


## REMARKS ABOUT THE DISCRETE PROFILES OF SHOCK WAVES

Denis Serre\* 

### Abstract

Discrete profiles for shock waves are important features in the numerical simulation for systems of conservation laws. One discusses the important role of the dimensionless parameter  $\eta := s\Delta t/\Delta x$ , where  $s$  is the shock velocity and  $\Delta t$ ,  $\Delta x$  are the time and space mesh sizes. It turns out that the rational case, the only one having being considered for systems by previous authors, is a rather special one. It is far from generic and does not share the properties that one should expect when  $\eta$  is irrational.

### Résumé

Les profils d'ondes de choc sont des outils essentiels dans l'étude des systèmes de lois de conservation. Il y en a principalement de deux sortes : les profils de viscosité et les profils discrets. Ces derniers peuvent être présents dans les simulations numériques par différences finies. On discute ici le rôle essentiel joué par le rapport sans dimension  $\eta := s\Delta t/\Delta x$ , où  $s$  est la vitesse du choc et  $\Delta t$ ,  $\Delta x$  sont les pas de temps et d'espace. Sauf pour une équation scalaire, seul le cas  $\eta \in \mathbb{Q}$  a été considéré auparavant. Il est pourtant loin d'être générique. On montre ici que les profils discrets pour  $\eta$  irrationnel, s'ils existent et ont une régularité raisonnables, ont des propriétés qui ne subsistent pas dans le cas rationnel. La question de l'existence de tels profils est donc beaucoup plus difficile que prévue.

One considers finite difference schemes which provide approximate solutions of the Cauchy problem for systems of conservation laws :

$$u_t + f(u)_x = 0, \quad x \in \mathbb{R}, \quad t > 0, \quad (1)$$

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R}. \quad (2)$$

---

\*Member of the Institut Universitaire de France.

Hereabove,  $f$  is a given smooth vector field defined on some open convex set  $\mathcal{U} \subset \mathbb{R}^n$ , which satisfies the strict hyperbolicity property : the Jacobian matrix  $df(u)$  is diagonalizable with real eigenvalues  $\lambda_1(u) \leq \lambda_2(u) \leq \dots$ , of constant multiplicities. One assumes that each eigenfield  $(\lambda, \ker(df - \lambda))$  is either linearly degenerate ( $d\lambda$  vanishes identically on  $\ker(df - \lambda)$ ) or genuinely nonlinear. In this last case,  $\lambda$  is a simple eigenvalue,  $\ker(df(u) - \lambda(u)) = \mathbb{R}r(u)$ , whereas  $d\lambda \cdot r \neq 0$  ; one then normalizes the eigenvector by imposing  $d\lambda \cdot r \equiv 1$ .

Let  $(\lambda, r)$  be an eigenfield of multiplicity  $m = 1$ . Following Lax [6], there exists a unique continuous function  $\phi$ , defined on a neighbourhood  $\mathcal{V}$  of  $\{(u, \lambda(u)); u \in \mathcal{U}\}$ , such that

1.  $\phi(u, s) = u$  if and only if  $s = \lambda(u)$ ,
2. for  $(u, v; s)$  such that  $(u, s) \in \mathcal{V}$  and  $(v, s) \in \mathcal{V}$ ,

$$f(v) - f(u) = s(v - u) \quad (3)$$

is equivalent to

$$\text{either } v = \phi(u, s) \quad \text{or } v = u.$$

The equation (3) is referred to as the *Rankine-Hugoniot condition* for  $(v, u; s)$ .

In the sequel, we shall pay attention to genuinely nonlinear fields only, so that the Lax's map satisfies, perhaps up to the restriction to a smaller set  $\mathcal{V}$ ,

$$(\lambda(\phi(u, s)) - s)(\lambda(u) - s) < 0 \quad \text{if } s \neq \lambda(u). \quad (4)$$

The function

$$\hat{u}(x, t) := \begin{cases} u_l, & x < st, \\ u_r := \phi(u_l, s), & x > st, \end{cases}$$

is a solution of (1) in the distributional sense. It is said to be admissible if moreover

$$\lambda(u_r) < s < \lambda(u_l).$$

In that case,  $(u_l, u_r; s)$  is called a *shock wave*. Let us remark that the two last inequalities are equivalent to each other and that, given  $(a, s) \in \mathcal{V}$ , either  $(a, \phi(a, s); s)$  or  $(\phi(a, s), a; s)$  is a shock wave, depending on the sign of  $\lambda(a) - s$ .

The stability of shock waves with respect to various kind of perturbations is of fundamental importance because only perturbations make a significant difference between admissible and non-admissible weak solutions. The most popular perturbation is the addition in (1) of a parabolic *viscous* term  $\epsilon(B(u)u_x)_x$ ; one then lets  $\epsilon$  tend to zero. But for practical purposes, the main perturbation is induced by space-time discretization, for instance when performing numerical computations with a finite difference scheme. It is essential to determine whether a shock wave is stable with respect to a given scheme, so that the approximate solution will mimic the one of the Cauchy problem.

