Large time behavior of numerical solutions of scalar conservation laws

September 8, 2005

1 Introduction

We consider *entropy solutions* of the scalar hyperbolic conservation law

$$\partial_t u(t,x) + \partial_x f(u)(t,x) = 0 \quad \forall (t,x) \in \mathbb{R}^{+*} \times \mathbb{T} \tag{1}$$

where $f \in \mathcal{C}^1(\mathbb{R})$, that is *weak solutions* of (1) satisfying furthermore

$$\partial_t S_k(u)(t,x) + \partial_x G_k(u)(t,x) \leq 0 \quad \forall (t,x) \in \mathbb{R}^{+*} \times \mathbb{T}$$

for every $k \in \mathbb{R}$ with $S_k(u) = |u-k|$ and $G_k(u) = \mathrm{sgn}(u-k)\,(f(u) - f(k))$. The discussion is limited to the one-dimensional case for the simplicity of the presentation but the results are valid in arbitrary dimension, as will be noted at the end of the paper.

More precisely, we are interested in the *large time behavior* of periodic *numerical solutions* of these scalar hyperbolic conservation laws. This behavior has been investigated for continuous solutions, under the genuine nonlinearity hypothesis, in [11] and, in the case of 2×2 systems, in [5]. It is there proved that the entropy solution asymptotically converges towards a constant in space function. The case where f'' has a finite number of zeros has been investigated, with similar results, in [3], [8], [1], [2] and [4] for example. This of course does not stand in the linear case when f'' vanishes on a whole interval: then, some profiles can be exactly translated for arbitrary time. This paper explores the large time behavior on the numerical point of view for a general flux f. It is "known" that entropic schemes are dissipative and that their numerical diffusion leads the periodical solutions to be spread over the period in large time, **even in the linear case**. On the other hand, some schemes are not dissipative, but they are not entropic. The present analysis proposes a way to explain these behaviors and to prove

a link between the numerical entropy property and the infinite time limit solution.

As an illustration of the two different kinds of schemes mentioned above (entropic but dissipative, or non-dissipative but non-entropic), let us present some classical numerical results. We here consider the linear advection equation

$$\partial_t u + \partial_x u = 0 \tag{2}$$

in dimension 1 with periodical initial condition of period 1

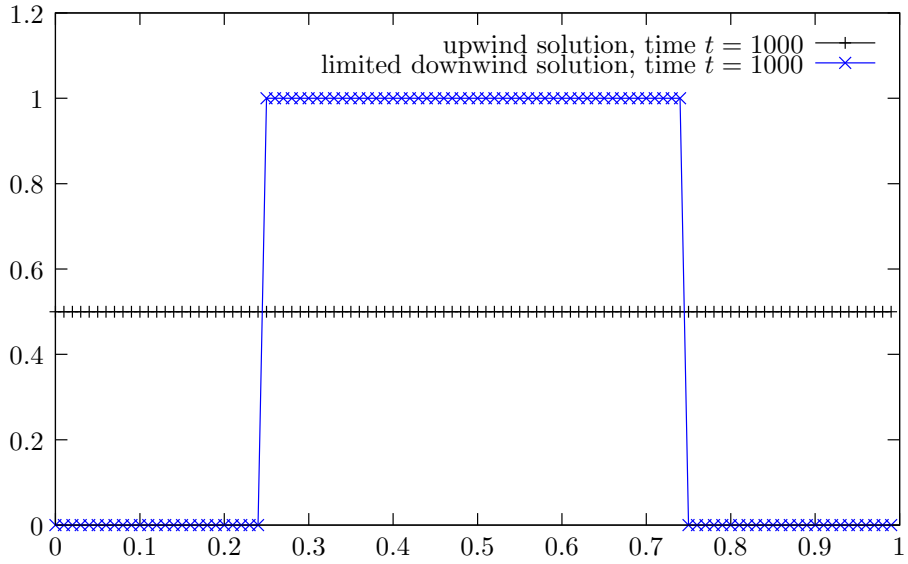
$$u^0(x) = \sum_{j\in\mathbb{Z}} \mathbb{1}_{[j+1/4,j+3/4)}(x).$$

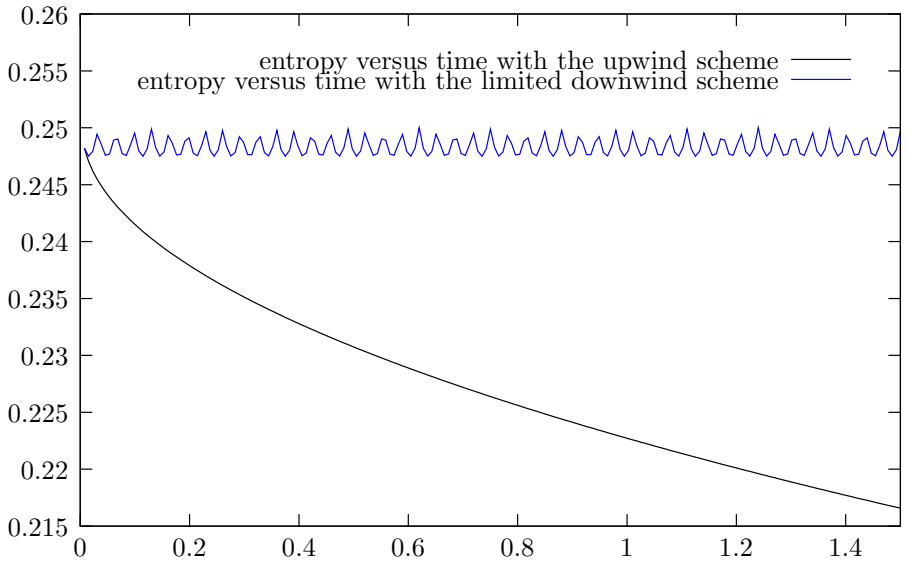
In the explicit finite volume framework, the exact transport PDE is replaced with the approximate one

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{u_{j+1/2}^n - u_{j-1/2}^n}{\Delta x} = 0 \quad \forall n \in \mathbb{N}, j \in \mathbb{Z},$$

where $\Delta t, \Delta x$ are the time and space steps and u_j^n stands for the approximate solution at time $n\Delta t$ in the cell $[(j-1)\Delta x, (j)\Delta x)$. The "numerical fluxes" $u_{j+1/2}^n$, computed with the help of the values of $\left(u_j^n\right)_{j\in\mathbb{Z}}$, define the finite volume scheme. The ratio $\Delta t/\Delta x$ is called the Courant number.

