Systems of conservation laws. Theory, Numerical approximation and Discrete shock profiles

Denis Serre
École Normale Supérieure
de Lyon[1]

September 6, 2007

[1]UMPA (UMR 5669 CNRS), ENS de Lyon, 46, allée d'Italie, F–69364 Lyon, cedex 07, FRANCE.

Acknowledgments

This text was written so as to serve as lecture notes for the students of the Shanghai Summer School (July 2007) at the Fudan University. As such, it contains a list of exercises, useful for training and practicing the field. I am happy to thank the organizers of the summer school, and especially my great friends Professors Li Ta-tsien and Gui-qiang Chen.

To Pascale

Chapter 1

Hyperbolic systems of conservation laws

A first-order system of conservation laws is a system of n partial differential equations in n unknowns $u_1, \dots, u_n$ that are functions of space $x = (x_1, \dots, x_d)$ and time t. It writes

$$\partial_t u_j + \mathrm{div}_x f_j(u) = 0, \qquad j = 1, \dots, n,$$

where the *fluxes* f_j^α are given smooth functions over the space of states $\mathscr{U}$. The latter is in general a convex set of $\mathbb{R}^n$ with a non-void interior.

In the sequel, we shall consider only the case of one space variable ($d = 1$), for which we rewrite

$$\partial_t u + \partial_x f(u) = 0. \tag{1.0.1}$$

The flux is thus a smooth map $f : \mathscr{U} \to \mathbb{R}^n$.

A typical example is gas dynamics, which writes

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + p(\rho, e)) &= 0, \\ \partial_t \left(\frac{1}{2}\rho v^2 + \rho e \right) + \partial_x \left(\left(\frac{1}{2}\rho v^2 + \rho e + p \right) v \right) &= 0. \end{aligned} \tag{1.0.2}$$

Hereabove, ρ, v, e denote the mass dentity, the velocity and the specific internal energy. The pressure p is determined through an equation of state $(\rho, e) \mapsto p(\rho, e)$ that characterizes the nature of the gas. For instance $p = \frac{2}{5}\rho e$ is an acceptable relation for the air in ordinary conditions. The state u has components

$$u_1 = \rho, \quad u_2 = \rho v, \quad u_3 = \frac{1}{2}\rho v^2 + \rho e.$$

The domain $\mathscr{U}$ is defined by

$$u_1 \geq 0, \qquad u_1 u_3 \geq \frac{1}{2} u_2^2.$$

About the terminology. The words 'conservation laws' are justified by the fact that if one has a reasonable solution, say a field of bounded variations satisfying (1.0.1) in the distributional sense, then one has

$$\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)\,dx + f(u(x_2,t)) - f(u(x_1,t)) = 0, \qquad \text{a.e. } t>0.$$

In particular, if u has constant limits $u_\pm$ as $x \to \pm\infty$, then

$$\int_{\mathbb{R}} (u(x,t) - a(x))\,dx + t(f(u_+) - f(u_-)) = 0.$$

The terminology thus comes from the case where a tends to a constant state $\bar{u}$ at infinity, for which we have[1]

$$\int_{\mathbb{R}} (u(x,t) - \bar{u})\,dx = \int_{\mathbb{R}} (a(x) - \bar{u})\,dx. \tag{1.0.3}$$

If $\bar{u} = 0$, the total mass is thus *conserved*.

1.1 The Cauchy problem: classical solutions

The Cauchy problem consists in solving (1.0.1) in $(0,T)\times\mathbb{R}$ under the condition (initial data) that u is prescribed at initial time:

$$u(0,x) = a(x), \tag{1.1.4}$$

where $a : \mathbb{R} \to \mathscr{U}$ is a given function with either smoothness or integrability properties.

1.1.1 Hyperbolicity

Let us denote $A(u) := Df(u)$ the Jacobian matrix of f. We say that the system (1.0.1) is *hyperbolic* if $A(u)$ is diagonalisable with real eigenvalues. For a linear system ($f(u) = Au$), this is the condition under which the Cauchy problem is well-posed in Sobolev spaces $H^s(\mathbb{R})$. We shall see that the situation is more intricate in the nonlinear case.

Exercise. Compute the Jacobian matrix for gas dynamics. Show that the system (1.0.2) is hyperbolic if, and only if, the equation of state $p = p(\rho, e)$ satisfies

$$p\frac{\partial p}{\partial e} + \rho^2 \frac{\partial p}{\partial \rho} > 0.$$

[1] When the Cauchy problem for a first-order system (1.0.1) is well-posed, the values at spatial infinity do not change with time.

The basic example of a hyperbolic system is that of a *scalar* equation, which means that $n = 1$. Then $A(u) = f'(u)$ is 1×1, thus diagonal. The best-known scalar equation is that of Burgers:

$$\partial_t u + \partial_x(u^2/2) = 0. \tag{1.1.5}$$

Hyperbolic systems cover a wide variety of applications:

- Compressible fluid dynamics,
- Electromagnetism (Maxwell's equations),
- Magnetohydrodynamics,
- Electrophoresis,
- Chromatography,
- Traffic flow,
- Elastodynamics,
- Singular limit of dispersive waves,
- Einstein equation of general relativity,
- ...,

as long as the diffusive or dispersive effects can be neglected.

1.1.2 Entropies

When a system (1.0.1) has a physical meaning, it is often compatible with an additional scalar conservation law

$$\partial_t \eta(u) + \partial_x q(u) = 0, \tag{1.1.6}$$

where the functions $\eta, q : \mathscr{U} \to \mathbb{R}$ are smooth and η is strongly convex, in the sense that its Hessian matrix $D^2\eta(u)$ at u is positive definite for every $u \in \mathscr{U}$. We say that η is a *convex entropy* and that q is its *entropy flux*.

The compatibility between (1.1.6) and (1.0.1) means that for every smooth field $u : \mathbb{R} \to \mathscr{U}$, one has

$$\partial_t \eta(u) + \partial_x q(u) = d\eta(u) \cdot (\partial_t u + \partial_x f(u)).$$

In other words, an entropy-entropy flux pair is a solution of the linear differential system

$$dq(u) = d\eta(u)A(u), \tag{1.1.7}$$

which can be written componentwise as

$$\frac{\partial q}{\partial u_i} = \sum_{j=1}^{n} \frac{\partial \eta}{\partial u_j} \frac{\partial f_j}{\partial u_i}, \qquad i = 1, \dots, n.$$

Differentiating (1.1.7), we find that $A(u)$ is self-adjoint with respect to the scalar product induced by $D^2\eta(u)$:

$$\langle D^2\eta(u)X, A(u)Y\rangle = \langle D^2\eta(u)Y, A(u)X\rangle, \qquad X, Y \in \mathbb{R}^n.$$

This identity is at the basis of the symmetrization result of Godunov and Friedrichs: there exist two symmetric matrices $S_0(u), S(u)$, depending smoothly on u, with S_0 positive definite, such that the system (1.0.1) rewrites

$$S_0(u)\partial_t u + S(u)\partial_x u = 0.$$

We say that (1.0.1) is *symmetrizable* in Friedrichs sense.

Warning. Our terminology "entropy" differs from that employed in Physics. For gas dynamics, our entropy will be $\eta = -\rho s$ with $s = s(\rho, e)$ the physical entropy. In particular η is convex in u if, and only if, s is a concave function of the quantities

$$\frac{1}{\rho}, \quad v, \quad \frac{1}{2}v^2 + e.$$

Exercise. Show that $\eta = -\rho s(\rho, e)$ is an entropy of gas dynamics whenever s is a solution of the transport equation

$$p\frac{\partial s}{\partial e} + \rho^2 \frac{\partial s}{\partial \rho} = 0.$$

1.1.3 Local well-posedness in $H^s(\mathbb{R}^d)$

The best result for classical solutions of the Cauchy problem is stated in Sobolev spaces. More precisely, we ask the derivatives to be in a Sobolev space, thus allowing the data and the solution to be non-zero at infinity. Because we deal with nonlinear systems, the well-posedness in Sobolev spaces H^s require that s be large enough.

Theorem 1.1 *We assume that the system (1.0.1) is symmetrizable in Friedrichs sense.*

Let $s > 3/2$ be a real number, and let K be a compact subset of $\mathscr{U}$. Let $a : \mathbb{R} \to K$ be an initial data such that $a' \in H^{s-1}(\mathbb{R})^n$.

Then there exists a time $T > 0$, and a unique classical solution u of the Cauchy problem (1.0.1,1.1.4) on $(0, T) \times \mathbb{R}$. The solution has the property that

$$\partial_t u, \partial_x u \in L^\infty(0, T; H^{s-1}).$$

The theorem above has a counterpart in several space dimensions, but with $\nabla_x a \in H^{s-1}(\mathbb{R}^d)$ and $s > 1 + d/2$. This condition and the Sobolev embedding ensure that the data and the solution are of class C^1. Whence the terminology of "classical solution". For a full proof, see either of [3, 6, 22].

We warn the reader that the existence time T of a classical solution depends on the data a. Thus the theorem cannot be used repeatedly to build a global solution.

This theorem applies in particular to systems (1.0.1) endowed with a strongly convex entropies, since such systems are symmetrizable.

1.2 The Cauchy problem: weak solutions

We show now that we may not expect a global-in-time classical solution of the Cauchy problem, for general data. We thus introduce a weakened notion of solution, which refers to the theory of distributions. This is coherent with the physical meaning of the equations. We then explain why the physically relevant solutions should display an irreversibility property. This is reminiscent to the second principle of thermodynamics.

1.2.1 Break-down of smooth solutions

Let us consider for instance the Burgers equation. There are at least two ways to see that the classical solution breaks down in finite time. For this to happen, we only need that a' takes a negative value somewhere.

The method of characteristics. Given a base point $x_0 \in \mathbb{R}$ we define a characteristic curve by solving the ODE

$$\frac{dx}{dt} = u(x,t), \qquad x(0) = x_0.$$

An elementary calculus gives the following results (**Exercise**: fill the details). The characteristic curve is the straight line $t \mapsto x_0 + ta(x_0)$, on which $u \equiv a(x_0)$.

Assume that a is not non-decreasing. There exist two points x_0, y_0 such that $(y_0 - x_0)(a(y_0) - a(x_0)) < 0$. The intersection (x,t) of the characteristics issued from x_0 and y_0 occurs at some point (x,t) with $t > 0$. We then have the contradiction $a(x_0) = u(x,t) = a(y_0)$.

Blow-up of derivatives. We can also calculate the x-derivative $v = \partial_x u$ along a characteristics. Differentiating (1.1.5), we have

$$\frac{dv}{dt} = (\partial_t + u\partial_x)v = -v^2.$$

This is a Ricatti equation, of which the solution blows up at a positive time if the initial data $v(0) = a'(x_0)$ is negative.

Exercise. Prove that the maximal classical solution is defined on the strip $(0, T^*) \times \mathbb{R}$, where $T^* = +\infty$ if a is non-decreasing, and

$$-\frac{1}{T^*} = \inf_{x \in \mathbb{R}} a'(x)$$

otherwise.

