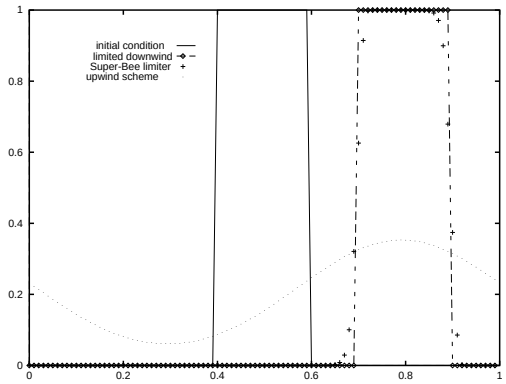

Figure 5: Initial condition and results for $t = 5.3$ (after 5.3 periods).

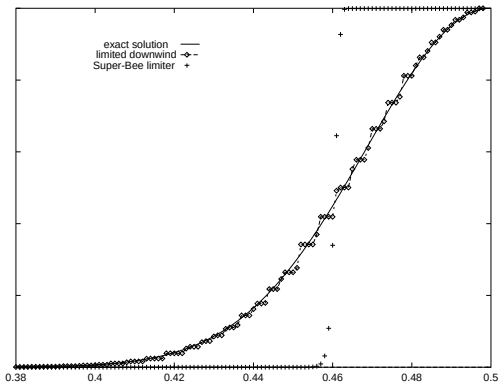

Figure 6: Initial condition and results for $t = 1000$ (after 1000 periods).

Remark 4

- All what we did here is for a non-negative velocity. Of course, a limited downwind scheme can be written equivalently for non-positive velocities.
- Furthermore, the formalism is able to treat of non-conservative non-constant velocity advection

$$\partial_t u + a(t,x)\partial_x u = 0,$$

discretized as follows :

$$u_j^{n+1} = u_j^n - a_j^n \frac{\Delta t}{\Delta x}(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n)$$

Because of theoretical difficulties about this kind of equations (existence of solutions...), we did not discuss about it and restricted ourselves to constant velocity advection. Nevertheless the following sections on compressible gas dynamics will show how to deal with this non-constant velocity transport.

4 Compressible gas dynamics

Now we discuss some application of the formalism previously introduced to compressible gas dynamics (Euler equations)

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t \rho u + \partial_x(\rho u^2 + p) = 0, \\ \partial_t \rho u e + \partial_x(\rho u e + p u) = 0. \end{cases} \tag{24}$$

The system is closed with a perfect gas law $p = (\gamma - 1)\left(\rho e - \frac{1}{2}\rho u^2\right)$, with $\gamma > 1$.

It is not simple to apply the non-dissipative approach to truly non-linear equations: the reason is the necessity of including some entropy constraints (see for example the approach developed in [17] for Burgers equation). So in the

following we would like to introduce a simplified approach. The non-dissipative scheme is applied only to the linearly degenerate part of (24). Let us lay stress on the fact that we do not know how to derive an Ultra-Bee limited scheme (see equation (21)) for such non-linear systems. The scheme we present in the following is not based on the Ultra-Bee formalism (21), but on the downwind formalism (17). We think that this new formalism is much more adapted to such non-trivial cases. The limited downwind scheme has the advantage of giving a contact-discontinuity-capturing scheme.

It is well known that (24) is made of 2 truly non-linear fields and one linear degenerate field. We would like to apply the very strong anti-dissipative scheme only to the degenerate linear field, and to apply another scheme (a diffusive one) to the truly non-linear part of the system. The hope is that the very strong anti-dissipative features of the scheme for the linear degenerate part will not interact with the truly non-linear part. It is important, since strong interaction might result in the capture of wrong shock rarefaction profiles.

Let us emphasize that there is *a priori* many solutions to numerically split (24) between truly non-linear parts and a linearly degenerate part : for example it is possible to use a Roe scheme as in [24]. However all our efforts to incorporate the previously discussed non-dissipative transport scheme in the Roe scheme failed, in the sense that it leaded to a very oscillating scheme, even for the calculation of pure contact discontinuities.

The approach that we present has the advantage of being "exact" and non-oscillating at contact discontinuities, at least for perfect gas laws.

Let us now present the scheme, which is split between 2 parts, one Lagrange part and one re-map part.

4.1 Lagrange part

Is is straightforward to prove that (24) is, for smooth solutions, equivalent to (25)

$$\left\{ \begin{array}{l} \rho D_t \tau - \partial_x u = 0, \\ \rho D_t u + \partial_x p = 0, \\ \rho D_t e + \partial_x p u = 0 \end{array} \right. \tag{25}$$

where $D_t = \partial_t + u\partial_x$ is the convective derivative, and $\tau = 1/\rho$ is the specific volume.

This is not exactly what is called the Lagrange reformulation of (24), but is quite similar to (see [12]).