The chosen entropic scheme is the upwind scheme, where $u_{j+1/2}^n = u_j^n$ $\forall n \in \mathbb{N}, j \in \mathbb{Z}$. The non-dissipative one is the limited downwind scheme of [6], which is in this linear case equivalent to the Ultrabee limiter described in [13]. This scheme has been shown to be exact on characteristic initial conditions (and on some patches of characteristic functions, see [6]). We shall compare the solutions at time $t = 1000$ and the evolutions of the integral over a period of an entropy, let us choose the quadratic entropy $S(u) = u^2/2$. The Courant number is 0.7654321.





It is clear that while the upwind scheme is globally entropic (the integral of the entropy decreases at each time step) and has a bad behavior in large time, the limited downwind one, with good large time behavior, is not entropic: the integral of the entropy is oscillating in time.

The following shall give a precise analysis of the mathematical links between these features.

In section 2 we recall some classical results in linear algebra and ergodic theory that will be used after. Section 3 presents the principal result in terms of linear algebra. This is an ergodic result dealing with inhomogeneous products of square bistochastic matrices. Section 4 presents the application of the preceding results to finite volume entropic schemes for scalar conservation laws and proposes concluding remarks.

2 Preliminaries

We here recall notions and results that can be found in [12] and that will be essential in section 3.

Definition 1 A square matrix $A = (A_{i,j})_{(i,j)\in\{1,\dots,J\}^2} \in \mathcal{M}_J(\mathbb{R})$ is bistochastic if and only if

$$\begin{array}{l} A_{i,j} \geq 0 \quad \forall (i,j) \in \{1,\dots,J\}^2, \\ \sum_{j=1}^{J} A_{i,j} = 1 \quad \forall i \in \{1,\dots,J\}, \\ \sum_{i=1}^{J} A_{i,j} = 1 \quad \forall j \in \{1,\dots,J\}. \end{array}$$

Theorem 1 Let $x = (x_j)_{j=1}^{J}, y = (x_j)_{j=1}^{J} \in u\mathbb{R}^J$. The two following properties are equivalent.

$$\sum_{j=1}^{J} |y_j - k| \leq \sum_{j=1}^{J} |x_j - k| \quad \forall k \in \mathbb{R}. \tag{3}$$

$$\text{There exists a bistochastic matrix } A \in \mathcal{M}_J(\mathbb{R}) \text{ such that } y = Ax. \tag{4}$$

Theorem 2 (Birkhoff) A square matrix $A = (A_{i,j})_{(i,j)\in\{1,\dots,J\}^2} \in \mathcal{M}_J(\mathbb{R})$ is bistochastic if and only if it is a convex combination of permutation matrices.

3 Results

This section lists some results that shall be commented in relation with scalar partial differential equations in section 4.

Theorem 3 Let $J \in \mathbb{N}$. Let $(U^n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^J$ defined by the induction

$$\begin{array}{l} U^0 \in \mathbb{R}^J, \\ U^{n+1} = F(U^n) \quad \forall n \in \mathbb{N}, \end{array} \tag{5}$$

where U^0 and $F : \mathbb{R}^J \longrightarrow \mathbb{R}^J$ are given. Assume the sequence is such that

$$\sum_{j=1}^{J} \left| U_j^{n+1} - k \right| \leq \sum_{j=1}^{J} \left| U_j^n - k \right| \quad \forall k \in \mathbb{R} \tag{6}$$

with $U^n = \left(U_j^n \right)_{j=1}^{J} \ \forall n \in \mathbb{N}$.
Then,
(1) The sequence $(U^n)_{n\in\mathbb{N}}$ has at least 1 and at most $J!$ limit points.
(2) Let U, V be two limit points. There exists a permutation matrix $P_{U,V}$ such that $V = P_{U,V}U$.
Assume moreover F is continuous. Then,
(3) Let U be a limit point. There exists a permutation matrix P_U such that $F(U) = P_U U$.
(4) If N is the number of different limit points and U is a limit point, $F^N(U) = U$.

We here propose 2 different proofs. The second one is more simple but we think that the first one brings more comprehensible material.

First proof At first, recall that according to theorem 1, the family of inequalities (6) is equivalent to the existence of a *bistochastic* matrix $A^{(n)}$ such that

$$U^{n+1} = A^{(n)} U^n.$$

Now, by Birkhoff's theorem, we know that a matrix is bistochastic if and only if it is a convex combination of permutation matrices: thus U^{n+1} is a convex combination of all the vectors of $\mathbb{R}^J$ obtained by permuting the entries of U^n. For any $U \in \mathbb{R}^J$, let us denote by $C(U)$ the set of convex combinations of all the vectors obtained by permuting the entries of U:

$$C(U) = \operatorname{conv} \{ PU \text{ s.t. } P \text{ is a permutation matrix} \}.$$

For every $U \in \mathbb{R}^J$, $C(U)$ is a compact convex subset of $\mathbb{R}^J$. Birkhoff's theorem implies that $U^{n+1} \in C(U^n)$. A rather straightforward consequence is that $C(U^{n+1}) \subset C(U^n)$: see lemma 1 in the following. Figure 1 illustrates this fact with $J = 3$.