1.2.2 Weak solutions

Since classical solutions are not global in general, although gas does flow ..., we need to accept solutions with less regularity. Typically, we consider fields $u(x,t)$ that are bounded measurable. Therefore $f(u)$ is also bounded and measurable, and the derivatives $\partial_t u$ and $\partial_x f(u)$ make sense, at least as distributions. The system (1.0.1) has to be rewritten by introducing test functions: We say that u is a *weak solution* of (1.0.1) over a domain $\Omega \in \mathbb{R} \times (0, +\infty)$ if there holds

$$\int_\Omega (u_j \partial_t \phi + f_j(u) \partial_x \phi)\, dx\, dt = 0, \tag{1.2.8}$$

for every $j = 1, \dots, n$ and every test function $\phi \in \mathscr{D}(\Omega)$.

For the Cauchy problem, we have a refined definition:

Definition 1.1 *We say that u is a* weak *solution of (1.0.1,1.1.4) over a strip $(0,T) \times \mathbb{R}$ if there holds*

$$\int_0^T \int_{\mathbb{R}} (u_j \partial_t \phi + f_j(u) \partial_x \phi)\, dx\, dt + \int_{\mathbb{R}} a(x) \phi(x,0)\, dx = 0, \tag{1.2.9}$$

for every $j = 1, \dots, n$ and every test function $\phi \in \mathscr{D}(\mathbb{R} \times (-\infty, T))$.

It is clear, from integration by parts, that a classical solution is also a solution in this weak sense. Conversely, a weak solution of class C^1 is also a classical solution. It is not hard to prove a little bit more, that a continuous field u that is piecewise C^1, is a classical solution if, and only if, it is a weak solution. We can view (1.2.9) as the most natural way to express the conservation laws, that is the underlying physical principles, for non-smooth flows.

1.2.3 The Rankine–Hugoniot condition

Immediately next to the piecewise-C^1 fields come the piecewise continuous ones. Thus let us consider a domain Ω, divided into two pieces $\Omega_\pm$ by a smooth curve Γ, and a weak solution of (1.0.1) u, such that its restrictions to each $\Omega_\pm$ is of class C^1 up to Γ. Each of these restrictions have limits along Γ, which we denote by $u_\pm$. When g is a function over $\mathscr{U}$, we define

$$[g(u)] := g(u_+) - g(u_-),$$

the *jump* of $g(u)$ across Γ. In the calculation below, we also denote ν the unit normal vector to Γ, with a given orientation.

Since u is a weak solution in each $\Omega_\pm$, where it is smooth, it is a classical solution away from Γ. Then we may integrate by parts (**Exercise**: fill the details) in each $\Omega_\pm$, to compute the left-hand side of (1.2.8). There remains the identity

$$\int_\Gamma ([u_j]\nu_t + [f_j(u)]\nu_x)\phi \, ds = 0.$$

Since the test function is arbitrary, this amounts to writing

$$[u_j]\nu_t + [f_j(u)]\nu_x = 0,$$

or in vectorial form,

$$[u]\nu_t + [f(u)]\nu_x = 0.$$

Since $[u] \neq 0$ by assumption, we see that $\nu_x \neq 0$, which means that the curve Γ can be parametrized by the time: $t \mapsto (X(t), t)$. Then the ratio $-\nu_t/\nu_x$ is nothing but the slope X'. Finally, we obtain the *Rankine-Hugoniot* relation

$$[f(u)] = \frac{dX}{dt}[u]. \tag{1.2.10}$$

These calculations can be made in the reverse order, and one obtains the following important result:

Proposition 1.1 *Let Ω and Γ (a smooth curve) be as above, and let $u : \Omega \to \mathscr{U}$ be a field such that the restriction of u to each of $\Omega_\pm$ is of class C^1 and extends as a C^1-field up to Γ.*

Then u is a solution of the system (1.0.1) in Ω if, and only if,

- *It is a classical solution away from Γ,*
- *It satisfies the Rankine-Hugoniot relation (1.2.10) across Γ.*

Example: For a scalar equation, one may rewrite (1.2.10) as

$$\frac{dX}{dt} = \frac{[f(u)]}{[u]},$$

which shows that the slope of Γ is $f'(\bar{u})$ for some $\bar{u}$ in the interval of extremities $u_\pm$. For the Burgers equation, we simply have

$$\frac{dX}{dt} = \frac{u_+ + u_-}{2}.$$

Convention. Since a discontinuity curve is parametrized by the time, we shall always choose the $\pm$ sides in the natural way:

$$u_-(X(t),t) = \lim_{x\uparrow X(t)} u(x,t), \qquad u_+(X(t),t) = \lim_{x\downarrow X(t)} u(x,t)$$

This amounts to choose the orientation of ν so that $\nu_x > 0$.

1.2.4 Non-uniqueness of weak solutions

The extension of the notion of solution resolves the lack of solution that we encountered in our study of classical solutions. However it introduces spurious, unphysical solutions. In particular, we have way too many solutions, typically an infinity, to the Cauchy problem.

To see this, we again consider the Burgers equation (1.1.5). We content ourselves with the null initial data $a \equiv 0$. Of course, there is a solution $u \equiv 0$, which is the physically relevant one. However, we can use discontinuities to build non-trivial solutions. For instance, the following definition yields a solution, for every choice of $b, c \in \mathbb{R}$ such that $b < 0 < c$:

$$u(x,t) := \begin{cases} 0, & x < bt, \\ 2b, & bt < x < (b+c)t, \\ 2c, & (b+c)t < x < ct, \\ 0, & ct < x., \end{cases}$$

since such a u is constant off lines, across which it satisfies the Rankine–Hugoniot condition.

Exercise. Build more general piecewise constant solutions to this Cauchy problem.

1.2.5 Entropy admissibility condition

Since we have replaced a non-existence trouble by a non-uniqueness one, we need to make a step backward and restrict the notion of weak solution. The clue is that, despite the apparent reversibility of systems (1.0.1), which is invariant under the space-time reversal $(x,t) \mapsto (X-x, T-t)$, the second principle of thermodynamics tells us that non-smooth flows of gas dynamics are irreversible. This is exactly saying that not all the discontinuities described by (1.2.10) are admissible.

To be more explicit, let $(u_-, u_+; s)$ be a triple that satisfies the Rankine-Hugoniot relation

$$[f(u)] = s[u]. \tag{1.2.11}$$

Then

$$u(x,t) := \begin{cases} u_-, & x < st, \\ u_+, & st < x \end{cases}$$

defines a weak solution of (1.0.1). However, since (1.2.11) is perfectly symmetric in $u_\pm$, we may exchange the role of u_- and u_+ in our construction, and we obtain an other solution v of (1.0.1). Irreversibility is that at least one of u and v is physically irrelevant.

One thus need a criterion in order to select the admissible discontinuities, presumably in the form of an inequality, not symmetric in $u_\pm$.

When the system is compatible with (1.1.6), where η is strongly convex, we remark that the Rankine-Hugoniot condition for (1.1.6)

$$[q(u)] = s[\eta(u)]$$

is not compatible with (1.2.10) in general. Because the elimination of s yields

$$[\eta(u)]\,[f(u)] = [q(u)]\,[u], \tag{1.2.12}$$

which is an equation in $\mathbb{R}^n$, with the obvious solution $u_+ = u_-$. It often happens that it has no other solution. For instance (1.2.12) gives, for the Burgers equation, $[u]^4 = 0$, from which we have $u_+ = u_-$.

Exercise. Prove that in the scalar case, with $f'' > 0$ and $\eta'' > 0$, (1.2.12) implies $u_+ = u_-$.

The calculation above tells that a discontinuous solution u of (1.0.1) cannot satistfy simultaneously (1.1.6) across discontinuities. Whence the idea to replace (1.1.6) by an inequality, in the sense of distributions:

$$\partial_t \eta(u) + \partial_x q(u) \leq 0. \tag{1.2.13}$$

It is important in this condition that we have chosen the pair (η, q) such that η is strongly convex. If it was concave, we should change the sense of the inequality in (1.2.13).

Since (1.1.6), hence (1.2.13), is automatically satisfied whenever u is a classical solution, (1.2.13) serves only across shocks. It can be reinterpreted as a jump inequality

$$[q(u)] \leq \frac{dX}{dt}[\eta(u)], \tag{1.2.14}$$

where we recall our convention of the $\pm$ sides and our definition $[g(u)] = g(u_+) - g(u_-)$.

Let us take the example of the Burgers equation, with $f(u) = \eta(u) = u^2/2$, thus $q(u) = u^3/3$. We already know that $s = (u_+ + u_-)/2$. Then (1.2.14) tells that

$$\left(\frac{u_+^2 + u_+ u_- + u_-^2}{3} - s \frac{u_+ + u_-}{2} \right) [u] \leq 0.$$

Since the parenthesis equals $[u]^2/12$, this amounts to saying $u_+ \leq u_-$. This is the admissibility condition that we were looking for.

Exercise. More generally, in the scalar case with a convex flux f, show that a discontinuity is admissible if and only if $u_+ \leq u_-$.

Terminology. The condition (1.2.13) is the Lax *entropy inequality*. Admissible discontinuities are called *shocks*, especially when (1.2.14) is a strict inequality. We shall see other selection criteria in the sequel, which are not all equivalent. Thus the notion of shock might differ, depending on which criterion we adopt to select admissible solutions. These criteria are however equivalent for many reasonable systems and in particular for scalar equations with convex fluxes.

1.2.6 The viscosity approach

A way to justify (1.2.13) is to say that a system like (1.0.1) describes only an idealized physical process, which would be better represented by the parabolic system of conservation laws

$$\partial_t u + \partial_x f(u) = \epsilon \partial_x^2 u, \tag{1.2.15}$$

where $\epsilon > 0$ is a small number. One may also have $\partial_x(B(u)\partial_x u)$ instead of $\partial_x^2 u$ in the right-hand side. Then $B(u) \in \mathbf{M}_n(\mathbb{R})$ is called the *viscosity tensor*.

What we expect is that the Cauchy problem (1.2.15, 1.1.4) admits a unique smooth solution u^ϵ, which converges boundedly almost everywhere to a field $u(x,t)$ as $\epsilon \to 0+$. If this is true, then one can pass to the limit in the sense of distributions in (1.2.15), and we find that u is a weak solution of (1.0.1). More precisely, we have

$$\int_0^T \int_{\mathbb{R}} (u_j^\epsilon \partial_t \phi + f_j(u^\epsilon)\partial_x \phi + \epsilon u_j^\epsilon \partial_x^2 \phi)\, dx\, dt + \int_{\mathbb{R}} a_j(x)\phi(x,0)\, dx = 0,$$

for every test function ϕ. Passing to the limit, we obtain that u is a weak solution of the Cauchy problem (1.0.1, 1.1.4).