One shall always consider conservative difference schemes, which write in the general form

$$u_j^{m+1} = u_j^m - \frac{\Delta t}{\Delta x}(f_{j+1/2}^m - f_{j-1/2}^m), \quad j \in \mathbb{Z}, m \in \mathbb{N}$$

$$f_{j+1/2}^m := F\left(\frac{\Delta t}{\Delta x}, u_{j-p+1}^m, \dots, u_{j+q}^m\right).$$

Hereabove, the consistency with (1) is ensured by  $F(\sigma, a, \dots, a) = f(a)$ . This iteration is called a  $(p+q+1)$ -points scheme because the computation of  $u_j^{m+1}$  needs the knowledge of *a priori*  $p+q+1$  values  $u_k^m$  for  $j-p \leq k \leq j+q$ . The simplest schemes involve only three points<sup>1</sup> ( $p=q=1$ ). For instance, the Lax-Friedrichs scheme is made with

$$F(\sigma, a, b) = \frac{1}{2}(f(a) + f(b)) + \frac{1}{2\sigma}(a - b),$$

whereas the Godunov's scheme comes from

$$F(\sigma, a, b) = f(R(a, b)),$$

$R(a, b)$  being the *middle point*  $v(0, t)$  in the solution of the Riemann problem

$$\begin{cases} v_t + f(v)_x = 0, \\ v(x, 0) = a, & x < 0, \\ v(x, 0) = b, & x > 0. \end{cases}$$

Let us remark that no ambiguity occurs when  $v$  is discontinuous at  $x=0$  since then  $f(v(0-, t)) = f(v(0+, t))$ .

---

<sup>1</sup>However, in special cases, the Godunov's scheme requires only two points.

The stability of shock waves is strongly related to the existence of an analogue of the shock wave at the discrete level : the so-called *discrete shock profile* (DSP). Let  $(u_r, u_l; s)$  be a shock wave for (1). A DSP looks very much to a traveling wave with velocity  $s$ , which achieves the values  $u_l, u_r$  as  $j \rightarrow \pm\infty$ . Since the ratio  $\Delta t/\Delta x$  is given, there is no reason why  $s$  should be equal to a slope of the lattice  $\mathbb{Z}\Delta x \times \mathbb{N}\Delta t$ . Thus a definition of a DSP cannot handle a discrete function (that is a function defined on the lattice) in general. One merely must consider functions defined on the entire plane  $\mathbb{R}_x \times \mathbb{R}_t$  with values in  $\mathcal{U}$ . Since it has to depend only on the traveling variable  $x - st$ , one may restrict to functions of one variable.

**Definition 1.** *Let us denote by  $\eta$  the dimensionless ratio  $s\Delta t/\Delta x$ . One says that a smooth function  $v : \mathbb{R} \rightarrow \mathcal{U}$  is a DSP of the shock wave  $(u_r, u_l; s)$  if it satisfies both*

$$\begin{aligned} v(z - \eta) = & v(z) - \frac{\Delta t}{\Delta x} \left[ F\left(\frac{\Delta t}{\Delta x}, v(z - p + 1), \dots, v(z + q)\right) \right. \\ & \left. - F\left(\frac{\Delta t}{\Delta x}, v(z - p), \dots, v(z + q - 1)\right) \right], \quad z \in \mathbb{R}, \end{aligned} \quad (5)$$

$$v(-\infty) = u_l, \quad v(+\infty) = u_r. \quad (6)$$

The meaning of (5) is that for any  $z_0 \in \mathbb{R}$ , the sequence

$$u_j^m := v(z_0 + j - m\eta)$$

satisfies the difference scheme. Thus  $u_{\Delta t, \Delta x}(x, t)$ , defined by interpolation from the values

$$u_{\Delta t, \Delta x}(j\Delta x, m\Delta t) = u_j^m,$$

is an approximate solution which converges to  $\hat{u}$  as  $\Delta x$  goes to zero, the ratio  $\Delta t/\Delta x$  being kept fixed. In particular it depends only on  $x - st$  up to  $\mathcal{O}(\Delta x)$ .

There are definitive reasons to consider DSP depending smoothly on a real parameter. From an algebraic point of view, the functional equation (5) needs only that  $v$  be defined on a lattice  $z_0 + \mathbb{Z} + \eta\mathbb{Z}$ . If, by chance,  $\eta$  is rational, say  $\eta = l/d$ , then this is a discrete lattice  $z_0 + \frac{1}{d}\mathbb{Z}$  and (5) becomes a discrete



dynamical system ; it becomes apparent, noting that  $z \mapsto d(z - z_0)$  transforms the lattice into  $\mathbb{Z}$ . Then a natural restriction of the size of the ratio<sup>2</sup>  $\Delta t/\Delta x$  makes  $F$  to be invertible with respect to either its second argument or to the last one. For instance, if  $d = 1$ , (5) can be rewritten as

$$v(j) = H(\Delta x/\Delta t, v(j-1), \dots, v(j-p-q)). \quad (7)$$

In this context, a DSP must be viewed as a heteroclinic orbit, between two rest points  $(u_l, \dots, u_l)$  ( $p+q$  times) and  $(u_r, \dots, u_r)$ , for the dynamical system

$$\begin{pmatrix} v_{-1} \\ v_{-2} \\ \vdots \\ v_{-p-q} \end{pmatrix} \mapsto \begin{pmatrix} H(v_{-1}, \dots, v_{-p-q}) \\ v_{-1} \\ \vdots \\ v_{-p-q+1} \end{pmatrix}.$$

There is a huge amount of theoretical tools and the study of such orbits is well understood, at least for small amplitude shock waves (see [8]). The main tool here is the center-manifold theorem. Let us give below the simplest application to our problem.