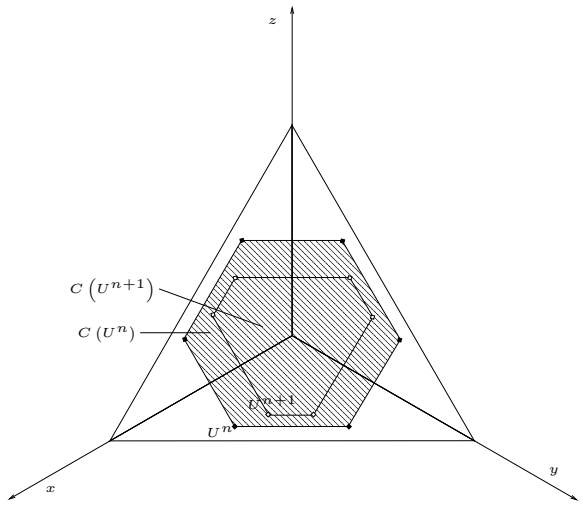


Figure 1: Inclusion of $C\left(U^{n+1}\right)$ in $C\left(U^n\right)$ in dimension 3.

From that, we deduce that $(C(U^n))_{n\in\mathbb{N}}$ is a non-increasing sequence of convex compact sets. Let us denote by C the limit: $C = \bigcap_{n\in N} C(U^n)$ is a convex compact subset of $\mathbb{R}^J$. On the other hand, the sequence $(U^n)_{n\in\mathbb{N}}$ is embedded in the compact set $C(U^0)$, so that there exists $U \in \mathbb{R}^J$ and a subsequence $\left(U^{k(n)}\right)_{n\in\mathbb{N}}$ converging towards U (there exists at least one limit point, this is the first claim in (1)). We shall show that $C(U) = C$.
Let us prove first that $C(U) \subset C$. From the fact that $(C(U^n))_{n\in\mathbb{N}}$ is a sequence of nested sets, it is clear that if $x \in C(U)$, $x \in C(U^n)$ $\forall n \in \mathbb{N}$. Thus $x \in C$.
Conversely, assume $x \in C$. It means that $x \in C(U^n)$ $\forall n \in \mathbb{N}$, so $x \in C(U^{k(n)})$ $\forall n \in \mathbb{N}$: x is a convex combination of the $J!$ vectors obtained by permutation of the coordinates of $U^{k(n)}$. In other words, $\forall n \in \mathbb{N}$, there exists $\alpha^{k(n)} = \left(\alpha_1^{k(n)}, \ldots, \alpha_{J!}^{k(n)}\right) \in [0,1]^{J!}$ verifying

$$x = \sum_{i=1}^{J!} \alpha_i^{k(n)} P_i U^{k(n)} \quad \forall n \in \mathbb{N},$$
$$\sum_{i=1}^{J!} \alpha_i^{k(n)} = 1 \quad \forall n \in \mathbb{N}$$

where $\{P_i \text{ s.t. } i \in \{1, \ldots, J!\}\}$ is the set of permutation matrices of $\mathcal{M}_J(\mathbb{R})$. The sequence $\left(\alpha^{k(n)}\right)_{n\in\mathbb{N}}$ stays in the compact $[0,1]^{J!} \in \mathbb{R}^{J!}$. One can thus extract a subsequence $\left(\alpha^{l(k(n))}\right)_{n\in\mathbb{N}}$ converging towards $\alpha = (\alpha_1, \ldots, \alpha_{J!}) \in$

$[0,1]^{J!}$ verifying $\sum_{i=1}^{J!} \alpha_i = 1$. Now, note that $U^{l(k(n))}$ converges to U as a subsequence of a converging sequence, and that, due to the continuity of the permutations, $P_i U^{l(k(n))}$ converges to $P_i U$ $\forall i \in \{1, \ldots, J!\}$. Finally,

$$x = \sum_{j=1}^{J!} \alpha_i P_i U,$$

x is a convex combination of the $J!$ vectors obtained by permutation of the coordinates of U, $x \in C(U)$.

Now, let $V \in \mathbb{R}^J$ be the limit of another converging subsequence of $(U^n)_{n\in\mathbb{N}}$. We then have $C(V) = C$, and consequently $C(U) = C(V)$. Lemma 2 then ensures that there exists a permutation matrix $P_{U,V}$ such that $V = P_{U,V}U$: point (2) is proved.
Thus there exists a most $J!$ different limit points of $(U^n)_{n\in\mathbb{N}}$. Point (1) of theorem is proved.

We now assume the continuity of F. Thus, if U is a limit point, $F(U)$ is a limit point too. Point (2) then ensures that there exists a permutation matrix $P_{U,F(U)}$ here denoted as P_U such that $F(U) = P_U U$. This proves point (3).

It remains to prove point (4). First remark that if U is a limit point, so are $\left(F^i(U)\right)_{i\in\mathbb{N}}$ (by the continuity of F). We shall prove that the sequence $\left(F^i(U)\right)_{i\in\mathbb{N}}$ contains all the limit points of $(U^n)_{n\in\mathbb{N}}$. Let L denote the set of limits points of $(U^n)_{n\in\mathbb{N}}$ ($1 \leq \text{card}(L) \leq J!$). Let us denote by $\omega(\eta)$ the modulus of uniform continuity of F in the compact $C(U^0)$: for $\eta > 0$,

$$\omega(\eta) = \sup_{(U,V)\in(C(U^0))^2} \left\{ ||F(U) - F(V)||_2 \text{ s.t. } ||U - V||_2 \leq \eta \right\}.$$

It is known that $\lim_{\eta\to 0, \eta>0} \omega(\eta) = 0$. If $\text{card}(L) = 1$, there is nothing to prove. Let us thus assume that $\text{card}(L) \geq 2$. We call d the minimal (Euclidean) distance between two different limit points:

$$d = \min_{(U,V)\in L^2 \text{ s.t. } V\neq U} ||U - V||_2 .$$

From the fact that $\text{card}(L)$ is finite, we have that $d > 0$. Let us choose $\eta > 0$ such that

$$\begin{array}{l} \omega(\eta) < \frac{d}{2}, \\ \eta < \frac{d}{2}. \end{array}$$