We now multiply (1.2.15) (with u^ϵ) by $d\eta(u^\epsilon)$. We obtain

$$\partial_t \eta(u^\epsilon) + \partial_x q(u^\epsilon) = \epsilon \partial_x^2 \eta(u^\epsilon) - \epsilon D^2\eta(u^\epsilon) : \partial_x u^\epsilon \otimes \partial_x u^\epsilon.$$

Since η is convex, this implies

$$\partial_t \eta(u^\epsilon) + \partial_x q(u^\epsilon) \leq \epsilon \partial_x^2 \eta(u^\epsilon).$$

Passing again to the limit as above, we obtain the entropy inequality (1.2.13) in the sense of distributions.

1.2.7 The scalar case

In the scalar case, every function η is an entropy, thus every convex function is a convex entropy. This yields as many entropy inequalities as there are convex functions. And all these inequalities must be written simultaneously. Since the set of convex funtions is a convex cone spanned by the affine functions and by the so-called *Kružkov entropies*

$$\eta_k(u) := |u-k|, \qquad q_k(u) = (f(u) - f(k))\mathrm{sign}(u-k),$$

we find that a discontinuity $(u_-, u_+; s)$ is admissible if, and only if

Rankine-Hugoniot: One has $s = [f(u)]/[u]$,

Oleinik condition: Either $u_- < u_+$ and the graph of the restriction of f to the interval $[u_-, u_+]$ is above its chord. Or $u_+ < u_-$ and the graph of the restriction of f to the interval $[u_+, u_-]$ is below its chord.

We leave the proof of that as an **Exercise**.

We also have a well-posedness result, due to S. Kružkov, for which we refer to [6, 22]:

Theorem 1.2 *Assume that the flux $f : \mathbb{R} \to \mathbb{R}$ is of class C^1. Let $a \in L^\infty(\mathbb{R})$ be given. Then there exists a unique bounded solution of the Cauchy problem satisfying the entropy inequality (1.2.13). It satisfies*

$$\sup_{x\in\mathbb{R},\, t>0} u(x,t) = \sup_{x\in\mathbb{R}} a(x), \qquad \inf_{x\in\mathbb{R},\, t>0} u(x,t) = \inf_{x\in\mathbb{R}} a(x).$$

If b is another bounded data, with associated solution v, then one has the contraction property

$$\int_{A+Mt}^{B-Mt} |u(x,t) - v(x,t)|\, dx \le \int_A^B |a(x) - b(x)|\, dx, \quad \forall A < B,\ \forall t < \frac{B-A}{M}, \tag{1.2.16}$$

where M is the supremum of f' over the interval $[\inf_{x\in\mathbb{R}} a(x), \sup_{x\in\mathbb{R}} a(x)]$.

The Cauchy problem is thus well-understood in the scalar case, and even in several space dimensions ($n = 1$ and $d \ge 1$). The situation is however much more open in case of systems ($n \ge 2$), for we have not any more a comparison or a contraction principle.

1.3 Shock waves

In this section, we investigate in more details the admissible discontinuities, and we introduce new admissibility criteria: the Lax shock inequality and the existence of viscous shock profile.

1.3.1 The Hugoniot locus

To begin with, we describe the set defined by the Rankine-Hugoniot condition (1.2.10), called the *Hugoniot locus* $\mathscr{H}$. Since it is an equation in $\mathbb{R}^n$ with $2n + 1$ parameters $(u_-, u_+; s) \in \mathscr{U} \times \mathscr{U} \times \mathbb{R}$, we expect that the Hugoniot locus be a piecewise smooth manifold of dimension $n + 1$. We observe that it contains the trivial elements $(w, w; s)$ where $w \in \mathscr{U}$ and $s \in \mathbb{R}$ are arbitrary. This exhausts $\mathscr{H}$ in the neighbourhood of $X_0 = (w_0, w_0; s_0)$, whenever the map

$$(w, z; s) \mapsto H(w, z; s) := f(z) - f(w) - s(z - w)$$

is a submersion at X_0, which means that $DH(X_0)$ is onto. Since

$$D_sH(X_0) = 0, \qquad D_zH(X_0)Z = (df(w_0) - s)Z, \qquad D_wH(X_0)Z = -(df(w_0) - s)Z,$$

we find that $\mathscr{H}$ is smooth at X_0 when s is not an eigenvalue of $df(w_0)$.

We already know that the eigenvalues of $df(w)$ are real numbers. From now on, we shall assume that they are *simple* (we say that the system (1.0.1) is *strictly hyperbolic*). We arrange the eigenvalues in increasing order

$$\lambda_1(w) < \cdots < \lambda_n(w).$$

We denote by $r_j(w)$ an eigenvector:

$$df(w)r_j(w) = \lambda_j(w)r_j(w).$$

We also denote $(\ell_1(w), \dots, \ell_n(w))$ a dual basis: each ℓ_j is a differential form and one has

$$\ell_j(w)r_k(w) = \delta_j^k.$$

At a point X_0 with $s_0 = \lambda_j(w_0)$, there is a bifurcation. Lax showed that $\mathscr{H}$ is locally the union of two smooth manifolds. The first one is the trivial one ($z = w$). The second one is parametrized by (w, s) with $z - w \sim (s - \lambda_j(w))r_j(w)$. It can be shown actually that at fixed w_0, the curve $s \mapsto z$ is tangent at second order to the integral curve of r_j, the solution of the differential equation

$$\frac{dZ}{ds} = r_j(Z), \qquad Z(s_0) = w_0.$$

To see this, we use the Taylor formula

$$f(v) - f(u) = \left(\int_0^1 df((1-\tau)u + \tau v)\, d\tau \right) (v - u) =: A(u, v)(v - u).$$

Since $A(u, u) = df(u)$ has simple real eigenvalues, the eigenvalues of $A(u, v)$ are still simple and real for u and v close to each other. We denote them $\Lambda_k(u, v)$. These are smooth functions in a neighbourhood of the diagonal, such that $\Lambda_k(u, u) = \lambda_k(u)$. Likewise, we have eigenfields $R_k(u, v)$, with $R_k(u, u) = r_k(u)$.

The Rankine-Hugoniot condition rewrites as

$$(A(w, z) - sI_n)[u] = 0.$$

Away from the diagonal, the Hugoniot locus is thus described as the union of sets $\mathscr{H}_k$, defined by

$$s = \Lambda_k(w, z), \qquad z - w \parallel R_k(w, z).$$

There remains to solve the equation

$$z - w = \alpha R_k(w, z).$$

To this end, we define $N(w,z;\alpha) := z - w - \alpha R_k(w,z)$. We have $N(w_0,w_0;0) = 0$. Since $D_sN(w_0,w_0;0) = I_n$, this function is a submersion, and its zero set is thus locally a submanifold of codimension n, thus of dimension $n+1$. Its tangent space at $(w_0,w_0;0)$ is given by

$$dz - dw = (d\alpha)R_k(w_0,w_0) = (d\alpha)r_k(w_0).$$

At fixed $w \equiv w_0$, this means that the Hugoniot locus is a curve $\alpha \mapsto (w_0, z(\alpha); s(\alpha))$ with

$$z = w_0 + \alpha r_k(w_0) + O(\alpha^2), \qquad s = \Lambda_k(w_0,z).$$

A refined estimate is that

$$z - w_0 = \alpha r_k\left(\frac{w_0+z}{2}\right) + O(\alpha^3).$$

This curve is denoted by $\mathscr{H}_k(w_0)$. We say that such a triple $(w_0,z;s)$ is a k-discontinuity.

1.3.2 Genuine nonlinearity

Let us investigate the inequality (1.2.14) when $(u_+,s) \in H_k(u_-)$. We again express the jumps $[q(u)]$ and $[\eta(u)]$ with the help of the Taylor formula. With the results of the previous paragraph, we obtain

Lemma 1.1 *Across a k-discontinuity, we have*

$$[q(u)] - s[\eta(u)] = \frac{\alpha^3}{12}(d\lambda_k r_k)D^2\eta(r_k,r_k) + O(\alpha^4).$$

It is time to introduce the following notion:

Definition 1.2 *We say that the k-th characteristic field (λ_k, r_k) is* genuinely nonlinear *at u_- if $d\lambda_k r_k \neq 0$ at u_-.*

When the k-th field is genuinely nonlinear (in short GNL), we can normalize r_k by $d\lambda_k r_k = 1$.

Under the genuine nonlinearity assumption, we see that

$$[q(u)] - s[\eta(u)] \sim \frac{\alpha^3}{12}D^2\eta(r_k,r_k),$$

so that, for small α, the sign of $[q(u)] - s[\eta(u)]$ is that of α. A small k-discontinuity is thus admissible for the entropy inequality if, and only if, α is negative.

1.3.3 The Lax shock inequality

Let us continue our study of small k-shocks under GNL assumption. We have

$$s = \lambda_k(u_-) + \alpha + O(\alpha^2) = \lambda_k(u_+) - \alpha + O(\alpha^2).$$

With α negative and small, this gives

$$\lambda_k(u_+) < s < \lambda_k(u_-), \tag{1.3.17}$$

while non-admissible small k-discontinuities satisfy the opposite inequalities. Therefore (1.3.17) is equivalent to $\alpha < 0$, thus to (1.2.13) for small k-discontinuities. Of course, since $[u]$ is small and the eigenvalues are simple, we also have

$$\lambda_{k-1}(u_-) < s < \lambda_{k+1}(u_+). \tag{1.3.18}$$

Definition 1.3 *Let $(u_-, u_+; s)$ satisfy the Rankine-Hugoniot condition. We say that it is a k-shock if it satisfies also the inequalities (1.3.17, 1.3.18).*

These inequalities are called the *Lax shock inequalities*. For small k-discontinuities, and if the k-th field is GNL, they are equivalent to the entropy inequality.

Shock *vs* entropy inequalities. As mentioned above, both admissibility conditions (shock/entropy inequalities) are equivalent when the discontinuity has a small strength, provided that the system admits a convex entropy, and that the k-th characteristic field is GNL. However they may not be equivalent when one of these assumptions is dropped, for instance the genuine nonlinearity or the smallness of the strength. It may even happen that one condition makes sense while the other one does not. The shock inequality makes sense even when the system does not admit a convex entropy, while the entropy condition does not need that the solution be piecewise smooth.

1.3.4 Viscous shock profiles

A third criterion consists in going back to the parabolic system (1.2.15) and looking for a travelling wave

$$u^\epsilon(x,t) := U\left(\frac{x-st}{\epsilon}\right).$$

Such a u^ϵ is a solution of (1.2.15) whenever U satisfies the ODE

$$U'' = (f(U) - sU)'. \tag{1.3.19}$$

If U has limits $u_\pm$ at $\pm\infty$, then u^ϵ converges boundedly almost everywhere towards

$$u(x,t) := \begin{cases} u_-, & x < st, \\ u_+, & st < x, \end{cases}$$

which will thus be declared admissible. The function U is called the *viscous shock profile* (VSP) of $(u_-, u_+; s)$. Integrating once (1.3.19), we obtain the first-order ODE

$$U' = f(U) - sU - q$$

where q is a constant of integration. This one can be calculated by letting $x \to \pm\infty$; we obtain on the one hand $q = f(u_-) - su_-$ and on the other hand $q = f(u_+) - su_+$. In particular, a necessary condition for the existence of such a U is the Rankine-Hugoniot condition (1.2.10). Finally, we obtain the ODE, called the *profile equation*

$$U' = f(U) - f(u_-) - s(u - u_-). \tag{1.3.20}$$

With a more general viscosity tensor $B(u)$, the profile equation is

$$B(U)U' = f(U) - f(u_-) - s(u - u_-). \tag{1.3.21}$$

Since the calculations of Paragraph 1.2.6 apply to our u^ϵ, we deduce that if the discontinuity $(u_-, u_+; s)$ admits a shock profile, then it satisfies the entropy inequality (1.2.14).