So let us consider (25) and the following numerical scheme (26) of order one in time and space (introduced in [6] and [9])

$$\left\{ \begin{array}{l} \rho_j \frac{\tilde{\tau}_j - \tau_j}{\Delta t} - \frac{u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}}{\Delta x} = 0, \\ \rho_j \frac{\tilde{u}_j - u_j}{\Delta t} + \frac{p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}}{\Delta x} = 0, \\ \rho_j \frac{\tilde{e}_j - e_j}{\Delta t} + \frac{p_{j+\frac{1}{2}} u_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} u_{j-\frac{1}{2}}}{\Delta x} = 0. \end{array} \right. \tag{26}$$

The values of $p_{j+\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$ are given by

$$\left\{ \begin{array}{l} p_{j+\frac{1}{2}} = \frac{p_j+p_{j+1}}{2} + \frac{(\rho c)_{j+\frac{1}{2}}}{2}(u_j - u_{j+1}), \\ u_{j+\frac{1}{2}} = \frac{u_j+u_{j+1}}{2} + \frac{1}{2(\rho c)_{j+\frac{1}{2}}}(p_j - p_{j+1}), \end{array} \right. \tag{27}$$

where $(\rho c)_{j+\frac{1}{2}}$ is some local approximation of the density multiplied by the (local) sound velocity. In the following we will take

$$(\rho c)_{j+\frac{1}{2}} = \sqrt{\max(\rho_j {c_j}^2, \rho_{j+1} {c_{j+1}}^2) \min(\rho_j, \rho_{j+1})}.$$

Recall that this scheme, as the following lemma says, is entropy increasing.

Lemma 4 There exists constants $c_j > 0$ such that if $c_j \Delta t \leq \Delta x$ for all j, we have $\tilde{S}_j \geq S_j$ for all j.

See [9] for the proof.

4.2 Transport part

In some sense solving the quasi-Lagrangian system (25) means solving (24) in a referential which moves with the matter. It is therefore easy to derive the discrete equations in the transport part. We first recall the upwind transport part proposed in [6]: in a second stage we will abandon it. For sake of simplicity, we assume here that the velocities are positive $u_{j+\frac{1}{2}} > 0 \ \forall j \in \mathbb{Z}$. We obtain

$$\left\{ \begin{array}{l} \overline{\rho}_j = \tilde{\rho}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_j - \tilde{\rho}_{j-1} \right), \\ \overline{\rho}_j \overline{u}_j = \tilde{\rho}_j \tilde{u}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_j \tilde{u}_j - \tilde{\rho}_{j-1} \tilde{u}_{j-1} \right), \\ \overline{\rho}_j \overline{e}_j = \tilde{\rho}_j \tilde{e}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_j \tilde{e}_j - \tilde{\rho}_{j-1} \tilde{e}_{j-1} \right), \end{array} \right. \tag{28}$$

where $\tilde{\rho}_j = \frac{1}{tilde\tau_j}$, $\tilde{u}_j$ and $\tilde{e}_j$ are given by (26). After little algebra we combine (26) and (28) to obtain the global scheme

$$\left\{ \begin{array}{l} \overline{\rho}_j = \rho_j - \frac{\Delta t}{\Delta x} \left(\tilde{\rho}_j u_{j+\frac{1}{2}} - \tilde{\rho}_{j-1} u_{j-\frac{1}{2}} \right), \\ \overline{\rho}_j \overline{u}_j = \rho_j u_j - \frac{\Delta t}{\Delta x} \left(\tilde{\rho}_j \tilde{u}_j u_{j+\frac{1}{2}} - \tilde{\rho}_{j-1} \tilde{u}_{j-1} u_{j-\frac{1}{2}} + p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} \right), \\ \overline{\rho}_j \overline{e}_j = \rho_j e_j - \frac{\Delta t}{\Delta x} u_{j+\frac{1}{2}} \left(\tilde{\rho}_j \tilde{e}_j u_{j+\frac{1}{2}} - \tilde{\rho}_{j-1} \tilde{e}_{j-1} u_{j+\frac{1}{2}} + p_{j+\frac{1}{2}} u_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} u_{j-\frac{1}{2}} \right). \end{array} \right. \tag{29}$$

Remark 5 System (29) is conservative.

Numerical system (29) is formally consistent with (24). Be careful that we have assumed that the velocities $u_{j+\frac{1}{2}}$ are positive for all j : it is the reason of the up-winding $\tilde{\rho}_j \tilde{u}_j u_{j+\frac{1}{2}}$ and not $\tilde{\rho}_{j+1} \tilde{u}_{j+1} u_{j+\frac{1}{2}}$. Considering (28), the transport part may be reinterpreted as pure transport (as in (5)) with velocities given at the interfaces $u_{j+\frac{1}{2}} > 0$. And we have an entropy property for this upwind transport (see [7]):

Lemma 5 If $0 < u_{j+\frac{1}{2}}\Delta t \leq \Delta x$ for all j, we have $\overline{S}_j \geq \min\left(\tilde{S}_j, \tilde{S}_{j-1}\right)$.

Even if this scheme has the advantage of being entropy consistent, it has the drawback of being extremely diffusive. So we now introduce another possibility for the transport part, based on the general non-diffusive transport scheme. Instead of (26) we consider the more general scheme

$$\left\{\begin{array}{l} \overline{\rho}_j = \tilde{\rho}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}}\right), \\ \overline{\rho}_j \overline{u}_j = \tilde{\rho}_j \tilde{u}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_{j+\frac{1}{2}} \tilde{u}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}} \tilde{u}_{j-\frac{1}{2}}\right), \\ \overline{\rho}_j \overline{e}_j = \tilde{\rho}_j \tilde{e}_j - \frac{\Delta t}{\Delta x} u_{j-\frac{1}{2}} \left(\tilde{\rho}_{j+\frac{1}{2}} \tilde{e}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}} \tilde{e}_{j-\frac{1}{2}}\right), \end{array}\right. \tag{30}$$

where the fluxes are to be defined. The idea here is to find some L^∞ stability and TVD conditions on these fluxes and to down-wind all fluxes in (30) as much as possible, as we did for linear advection. A problem here is the correlation between ρ, ρu and ρe, leading to tricky computations. In order to deal with a simpler de-correlated problem, we impose that the the three fluxes are linked by

$$\left\{\begin{array}{l} \tilde{\rho}_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \tilde{\rho}_j + (1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1}, \\ \tilde{\rho}_{j+\frac{1}{2}} \tilde{u}_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \tilde{\rho}_j \tilde{u}_j + (1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1} \tilde{u}_{j+1}, \\ \tilde{\rho}_{j+\frac{1}{2}} \tilde{e}_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \tilde{\rho}_j \tilde{e}_j + (1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1} \tilde{e}_{j+1}. \end{array}\right. \tag{31}$$