There exists $m(\eta) \in \mathbb{N}$ such that $\forall m \in \mathbb{N}$ verifying $m \geq m(\eta)$, $\exists U(\eta, m) \in L$ with

$$||U^m - U(\eta, m)||_2 \leq \eta$$

(see lemma 3 for a proof of this; this means that for n large enough each element U^n of the sequence is near a limit point). In particular, there exists $U(\eta, m(\eta)) \in L$ verifying

$$\left|\left|U^{m(\eta)} - U(\eta, m(\eta))\right|\right|_2 \leq \eta.$$

By the definition of ω, we then have

$$\left|\left|F(U^{m(\eta)}) - F(U(\eta, m(\eta)))\right|\right|_2 \leq \omega(\eta)$$

and consequently

$$\left|\left|F(U^{m(\eta)}) - F(U(\eta, m(\eta)))\right|\right|_2 < \frac{d}{2}.$$

On the other hand, $F(U^{m(\eta)}) = U^{m(\eta)+1}$, so there exists $U(\eta, m(\eta)+1) \in L$ such that

$$\left|\left|F(U^{m(\eta)}) - U(\eta, m(\eta)+1)\right|\right|_2 \leq \eta < d/2$$

(due to lemma 3). Thus

$$||F(U(\eta, m(\eta))) - U(\eta, m(\eta)+1)||_2 < d.$$

But $F(U(\eta, m(\eta))) \in L$, so that

$$\inf_{V \in L \setminus \{F(U(\eta, m(\eta)))\}} ||V - F(U(\eta, m(\eta)))||_2 \geq d.$$

Consequently, $U(\eta, m(\eta)+1) = F(U(\eta, m(\eta)))$. We have proved that

$$\left|\left|U^{m(\eta)+1} - F(U(\eta, m(\eta)))\right|\right|_2 \leq \eta.$$

By induction, it follows that

$$\left|\left|U^{m(\eta)+p} - F^p(U(\eta, m(\eta)))\right|\right|_2 \leq \eta \quad \forall p \in \mathbb{N}.$$

Thus $L = \{F^p(U) \text{ s.t. } p \in \mathbb{N}\}$. But we know that $\text{card}(L) \leq J!$. Then, there are at least two identical vectors in $\{F^0(U), \ldots, F^{J!}\}$, let us say $F^p(U)$ and $F^{p+N}(U)$, where $p < J!$ and $N \leq J! - p$ is the minimal integer such that $F^{p+N}(U) = F^p(U)$. It follows straightforwardly that $F^{i+N}(U) = F^i(U)$ $\forall i \in \mathbb{N}$ and that $N = \text{card}(L)$.
(End of the first proof.)

Lemma 1 For $U \in \mathbb{R}^J$, define $C(U)$ by

$$C(U) = \text{conv}\left\{PU \text{ s.t. } P \text{ is a permutation matrix}\right\}.$$

Assume $V \in C(U)$. Then, $C(V) \subset C(U)$.

Proof Let $x \in C(V)$: we have to prove that $x \in C(U)$. There exist $(\alpha_i)_{i=1}^{J!} \in [0,1]^{J!}$ such that

$$x = \sum_{i=1}^{J!} \alpha_i P_i V,$$
$$\sum_{i=1}^{J!} \alpha_i = 1$$

where $\{P_i \text{ s.t. } i \in \{1,\ldots,J!\}\}$ is the set of permutations matrices of $\mathcal{M}_J(\mathbb{R})$. Now, since $V \in C(U)$, there exist $(\beta_i)_{i=1}^{J!} \in [0,1]^{J!}$ such that

$$V = \sum_{i=1}^{J!} \beta_i P_i U,$$
$$\sum_{i=1}^{J!} \beta_i = 1.$$

So

$$P_j V = \sum_{i=1}^{J!} \beta_i P_j P_i U \quad \forall j \in \{1,\ldots,J!\}$$

and, because $P_j P_i$ is a permutation matrix, we see that $P_j V \in C(U)\ \forall j \in \{1,\ldots,J!\}$. Since $C(U)$ is convex, a convex combination of vectors $P_j V \in C(U)$ belongs to $C(U)$: thus $x \in C(U)$. Finally, $C(V) \subset C(U)$.

Lemma 2 For $U \in \mathbb{R}^J$, define $C(U)$ by

$$C(U) = \text{conv}\left\{PU \text{ s.t. } P \text{ is a permutation matrix}\right\}.$$

Let $U, V \in \mathbb{R}^J$ and assume that $U \in C(V)$ and that $V \in C(U)$. Then, there exists a permutation matrix $P_{U,V}$ such that $V = P_{U,V} U$.

Proof $V \in C(U)$: $\exists\, (\beta_i)_{i=1}^{J!} \in [0,1]^{J!}$ such that

$$V = \sum_{i=1}^{J!} \beta_i P_i U,$$
$$\sum_{i=1}^{J!} \beta_i = 1.$$

If we assume that V *is not* a permutation of the coordinates of U, it is a strict convex combination of its permutations and there exists at least two *positive* coefficients β_i, for example β_1 and β_2, for which

$$V = \beta_1 P_1 U + \beta_2 P_2 U + \sum_{i=3}^{J!} \beta_i P_i U = \\ (\beta_1 + \beta_2) \left(\frac{\beta_1}{\beta_1 + \beta_2} P_1 U + \frac{\beta_2}{\beta_1 + \beta_2} P_2 U \right) + \sum_{i=3}^{J!} \beta_i P_i U$$

with furthermore $P_1 U \neq P_2 U$. From the convexity of the norm $||\cdot||_2$ one gets

$$||V||_2 \leq (\beta_1 + \beta_2) \left|\left| \frac{\beta_1}{\beta_1 + \beta_2} P_1 U + \frac{\beta_2}{\beta_1 + \beta_2} P_2 U \right|\right|_2 + \sum_{i=3}^{J!} \beta_i \, ||P_i U||_2$$

and, because $||P_i U||_2 = ||U||_2$ for any permutation matrix P_i,

$$||V||_2 \leq (\beta_1 + \beta_2) \left|\left| \frac{\beta_1}{\beta_1 + \beta_2} P_1 U + \frac{\beta_2}{\beta_1 + \beta_2} P_2 U \right|\right|_2 + \sum_{i=3}^{J!} \beta_i \, ||U||_2 \, .$$