1.4 The Riemann problem

The system (1.0.1), as well as various admissibility criteria, are invariant under the rescaling $(x, t) \mapsto (\mu x, \mu t)$. If the initial data is of the form

$$a(x) = \begin{cases} a_-, & x < 0, \\ a_+, & x > 0, \end{cases}$$

we therefore expect that the solution be homogenous of degree zero: $u(x, t) = V(x/t)$. This must be so if the solution is unique. Finding V when a_- and a_+ are given is the *Riemann problem*. Its solution is used to design some efficient numerical schemes, among which the Godunov scheme (see Section 2.2.4) and the Glimm scheme.

To solve the Riemann problem, we need first to know what are elementary centered waves. We already have encountered the shock waves. We have yet to discover the rarefaction waves and the contact discontinuities.

1.4.1 Rarefaction waves

The rarefaction waves are smooth solutions of the form $u(x, t) = V(x/t)$. They arise when a field is GNL. For such a solution, we have

$$u_t = -\frac{x}{t^2}V', \qquad f(u)_x = \frac{1}{t}(f(V))'.$$

Therefore the system (1.0.1) reduces to

$$\frac{d}{d\xi}f(V(\xi)) - \xi\frac{dV}{d\xi} = 0,$$

that is

$$(df(V(\xi)) - \xi)\frac{dV}{d\xi} = 0. \tag{1.4.22}$$

With $V' \neq 0$, this means that ξ is an eigenvalue:

$$\xi = \lambda_k(V(\xi)). \tag{1.4.23}$$

In addition,

$$V'(\xi) \parallel r_k(V(\xi)). \tag{1.4.24}$$

Differentiating (1.4.23), we obtain

$$1 = d\lambda_k(V)V'.$$

Using (1.4.24), we deduce that the k-th characteristic field must be GNL. We then have

$$\frac{dV}{d\xi} = r_k(V). \tag{1.4.25}$$

An integral curve of (1.4.25) is called a *k-rarefaction curve*. Two states $V(\xi_1)$ and $V(\xi_2)$ on the same curve can be linked by a rarefaction wave in the wedge $t\xi_1 < x < t\xi_2$. Because of (1.4.23), we see that the state with the smallest value of $\lambda_k(V)$ is at left and the other one is at right.

1.4.2 Contact discontinuities

When a field is not genuinely non linear, it may happen the opposite:

Definition 1.4 *The k-th characteristic field is said to be* Linearly degenerate *if $d\lambda_k r_k \equiv 0$.*

When a field is linearly degenerate (LD), the eigenfield r_k does not have a canonical normalization, contrary to the GNL case.

The important point is that linear degeneracy implies a coincidence between rarefaction (notice that rarefaction waves do not exist in this case) and discontinuities, as well as some kind of reversibility:

Theorem 1.3 *Assume that the k-th field is LD. Let γ be a k-rarefaction curve. Then*

i). λ_k is constant along γ,

ii). If $u_\pm \in \gamma$ and $s = \lambda_k|_\gamma$, then $(u_-, u_+; s)$ satisfies the Rankine-Hugoniot relation (1.2.10),

iii). The triple $(u_-, u_+; s)$ also satisfies the identity $[q(u)] = s[\eta(u)]$.

In particular, γ is contained (and in general equals to) the u-projection of $\mathscr{H}_k(u_-)$, for every of its points u_-.

Such triples $(u_-, u_+; s)$ are called *contact discontinuities*. They are always declared admissible. It is clear from above that the triple $(u_+, u_-; s)$ is admissible too: contact discontinuities are reversible.

1.4.3 The theorem of Lax

The general solution of the Riemann problem is made of $n+1$ constant states $u_0 = a_-, u_1, \dots, u_n = a_+$, separated by simple (or composite) waves. Typically, u_{j-1} is separated from u_j by a j-th wave. If the j-th field is LD, the wave is a CD and u_{j-1}, u_j must belong to the same integral curve of r_j. If the j-th field is GNL, the wave is a rarefaction or a shock. When a_- and a_+ are far apart, or when the fields are neither LD nor GNL, then one can find either composite waves, where shocks are embeded in rarefactions, or shocks that are not Lax shocks. Terminology lists under-compressive and over-compressive shocks besides Lax shocks.

Let us assume for instance that the k-th field is GNL at u_-. Then the set of states u_+ (at right) that can be reached from u_- (at left) through a rarefaction is the forward part of the integral curve of r_k starting from u_-. Call it $R_k(u_-)$. The set of states u_+ (at right) that can be reached from u_- (at left) is the backward part ($s < \lambda_k(u_-)$) of the Hugoniot curve $\mathscr{H}_k(u_-)$. Call it $S_k(u_-)$; it is tangent at second order to $R_k(u_-)$. Therefore the union $W_k^f(u_-) = S_k(u_-) \cup R_k(u_-)$ is locally a C^2-curve, tangent to r_k at u_-. It is the k-th *forward wave curve*, named that way because the left state u_- is specified. If we fix u_+ instead and consider the set of states u_- that can be connected to u_+, we find the *backward* wave curve $W_k^b(u_+)$. By definition, we have

$$(u_+ \in W_k^f(u_-)) \Longleftrightarrow (u_- \in W_k^b(u_+)).$$

For a LD field, the forward and backward wave curves coincide and consists in the integral curve of r_k.

Solving the Riemann problem amounts to find the collection of intermediate states $u_1, \dots, u_{n-1}$ such that

$$u_1 \in W_1^f(u_0), \dots, u_j \in W_j^f(u_{j-1}), \dots, u_n \in W_n^f(u_{n-1}). \tag{1.4.26}$$

Using a parametrization of the wave curves, and the fact that they are tangent to r_k at their base point, Lax established the following fundamental result. Again, we refer to [6, 22] for a full proof.

Theorem 1.4 *Assume that the characteristic fields are either GNL or LD. let $a_- \in \mathscr{U}$ be given. Then there exists two neighbourhoods $\mathscr{V} \subset \mathscr{W}$ of a_- such that, if $a_+ \in \mathscr{V}$, then there exists a unique solution of the Riemann Problem in the form of (1.4.26), where all the waves take values in $\mathscr{W}$.*

We point out that a solution to the RP still satisfies $u = U(x/t)$ with

$$\frac{d}{d\xi} f(U) = \xi \frac{dU}{d\xi}$$

in the distributional sense. In particular

$$\frac{d}{d\xi}(f(U) - \xi U) = -U$$

shows that $\xi \mapsto f(U) - \xi U$ is Lipschitz continuous, despite the fact that the solution itself may be discontinuous. This remark is at the basis of the Godunov scheme.

1.5 Existence of viscous shock profiles

We now consider whether or not a given discontinuity $(u_-, u_+; s)$ admits a shock profile. The profile equation, here with a general viscous tensor B, is

$$B(U)U' = f(U) - f(u_-) - s(U - u_-), \tag{1.5.27}$$

We recall that we have assumed the RH condition (1.2.10). Our first assumption about B is that it is invertible, although in realistic examples (like gas dynamics), it would not be. We shall also need an assumption that makes (1.2.15) a parabolic system that stabilizes the constant states, see (**Stab**) below. It will follow from dissipativeness, a rather natural assumption.

The profile equation can be recast as an ODE $U' = G(U; s)$, where we know that $G(u_\pm; s) = 0$. Since we look for a heteroclinic orbit from u_- to u_+ (meaning that $U(\pm\infty) = u_\pm$), a shock profile exists if, and only if, the unstable manifold $W^u(u_-)$ and the stable manifold $W^s(u^+)$, at $u_\pm$ respectively, of this dynamical system have a non-trivial intersection. As a matter of fact, every value $U(x)$ of a shock profile belongs to this intersection, and conversely, if a is in this intersection, then the solution of the Cauchy problem

$$U' = G(U; s), \qquad U(0) = a$$

is a shock profile, because $\lim_{x\to\pm\infty} U(x) = u_\pm$ by assumption.

Structural stability. Assume that s is not an eigenvalue of neither $A(u_-)$ nor $A(u_+)$. Thus $\lambda_j(u_+) < s < \lambda_{j+1}(u_+)$ and $\lambda_{k-1}(u_-) < s < \lambda_k(u_-)$ for some j, k. If $B \equiv I_n$, then the dimensions of $W^u(u_-)$ and $W^s(u_+)$ are respectively $n - k + 1$ and j. It is well-known that the intersection persists under small perturbations of the data (one says that this intersection is *structurally stable*) if, and only if, the tangent spaces to these manifolds add up to $\mathbb{R}^n$; this property is called *transversality*. Since the intersection of these tangent spaces must contain U', thus must be of dimension one at least, this implies that the sum $(n - k + 1) + j$ be at least $n + 1$. Whence the necessary condition for this structural stability: $j \geq k$. The limit case $j = k$ is nothing but the Lax shock condition. Other cases where $j > k$ are called *over-compressive* shocks. When $j < k$, the transversality always fails and the shock is called *under-compressive*.

The discussion above remains valid whenever the system (1.0.1) admits a strongly convex entropy η and the tensor B is *dissipative*, in the sense that

$$X \mapsto D^2\eta(U)(B(U)X, X) \tag{1.5.28}$$

is positive definite for every U.

Since small shocks for GNL fields are Lax shocks, we shall only consider the case $j = k$. More generally, we shall discuss the case where $u_+ \in \mathscr{H}_k(u_-)$.

Exercise. Assume that B is dissipative for some U and that s is not an eigenvalue of $A(u_-)$. Prove that $D_U G(u_-; s)$ does not have a purely imaginary eigenvalue. One says that u_- is a *hyperbolic* fixed point of the ODE. Here *hyperbolic* is in the sense of dynamical systems; it has nothing to do with hyperbolic PDEs.

1.5.1 The scalar case

In the scalar case, B must be positive in order that (1.2.15) be parabolic. Then

$$G(u; s) = \frac{f(u) - f(u_-) - s(u - u_-)}{B(u)}.$$

Every solution of the ODE is monotonous and tends at $\pm\infty$ towards consecutive zeroes of $G(\cdot; s)$. A shock profile exists if, and only if, $G(\cdot; s)$ is strictly of the sign of $u_+ - u_-$ over the interval between u_- and u_+. This amounts to saying that the Oleinik condition is satisfied in a strict sense, with the graph of f strictly above or below its chord.

Therefore, in the scalar case, the existence of a shock profile is, apart from borderline cases, equivalent to the entropy criterion. For instance, if f is convex, both criteria give the same constraint $u_+ < u_-$.