In this system $\alpha_{j+\frac{1}{2}}$ is a linear combination coefficient, **the same for each of the three quantities ρ, ρu and ρe.** It remains to define $0 \leq \alpha_{j+\frac{1}{2}} \leq 1$ (this constraint being equivalent to consistency). Of course (28) corresponds to $\alpha_{j+\frac{1}{2}} = 0$ for all j. We choose this coefficient to be the largest as possible, provided that (16) is true for all variables, that is for u in (16) being replaced by ρ, ρu and ρe.

We do not write the complete algorithm, leaving all straightforward calculus to the reader, but prefer to explain what it consists in. We take the first equation of (28), and, as for advection equation, find a stability condition on ρ:

$$\left\{\begin{array}{l} {m^\rho}_j = \min(\tilde{\rho}_{j-1}, \tilde{\rho}_j), \\ {M^\rho}_j = \max(\tilde{\rho}_{j-1}, \tilde{\rho}_j), \end{array}\right.$$

and

$$\frac{\tilde{\rho}_j - {m^\rho}_j}{u_{j-\frac{1}{2}} \Delta t / \Delta x} + \tilde{\rho}_{j-\frac{1}{2}} \leq \tilde{\rho}_{j+\frac{1}{2}} \leq \frac{\tilde{\rho}_j - {M^\rho}_j}{u_{j-\frac{1}{2}} \Delta t / \Delta x} + \tilde{\rho}_{j-\frac{1}{2}}.$$

Now, adding a consistency condition ${m^\rho}_{j-1} \leq \tilde{\rho}_{j-\frac{1}{2}} \leq {M^\rho}_j \ \forall j \in \mathbb{Z}$, we obtain a sufficient condition,

$$\frac{\tilde{\rho}_j - {m^\rho}_j}{u_{j-\frac{1}{2}} \Delta t / \Delta x} + {m^\rho}_j \leq \tilde{\rho}_{j+\frac{1}{2}} \leq \frac{\tilde{\rho}_j - {M^\rho}_j}{u_{j-\frac{1}{2}} \Delta t / \Delta x} + {M^\rho}_j.$$

We do the same for each quantity, ρu and ρe. All inequalities (two for ρ, two for ρu and two for ρe) have their equivalent form on $\alpha_{j+\frac{1}{2}}$ thanks to (31) (the

consistency conditions leading to the only equation $0 \leq \alpha_{j+\frac{1}{2}} \leq 1$). Thus we can prove a result which is a generalization of theorem 1. This result states that there exists a non-empty interval $I_{j+\frac{1}{2}}$, containing 0, such that all fluxes in the interval $\alpha_{j+\frac{1}{2}} \in I_{j+\frac{1}{2}}$ define new values with $m^{\rho}{}_j \leq \overline{\rho}_j \leq M^{\rho}{}_j$ (with similar inequalities for $\overline{\rho u}_j$ and $\overline{\rho e}_j$). Of course a necessary condition is the CFL condition $\max_{j\in\mathbb{Z}} u_{j+\frac{1}{2}}\Delta t \leq \Delta x$, as in theorem 1. Now that the stability interval is found, we take the nearest value to 1 in this interval, that is we choose the largest $\alpha_{j+\frac{1}{2}}$ (following the idea of down-winding fluxes as most as possible).

For perfect gas law, the following result explains why this is a very attractive procedure.

Lemma 6 Assume that there exists 2 constants $u \in \mathbb{R}$ and $p \in \mathbb{R}^{+\star}$ such that $u_j = u$ and $p_j = p$ (it means that the fluid has a constant pressure and a constant velocity), and that $\rho_j > 0\ \forall j \in \mathbb{Z}$. Assume the CFL condition $|u|\Delta t/\Delta x \leq 1$ is verified.

Then whatever $0 \leq \alpha_{j+\frac{1}{2}} \leq 1$ is, the numerical solution of the scheme (26), (30) satisfies

$$\left\{ \begin{array}{l} \overline{\rho}_j = \rho_j - \frac{\Delta t}{\Delta x} u \left(\rho_{j+\frac{1}{2}} - \rho_{j-\frac{1}{2}} \right), \\ \overline{u}_j = \tilde{u}_j = u, \\ \overline{p}_j = \tilde{p}_j = p. \end{array} \right. \tag{32}$$

It means that if the initial condition corresponds to pure transport, then the numerical scheme (26)-(28) reduces to pure transport. Furthermore, in this case, the stability conditions on $\rho_{j+\frac{1}{2}} u_{j+\frac{1}{2}}$ and $\rho_{j+\frac{1}{2}} e_{j+\frac{1}{2}}$ are automatically verified if the conditions on $\rho_{j+\frac{1}{2}}$ are, so that the resulting scheme is the same as the one presented for advection at constant velocity. Consequently theorems 3 and 4 are valid.

In the case of a contact discontinuity characterized by equality of velocity and pressure from one side to the other of the discontinuity, it explains why contact discontinuities are not smeared over more than one cell by the downwind scheme proposed in this work. Note that one cell is of course the optimum.