One supposes that the (genuinely nonlinear) eigenvalue  $\lambda$  vanishes somewhere in  $\mathcal{U}$ . Then, because of genuine nonlinearity, both the set  $\Gamma := \{u \in \mathcal{U}; \lambda(u) = 0\}$  and its image  $f(\Gamma) =: \mathcal{C}$  are codimensional-one submanifolds. The mapping  $f$  is a fold : the equation  $f(u) = z$  has one solution if  $z \in \mathcal{C}$  but it has two or zero otherwise, depending on which side of  $\mathcal{C}$  the point  $z$  belongs to. Pairs of solutions are of the form  $(a, \phi(a, 0))$ . Let us consider three-points schemes ( $p = q = 1$ ). Then a DSP for a steady shock wave ( $s = 0$ ) is a solution of the equation

$$F(\sigma, u_j, u_{j+1}) = F(\sigma, u_{j-1}, u_j), \quad j \in \mathbb{Z}.$$

Since moreover  $u_j$  tends to  $u_{r,l}$  as  $j$  goes to  $\pm\infty$ , then one *integrates* once the profile equation by

$$F(\sigma, u_j, u_{j+1}) = f(u_{r,l}). \quad (8)$$

---

<sup>2</sup>the so-called *Courant-Friedrichs-Levy condition*.. It usually implies  $-p < d < q$ .

Hereabove, the CFL condition frequently allows us to invert the map  $b \mapsto F(\sigma, a, b)$  as a smooth function  $z \mapsto G(\sigma, a, z)$ , so that

$$u_{j+1} = G(\sigma, u_j, z), \quad z := f(u_{r,l}). \quad (9)$$

We know that  $d_a F$  and  $d_a G$  are invertible. On one hand  $d_a F + d_b F = df$  whenever  $b = a$ ; on the other hand  $d_a F + d_b F d_a G = 0$ . It follows

$$d_a F(I - d_a G) + df d_a G = 0, \quad (10)$$

when  $b = a$ . Thus the kernel of  $d_a G(\sigma, a, f(a)) - I$  is isomorphic to the one of  $df(a)$ : it is a line if  $a \in \Gamma$  and it is trivial otherwise. We now consider the extended iteration  $U_{j+1} = g(U_j)$ , where

$$U_j := \begin{pmatrix} u_j \\ v_j \end{pmatrix}$$

and

$$g \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} G(\sigma, u, f(v)) \\ v \end{pmatrix}.$$

In this formulation, a DSP for the shock wave  $(u_r, u_l; s = 0)$  corresponds to a heteroclinic orbit between the rest points  $(u_l, u_l)^t$  and  $(u_l, u_r)^t$ .

Let  $a$  belong to  $\Gamma$  and  $A := (a, a)^t$ . The differential of  $g$  at  $A$  is

$$dg(A) = \begin{pmatrix} d_a G(a, f(a)) & d_b G(a, f(a))df(a) \\ 0 & I \end{pmatrix}.$$

For most schemes,  $\mu = 1$  is the only eigenvalue of  $dg(A)$  on the unit circle<sup>3</sup>; this property is the non-resonance condition of Majda-Ralston [8]. The multiplicity of  $\mu = 1$  depends of the accuracy of the scheme; it should be  $n + k$  for a  $k$ -th order scheme (see Michelson [9]). Let us consider for instance a first-order scheme. Then the central manifold  $\mathcal{M}$  of  $g$  at  $A$  is a  $(n + 1)$ -dimensional manifold. By definition :

- it is locally invariant by  $g$ ,

---

<sup>3</sup>but the Lax-Friedrichs scheme is exceptionnal : both  $\mu = \pm 1$  are eigenvalues of  $dg(A)$ .

- it contains every globally defined sequence  $(g^m(U_0))_{m \in \mathbb{Z}}$  which remains in a small neighbourhood of  $A$ .

In particular, it contains all the fixed points of  $g$ . These are of two kinds : the trivial pairs  $(u, u)^t$  and the pairs  $(\phi(u, 0), u)$ . Thus the fixed points define two codimensional-one submanifolds  $\mathcal{M}_0$  and  $\mathcal{M}_1$  of  $\mathcal{M}$ , transversal to each other. Their intersection is nothing but the set of pairs  $(u, u)$  such that  $u \in \Gamma$ . Because of the nonlinear hypothesis, the fibers  $v = \text{constant}$  define curves in  $\mathcal{M}$ , which are transversal to both  $\mathcal{M}_0$  and  $\mathcal{M}_1$ . Let  $u_l$  be close to  $a$ . The curve  $\gamma(u_l)$  defined in  $\mathcal{M}$  by  $v = u_l$  meets  $\mathcal{M}_0$  in  $(u_l, u_l)$  and  $\mathcal{M}_1$  in  $(u_r := \phi(u_l, 0), u_l)$ , which is different from the former if  $\lambda(u_l) \neq 0$ . The diffeomorphism  $g$  acts on  $\gamma(u_l)$ , preserving the orientation<sup>4</sup>, with exactly two fixed points. One concludes that there exists a one-parameter family of heteroclinic orbits between  $(u_l, u_l)$  and  $(u_r, u_l)$ . The orientation of these orbits is determined by an analysis of the stability property of these fixed points : it goes from  $(u_l, u_l)$  towards  $(u_r, u_l)$  if and only if  $\lambda(u_l) > 0$ .