Now, the strict convexity of the norm $||\cdot||_2$ and the fact that $P_1 U \neq P_2 U$ implies

$$||V||_2 < \beta_1 \, ||P_1 U||_2 + \beta_2 \, ||P_2 U||_2 + \sum_{i=3}^{J!} \beta_i \, ||U||_2 = \\ \beta_1 \, ||U||_2 + \beta_2 \, ||U||_2 + \sum_{i=3}^{J!} \beta_i \, ||U||_2$$

finally leading to

$$||V||_2 < ||U||_2 \, .$$

But $U \in C(V)$, so $\exists \, (\alpha_i)_{i=1}^{J!} \in [0,1]^{J!}$ such that

$$U = \sum_{i=1}^{J!} \alpha_i P_i V,$$
$$\sum_{i=1}^{J!} \alpha_i = 1.$$

So

$$||U||_2 \leq ||V||_2 ,$$

which is in contradiction with $||V||_2 < ||U||_2$. Thus V is necessarily a permutation of the coordinates of U.

Lemma 3 Let $(U^n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^J$ embedded in a compact set D. Denote L the set of limit points of $(U^n)_{n\in\mathbb{N}}$. Then, $\forall\epsilon > 0$, $\exists m(\epsilon) \in \mathbb{N}$ such that $\forall n \geq m(\epsilon)$, $\exists U(n) \in L$ with

$$||U^n - U(\epsilon, n)||_2 \leq \epsilon.$$

Proof Assume the result is false: $\exists\epsilon > 0$ such that $\forall m \in \mathbb{N}$, $\exists n(m) \geq m$ such that

$$\inf_{U\in L} \left|\left|U^{n(m)} - U\right|\right|_2 > \epsilon.$$

Let us consider the sequence $\left(U^{n(m)}\right)_{m\in\mathbb{N}}$. It is embedded in the compact D, so one can extract a subsequence converging toward a limit that we shall denote V. We have

$$\inf_{U\in L} ||V - U||_2 > \epsilon,$$

and this is in contradiction with the fact that V, as the limit of a subsequence of $(U^n)_{n\in\mathbb{N}}$, belongs to L.

We now turn to the second proposed proof.

Second proof Again, the sequence $(U^n)_{n\in\mathbb{N}}$ being bounded, it has at least one limit point, let us say $U = (U_j)_{j=1}^J$. For $k \in \mathbb{R}$, let us consider the sequence $(S_k^n)_{n\in\mathbb{N}}$ of $\mathbb{R}^J$ defined by

$$S_k^n = \sum_{j=1}^{J} \left|U_j^n - k\right| .$$

It is a nonnegative decreasing sequence, thus is converging toward, let us write, S_k. Thus every limit point $V = (V_j)_{j=1}^J$ is such that

$$\sum_{j=1}^{J} |V_j - k| = S_k = \sum_{j=1}^{J} |U_j - k| .$$

Lemma 4 allows to conclude that V is a permuted vector of U. This proves points (1) and (2) of the theorem. The proof of the other propositions in the theorem remains the same. *(End of the second proof.)*

Lemma 4 Let $U = (U_j)_{j=1}^J \in \mathbb{R}^J$, $V = (V_j)_{j=1}^J \in \mathbb{R}^J$ be such that

$$\sum_{j=1}^{J} |V_j - k| = \sum_{j=1}^{J} |U_j - k| \quad \forall k \in \mathbb{R}.$$

Then, there exists a permutation matrix P such that $V = PU$.

Note that the converse is trivial.

Proof This is a quite direct consequence of lemma 5, that can be found in [9] or [12]). We have both

$$\sum_{j=1}^{J} |V_j - k| \leq \sum_{j=1}^{J} |U_j - k| \quad \forall k \in \mathbb{R}$$

and

$$\sum_{j=1}^{J} |V_j - k| \geq \sum_{j=1}^{J} |U_j - k| \quad \forall k \in \mathbb{R},$$

so that applying lemma 5 twice, we have

$$\sum_{j=1}^{l} V_{\downarrow j} = \sum_{j=1}^{l} U_{\downarrow j} \quad \forall l \in \{1, \ldots, J\}$$

($U_\downarrow$ is the decreasing rearrangement of U, see definition 2). Thus $V_{\downarrow 1} = U_{\downarrow 1}$, and by induction $V_{\downarrow j} = U_{\downarrow j}$ $\forall j \in \{1, \ldots, J\}$, $V_\downarrow = U_\downarrow$. Now, since V is a permutation of $V_\downarrow$ and U is a permutation of $U_\downarrow$, we see that V is a permutation of U.

Lemma 5 Let $U = (U_j)_{j=1}^J \in \mathbb{R}^J$, $V = (V_j)_{j=1}^J \in \mathbb{R}^J$. Let us denote by $U_\downarrow$ and $V_\downarrow$ their respective decreasing rearrangements (see definition below). Then we have

$$\sum_{j=1}^{J} |V_j - k| \leq \sum_{j=1}^{J} |U_j - k| \quad \forall k \in \mathbb{R}$$

if and only if

$$\begin{cases} \sum_{j=1}^{l} V_j \leq \sum_{j=1}^{l} U_j & \forall l \in \{1, \ldots, J\}, \\ \sum_{j=1}^{J} V_j = \sum_{j=1}^{J} U_j. \end{cases}$$

Definition 2 Let $U = (U_j)_{j=1}^J \in \mathbb{R}^J$. The decreasing rearrangement of U, denoted as $U_\downarrow$, is the vector of $\mathbb{R}^J$ defined by

$$\begin{array}{l} \exists P \text{ permutation matrix such that } U_\downarrow = PU \\ U_{\downarrow j+1} \leq U_{\downarrow j} \quad \forall j \in \{1, \ldots, J-1\}. \end{array}$$

This is the end of the part devoted to the main result of this paper, theorem 3.