Proof: the proof of lemma 6 is split in two steps.

Concerning the Lagrange step (26) there is no difficulty, since formulae (27) give $p_{j+\frac{1}{2}} = p$ and $u_{j+\frac{1}{2}} = u$.

The second step is the transport part (28). For sake of simplicity we assume that $u > 0$ as in the preceding. Of course the result is true even if $u \leq 0$.

Firstly let us prove that $\overline{u}_j = u$. We assumed that

$$\left\{ \begin{array}{l} \tilde{\rho}_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \tilde{\rho}_j + (1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1}, \\ \tilde{\rho}_{j+\frac{1}{2}} \tilde{u}_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \tilde{\rho}_j \tilde{u}_j + (1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1} \tilde{u}_{j+1}. \end{array} \right.$$

As a consequence of this,

$$\tilde{u}_{j+\frac{1}{2}} = \frac{[\alpha_{j+\frac{1}{2}} \tilde{\rho}_j] \tilde{u}_j + [(1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1}] \tilde{u}_{j+1}}{[\alpha_{j+\frac{1}{2}} \tilde{\rho}_j] + [(1 - \alpha_{j+\frac{1}{2}}) \tilde{\rho}_{j+1}]},$$

and this means that $\tilde{u}_{j+\frac{1}{2}}$ is a linear combination of $\tilde{u}_j$ and $\tilde{u}_{j+1}$, so that in the case of pure transport, $\tilde{u}_{j+\frac{1}{2}} = u$. Now taking the numerical transport equation (28), we write

$$\overline{u}_j(\tilde{\rho}_j - u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}})) = u(\tilde{\rho}_j - u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}})),$$

or

$$\overline{u}_j\overline{\rho}_j = u\overline{\rho}_j. \tag{33}$$

Under the hypothesis that $\rho_j > 0\ \forall j \in \mathbb{Z}$, we have $\overline{\rho}_j > 0\ \forall j \in \mathbb{Z}$ and then (33) reduces to

$$\overline{u}_j = u.$$

Now, to prove that $\overline{p}_j = p$, we recall that

$$\overline{p}_j = (\gamma-1)\overline{\rho}_j(\overline{e}_j - {\overline{u}_j}^2/2) = (\gamma-1)\overline{\rho}_j(\overline{e}_j - u^2/2).$$

It implies that

$$\begin{aligned}
\overline{p}_j &= (\gamma-1)(\rho_j e_j - u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}}\tilde{e}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}}\tilde{e}_{j-\frac{1}{2}}) \\
&\qquad - u^2/2(\rho_j - u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}} - \tilde{\rho}_{j-\frac{1}{2}}))) \\
&= (\gamma-1)(\rho_j(e_j - u^2/2) \\
&\qquad - u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}}(\tilde{e}_{j+\frac{1}{2}} - u^2/2) - \tilde{\rho}_{j-\frac{1}{2}}(\tilde{e}_{j-\frac{1}{2}} - u^2/2))) \\
&= p - (\gamma-1)u\Delta t/\Delta x(\tilde{\rho}_{j+\frac{1}{2}}(\tilde{e}_{j+\frac{1}{2}} - u^2/2) - \tilde{\rho}_{j-\frac{1}{2}}(\tilde{e}_{j-\frac{1}{2}} - u^2/2)).
\end{aligned}$$

With the linear combination hypothesis this means

$$\begin{aligned}
\overline{p}_j = p - (\gamma-1)u\Delta t/\Delta x(&\alpha_{j+\frac{1}{2}}(\tilde{\rho}_j\tilde{e}_j - u^2/2) + (1-\alpha_{j+\frac{1}{2}})(\tilde{\rho}_{j+1}\tilde{e}_{j+1} - u^2/2) \\
&- \alpha_{j-\frac{1}{2}}(\tilde{\rho}_{j-1}\tilde{e}_{j-1} - u^2/2) + (1-\alpha_{j-\frac{1}{2}})(\tilde{\rho}_j\tilde{e}_j - u^2/2)),
\end{aligned}$$

and finally

$$\overline{p}_j = p - (\gamma-1)u\Delta t/\Delta x(\alpha_{j+\frac{1}{2}}p + (1-\alpha_{j+\frac{1}{2}})p - \alpha_{j-\frac{1}{2}}p - (1-\alpha_{j-\frac{1}{2}})p) = p$$

as expected.

Remark 6 In this section as in the section for advection, we make the assumption that $u_{j+\frac{1}{2}} > 0\ \forall j \in \mathbb{Z}$. Of course it is possible to derive an equivalent algorithm in the case $u_{j+\frac{1}{2}} < 0\ \forall j \in \mathbb{Z}$. Furthermore, the stability constraints are local, and consequently it is possible to define the fluxes for every velocity repartition (even of non-constant sign).

4.3 Numerical results

We here give two numerical results, obtained for Riemann problems. In all cases we use a perfect gas law $p = (\gamma-1)\rho\varepsilon$ with $\gamma = 1.4$. Numerical results in 1D are computed with two meshes, the first one is a 100-cell mesh, the other is 500-cell mesh. The space interval is $[0, 1]$.

The first test case is the so-called Sod shock tube, with a discontinuous initial condition

$$\begin{cases} \rho^0(x) = 1 \text{ if } x \leq 0.5, & 0.125 \text{ otherwise,} \\ p^0(x) = 1 \text{ if } x \leq 0.5, & 0.1 \text{ otherwise,} \\ u^0(x) = 0 \ \forall x. \end{cases}$$

The time is $t = 0.14$.