A similar analysis works for any rational values of  $\eta$ , but the phase space is then of a large dimension  $N = (p + q + 1)d$ . It has been done by Majda and Ralston [8]. Since this is a local analysis, the existence of a DSP is proved only for weak shocks, that is under the hypothesis  $\|u_r - u_l\| \leq \epsilon(d, a)$ . Here,  $\lambda(a)\Delta t/\Delta x = l/d$ . Unfortunately,  $\epsilon(d, a)$  shrinks to zero as  $d$  goes to infinity, which prevents to use the Majda-Ralston result in order to prove the existence of DSP for irrational  $\eta$ 's by continuation. This is not due to a bad choice of mathematical tools but rather to a deep mathematical reason : the spectral properties of the functional equation (5) make the rational and irrational cases very different from each other<sup>5</sup>.

A one-parameter family of DSP's for a given shock wave  $(u_l, u_r; s)$  appears in the study of the stability of a DSP under an initial disturbance. Let

$$v_j^n := v(j - n\eta)$$

---

<sup>4</sup>this is a consequence of the CFL hypothesis.

<sup>5</sup>this is not true for scalar equations ( $n = 1$ ). In this case, there is an accurate existence theory of DSP for any value of  $\eta$ . See the work of Jennings [4].

and  $u_j^0 = v_j^0 + w_j$  a perturbed initial data for the difference scheme. Assuming that  $(w_j)_j$  is summable and small, one expects an asymptotic behaviour which mimics the one arising for viscous shock waves (see Liu [7]). The default mass  $M := \sum_{j \in \mathbb{Z}} (u_j^n - v_j^n) = \sum_j w_j$  does not depend on  $n$  because the scheme is conservative. The sequence  $(u_j^n)_j$  will be asymptotic in the  $l^\infty$ -norm to another DSP  $(v(z_0 + j - n\eta))_j$ . The shift  $z_0$  is determined by the following procedure : the default mass splits into three parts

$$M = \sum_{\mu < \lambda(u_l)} r_\mu(u_l) + \sum_{\mu > \lambda(u_r)} r_\mu(u_r) + m(z_0), \quad (11)$$

where  $r_\mu(u) \in \ker(df(u) - \mu)$  and

$$m(z_0) := \sum_{j \in \mathbb{Z}} (v(z_0 + j) - v(j)).$$

The first sum in (11) is a vector of an invariant subspace  $E^-(u_l)$  of  $df(u_l)$  whereas the second is a vector of an invariant subspace  $E^+(u_r)$  of  $df(u_r)$ . For weak shock waves, these spaces are transversal to each other and the sum of their dimensions is<sup>6</sup>  $n - 1$ . Thus a good decomposition in (11) needs the image of  $z \mapsto m(z)$  to run over a one-dimensional curve of  $\mathbb{R}^n$ . The analogue of  $m$  for viscous shock waves is

$$m(z) = \int_{\mathbb{R}} (v(z + y) - v(y)) dy = z(u_r - u_l),$$

which runs over a straight line. We shall show below that this might not be the case for DSP, at least when  $\eta$  is irrational. One only knows in general that  $m(z + 1) = m(z) + u_r - u_l$ .

**Proposition 1.** *Let us consider the Godunov's scheme and assumes that  $\lambda(u)$  is the smallest eigenvalue of  $df(u)$  (that is  $\lambda = \lambda_1$ ). One assumes that the Riemann problem has a unique admissible solution.*

*Let us consider a DSP for a steady shock wave  $(u_r, u_l; s = 0)$ . Then  $v$  is locally constant off a unit interval which may be fixed as  $]0, 1[$  :*

---

<sup>6</sup>this is the so-called Lax shock condition.

- $v(x) = u_l$  for  $x < 0$  and  $v(x) = u_r$  for  $x > 1$ ,
- let  $S(u_r)$  be the shock curve towards  $u_r$  associated to the eigenvalue  $\lambda$ . It contains  $u_l$ . For  $x \in ]0, 1[$ ,  $v(x)$  runs over  $S(u_r)$  from  $u_l$  to  $u_r$ .

From this, one finds that  $m(x) = v(x - [x]) - v(0) + [x](u_r - u_l)$ , where  $[x]$  denotes the integral part of  $x$ . For general systems (the full gas dynamics is a good example), the shock curve is not a straight line, so that  $m(z)$  does not run along a straight line. On the other hand,  $m$  trivially runs over the whole line  $\mathbb{R}$  for a scalar conservation law ( $n = 1$ ) and this makes a major difference between the general case and the scalar one.

**Proof.** As before, a DSP for a steady shock satisfies  $F(\sigma, u_j, u_{j+1}) = f(u_{r,i})$ . Here,  $F(\sigma, a, b)$  is nothing but  $f(R(a, b))$ . Since  $f$  is a fold, the equality  $f(R(u_j, u_{j+1})) = f(u_{r,i})$  means either  $R(u_j, u_{j+1}) = u_r$  or  $R(u_j, u_{j+1}) = u_l$ . In the first case, the Riemann problem from  $u_r$  to  $u_{j+1}$  involves only forward waves whereas the Riemann problem from  $u_j$  to  $u_r$  involves only backward waves. In the second case, the Riemann problem from  $u_l$  to  $u_{j+1}$  involves only forward waves whereas the Riemann problem from  $u_j$  to  $u_l$  involves only backward waves.