In the case where F has an homogeneity property, one has a more precise result, described in the following corollary.

Corollary 1 As in theorem 3, let $(U^n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^J$ verifying (5) and (6). Assume moreover F commutes with every translation matrix:

$$F(TU) = TF(U) \quad \forall U \in \mathbb{R}^J, \forall T \in \mathcal{M}_J(\mathbb{R}) \text{ translation matrix.}$$

Then,
(1) The sequence $(U^n)_{n\in\mathbb{N}}$ has at least 1 and at most J limit points.
(2) Let U, V be two limit points. There exists a translation matrix $T_{U,V}$ such that $V = T_{U,V}U$.
Assume moreover F is continuous. Then,
(3) Let U be a limit point. There exists a translation matrix T_U such that $F(U) = T_U U$.

Every permutation matrix involved in theorem 3 can be replaced by a translation matrix.

Proof Let U be a limit point. Then, from theorem 3, $\exists P_U$, permutation matrix, such that $F(U) = P_U U$. The homogeneity property on F implies that P_U commutes with every translation matrix. Lemma 6 gives that P_U is a translation matrix, this proves point (3). Now recall that the set of limit points is $\left\{F^0(U), \ldots, F^{N-1}(U)\right\}$ if N is the number of limit points. From the fact that the product of translation matrices is a translation matrix, one gets points (1) and (2).

Lemma 6 Let $P \in \mathcal{M}_J(\mathbb{R})$ be a permutation matrix that commutes with every translation matrix. P is a translation matrix.

This is a very classical result but we shall provide a proof for the completeness of the discussion.

Proof $T \in \mathcal{M}_J(\mathbb{R})$ is a translation matrix if and only it is a permutation matrix verifying furthermore

$$T_{(i+k)\%J,(j+k)\%J} = T_{i,j} \quad \forall i,j,k \in \{1,\dots,J\}. \tag{7}$$

There are J different translation matrices; let us denote, for $k \in \{1,\dots,J\}$, T_k the translation matrix such that

$$(T_k)_{i,j} = \delta_{i,(j+k)\%J} \quad \forall i,j \in \{1,\dots,J\}.$$

P is such that $\forall k \in \{1,\dots,J\}$, $T_k P = P T_k$, or equivalently $P = (T_k)^{-1} P T_k$, where we notice that $(T_k)^{-1}$ is the translation matrix $(T_k)^T$. Now let us compute PT_k:

$$(PT_k)_{i,j} = P_{i,(j+k)\%J} \quad \forall i,j \in \{1,\dots,J\}.$$

Symmetrically, for any $Q \in \mathcal{M}_J(\mathbb{R})$,

$$\left((T_k)^{-1} Q\right)_{i,j} = Q_{(i+k)\%J,j},$$

so that $P = (T_k)^{-1} P T_k$ writes

$$P_{i,j} = P_{(i+k)\%J,(j+k)\%J} \quad \forall i,j \in \{1,\dots,J\}.$$

This has to be for any $k \in \{1,\dots,J\}$, so P is a permutation matrix verifying (7). Thus P is a translation matrix.

4 Comments

We here investigate the consequences of the previous exposition in terms of numerical resolution of scalar conservation laws with periodical initial conditions.

We first, in subsection 4.1, propose some comments on the *hypothesis* that are done. Then, subsection 4.2 interprets the *results* of the preceding.

4.1 Comments on the assumptions

Let us discuss a little on the assumptions in theorem 3 and corollary 1 regarding the numerical approximations of the scalar partial differential problem

$$\begin{cases} \partial_t u(t,x) + \partial_x f(u)(t,x) = 0 \quad \forall (t,x) \in \mathbb{R}^{+*} \times \mathbb{T}, \\ u(0,x) = u^0(x) \quad \forall x \in \mathbb{T} \end{cases} \tag{8}$$

where $\mathbb{T}$ is the torus $\mathbb{R}/\mathbb{Z}$ and $u^0 \in L^\infty(\mathbb{T})$ is given. The conclusion of this section is that (5) with (6) can stand for a very general discretization of this problem where the discrete solution is such that each of its Kruzkov is *globally* non-increasing in time.

Let J be the number of cells discretizing $[0,1[$ and denote $\Delta x = 1/J$. Usual (finite volume) approximate solutions to (8) are obtained by induction of the type

$$\begin{cases} u_j^0 = J \displaystyle\int_{(j-1)\Delta x}^{(j)\Delta x} u^0(x)\,dx \quad \forall j \in \{1,\dots,J\}, \\ \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + \dfrac{f_{j+1/2}^{n+1/2} - f_{j-1/2}^{n+1/2}}{\Delta x} = 0 \quad \forall (n,j) \in \mathbb{N} \times \{1,\dots,J\} \end{cases} \tag{9}$$

with $f_{J+1/2}^{n+1/2} = f_{1/2}^{n+1/2}$ $\forall n \in \mathbb{N}$ for periodicity reasons. The computation way of every flux $f_{j+1/2}^{n+1/2}$ defines the numerical scheme (9). Remark that form (5) is more general than (9).

Remark 1 Although form (5) seems to reduce to one-step schemes, it covers more general ones, explicit and implicit, of type

$$U^{n+1} = F(U^{n-k},\dots,U^n,\dots,U^{n+l})$$

where $k,l \in \mathbb{N}$. To see this, just write

$$\overline{U}^n = \begin{pmatrix} U^{n-k} \\ \vdots \\ U^{n+1} \\ \vdots \\ U^{n+l} \end{pmatrix} = \begin{pmatrix} U^{n-k} \\ \vdots \\ F(\overline{U}^n) \\ \vdots \\ U^{n+l} \end{pmatrix} = \overline{F}(U^{n-k},\dots,U^n,\dots,U^{n+l}) = \overline{F}(\overline{U}^n).$$

If F is continuous, $\overline{F}$ is, and if

$$\sum_{j=1}^{J} \left| U_j^{n+1} - k \right| \le \sum_{j=1}^{J} \left| U_j^n - k \right| \quad \forall k \in \mathbb{R},$$

$$\sum_{j=1}^{J\times(l+k+1)} \left| \overline{U}_j^{n+1} - k \right| \le \sum_{j=1}^{J\times(l+k+1)} \left| \overline{U}_j^n - k \right| \quad \forall k \in \mathbb{R}$$

where $\left(\overline{U}_j^n\right)_{j=1}^{J\times(l+k+1)}$ are the coordinates of $\overline{U}$.