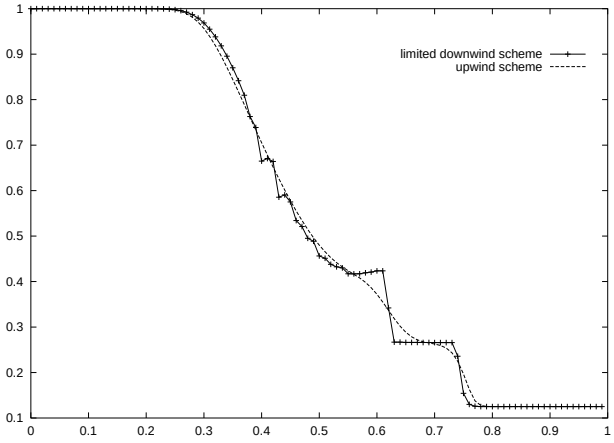

Figure 7: Density for Sod tube, with 100 cells.

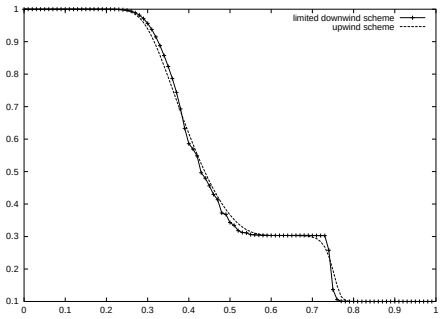

Figure 8: Pressure for Sod tube, with 100 cells.

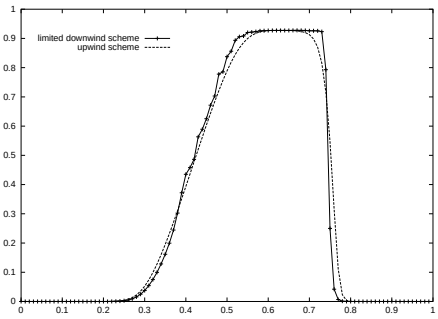

Figure 9: Velocity for Sod tube, with 100 cells.

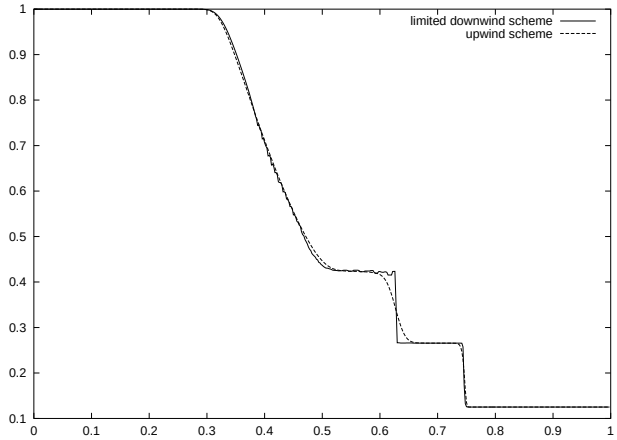

Figure 10: Density for Sod tube, with 500 cells.

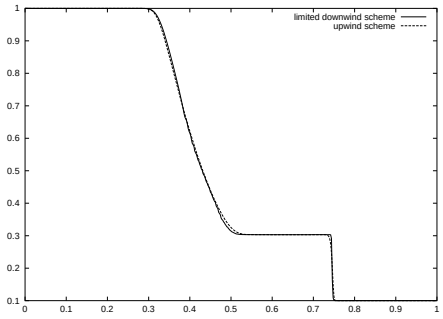

Figure 11: Pressure for Sod tube, with 500 cells.

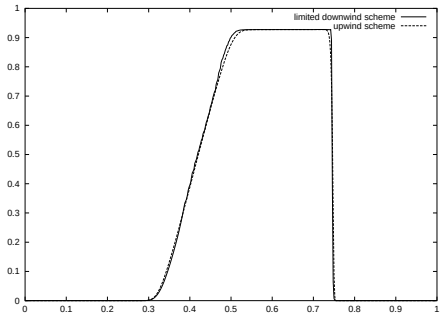

Figure 12: Velocity for Sod tube, with 500 cells.

We see (figures 7, 8 and 9 for upwind scheme and limited downwind scheme in the transport part) that the downwind scheme ($\alpha_{j+\frac{1}{2}} = \max$) "exactly" propagates the contact discontinuity; this discontinuity is located on one cell only and does not create velocity neither pressure oscillations. On the contrary the numerical results obtained with the upwind scheme (that is the Lagrange+remap scheme with upwind re-mapping $\alpha_{j+\frac{1}{2}} = 0$) is smeared at the contact discontinuity. This is a well known numerical diffusion phenomenon. To our knowledge almost all high order schemes also give this smearing effect. What is remarkable here with the downwind scheme is that the contact discontinuity is optimal and also that it remains optimal for $t \geq .14$. This optimality result is of course a consequence of lemma 6.

On the other hand, the shock is a little dissipated. The reason of this is that it is transported by a truly non-linear field, and the truly non-linear fields are solved with a classical algorithm (in the Lagrange part).

However we note here that the rarefaction is quite perturbed, being approx-

imatively turned into a step function. Recall that for this scheme we do not have any entropy property in the transport part.

On the figures 10, 11 and 12, we have the same quantities but on a refined mesh. It shows that the scheme seems to converge (even in the rarefaction wave), the contact discontinuity is optimal, with very few oscillations, for ρ. There are absolutely no oscillations for p and u. Most of oscillations in the rarefaction fan disappear with this refined mesh.