Since  $R$  is continuous,  $R(u_j, u_{j+1})$  tends to  $u_r$  as  $j$  goes to  $+\infty$ . Thus  $R(u_j, u_{j+1}) = u_r$  for  $j \geq J_1$ . For the same reason,  $R(u_j, u_{j+1}) = u_l$  for  $j \leq J_0$ .

Let  $j$  be such that  $R(u_{j-1}, u_j)$  and  $R(u_j, u_{j+1})$  are equal, say for instance equal to  $a$ . Then gluing together the Riemann problems from  $a$  to  $u_j$  (with only backward waves) and from  $u_j$  to  $a$  (with only forward waves), one sees that  $a = R(u_j, u_j)$ , so that  $a = u_j$ . A first consequence is that  $u_j = u_l$  for  $j \leq J_0$  and  $u_j = u_r$  for  $j > J_1$ .

Now let us suppose that there is a integer  $j$  so that  $R(u_{j-1}, u_j) = u_r$  and  $R(u_j, u_{j+1}) = u_l$ . Then one goes from  $u_j$  to  $u_l$  by backward waves and from  $u_r$  to  $u_j$  by forward waves. Since one passes from  $u_l$  to  $u_r$  by a steady shock, one obtains a solution of the Riemann problem from  $u_j$  to itself by gluing all these waves together. Since it is not a constant solution, this contradicts the

uniqueness hypothesis.

This shows that  $J_1 = J_0 + 1$ . Except for  $j = J_1$ ,  $u_j$  is either  $u_l$  or  $u_r$ . Finally, one goes from  $u_{J_1}$  to  $u_r$  by backward waves. Since only  $\lambda(u)$  might be negative among the eigenvalues of  $df(u)$ , there is at most one wave, of the first family. Since the shock velocity  $s(u, u_r)$  is monotonous along  $S(u_r)$  as a function of  $u$ , the wave is backward if and only if  $u_{J_1}$  lies on the same side than  $u_l$  of the wave curve<sup>78</sup>  $\mathcal{O}_1(u_r)$ . On the other hand the Riemann problem from  $u_l$  to  $u_{J_1}$  is solvable by forward waves and this means that  $u_{J_1}$  belongs to a half-space bounded by the following codimensional-one submanifold  $K(u_l)$ . This is the set of states  $a$  such that the first wave in the Riemann problem from  $u_l$  to  $a$  is a steady shock. The wave curve  $\mathcal{O}_1(u_r)$  meets transversally  $K(u_l)$  at  $u_r$  and the aforementioned half-space does not contain the rarefaction part of  $\mathcal{O}_1(u_r)$ . Thus  $u_{J_1}$  belongs to the shock curve and lies between  $u_l$  and  $u_r$ .

□

We now turn toward the irrational case. The functional equation (5) is no longer a finitely dimensional dynamical system. Although central manifold theory has been developped in infinite dimensional contexts, it does not apply to our problem because it always needs a Fredholm alternative for the linearized operator arising from (5) :

$$\mathcal{L}V(z) := V(z - \eta) - V(z) + \sigma \sum_{k=-p+1}^q d_{u_k} F(\sigma, a, \dots, a)(V(z+k) - V(z+k-1)).$$

In Fourier variable, it becomes

$$\mathcal{F}(\mathcal{L}V)(\xi) = X(\xi)\mathcal{F}V(\xi),$$

where

$$X(\xi) := (e^{-i\xi\eta} - 1)I_n + \sigma(1 - e^{-i\xi}) \sum_{k=-p+1}^q e^{ik\xi} d_{u_k} F(\sigma, a, \dots, a).$$

---

<sup>7</sup>all the arguments of this last paragraph of the proof are valid at least if the shock curve satisfies Liu's admissibility criterion. In particular, they work for moderate shock strength.

<sup>8</sup>one uses unusual notations :  $\mathcal{O}_1(u_r)$  is the set of states  $u$  such that one goes from  $u$  to  $u_r$  by a 1-wave.

If  $\eta$  is irrational,  $\det X(2m\pi) = (e^{-2im\pi\eta} - 1)^n$  approaches zero for arbitrarily large values of the integer  $m$  although it does not necessarily vanish. This shows that  $\mathcal{L}$  is not an invertible endomorphism, even up to a finitely dimensional subspace. The only hope is to invert  $\mathcal{L}$  by allowing a *loss of derivatives*, providing  $\eta$  is well approximated by rational numbers. If so, one should try a Nash-Moser procedure in the spirit of a famous problem: the linearization of  $S^1$ -diffeomorphisms (see for instance [1]). Blood, sweat and tears,...

Let us mention an interesting analysis by S. Benzoni-Gavage [2] in an infinite dimensional context. It concerns an upstream semi-discretization when the spectrum of  $df$  is non-negative :

$$\frac{du_j}{dt} = -\frac{f(u_j) - f(u_{j-1})}{\Delta x}. \quad (12)$$

Then a semi-discrete shock profile is a function  $v : \mathbb{R} \rightarrow \mathcal{U}$  such that  $v(\pm\infty) = u_{r,l}$  and  $u_j(t) := v(j - st/\Delta x)$  solves (12). It is thus a heteroclinic orbit of the retarded functional difference equation in the sense of Hale :

$$sv'(x) = f(v(x)) - f(v(x-1)). \quad (13)$$

S. Benzoni-Gavage proves the existence of such semi-discrete profiles for small shocks. One remarks that there is not any dimensionless parameter in this problem.