1.5.2 Reduction to a center manifold (bifurcation analysis)

We now develop a strategy for proving the existence of discrete shock profiles associated to discontinuities of small strength. More precisely, we look for profiles of small amplitude, althouh we do not exclude that large amplitude profiles could exist for some small shocks in pathological situations.

The idea is to augment the profile equation (1.5.27) with the obvious one $s' = 0$. Thus we deal with the dynamical system

$$\begin{pmatrix} U \\ s \end{pmatrix}' = H(U; s) := \begin{pmatrix} G(U; s) \\ 0 \end{pmatrix}. \tag{1.5.29}$$

This is the reason why we kept trace of s in the abstract form $U' = G(U; s)$ of the profile equation.

In the following analysis, we choose an index $1 \leq k \leq n$ and we keep u_- fixed. Then we investigate the flow of (1.5.29) in a small enough neighbourhood $\mathscr{V}$ of

$$X_- := \begin{pmatrix} u_- \\ \lambda_k(u_-) \end{pmatrix}.$$

To begin with, we notice that the rest points ($H(X) = 0$) in $\mathscr{V}$ fall into two categories:

i). The pairs $X = (u_-; s)$ for every s close to $\lambda_k(u_-)$,

ii). The points $(u_+; s)$ corresponding to triples $(u_-, u_+; s)$ on the Hugoniot curve $\mathscr{H}_k(u_-)$.

We point out that these two curves are transversal to each other since their tangents at X_- are linearly independent.

We now make a reduction to the *center manifold* $\mathscr{M}_-$ of (1.5.29) at the point X_-. This is a smooth manifold with several properties, among which the local invariance under the flow of the ODE (1.5.29). For a thorough account of what is a center manifold and how one proves its existence, we refer to [4]. The reason why we use it is that $\mathscr{M}_-$ contains every trajectory that is globally defined and is contained in $\mathscr{V}$. In particular, it contains

- The rest points in $\mathscr{V}$,
- The homoclinic and heteroclinic orbits that remain in $\mathscr{V}$.

According to the latter point, $\mathscr{M}_-$ contains all the shock profiles that are associated to discontinuities $(u_-, u_+; s)$ as long as they are entirely contained in $\mathscr{V}$ (in particular the final state u_+ belongs to $\mathscr{V}$).

The center manifold is not always unique, but it has some amount of uniqueness since it necessarily contains some specific orbits (see above). Its dimension c and its tangent space T at X_- are given by the linearized system

$$\begin{pmatrix} U \\ s \end{pmatrix}' = DH(X_-) \begin{pmatrix} U \\ s \end{pmatrix}. \tag{1.5.30}$$

The tangent space T is nothing but the central invariant subspace of $DH(X_-)$, that is the sum of the characteristic spaces associated to the eigenvalues with zero real part. Whence $c = \dim T$ is the sum of the multiplicities of these eigenvalues.

Thus let us investigate the spectrum of

$$DH(X_-) = \begin{pmatrix} D_U G(X_-) & D_s G(X_-) \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} B(u_-)^{-1}(df(u_-) - \lambda_k(u_-) I_n) & 0 \\ 0 & 0 \end{pmatrix}.$$

We shall assume in the sequel that the state u_- is linearly stable for (1.2.15), which means

> (**Stab**) There exists a number $\theta > 0$ such that for every $\xi \in \mathbb{R}$, the eigenvalues of $\xi^2 B(u_-) + i\xi df(u_-)$ have a real part larger than or equal to $\theta\xi^2$.

Exercise. Prove that the dissipativeness of B with respect to a strongly convex entropy implies (**Stab**).

With (**Stab**), we now that $B(u_-)^{-1}(df(u_-) - sI_n)$ is almost a hyperbolic matrix: its only purely imaginary eigenvalue is $\mu = 0$. Actually, one can prove a little bit more:

Lemma 1.2 *Under the stability assumption (***Stab***), one has*

- *For every $k = 1, \dots, n$, $\ell_k B r_k > 0$ at u_-,*
- *And $\mu = 0$ is a simple eigenvalue of $B(u_-)^{-1}(df(u_-) - \lambda_k(u_-))$.*

Under (**Stab**), the dimension of the center manifold is thus $c = 2$. Moreover, since the tangent space at X_- is spanned by the vectors

$$\begin{pmatrix} r_k(u_-) \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

we can take $X \mapsto y := (\ell_k(u_-)(U - u_-), s)$ as local coordinates on $\mathscr{M}_-$. The flow of (1.5.29) over the center manifold can be rewritten as the flow of a tangent vector field h:

$$y' = h(y) := \begin{pmatrix} \ell_k(u_-)G(u(y); y_2) \\ 0 \end{pmatrix}. \tag{1.5.31}$$

Because s is constant along trajectories, the second component h_2 vanishes identically.

We are now in a very simple situation. All the interesting dynamics (small and global trajectories) near X_- is contained in the surface $\mathscr{M}_-$, and the dynamics on $\mathscr{M}_-$ is horizontal ($y_2 = s$ constant). On every horizontal line $y_2 = s$, the dynamics is given by $y_1' = h_1(y_1, s)$, an autonomous differential equation on an interval. Between two consecutive zeroes of $h_1(\cdot, s)$, there is a heteroclinic orbit, which is nothing but a shock profile.

There remains to describe the zero set of h and of $h_1(\cdot, s)$, and to check the sign of the latter function between consecutive zeroes. The zero set is precisely the set of rest points of (1.5.29), already described. At fixed s, the zeroes of $h_1(\cdot, s)$ are therefore of two types:

- either $y_1 = 0$, corresponding to the pair (u_-, s),
- or the non-zero values y_1 corresponding to the states $u_+ \neq u_-$ for which $(u_-, u_+; s)$ is in the Hugoniot set.

1.5.3 Lax shocks

Let us assume in this paragraph that the k-th field is GNL. We recall that the k-th Hugoniot curve at u_- has the property that

$$u = u_- + \alpha r_k(u_-) + O(\alpha^2) \quad (\text{thus } y_1 \sim \alpha), \qquad s = \lambda_k(u_-) + \alpha + O(\alpha^2).$$

Its tangent is therefore transversal to the s-axis. In a neighbourhood of X_-, this curve intersects an horizontal line $y_2 = s$ in exactly one point which we denote $(u_+(s), s)$. Thus the ODE $y_1' = h_1(y_1, s)$ has exactly two fixed points, namely 0 and $\ell_k(u_-)(u_+(s) - u_-)$. Therefore there exists one, and only one, heteroclinic orbit of (1.5.27) between u_- and $u_+(s)$.

When $s \neq \lambda_k(u_-)$, a Taylor expansion using the fact that $u - u_- \sim y_1 r_k(u_-)$ on $\mathscr{M}_-$ gives

$$h_1(y_1, s) \sim (\lambda_k(u_-) - s)y_1.$$

In other words,

$$\frac{\partial h_1}{\partial y_1}(0, s) = \lambda_k(u_-) - s.$$

Therefore $(0, s)$ is an unstable rest point of (1.5.31) if and only if $\lambda_k(u_-) < s$.

The orbit mentioned above thus goes from u_- to $u_+(s)$ if and only if $\lambda_k(u_-) < s$. Since a discontinuity $(u_-, u_+(s); s)$ is either a Lax shock or an anti-Lax one (the latter terminology means that $(u_+, u_-; s)$ is a Lax shock), we have the following conclusion.

Theorem 1.5 *Assume that u_- is linearly stable (property (**Stab**)) for the parabolic equation (1.2.15). Let us assume also that the k-th characteristic field is genuinely nonlinear at u_-.*

Then there are neighbourhoods $\mathscr{V}_- \subset \mathscr{W}_-$ of u_- such that, for every triple $(u_-, u_+; s)$ satisfying the Rankine-Hugoniot condition and $u_+ \in \mathscr{V}_-$, there exists a viscous shock profile from u_- (at left) to u_+ (at right), entirely contained in $\mathscr{W}_-$, if and only if, this discontinuity is a Lax shock.

Among all fields taking values in $\mathscr{W}_-$, this profile is unique, up to the shift $U \mapsto U(\cdot - \xi_0)$.

Remarks.

- Under the assumptions of Theorem 1.5, and if the discontinuity $(u_-, u_+; s)$ does not admit a profile, then it is not a Lax shock. Since the k-th field is GNL and $(u_+, u_-; s)$ is in $\mathscr{H}_k$ with $u_+ \in \mathscr{V}_-$, the latter discontinuity is a Lax shock and therefore it admits a viscous shock profile.

- An important fact is that the existence or the non-existence of a viscous shock profile does not really depend on the choice of a tensor B satisfying (**Stab**). What could depend on this choice is the size of the neighbourhood $\mathscr{V}_-$.

Exercise. What can be said if $d\lambda_k r_k$ vanishes at u_-, but the second derivative $d(d\lambda_k r_k)r_k$ is non-zero ?

1.5.4 Under-compressive shocks

Let $(u_-, u_+; s)$ satisfy the Rankine-Hugoniot condition, and let us assume that

$$\lambda_{k-1}(u_-) < s < \lambda_k(u_-), \qquad \lambda_{k-1}(u_+) < s < \lambda_k(u_+) \tag{1.5.32}$$

for some index k. We say that the discontinuity is an under-compressive shock with *defect index one.*

Since the stable manifold $W^s(u_+)$ has dimension $k-1$ and the unstable $W^u(u_-)$ has dimension $n-k+1$, which sum up to only n, and since the intersection is invariant by the flow of (1.5.27), it is unlikely that these manifolds intersect. Therefore there does not exist a heteroclinic orbit from u_- to u_+ in general. The same is true from u_+ to u_-.

Exercise. Figure out what happens for a differential equation in the plane, where the vector field has two hyperbolic zeroes. The cases saddle–sink or spring–saddle correspond to Lax shocks, while the case saddle–saddle corresponds to the under-compressive situation.

Let us assume however that the map $(a, b; \sigma) \mapsto f(b) - f(a) - \sigma(b-a)$ is a submersion at $(u_-, u_+; s)$. Then the Hugoniot locus is locally a submanifold of dimension $n+1$. It is now a generic fact that the set $\mathscr{P}$ of under-compressive shocks $(a, b; \sigma)$ for which there exists a viscous profile from a to b is a submanifold of codimension one, thus a manifold of dimension n. In favourable cases, $\mathscr{P}$ can be parametrized by either a or b: for every a in some open region, there exists a state $b = \beta(a)$ and a velocity $s = \sigma(a)$, such that $(a, b; s)$ is an undercompressive shock admitting a viscous shock profile. The maps β, σ are smooth.

We warn the reader that, contrary to the Lax case, the manifold $\mathscr{P}$ does depend on the choice of B. It varies smoothly, in general, when B changes. Thus it becomes crucial to have a physically relevant viscosity tensor.

Chapter 2

Finite difference schemes for conservation laws

We now turn to the numerical analysis of first-order systems of conservation laws (1.0.1). Given an initial data (1.1.4), we wish to compute an accurate approximation of the solution of the Cauchy problem. We warn the reader that since we do not have yet shown that the Cauchy problem is well-posed, except in either the scalar case or locally in time for smooth data, it is hard to say that an approximate solution given by a numerical scheme is accurate. The accuracy of a numerical scheme in presence of shocks is an outstanding problem, which has not been fully resolved so far. This problem is at the origin of some questions addressed in the next Chapter.