The second case we propose is the Lax shock tube:

$$\begin{cases} \rho^0(x) = 0.445 \text{ if } x \leq 0.5, & 0.5 \text{ otherwise,} \\ p^0(x) = 3.528 \text{ if } x \leq 0.5, & 0.571 \text{ otherwise,} \\ u^0(x) = 0.698 \text{ if } x \leq 0.5, & 0. \text{ otherwise.} \end{cases}$$

The time is $t = 0.15$.

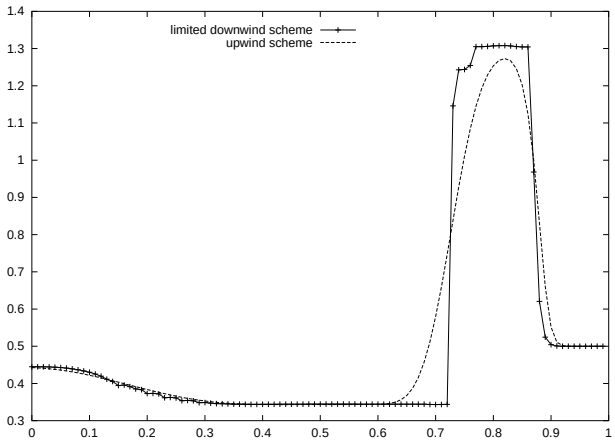

Figure 13: Density for Lax tube, with 100 cells.

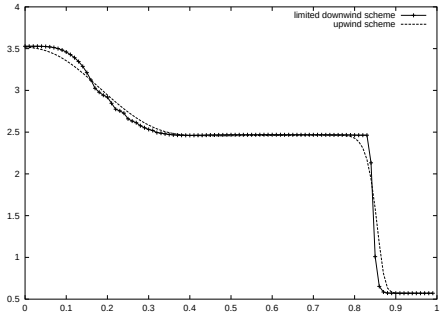

Figure 14: Pressure for Lax tube, with 100 cells.

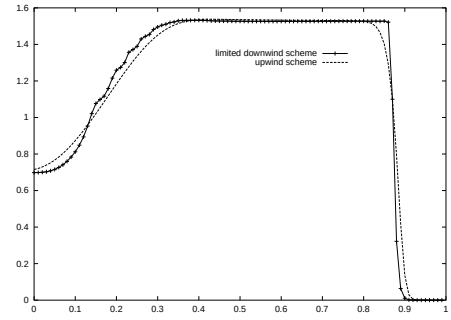

Figure 15: Velocity for Lax tube, with 100 cells.

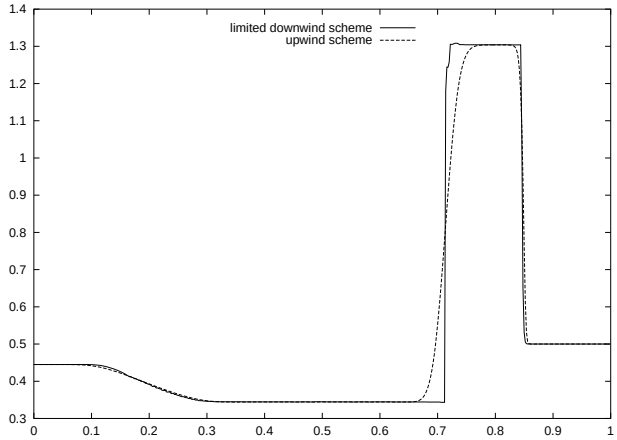

Figure 16: Density for Lax tube, with 500 cells.

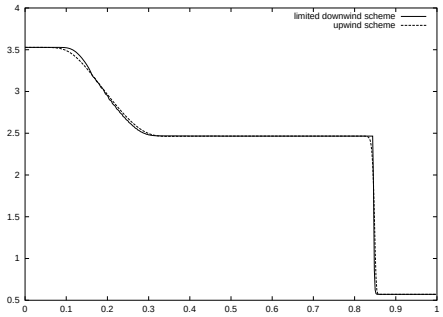

Figure 17: Pressure for Lax tube, with 500 cells.

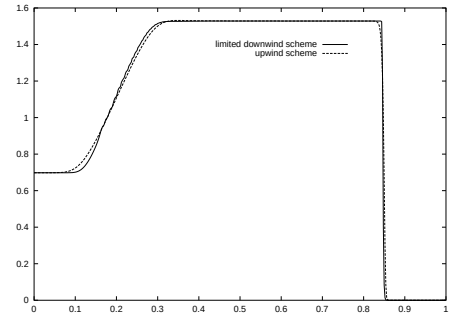

Figure 18: Velocity for Lax tube, with 500 cells.

We observe the same type of result. The contact discontinuity is well created and then exactly advected in the projection steps. The shock is here very fine and the rarefaction is quite good.

The results presented in this section for 1-D Euler equations are quite similar to those of Harten and Hyman in [14]. Using subcell computations they obtained for test cases exact contact discontinuities and shocks, but discontinuities in rarefaction profiles. The entropy modifications they proposed then has the drawback of spreading the contact discontinuity. Furthermore the present downwind scheme, as a finite-volume one, is much more simple, and very easy to transpose in 2-D.