Let us point out another trouble concerning DSP. It will be sufficient to consider only three-points schemes and we shall forget about the dependence of  $F$  on  $\sigma$ . The profile equation is

$$v(z - \eta) = v(z) - \sigma(F(v(z), v(z+1)) - F(v(z-1), v(z))).$$

Let us define the *reduced* numerical flux  $\theta(a, b)$  by

$$\theta(a, b) := F(a, b) - \frac{s}{2}(a + b) - f(u_l) + su_l.$$

From the Rankine-Hugoniot condition :

$$\theta(u_r, u_r) = \theta(u_l, u_l) = 0.$$

The DSP of  $(u_l, u_r; s)$  (let us remind that we do not know about its existence if  $n \geq 2$ ) obeys the equation

$$v(z - \eta) = v(z) + \frac{\eta}{2}(v(z - 1) - v(z + 1)) - \sigma(\Theta(z + \frac{1}{2}) - \Theta(z - \frac{1}{2})), \quad (14)$$

where we use the notation  $\Theta(z + \frac{1}{2}) = \theta(v(z), v(z + 1))$ . Being reasonably optimistic, one assumes  $v$  to be of bounded variation and define

$$V(y) := \sum_{j \in \mathbb{Z}} (v(j + y) - v(j)),$$

the series being summable. One shall assume moreover that  $V$  is continuous. By definition,  $V(z + 1) = V(z) + u_r - u_l$  and  $V(0) = 0$ . On the other hand, (14) implies  $V(z + \eta) = V(z) + \eta(u_r - u_l)$ , because  $\Theta$  tends to zero at infinity. Thus  $V(z + y) - V(z) = y(u_r - u_l)$  for all  $y$  in  $\mathbb{Z} + \eta\mathbb{Z}$ . Since  $\eta$  is irrational, this subgroup is dense in  $\mathbb{R}$  and this identity holds for every real number  $y$ . Finally:

**Proposition 2.** *Let  $v$  be a smooth enough DSP for a shock wave  $(u_r, u_l; s)$ , with an irrational value of  $s\Delta t/\Delta x$ . Then*

$$\sum_{j \in \mathbb{Z}} (v(j + y) - v(j)) = y(u_r - u_l), \quad y \in \mathbb{R}. \quad (15)$$

The formula (15) cannot be extended to rational values of  $\eta$ , as shown by the example constructed above with the Godunov's scheme<sup>9</sup> and a steady shock wave ( $\eta = 0$ ). Indeed the sum in the left-hand side of (15) reduces, thanks to proposition 1, to  $v(y) - u_l$  if  $0 < y < 1$ . Since  $v(y) \in \mathcal{O}_1(u_r)$ , this cannot be equal to  $y(u_r - u_l)$ , unless  $\mathcal{O}_1(u_r)$  is a straight line, which is false in general.

This conclusion prevents the theory of DSP's to have nice results. Something must be wrong in the following list :

- existence of DSP's for irrational values of  $\eta$ ,

---

<sup>9</sup>as pointed out by the referee, the choice of the Godunov scheme could be controversial, since it really displays no viscosity for steady shocks. However, it has been proved that this scheme has a good behaviour in the scalar case, regarding DSPs ; this is a part of Jennings' work. Thus the bad behaviour that we describe here is actually related to nonlinear interactions in a systems. Thus we still believe that it is faithful to general facts about schemes.



- smoothness of these DSP's with respect to the space variable  $z$ ,
- continuity of these DSP's with respect to the parameter  $\eta$ .

Once again, a comparison should be made with the linearization of  $S^1$ -diffeomorphisms, since this last problem is solvable for well approximated irrational rotation numbers, whereas it is not for rational ones. This displays, as in our problem, a discontinuous behaviour on  $\mathbb{Q}$  when looking at the dependence on the parameter.

Our last discussion concerns *a priori* estimates for DSP's. The first one is rather classical since it is a reformulation of the maximum principle. At last, one will give an identity which is sometimes a  $L^1$ -estimate.

Regarding the maximum principle, we assume that the phase space  $\mathcal{U}$  is the union of a family  $(D_\alpha)_\alpha$  of convex compact subsets. The index runs over a set  $(\mathbb{R}^+)^q$  and the family is monotonous, continuous and satisfies

$$D_{\alpha \vee \beta} = D_\alpha \cup D_\beta.$$

Indeed, what we need is only to find a unique smaller  $D_\gamma$  containing a given compact subset of  $\mathcal{U}$ . For practical purposes, these  $D_\alpha$  are positively invariant domains in the sense of Chuey, Conley and Smoller [3] or in the sense of Hoff [5]. We shall assume that the finite difference scheme is *monotonous* with respect to  $(D_\alpha)_\alpha$  and for  $0 < \sigma < \sigma_0$  (this last condition is the CFL one) :