At the beginning, we give ourselves a mesh length $\Delta x > 0$ and a time step $\Delta t > 0$. We then approximate the space time domain $(0, +\infty) \times \mathbb{R}$ by the grid of points $(t_m := m\Delta t, x_j := j\Delta x)$ for $m \in \mathbb{N}$ and $j \in \mathbb{Z}$. We also use the points $x_{j+1/2}$ defined in the same way.

A discretization of the initial data provides initial values u_j^0. Several choices are possible. For instance, we could take

$$u_j^0 := \int_{x_{j-1/2}}^{x_{j+1/2}} a(x)\, dx, \tag{2.0.1}$$

although it can be easier to set $u_j^0 := a(x_j)$ if a is continuous. We denote $U^0 := (u_j^0)_{j\in\mathbb{Z}}$. It is an element of $\ell^\infty(\mathbb{Z})$ if $a \in L^\infty(\mathbb{R})$. More generally, we have $U^0 \in \ell^p$ provided $a \in L^p$ and we make the choice (2.0.1). When $p = 2$, (2.0.1) amounts to projecting orthogonally over the subspace of piecewise constant elements of $L^2(\mathbb{R})$.

An *explicit* numerical scheme is a mapping $H_\Delta : \ell^\infty \to \ell^\infty$. We use the scheme to define vectors $U^m = (u_j^m)_{j\in\mathbb{Z}}$ inductively by

$$U^{m+1} = H_\Delta(U^m).$$

If the scheme is appropriately chosen, we expect that u_j^m is a good approximation of $u(t_m, x_j)$ where u is the supposed-to-be solution of the Cauchy problem (1.0.1,1.1.4). By a solution, we mean an admissible one, with respect to appropriate entropy conditions.

We warn the reader that since we expect the solution to have discontinuities, a pointwise convergence of the approximate solution towards u might be too ambitious. We should merely ask for a boundedly almost everywhere convergence. This requires to extend the approximate solution to the whole domain $(0,\infty)\times\mathbb{R}$. This is usually done by interpolation. For instance, we may define $u^{\Delta x}\equiv u_j^m$ over the cell $(x_{j-1/2},x_{j+1/2})\times(t_m,t_{m+1})$.

2.1 Conservative schemes

Since the system (1.0.1) commutes with translations, we ask that a scheme has the same property. Also, for a scheme to be of practical interest, we wish that the component $H_\Delta(U)_j$ depends only on finitely many components of U. Therefore a scheme has the general form

$$H_\Delta(U)_j = h(u_{j-p},u_{j-p+1},\dots,u_{j+q}).$$

We say that such a scheme is a $(p+q+1)$-point scheme. For instance, a scheme

$$H_\Delta(U)_j = h(u_{j-1},u_j,u_{j+1})$$

is a 3-point scheme.

We warn the reader that the function h depends also on Δx and Δt. See the examples given in Section 2.2.

The induction defined by a scheme writes

$$u_j^{m+1} = h(u_{j-p}^m,u_{j-p+1}^m,\dots,u_{j+q}^m), \tag{2.1.2}$$

where the initial data u_j^0 is computed from a, as explained above.

Another natural requirement is *conservativity*, in order to mimic the conservation property of (1.0.1). We therefore ask that there is a function F of $p+q$ arguments in $\mathscr{U}$, called the *numerical flux*, such that

$$h(v_{-p},\dots,v_q) = v_0 + \frac{\Delta t}{\Delta x}(F(v_{-p},\dots,v_{q-1}) - F(v_{1-p},\dots,v_q)).$$

Then the scheme rewrites in the natural way

$$\frac{u_j^{m+1}-u_j^m}{\Delta t} + \frac{f_{j+1/2}^m - f_{j-1/2}^m}{\Delta x} = 0, \tag{2.1.3}$$

with

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

For instance, in a 3-point scheme, one has $f_{j+1/2}^m = F(u_j^m,u_{j+1}^m)$.

Once again, the numerical flux may depend on Δx and/or Δt. In practice, it depends only on the ratio $\lambda := \Delta t/\Delta x$, in order to reflect the scale invariance of the PDEs (1.0.1) under the dilations $(x,t)\mapsto(\mu x,\mu t)$.

2.1.1 Consistency

For a scheme to be *consistent* with (1.0.1), it is natural to assume that the numerical flux equals f on the diagonal:

$$F(v, \dots, v) = f(v). \tag{2.1.4}$$

For a consistent scheme, we have the Lax–Wendroff Theorem:

Theorem 2.1 *Assume that the finite difference scheme (2.1.3) is consistent. Let $a \in L^\infty(\mathbb{R})$ be given. Given a sequence $\epsilon_k \to 0$ and a number $\lambda > 0$, denote u^{ϵ_k} the approximate solution associated to $\Delta x = \epsilon_k$ and $\Delta t = \lambda \epsilon_k$ (one interpolates u^{ϵ_k} in order to have it defined on $(0, \infty) \times \mathbb{R}$).*

*Let us assume that the sequence u^{ϵ_k} converges boundedly almost everywhere towards a field u. Then u is a weak (*i.e. *distributional) solution of the Cauchy problem (1.0.1, 1.1.4).*

One may ask whether the following partial converse of Theorem 2.1 holds: if the Cauchy problem (1.0.1,1.1.4) admits a smooth solution u over $(0, T) \times \mathbb{R}$, then the approximate solution converges towards u as $\Delta x \to 0$. It turns out that the answer is negative in general under the assumption of consistency only. Such a statement needs an extra assumption, of stability. This is a well-known general fact in numerical analysis, at least in the realm of linear evolution problems.

2.1.2 Order of accuracy

Let us assume that u is a smooth solution of (1.0.1). We know that such solutions exist for rather general smooth data, at least locally in time (Theorem 1.1). Let us fix the grid ratio $\lambda = \Delta t / \Delta x$. We Taylor expand the following expression in terms of Δt:

$$u(x, t + \Delta t) - h(u(x - p\Delta x, t), \dots, u(x + q\Delta x, t)).$$

If u was also a solution of the difference scheme, this should be zero. Since the scheme is consistent and u is smooth, it is certainly an $O(\Delta t^2)$. We say that the scheme is of order ℓ at least if this expression is an $O(t^{\ell+1})$.

This notion of accuracy refers only to the approximation of smooth solutions. In presence of shocks, the situation is not so nice in practice. One observes that second-order or higher-order schemes generate wild oscillations around discontinuities, a kind of *Gibbs phenomenon*. For this reason, second-order schemes are usually completed by *flux limiters* which have the role to cancel such oscillations. The price to pay is the loss of accuracy: the location of the shock waves is computed at first order only. Besides, flux limiters destroy the abstract form (2.1.3) and it becomes almost impossible to make a theoretical analysis of the scheme under consideration.

Numerical viscosity. Let us consider a first-order scheme in conservative form (2.1.3). If u is a smooth solution, we have

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + \Delta t\, \partial_t u(x,t) + \frac{\Delta t^2}{2}\partial_t^2 u(x,t) + O(\Delta t^3) \\ &= u(x,t) - \Delta t\, \partial_x f(u(x,t)) + \frac{\Delta t^2}{2}\partial_x \left(df(u(x,t))\partial_x f(u(x,t))\right) + O(\Delta x^3). \end{aligned}$$

Likewise, we have

$$\begin{aligned} F_{j+1/2} - F_{j-1/2} &= F(u(x_{j+1-p}),\dots,u(x_{j+q})) - F(u(x_{j-p}),\dots,u(x_{j+q-1})) \\ &= \Delta x \sum_{k=1-p}^{q} d_k F(u,\dots,u)\partial_x u + \Delta x^2 \sum_k (k-1/2) d_k F(u,\dots,u)\partial_x^2 u \\ &\quad + \frac{\Delta x^2}{2}\sum_{k,l}(k+l-1) D^2_{k,l} F(u,\dots,u)(\partial_x u, \partial_x u) + O(\Delta x^3) \\ &= \Delta x\, df(u(x))\partial_x u + \Delta x^2 \sum_k (k-1/2) d_k F(u,\dots,u)\partial_x^2 u \\ &\quad + \frac{\Delta x^2}{2}\sum_{k,l}(k+l-1) D^2_{k,l} F(u,\dots,u)(\partial_x u, \partial_x u) + O(\Delta x^3), \end{aligned}$$

where we have denoted $d_k F$ the differential of $F(u_{1-p},\dots,u_q)$ with respect to u_k, and we have used the identity

$$\sum_{k=1-p}^{q} d_k F(u,\dots,u) = df(u),$$

which follows from consistency.

In conclusion, we obtain

$$u(x,t+\Delta t) - h(u(x-p\Delta x,t),\dots,u(x+q\Delta x,t)) = \lambda \frac{\Delta x^2}{2} D + O(\Delta x^3)$$

with

$$\begin{aligned} D &:= \lambda \partial_x (df(u)\partial_x f(u)) - \sum_k (2k-1) d_k F(u,\dots,u)\partial_x^2 u \\ &\quad + \sum_{k,l}(k+l-1) D^2_{k,l} F(u,\dots,u)(\partial_x u, \partial_x u). \end{aligned}$$

We use again consistency to obtain

$$\sum_{k,l=1-p}^{q} D^2_{k,l} F(u,\dots,u) = D^2 f(u).$$

This allows us to simplify the formula for D:

$$D = \partial_x \left(\lambda df(u)\partial_x f(u) + \sum_k (2k-1) d_k F(u,\dots,u)\partial_x u \right) =: -\partial_x (B(u)\partial_x u). \tag{2.1.5}$$

The tensor B, given by

$$B(u) = -\lambda df(u)^2 - \sum_k (2k-1) d_k F(u,\dots,u),$$

is called the *numerical viscosity*.

The fact that the scheme be of order one tells us that B does not vanish. We point out that if v is a smooth solution of the second-order system in conservation form

$$\partial_t v + \partial_x f(v) = \Delta x \partial_x (B(v)\partial_x v), \tag{2.1.6}$$

instead of (1.0.1), then

$$v(x, t+\Delta t) - h(v(x-p\Delta x, t), \dots, v(x+q\Delta x, t)) = O(\Delta x^3).$$

We point out that this v does depend on Δx and is expected to tend towards u as $\Delta x \to 0$. The numerical scheme thus approximates (2.1.6) in a better way than (1.0.1). One says that (2.1.6) is the *equivalent equation* of the difference scheme.

Recall that for the Cauchy problem for (2.1.6) being well-posed, one needs that the spectrum of $B(u)$ be of non-negative real part. This comes naturally as a necessary condition for the stability of the difference scheme.