5 Extension in dimension two

It is possible to derive 2-D algorithms from the previous 1D schemes very simply, using an alternating direction method (cf. [22]). We here do not report theory neither numerics for advection equation in 2-D but only present one numerical result for a divergent Sod tube for compressible Euler equations in 2-D,

$$\left\{\begin{array}{l} \partial_t \rho + \partial_x(\rho u) + \partial_y(\rho v) = 0, \\ \partial_t \rho u + \partial_x(\rho u^2 + p) + \partial_y(\rho u v) = 0, \\ \partial_t \rho v + \partial_x(\rho u v) + \partial_y(\rho v^2 + p) = 0, \\ \partial_t \rho u e + \partial_x(\rho u e + p u) + \partial_y(\rho v e + p v) = 0, \end{array}\right. \tag{34}$$

the system being closed with a perfect gas law $p = (\gamma - 1)\rho\left(e - \frac{1}{2}u^2 - \frac{1}{2}v^2\right)$, with $\gamma = 1.4$.

We do not write the scheme, which is a straightforward alternating direction extension of the one given in the previous section.

The considered test is a divergent Sod shock tube with circular initial discontinuity in $[0,1] \times [0,1]$:

$$\left\{\begin{array}{ll} \rho^0(x) = 1 \text{ if } r \leq 0.5, & 0.125 \text{ otherwise}, \\ p^0(x) = 1 \text{ if } r \leq 0.5, & 0.1 \text{ otherwise}, \\ u^0(x) = 0 \ \forall r, & \end{array}\right.$$

with $r = \sqrt{x^2 + y^2}$. The time is $t = 0.2$. We compare (on figure 19 and 20) the results given by the upwind and the limited downwind discretizations of the re-map part and observe the same behavior as in 1-D. The mesh is of 50×50 cells.

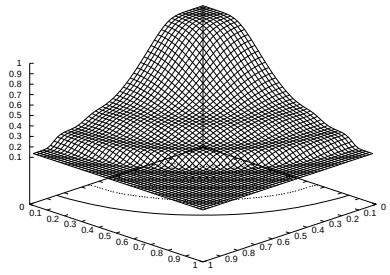

Figure 19: Density with upwind re-map part.

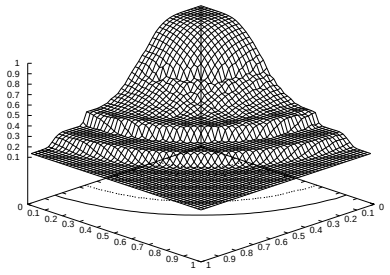

Figure 20: Density with downwind re-map part.

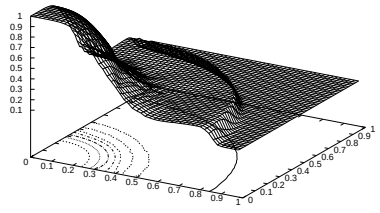

Figure 21: Pressure with downwind remap part (different view angle than for ρ).

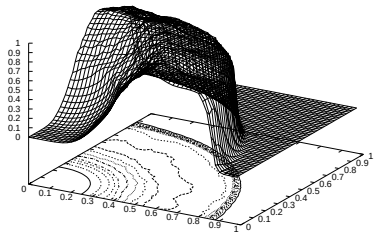

Figure 22: Velocity with downwind remap part (same view angle as for p).

While there is no oscillations on pressure and velocity, we note that, as in 1-D, the contact-discontinuity is optimal, in the sense that in every radial direction, it is spread on only one cell.

6 Conclusion

This work deals with numerical dissipation, an important question in general numerics for time-evolution problems. We proposed a new formalism giving some very natural and explicit conditions for a finite-volume scheme to be stable and convergent in the linear case of advection. Among all the convergent schemes that are taken into account in this formalism, we proposed to take the closest to the downwind scheme. We then analyze this given scheme, highlighting a very unusual property: an exact computation of advection for a set of initial condition that is dense in L^1 functions. What is even more amazing is the behavior of this scheme for any initial condition: the scheme seems to project this initial condition on that set of function, and then to advect it exactly. That is why we conjecture a result of infinite-time convergence.

Then we evaluate the possibility of transposing this analysis to any non-linear hyperbolic system. We here took the case of Euler equations for an ideal gas. We proposed a way to re-derive the same type of scheme, extracting the linear effects (due to the linearly degenerate field) with a Lagrange-projection splitting and state some good properties of this scheme, including the exact advection of (non-stationary) contact discontinuities. These results are illustrated on two classical shock tubes computations: the Sod and the Lax test cases.

The present analysis seems to us to be powerful. Extension of this approach to linear systems, multi-fluid flows and kinetic equations are in progress: see [16] for preliminary results on multi-fluids with interfaces.

References

[1] H. Wang, R. E. Ewing, G. Qin, S. L. Lyons, M. Al-Lawatia and S. Man. A family of eulerian-lagrangian localized adjoint methods for multi-dimensional advection-reaction equations. *Journal of Computational Physics*, 152:120–163, 1999.

[2] M. Arora and P. L. Roe. A well-behaved tvd limiter for high-resolution calculus of unsteady flow. *Journal of Computational Physics*, 132(5):3–11, 1997.

[3] G. E. Karniadakis B. Cockburn and C. W. Shu. The development of discontinuous galerkin methods. In Shu Cockburn, Karniadakis, editor, *Discontinuous Galerkin Methods*. Springer, Lecture Notes in Computational Science and Engineering, 1999.

[4] S. Hou B. Cockburn and C. W. Shu. TVB runge-kutta local projection discontinuous galerkin finite element method for conservation laws IV: the multidimensionnal case. *Math. of Comp.*, 1990.

[5] C. Chainais-Hillairet. First and second order schemes for a hyperbolic equation: convergence and error estimate. In Benkhaldoun and Vilsmeier editors, editors, *Finite volume for complex applications Problems and perspectives*, pages 137–144. Hermes Paris, 1997.