Hypothesis (6) means that F is Kruzkov entropies non-increasing,

$$\{u \mapsto |u-k| \text{ s.t. } k \in \mathbb{R}\}$$

being the set of Kruzkov entropy functions. The theory of scalar partial differential equations ensures the existence and uniqueness of an *entropy solution* to (8), that is to say a weak solution to (8) verifying furthermore

$$\partial_t |u-k|(t,x) + \partial_x \left(\operatorname{sgn}(u-k)\left|f(u)-f(k)\right|\right)(t,x) \leq 0 \qquad \forall (t,x) \in \mathbb{R}^{+*} \times \mathbb{T}, \quad \forall k \in \mathbb{R}$$

(see [7] and [10] for example). Integrating the preceding inequalities over $\mathbb{T}$, one gets

$$\int_{\mathbb{T}} |u(t,x)-k| \, dx \leq \int_{\mathbb{T}} \left|u^0(x)-k\right| \, dx \quad \forall k \in \mathbb{R},$$

whose discrete form is (6)[1]. This discrete entropy assumption is much weaker than the local approximate entropy assumption usually made, i.e. the existence of consistent entropy fluxes $G^n_{j+1/2}$ such that

$$\frac{\left|u_j^{n+1}-k\right| - \left|u_j^n-k\right|}{\Delta t} + \frac{G^n_{j+1/2} - G^n_{j-1/2}}{\Delta x} \leq 0 \qquad \forall (n,j) \in \mathbb{N} \times \{1,\ldots,J\}, \quad \forall k \in \mathbb{R}$$

with $G^n_{J+1/2} = G^n_{1/2}$ $\forall n \in \mathbb{N}$ for periodicity reasons.

Now, let us remark that entropy condition (6) implies that $\sum_{j=1}^J u_j^{n+1} = \sum_{j=1}^J u_j^n$ $\forall n \in \mathbb{N}$. Indeed, choose $k \leq \min_{j\in\{1,\ldots,J\}} \left(u_j^n, u_j^{n+1}\right)$: inequality (6) reads

$$\sum_{j=1}^J \left(u_j^{n+1} - k\right) \leq \sum_{j=1}^J \left(u_j^n - k\right),$$

that is

$$\sum_{j=1}^J u_j^{n+1} \leq \sum_{j=1}^J u_j^n.$$

[1]It is worth noticing that equation (5) is much more general than the discrete version of a one-order scalar partial differential equation, but equation (6) is strongly related to entropy solutions of this type of equations.

Conversely, choosing $k \geq \max_{j\in\{1,\dots,J\}} \left(u_j^n, u_j^{n+1}\right)$ leads to

$$-\sum_{j=1}^{J}\left(u_j^{n+1}-k\right) \leq -\sum_{j=1}^{J}\left(u_j^n-k\right),$$

and finally

$$\sum_{j=1}^{J} u_j^{n+1} \geq \sum_{j=1}^{J} u_j^n,$$

so that the two quantities are equal. Thus, for every $n \in \mathbb{N}$, U^n lies on the hyperplane $H = \left\{(x_j)_{j=1}^J \in \mathbb{R}^J \text{ s.t. } \sum_{j=1}^J x_j = \sum_{j=1}^J u_j^0\right\}$.

4.2 Comments on the results

Let us now interpret theorem 3 and corollary 1. According to the theorem, the numerical solution of scalar equation (8) given by (5) with (6) and F continuous should attain in infinite time a periodic orbit with a finite number of limit points, all these limit points being permutations of the coordinates of each other. Thus the numerical solution $\left(U^n\right)_{n\in\mathbb{N}}$ of (8) given by a scheme of the form $U^{n+1} = F(U^n)$ where F is continuous and makes the integral of every Kruzkov entropy decrease asymptotically converges toward an orbit. Doing the supplementary assumption that the scheme commutes with translations of the space indices, the limit orbit is composed of vectors that are translations the ones from the others.

Let us do the complementary assumption that F does not consist in a permutation except on vectors other than multiple to $(1,\dots,1)^T \in \mathbb{R}^J$. Then, the only limit point of the numerical solution is proportional to $(1,\dots,1)^T$. Now invoking that $\sum_{j=1}^J u_j^n = \sum_{j=1}^J u_j^0 \ \forall n \in \mathbb{N}$ because of (6), we get that $\lim_{n\to+\infty} U^n = \overline{U}$ with

$$\overline{U} = \frac{\sum_{j=1}^J u_j^0}{J}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

This complementary assumption is clearer when we do the homogeneity assumption of corollary 1 on F. Indeed, assumption that F commutes with every translation, which seems very natural in the case of the homogeneous in space partial differential equation (8), leads to the following conclusion: if we do the complementary assumption that F does not consist in a permutation of the coordinates on any point not proportional to $(1,\dots,1)^T$, then

U^n converges towards $\frac{\sum_{j=1}^{J} u_j^0}{J}(1\ldots,1)^T$. And hypothesis that F does not consist in a permutation of the coordinates is now clearly related (at least in the linear case (2)) to the fact the the Courant number ($\Delta t/\Delta x$ in the above mentioned linear case) is not an integer.