2.1.3 Linearized L^2-stability

When approximating smooth solutions, it is useful to make a linear analysis, by linearizing both the system (1.0.1) and the difference scheme. We are thus led to the study of the linear system

$$\partial_t u + A(\bar{u})\partial_x u = 0, \tag{2.1.7}$$

together with the linear scheme

$$u_j^{m+1} = u_j^m + \lambda \sum_{k=1-p}^{q} d_k F(\bar{u},\dots,\bar{u})(u_{j+k-1}^m - u_{j+k}^m). \tag{2.1.8}$$

The Cauchy problem for (2.1.7) is well-posed in every space $L^p(\mathbb{R})^n$ under hyperbolicity. Actually the system can be recast as a list of decoupled transport equations and the solution can be computed explicitly. The situation is not so nice in several space dimensions, where we do have well-posedness in L^2, but not in L^p for $p \neq 2$. For this reason, we shall work only in L^2 within this section.

The scheme (2.1.8) defines a linear operator $S_\Delta : \ell^2 \to \ell^2$. The approximate solution at time t_m is given by

$$U^m = (S_\Delta)^m U^0.$$

Recall that the ratio λ is kept fixed. Let us say that $\Delta t = 1/N$ for some large integer N. Then the approximate solution at time $t = 1$ is $(S_\Delta)^N U^0$. If the difference scheme converges in this linear setting, then the Principle of Uniform Boundedness (sometimes called the Banach–Steinhaus Theorem) implies that $\| (S_\Delta)^N \|_{\mathscr{L}(\ell^2)}$ remains bounded as $N \to +\infty$. In other words, a necessary condition for convergence is the (linear) *stability*:

$$\exists\, C < \infty \quad \text{s.t.} \quad \| (S_\Delta)^N \|_{\mathscr{L}(\ell^2)} < C \quad \text{for} \quad N\Delta t \sim 1. \tag{2.1.9}$$

We notice that the stability of the scheme implies

$$\exists\, C < \infty \quad \text{s.t.} \quad \rho(S_\Delta) \leq 1 + C\Delta t, \tag{2.1.10}$$

as $\Delta t \to 0$, where ρ denotes the spectral radius of the linear operator S_Δ in ℓ^2.

Thanks to linearity and translation invariance, we can perform a Fourier transform, a continuous one for (2.1.7) and a discrete one for (2.1.8):

$$\hat{U}^m(\xi) := \sum_{j \in \mathbb{Z}} e^{-ij\xi} u_j^m.$$

We find that

$$\hat{U}^m(\xi) = \sigma(\xi)^m \hat{U}^0, \qquad \forall \xi \in \mathbb{R},$$

for some matrix $\sigma(\xi) \in \mathbf{M}_n(\mathbb{C})$ defined by

$$\sigma(\xi) = I_n + \lambda \sum_{k=1-p}^{q} \left(e^{i(k-1)\xi} - e^{ik\xi} \right) d_k F(\bar{u}, \ldots, \bar{u}).$$

The linear stability thus amounts to saying that

$$\exists\, C < \infty \quad \text{s.t.} \quad \|\sigma(\xi)^N\|_{\ell^2} < C, \quad \forall\, \xi \in \mathbb{R} \text{ and } \quad N\Delta t \sim 1. \tag{2.1.11}$$

This requires

$$\exists\, C < \infty \quad \text{s.t.} \quad \rho(\sigma(\xi)) \leq 1 + C\Delta t, \qquad \forall \xi \in \mathbb{R}. \tag{2.1.12}$$

It may happen that $\sigma(\xi)$ does not depend at all upon Δt, but only on the grid ratio λ. In this particular case (the Godunov scheme for instance), then the stability condition becomes that the set of matrices $\sigma(\xi)^m$ is uniformly bounded in $\xi \in \mathbb{R}/2\pi\mathbb{Z}$ and $m \in \mathbb{N}$. In this situation, (2.1.12) is equivalent to

$$\rho(\sigma(\xi)) \leq 1, \qquad \forall \xi \in \mathbb{R}. \tag{2.1.13}$$

We warn the reader that (2.1.11) is a necessary and sufficient condition for linearized stability, but (2.1.12) or (2.1.13) is only a necessary condition.

Small frequencies. Let us Taylor expand $\sigma(\xi)$ about $\xi = 0$:

$$\sigma(\xi) = I_n - i\xi\lambda df(\bar{u}) + \lambda\xi^2 \sum_{k=1-p}^{q} (k - 1/2) d_k F(\bar{u}, \dots, \bar{u}) + O(\xi^3). \tag{2.1.14}$$

The spectrum of $\sigma(\xi)$ obeys a Taylor expansion, which can be found from (2.1.14) by using the Implicit Function Theorem. For each j, there is a smooth eigenvalue $\xi \mapsto \Lambda_j(\xi)$, such that

$$\Lambda_j(\xi) = 1 - i\xi\lambda\lambda_j + \lambda\xi^2 \ell_j \left(\sum_{k=1-p}^{q} (k - 1/2) d_k F(\bar{u}, \dots, \bar{u}) \right) r_j + O(\xi^3),$$

where we recall that $\ell_j(u)$ and $r_j(u)$ are the eigenform and the eigenvector of $df(u)$, respectively, associated to $\lambda_j(u)$ and normalized by $\ell_j r_j = 1$.

The modulus of Λ_j equals

$$1 + \xi^2 \left(\lambda^2\lambda_j^2 + \lambda \sum_{k=1-p}^{q} (2k-1) \ell_j(\bar{u}) d_k F(\bar{u}, \dots, \bar{u}) r_j(\bar{u}) \right) + O(\xi^3).$$

The necessary condition (2.1.12) thus implies

$$\lambda\lambda_j^2 + \sum_{k=1-p}^{q} (2k-1) \ell_j(\bar{u}) d_k F(\bar{u}, \dots, \bar{u}) r_j(\bar{u}) \le 0, \qquad \forall j = 1, \dots, n. \tag{2.1.15}$$

This is equivalent to saying that

$$\ell_j(\bar{u}) B(\bar{u}) r_j(\bar{u}) \ge 0, \qquad \forall j = 1, \dots, n. \tag{2.1.16}$$

Exercise. In general, the inequality (2.1.16) is independent of the positivity of the spectrum of $B(u)$. Prove however that this positivity implies (2.1.16) when both matrices $df(u)$ and $B(u)$ are symmetric. More generally, assume that (1.0.1) admits a strongly convex entropy η, and that B is η-dissipative in the sense of (1.5.28). Prove that $D^2\eta(r_j, \cdot) = D^2\eta(r_j, r_j)\ell_j$. Deduce that (2.1.16) holds true.

The asymptotic analysis at small frequency has the interest of relying the numerical viscosity B to the linearized stability of the scheme. However it is far from satisfactory in general. As we shall see in the examples below, the stronger constraints that stability imposes often come because of not-to-small values of ξ.

2.1.4 The Courant-Friedrichs-Lewy condition

Since λ is positive, the necessary condition (2.1.15) tells us that that the sum over k has to be negative. Actually, it tells us more than that: λ has to be small enough. Imagine

for instance that the derivatives $d_k F$ of the numerical flux are proportional to λ^{-1}. This sounds reasonnable since both quantities have the dimension of a velocity. Then (2.1.15) provides an upper bound for λ, which is proportional to the inverse of the spectral radius $\rho(df(u))$. An explicit condition relating $df(u)$ and λ will be given on specific examples.

The condition (2.1.15), expressed as an upper bound of $\lambda = \Delta t/\Delta x$, is called the *Courant-Friedrichs-Lewy condition* (CFL). It tells us that Δt must be smaller than a given constant times Δx. When $\Delta t/\Delta x$ is too large, violent instabilities take place, in general in the form of wild oscillations. The numerical solution $u^{\Delta x}$ then does not converge at all and it is impossible to guess what the exact solution looks like.

A practical explanation of (CFL), viewed as a limitation of λ, is given by the propagation with finite velocity. Say for instance that there is a finite constant V such that $\rho(df(u)) < V$ for every relevant value u of the state. Let us consider an initial data a that is constant ($\equiv \bar{u}$) outside a compact interval $[-L, L]$. Then it can be proved that the solution of the Cauchy problem has the property

$$\mathrm{Supp}[u(\cdot,t) - \bar{u}] \subset [-L - Vt, L + Vt]. \tag{2.1.17}$$

This estimate is accurate in the sense that for most initial data of this form, the solution does vary (*i.e.* is not constant) in the domain $-L + t\lambda_1(\bar{u}) < x < L + t\lambda_n(\bar{u})$.

On an other hand, the numerical solution is such that $u^{\Delta x}(x,t) \equiv \bar{u}$ when $x > L + pt/\lambda + O(\Delta t)$ and also when $x < -L - qt/\lambda + O(\Delta t)$. If $u^{\Delta x}$ is going to converge pointwise towards u, then one needs obviously that

$$-L - qt/\lambda \leq -L + t\lambda_1(\bar{u}), \qquad L + t\lambda_n(\bar{u}) \leq L + pt/\lambda$$

for positive t, that is

$$\lambda\lambda_1(\bar{u}) \geq -q, \qquad \lambda\lambda_n(\bar{u}) \leq p. \tag{2.1.18}$$

For instance, in the case of a three-point scheme ($p = q = 1$), one finds the well-known CFL condition

$$\lambda\, \rho(df(\bar{u})) \leq 1, \qquad \forall \bar{u}. \tag{2.1.19}$$

2.1.5 Entropy-consistent schemes

Let us assume that the system (1.0.1) admits an entropy-flux pair (η, q) with as usual $D^2\eta > 0$. One may wander whether the limit of approximate solutions, assuming that it exists, satifies the entropy inequality (1.2.13). The way to ensure that is to ask the scheme to be *entropy-consistent*.

We say that a conservative difference scheme is entropy-consistent if there exists a numerical entropy flux $Q = Q(u_{1-p}, \ldots, u_q)$, consistent in the sense that $G(u, \ldots, u) \equiv q(u)$ and such that, whenever

$$v := u_0 + \lambda(F(u_{-p}, \ldots, u_{q-1}) - F(u_{1-p}, \ldots, u_q)),$$

one has

$$\eta(v) \leq \eta(u_0) + \lambda(G(u_{-p}, \dots, u_{q-1}) - G(u_{1-p}, \dots, u_q)). \tag{2.1.20}$$

This inequality is the discrete counterpart of (1.2.13). The Lax-Wendroff Theorem 2.1 extends to the context of entropy-consistent schemes, in the sense that if $u^{\Delta x}$ converges boundedly almost everywhere, then its limit is not only a weak solution, but it is an entropy solution: It satisfies (1.2.13) in the distributional sense.

Entropy consistency provides a non-linear form of stability. If a is constant ($\equiv \bar{u}$) outside of a compact interval, we may assume (up to the addition of an affine function to η) that $\eta(\bar{u}) = 0$ and that η is positive otherwise. Then

$$\sum_{j\in\mathbb{Z}} \eta(u_j^0)$$

is finite. Because of (2.1.20), this remains true for the approximate solution:

$$\sum_{j\in\mathbb{Z}} \eta(u_j^0) \leq \sum_{j\in\mathbb{Z}} \eta(u_j^0) \leq \Delta x \int_{\mathbb{R}} \eta(a(x))\, dx.$$

This provides an *a priori* estimate of $u^{\Delta x}$ in some Lebesgue space $L^\infty(0, +\infty; L^p(\mathbb{R}))$ or in an Orlicz space.