[6] B. Després. Inégalité entropique pour un solveur conservatif du système de la dynamique des gaz en coordonnées de Lagrange. *Comptes Rendus de l'Académie des Sciences, série 1*, 324:1301–1306, 1997.

[7] B. Després. Inégalités entropiques pour un solveur de type Lagrange + convection des équations de l'hydrodynamique. Technical report, CEA, 1997.

[8] B. Després. Discontinuous galerkin method for the numerical solution of euler equations in axisymmetric geometry. In Shu Cockburn, Karniadakis, editor, *Discontinuous Galerkin Methods*. Springer, Lecture Notes in Computational Science and Engineering, 1999.

[9] B. Després. Invariance properties of lagrangian systems of conservation laws, approximate Riemann solvers and the entropy condition. To appear in *Nümerische Mathematik*, 2001.

[10] J. Steinhoff, M. Fan and L. Wang. A new eulerian method for the computation of propagating sjort acoustic and electromagnetic pulses. *Journal of Computational Physics*, 157:683–706, 2000.

[11] E. Godlewski and P.-A. Raviart. *Hyperbolic systems of conservation laws*. Ellipses, 1991.

[12] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*. Springer, 1995.

[13] A. Harten. On a class of high resolution total-variation-stable finite-difference schemes. *SIAM Journal of Numerical Analysis*, 21(1):1–23, 1984.

[14] A. Harten and J. M. Hyman. Self adjusting grid methods for one-dimensional hyperbolic conservation laws. *Journal of Computational Physics*, 50:235–269, 1983.

[15] A. Kurnagov and G. Petrova. Central schemes and contact discontinuities. *M2AN*, 34(6):1259–1275, 2000.

[16] F. Lagoutière. *Modélisation mathématique et résolution numérique de problèmes de fluides compressibles à plusieurs constituants*. PhD thesis, Université Paris-VI, 2000.

[17] F. Lagoutière. Numerical resolution of scalar convex equations: explicit stability, entropy and convergence conditions. *submitted to ESAIM, proceedings of CEMRACS 1999*, 2001.

[18] B. Van Leer. Towards the ultimate conservative difference scheme, II. Monotonicity and conservation combined in a second-order scheme. *Journal of Computational Physics*, 14:361–376, 1974.

[19] R. J. LeVeque. *Numerical methods for conservation laws*. Birkhäuser, 1992.

[20] P. L. Roe. Some contribution to the modelling of discontinuous flows. *Lectures in Applied Mathematics*, 22:163–193, 1985.

[21] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77:439–471, 1988.

[22] J. C. Strikwerda. *Finite difference schemes and partial differential equations*. Wadsworth & Brooks, 1989.

[23] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal of Numerical Analysis*, 21(5):995–1011, 1984.

[24] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics. A practical introduction*. Springer, 1997.

[25] J.-P. Vila. High-order schemes and entropy condition for nonlinear hyperbolic systems of conservation laws. *Mathematics of Computation*, 50(181):53–73, 1988.

A Linear instability of the Ultra-Bee scheme

The scheme (16) is not linearly stable. In order to state such a property we consider an initial condition such that $r_{j+\frac{1}{2}} \approx 1$ or $r_{j+\frac{1}{2}} = 1$ (recall definitions introduced in subsection 2.3 for the limiters approach). This correspond to an almost affine initial condition.

Theorem 5 Consider the scheme (16), equivalent to the Ultra-Bee limiter. The following properties hold

Let us assume that $\nu < \frac{1}{2}$. Let us consider the linear scheme defined by $\varphi_{j+\frac{1}{2}} = \frac{2}{1-\nu}$ which is approximatively the value of the Ultra-Bee limiter for $r_{j+\frac{1}{2}} \approx 1$,

$$\bar{u}_j = u_j - \nu(u_j - u_{j-1}) - \nu((u_{j+1} - u_j) - (u_j - u_{j-1})). \qquad (35)$$

This scheme (35) is linearly unstable.

Let us assume that $\nu > \frac{1}{2}$. Let us consider the linear scheme defined by $\varphi_{j+\frac{1}{2}} = \frac{2r_{j+\frac{1}{2}}}{\nu}$ which is approximatively the value of the Ultra-Bee limiter for $r_{j+\frac{1}{2}} \approx 1$,

$$\bar{u}_j = u_j - \nu(u_j - u_{j-1}) - (1-\nu)((u_1 - u_{j-1}) - (u_{j-1} - u_{j-2})). \qquad (36)$$

This scheme (36) is linearly unstable.

In other words the scheme is linearly unstable around all smooth profiles characterized by $r_{j+\frac{1}{2}} \approx 1$.

Proof: In the first case ($\nu < \frac{1}{2}$), the scheme reduces (straightforward manipulations) to

$$\bar{u}_j = u_j - \nu(u_{j+1} - u_j),$$

what is the non-limited downwind scheme, and what is well-known to be unstable.

In the second case ($\nu > \frac{1}{2}$), the scheme reduces to

$$\bar{u}_j = u_{j-1} + (1-\nu)(u_{j-1} - u_{j-2}).$$

This scheme can be viewed as a 2-step scheme, the first step being up-winded with a Courant number $\nu_1 = 1$, what is known to be an exact scheme (the solution is right-shifted of one cell at every time step); the second step being a non-limited downwind scheme with Courant number $\nu_2 = -\nu$ (with a negative velocity), this step is of course highly linearly unstable.