1. Let  $\alpha$  and  $a_{-p}, \dots, a_q \in D_\alpha$  be given. Then

$$b := a_0 - \sigma(F(\sigma, a_{-p+1}, \dots, a_q) - F(\sigma, a_{-p}, \dots, a_{q-1}))$$

belongs to  $D_\alpha$ ,

2. If moreover  $b \in \partial D_\alpha$ , then  $a_{-p} = \dots = a_q$ .

Let us give a few examples of such a situation. For scalar equations,  $D_\alpha = [-\alpha_1, \alpha_2]$  and this definition of monotonicity fits with the usual one:

$$(a_{-p}, \dots, a_q) \mapsto a_0 - \sigma(F(\sigma, a_{-p+1}, \dots, a_q) - F(\sigma, a_{-p}, \dots, a_{q-1}))$$

is monotonous with respect to each of its arguments. For general systems, the Godunov's scheme is monotonous with respect to domains  $D_\alpha$  which are strictly convex and positively invariant under the resolution of the Riemann problem. One still needs the CFL condition be satisfied. The Lax-Friedrichs' scheme is monotonous in a weaker sense because the computation of  $b$  does not involve  $a_0$  : if  $b \in \partial D_\alpha$  and  $a_{\pm 1} \in D_\alpha$ , then  $a_{-1} = a_1 = b$ .

**Proposition 3.** *Let the finite difference scheme be monotonous with respect to  $(D_\alpha)_\alpha$  as above. Let  $(u_l, u_r; s)$  be a shock wave and let  $D_\beta$  be the smaller element of the family, containing both  $u_l$  and  $u_r$ .*

*Then any continuous DSP for  $(u_l, u_r; s)$  maps  $\mathbb{R}$  into  $D_\beta$ .*

**Proof.** Let  $v$  be a DSP for  $(u_l, u_r; s)$ . By continuity,  $v$  is bounded. Let  $D_\gamma$  be the smallest domain containing all the values of  $v$ . Certainly,  $D_\beta \subset D_\gamma$ . If  $D_\beta \neq D_\gamma$ , then there is a  $z_0 \in \mathbb{R}$  such that  $v(z_0) \in \partial D_\gamma$ , by compactness. The monotonicity implies that  $v(z_0 + \eta + k) = v(z_0)$  for  $-p \leq k \leq q$ . By recursion,  $v(z_0 + m\eta) = v(z_0)$  for all  $m \in \mathbb{N}$ , so that  $v(z_0) \in \{u_l, u_r\}$ . An obvious contradiction. □

Let us remark that the proposition 3 holds true even for the Lax-Friedrichs' scheme. The proof is straightforward.

The maximum principle does not prevent from bad oscillations of  $v$  at infinity<sup>10</sup>. This is certainly the core of the difficulties mentionned above. However, as Jennings pointed out [4], the scalar case ( $n = 1$ ) is more favourable. We shall give below a  $L^1$ -type estimate in this context.

We first apply proposition 3 to see that a DSP for a scalar shock wave satisfies

$$(v(x) - u_r)(v(x) - u_l) < 0, \quad x \in \mathbb{R}. \quad (16)$$

We know restrict our attention to three-points schemes ( $p = q = 1$ ) and use the

---

<sup>10</sup>for rational  $\eta$  and small shocks, Michelson [9] proved that the rescaled profile converges uniformly toward those of a viscous Burgers' equation, as  $\|u_r - u_l\|$  goes to zero. This is compatible with oscillations as  $x$  goes to  $\pm\infty$  for a fixed shock with an irrational  $\eta$ .

form (14) of the profile equation. Following Jennings, the profile is monotonous, so that it is of bounded variation.

**Lemma 1.** *Let us assume that*

$$\lim_{z \rightarrow +\infty} z(v(z) - u_r) = \lim_{z \rightarrow -\infty} z(v(z) - u_l) = 0. \quad (17)$$

*Then the following integral is well defined and is equal to  $\frac{\eta^2}{2}(u_l - u_r)$  :*

$$\int_{-\infty}^{+\infty} z \left( v(z - \eta) - v(z) + \frac{\eta}{2}(v(z + 1) - v(z - 1)) \right) dz.$$

**Proof.** Let  $\rho(A, B)$  be the integral, restricted to the interval  $(A, B)$ . It splits into two parts  $\rho_+(B) + \rho_-(A)$ , with

$$\rho_+(B) := \int_{B-1}^{B-\eta} (z + \eta)v(z)dz - \int_{B-1}^B zv(z)dz + \frac{\eta}{2} \int_{B-1}^{B+1} (z - 1)v(z)dz.$$

An integration by parts gives

$$\rho_+(B) = - \int_{B-1}^{B+1} Q(z, \eta, B)dv(z) - \frac{\eta^2}{2}v(B + 1),$$

where  $Q$  is continous and satisfies uniformly  $Q = \mathcal{O}(B)$ . Since the Stieljes measure  $dv$  has a constant sign, the integral is bounded by  $\mathcal{O}(B(v(B + 1) - v(B - 1)))$ , which decreases to zero at infinity because of (17). Thus

$$\lim_{B \rightarrow +\infty} \rho_+(B) = -\frac{\eta^2}{2}u_r.$$

Similarly,

$$\lim_{A \rightarrow -\infty} \rho_-(A) = \frac{\eta^2}{2}u_l.$$

On the other hand,

$$\begin{aligned} \int_A^B z \left( \Theta \left( z - \frac{1}{2} \right) - \Theta \left( z + \frac{1}{2} \right) \right) dz &= \int_A^{A+1} z \Theta \left( z - \frac{1}{2} \right) dz \\ &\quad - \int_{B-1}^B z \Theta \left( z + \frac{1}{2} \right) dz + \int_{A+1/2}^{B-1/2} \Theta(z) dz. \end{aligned}$$

Since  $\theta(u_r, u_r) = \theta(u_l, u_l) = 0$ , the hypothesis (17) implies

$$z\Theta(z) \xrightarrow{z \rightarrow +\infty} 0.$$

Hence the two first integrals of the right-hand side tend to zero as  $A$  and  $B$  go respectively to  $\pm\infty$ .