Such a result is known in the continuous case when flux f in equation (8) is either strictly convex or strictly concave: see [11]. Thus a similar result in the discrete case is not surprising. Nevertheless, the genuine nonlinearity, that is to say, the strict convexity or concavity, has not been done in the discrete case: this means that, even for a linear degenerate equation, the numerical solution given by a scheme such as in theorem 3 converges asymptotically towards its mean value over $\mathbb{T}$ if the Courant number is not an integer. This is not the case for the continuous solution, for example in a linear case: if $f(u) = au$ with $a \in \mathbb{R}$, the unique weak solution is $u(t,x) = u^0(x - at)\ \forall (t,x) \in \mathbb{R}^+ \times \mathbb{T}$. Thus the numerical solution of a globally entropic and continuous scheme with Courant number not integer has a bad behavior in infinite time. Let us at last remark that this result can be obtained via a much simpler way for the linear advection equation with the linear upwind scheme. Indeed, let us consider $f(u) = au$ with $a > 0$ for example. Then the upwind scheme writes

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad \forall (n,j) \in \mathbb{N} \times \{1,\ldots,J\}$$

with, by definition, $u_{-1}^n = u_J^n\ \forall n \in \mathbb{N}$ to respect the periodicity condition. This has a matrix expression:

$$U^{n+1} = AU^n$$

where

$$A = \begin{pmatrix} 1-\lambda & 0 & \cdots & 0 & \lambda \\ \lambda & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda & 1-\lambda \end{pmatrix}$$

where λ is the Courant number, defined as $\lambda = a\Delta t/\Delta x$. It is known that under the Courant-Friedrichs-Levy condition $0 < \lambda \leq 1$, the scheme is convergent toward the unique weak solution of the PDE. Moreover, we see that under this condition, matrix A is bistochastic. Under the stronger hypothesis that $0 < \lambda < 1$, A is bistochastic and irreversible. One can show,

too, that the eigenvalues of A are the

$$(1-\lambda)+\lambda e^{\frac{2\imath j\pi}{J}}, \quad j=1,\dots,J.$$

Thus 1 is the only eigenvalue of modulus $\rho(A)=1$. It is then an exercise (using the Perron-Frobenius theorem for example, see [12], exercise 5.9 (b) (iv)) to show that

$$A^{(n)} \longrightarrow_{n\to+\infty} \frac{1}{J}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

Thus U^n converges toward $\overline{U}$ with the definition above. Note the need of the assumption $\lambda \neq 1$, which is much similar to the one done in the general case where $F(U)$ should not be a permutation of U except if U is proportional to $(1,\dots,1)$.

Remark 2 Let us consider a continuous in time scheme of the form

$$\partial_t U(t) = F(U(t))$$

with $U(0)=U^0 \in \mathbb{R}^J$ and F mapping $\mathbb{R}^J$ into $\mathbb{R}^J$, while doing the assumptions that F is Lipschitz-continuous and that

$$\partial_t \sum_{j=1}^{J} |U_j(t)-k| \leq 0 \quad \forall k \in \mathbb{R}.$$

Then, the unique solution U *converges* in infinite time toward a limit in $\mathbb{R}^J$. The demonstration follows the second one of theorem 3. For every $k \in \mathbb{R}$,

$$S_k(t) = \sum_{j=1}^{J} |u_j(t)-k|$$

is a lower bounded decreasing function, thus it is converging. Furthermore, $U(t)$ is embedded in a compact of $\mathbb{R}^J$, thus it has at least a limit point U. Every limit point V of $\{U(t) \text{ s.t. } t \in \mathbb{R}\}$ is such that

$$\sum_{j=1}^{J} |V_j-k| = \sum_{j=1}^{J} |U_j-k| \quad \forall k \in \mathbb{R}$$

and applying lemma 4 we conclude that all limit points are permutation the ones from the others. Finally recalling that $U(t)$ is continuous in time (thus $U(t)$ cannot "jump" from a neighborhood of a limit point to one of its permutations), we see that the limit point is unique.

To conclude this paper, let us observe that all the present results are valid in any spatial dimension $d \in \mathbb{N}$ for equation

$$\partial_t u + \sum_{i=1}^{d} \partial_{x_i} f_i(u) = 0 \quad \forall (t,x) \in \mathbb{R}^{+*} \times \mathbb{T}^d$$

Indeed, the very general numerical scheme formalism $U^{n+1} = F(U^n)$ used along the presentation allows to treat the multidimensional case.

References

[1] K.-S. Cheng, Asymptotic behavior of solutions of a conservation law without convexity conditions, J. of Diff. Eq. 40 (1981), 343-376.

[2] K.-S. Cheng, Decay rate of periodic solutions for a conservation law, J. of Diff. Eq. 42 (1981), 390-399.

[3] C. Dafermos, Applications of the Invariance Principle for Compact Processes II. Asymptotic Behavior of Solutions of a Hyperbolic Conservation Law, J. of Diff. Eq. 11 (1971), 416-424.

[4] C. Dafermos, Regularity and large time behavior of solutions of a conservation law without convexity, Proc. of the Roy. Soc. of Edinburgh 99A (1985), 201–239.

[5] C. Dafermos, Large Time Behavior of Periodic Solutions of Hyperbolic Systems of Conservation Laws, J. of Diff. Eq. 121 (1994), 183-202.

[6] B. Després, F. Lagoutière, Contact discontinuity capturing schemes for linear advection and compressible gas dynamics, J. Sci. Comput,. 16 (2001), no. 4: 479–524 (2002)

[7] E. Godlewski, P.-A. Raviart, Hyperbolic systems of conservation laws, Ellipses (1991).

[8] J. M. Greenberg and D. D. M. Tong, Decay of solutions of $\partial u/\partial t + \partial f(u)/\partial x = 0$, J. of Math. Anal. and App. 43 (1973), 56–71.

[9] G. H. Hardy, J. E. Littlewood, G. Pólya, Inequalities, Cambridge Univ. Press, London.

[10] S. Kruzkov, First-order quasilinear equations in several independent variables, Math. USSR Sb. 10 (1970), 217–243.

[11] P. D. Lax, Hyperbolic Systems of Conservation Laws II, CPAM X (1957), 537–566.

[12] D. Serre, Matrices: Theory and Applications, Springer (2002).

[13] E. F. Toro, Riemann solvers and numerical methods for fluid dynamics, Springer-Verlag (1997).