We now multiply (14) by  $z$  and integrate on the whole real line. We conclude that the integral of  $\Theta$  is well defined and

$$\sigma \int_{\mathbf{R}} \Theta(z) dz = \frac{\eta^2}{2}(u_l - u_r). \quad (18)$$

Let us remark that this formula holds also for systems ( $n \geq 2$ ) provided a DSP converges fast enough towards its limits  $u_{r,l}$  at infinity.

Let us apply this formula to our beloved schemes. First the Lax-Friedrichs':

$$\theta_{LF}(a, b) = \frac{1}{2}(f(a) + f(b)) + \frac{1}{2\sigma}(a - b) - f(u_l) + su_l.$$

It can be rewritten as

$$\frac{1}{2}(f_0(a) + f_0(b)) + \frac{1}{2\sigma}(a - b),$$

where

$$f_0(u) := f(u) - f(u_l) - s(u - u_l) = f(u) - f(u_r) - s(u - u_r).$$

Thus a scalar Lax-Friedrichs' DSP satisfying the decay (17) will have the property

$$\int_{\mathbf{R}} f_0(v(z)) dz = \frac{1 - \eta^2}{2\sigma}(u_r - u_l). \quad (19)$$

From Oleinik's shock inequality, the function  $f_0$  has the same sign between  $u_l$  and  $u_r$  than  $u_r - u_l$ ; indeed  $(u_r - u_l)f_0(u) > 0$ . Because  $v$  takes values in this interval, we must view (19) as an *a priori* estimate. In most cases, say when the Lax's shock condition  $f'(u_r) < s < f'(u_l)$  is satisfied, there is a positive constant  $C$  such that  $(u_r - u_l)f_0(u) \geq C|(u_r - u)(u_l - u)|$  for  $u$  between  $u_r$  and  $u_l$ . Then (19) is a  $L^1$ -type estimate :

$$\int_{\mathbf{R}} |(v(z) - u_r)(v(z) - u_l)| dz \leq \frac{1 - \eta^2}{2C\sigma}(u_r - u_l)^2.$$

We end by establishing an analogous result for the upstream scheme

$$u_j^{n+1} = u_j^n - \sigma(f(u_j^n) - f(u_{j-1}^n)),$$

which is relevant when the spectrum of  $df$  is everywhere non-negative. The shock velocity must be non-negative and one has  $F(a, b) = f(a)$ , so that  $\theta(a, b) = f_0(a) - \frac{s}{2}(b - a)$ . The identity (18) takes the form

$$\int_{\mathbf{R}} f_0(v(z)) dz = \frac{\eta(1 - \eta)}{2\sigma} (u_r - u_l). \quad (20)$$

It yields again an *a priori* estimate in the scalar case.

**Acknowledgement.** this article got benefit of discussions with S. Benzoni-Gavage. The author wishes to express his gratitude to her. The author is also much indebted to the referee, who made important remarks on the topic.

## References

- [1] Alinhac, S.; Gérard, P. *Opérateurs pseudo-différentiels et théorème de Nash-Moser*. Editions du CNRS, Paris, (1991).
- [2] Benzoni-Gavage, S., *Semi-discrete shock profiles for hyperbolic systems of conservation laws*, Preprint ENS de Lyon, France, (1995).
- [3] Chuey, K.; Conley, C. and Smoller, J., *Positively invariant regions of nonlinear diffusion equations*, Indiana Univ. Math. J., vol 26, (1977), 373-392.
- [4] Hoff, D, *Invariant regions for systems of conservation laws*, Trans. of AMS., vol 289, (1985), 591-610.
- [5] Jennings, G., *Discrete shocks*, Comm. Pure & Appl. Math., vol 27, (1974), 25-37.
- [6] Lax, P. D., *Hyperbolic systems of conservation laws II*, Comm. Pure & Appl. Math., vol 10, (1957), 537-566.

- [7] Liu, T.-P., *Nonlinear stability of shock waves for viscous conservation laws*.  
Memoir of the Amer. Math. Soc., # 328. Providence, (1985).
- [8] Majda, A.; Ralston, J., *Discrete shock profiles for systems of conservation laws*, Comm. Pure & Appl. Math., vol 32, (1979), 445-482.
- [9] Michelson, D., *Discrete shocks for difference approximations to systems of conservation laws*, Adv. in Appl. Math., vol 5, (1984), 433-469.

Unité de Mathématiques Pures et Appliquées  
(CNRS UMR #128)  
ENS Lyon  
46, Allée d'Italie  
F-69364 LYON Cedex 07

Received September 14, 1995

Revised December 19, 1995