By a linearization procedure, it can be proved that an entropy-consistent scheme is linearly L^2-stable. Such a scheme does have numerical viscosity: it cannot be second-order accurate or more.

2.2 Examples

2.2.1 The naive centered scheme

The simplest way to approximate (1.0.1) is to replace the time derivative by a backward difference and space derivative by a centered difference:

$$\partial_t u \mapsto \frac{u_j^{m+1} - u_j^m}{\Delta t}, \qquad \partial_x f(u) \mapsto \frac{f(u_{j+1}^m) - f(u_{j-1}^m)}{2\Delta x}.$$

This yields a three-point scheme with the numerical flux

$$F(u_0, u_1) = \frac{1}{2}(f(u_0) + f(u_1)).$$

We find easily the numerical viscosity

$$B(u) = -\frac{1}{2}(df(u))^2.$$

Since $\ell_j B r_j = -\lambda_j^2/2$ is negative, the linearized stability condition is violated. The centered scheme suffers violent instabilities of Hadamard type, which make it useless in the approximation of the Cauchy problem. This instability was observed by von Neumann in the very first attempt to calculate solutions of gas dynamics.

Exercise. Let us approximate the time derivative by a centered difference too:

$$\partial_t u \mapsto \frac{u_j^{m+1} - u_j^{m-1}}{2\Delta t}.$$

This is the *leap-frog*[1] scheme. Show that it is linearly stable under the CFL condition (2.1.19). You are warned that the leap-frog scheme involves two time steps instead of one. Thus you must rewrite it in the form

$$\begin{pmatrix} U^{m+1}(\xi) \\ U^m(\xi) \end{pmatrix} = \Sigma(\xi) \begin{pmatrix} U^m(\xi) \\ U^{m-1}(\xi) \end{pmatrix}.$$

2.2.2 The Lax–Friedrichs scheme

The Lax-Friedrichs scheme uses the approximation

$$\partial_t u \mapsto \frac{1}{\Delta t}\left(u_j^{m+1} - \frac{u_{j+1}^m + u_{j-1}^m}{2}\right),$$

still with the centered difference for space derivative. It thus writes

$$u_j^{m+1} = \frac{1}{2}(u_{j+1}^m + u_{j-1}^m) + \frac{\lambda}{2}(f(u_{j-1}^m) - f(u_{j+1}^m)).$$

It is a three-point scheme (although u_0 is not present in H_Δ) with numerical flux

$$F_{LF}(u_0, u_1) = \frac{1}{2\lambda}(u_0 - u_1) + \frac{1}{2}(f(u_0) + f(u_1)).$$

The numerical viscosity is

$$B_{LF}(u) = \frac{1}{2}(\lambda^{-2} I_n - (df(u))^2).$$

The stability condition (2.1.15) at small frequencies thus writes as (2.1.19).

Let us now investigate the more precise condition (2.1.13). We first compute the matrix $\sigma(\xi)$:

$$\sigma_{LF}(\xi) = (\cos\xi) I_n - i\lambda(\sin\xi) df(\bar{u}).$$

For the spectral radius of $\sigma_{LF}(\xi)$ to be less than one at every $\xi \in \mathbb{R}$, it is necessary and sufficient to have (2.1.15). Therefore this CFL condition ensures the linearized stability of the Lax-Friedrichs scheme.

In practice, one observes acceptable numerical results with the Lax-Friedrichs scheme under the CFL condition. The drawback is that the shock waves are smeared because of a rather high numerical viscosity. Understanding this phenomenon is one task of the next chapter.

The Lax-Friedrichs scheme is entropy-consistent, with the numerical entropy flux

$$Q_{LF} = \frac{1}{2\lambda}(\eta(u_0) - \eta(u_1)) + \frac{1}{2}(q(u_0) + q(u_1)).$$

[1] The French expression *saute-mouton* translates as *leap-sheep*, although Brittons are fond of lamb meat and Frenchs are fond of frogs.

Exercise. Because the formula $u_j^{m+1} = h(u_{j-1}^m, u_{j+1}^m)$ does not involve u_j^m itself, one can express $u_j^{m+2} = \hat{h}(u_{j-2}^m, u_j^m, u_{j+2}^m)$. Show that $\hat{h}$ defines a conservative difference scheme. Write explicitly its numerical flux $\hat{F}$, then the numerical viscosity $\hat{B}$.

2.2.3 The Lax–Wendroff scheme

The Lax-Friedrichs scheme is only first order, since its numerical viscosity is non-zero. We now present a second-order scheme, due to Lax and Wendroff. It has several variants, depending on the choice for $A_{j\pm1/2}^m$ in the formula below:

$$\begin{aligned} u_j^{m+1} &= u_j^m + \frac{\lambda}{2}(f(u_{j-1}^m) - f(u_{j+1}^m)) \\ &\quad + \frac{\lambda^2}{2}\left(A_{j+1/2}^m(f(u_{j+1}^m) - f(u_j^m)) + A_{j-1/2}^m(f(u_{j-1}^m) - f(u_j^m))\right). \end{aligned}$$

The purpose of the matrix $A_{j+1/2}^m$ is to approximate $df(u)$ at $(x_{j+1/2}, t_m)$. Convenient choices are

$$A_{j+1/2}^m = df\left(\frac{u_{j+1}^m + u_j^m}{2}\right), \quad A_{j+1/2}^m = \frac{1}{2}(df(u_{j+1}^m) + df(u_j^m)), \quad A_{j+1/2}^m = A(u_j^m, u_{j+1}^m).$$

What we always need is that

$$(u_j = u_{j+1}) \Longrightarrow (A_{j+1/2} = df(u_j)).$$

Exercise. Write the numerical flux F_{LW}. Check that the numerical viscosity vanishes identically. Interestingly enough, this property does not depend upon the choice of $A_{j\pm1/2}^m$. At last, show that the Lax-Wendroff scheme is linearly stable under the CFL condition (2.1.19).

The Lax-Wendroff is not entropy-consistent, since the numerical viscosity vanishes identically.

2.2.4 The Godunov scheme

The Godunov scheme is a little bit more elaborated. To some extent, it is a finite volume scheme. Once $(u_j^m)_{j\in\mathbb{Z}}$ has been computed, one defines $u^{\Delta x}$ at time $t_m = m\Delta t$ by

$$u(x, t^m) := u_j^m \text{ over } \left(x_{j-1/2}, x_{j+1/2}\right).$$

Thus u is piecewise constant at time t^m. This allows us to solve explicitly the Cauchy problem for (1.0.1) over a time interval $(t^m, t^m + T)$ by gluing the solutions of the Riemann problems between consecutive state (u_j^m, u_{j+1}^m). These solutions agree as long as the waves emanating from the discontinuity do not reach the walls of the cell $(j\Delta x, (j+1)\Delta x)$. Since

the waves typically travel at a velocity $\lambda_k(\bar{u})$ for some k and some $\bar{u}$, the time interval during which we may glue the Riemann problems is at least

$$\frac{\Delta x}{2 \max \rho(df(\bar{u}))}.$$

This can be improved, by remarking that we shall need only to know the values of the solution on the vertical lines defined by $x = x_j$ for $j \in \mathbb{Z}$. Thus it is sufficient that this value is unchanged, even though consecutive Riemann problems interact. It is thus enough that the waves emanating from the points $x_{j\pm1}$ do not reach the line $x = x_j$. This must be true on a time interval twice as big as our first estimate above. We thus allow a time interval

$$\Delta t = \frac{\Delta x}{\max \rho(df(\bar{u}))}.$$

In other words, the Godunov scheme can be used under the CFL condition (2.1.19).

We now construct the values u_j^{m+1}. To do so, we consider the solution U^m constructed above, at time $t_{m+1} - 0$. Then we make an average in each cell $\big(x_{j-1/2}, x_{j+1/2}\big)$:

$$u_j^{m+1} := \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} U^m(y, t_{m+1})\, dy.$$

Integrating over the domain $\big(x_{j-1/2}, x_{j+1/2}\big) \times (t_m, t_{m+1})$ the conservation law (1.0.1) satisfied by U^m, we obtain

$$u_j^{m+1} = u_j^m + \lambda(F_{j-1/2}^m - F_{j+1/2}^m),$$

where the numerical flux is given by

$$F_{j+1/2}^m = F_G(u_j^m, u_{j+1}^m), \qquad F_G(a, b) := f(R(a, b; x/t = 0)).$$

Hereabove, we have denoted $R(a, b; x/t)$ the solution of the Riemann problem between the left state a and the right state b. We recall that this solution depends only on x/t.

We point out that there may be an ambiguity in the value $R(a, b; 0)$, in the case where the Riemann problem admits a discontinuity across $x = 0$. However this is harmless because then the Rankine-Hugoniot condition $[f(u)] = 0$ tells us that the value $f(R(a, b; 0))$ is not ambiguous.

Stability analysis. At a linearized level, $F_G(a, b)$ is replaced by $f(\bar{u}) + df(\bar{u})_+(a - \bar{u}) + df(\bar{u})_-(b - \bar{u}) = df(\bar{u})_+ a + df(\bar{u})_- b$, where $df(\bar{u})_\pm X$ are the projections of $df(\bar{u})X$ on the stable/unstable subspaces of $df(\bar{u})$. We therefore have $d_0 F_G(\bar{u}, \bar{u}) = df(\bar{u})_+$ and $d_1 F_G(\bar{u}, \bar{u}) = df(\bar{u})_-$. Whence the formula

$$\sigma_G(\xi) = I_n + \lambda(e^{-i\xi} - 1)df(\bar{u})_+ + \lambda(1 - e^{i\xi})df(\bar{u})_-.$$

The eigenvalues of $\sigma_G(\xi)$ are the numbers

$$1 + 2(1 - \cos\xi)\lambda|\lambda_j|(\lambda|\lambda_j| - 1), \qquad j = 1, \dots, n.$$

One deduces that the condition (2.1.13) of linearized stability is equivalent to the CFL condition (2.1.19).

Once again, the Godunov scheme gives an acceptable approximation whenever the CFL condition holds for relevant states. Shocks are smeared because of the numerical viscosity. We observe however that the viscous tensor

$$B(u) = |df(u)|(I_n - \lambda|df(u)|) \qquad (\text{with } |df(u)| := df(u)_+ - df(u_-))$$

vanishes in the direction of the kernel of $df(u)$. This suggests that steady discontinuities are not smeared. This point will be studied in the next Chapter.

The Godunov scheme is entropy-consistent, with numerical entropy flux

$$Q_G = q(R(u_0, u_1; 0)).$$

We point out that this flux is not that well defined if the Riemann problem admits a steady discontinuity. In this case, Q_G should be multi-valued:

$$Q_G(u_0, u_1) = [q(R(u_0, u_1; 0+)), q(R(u_0, u_1; 0-))].$$

2.3 Schemes for scalar equations

We have seen in Paragraph 1.2.7 that for scalar equations ($n = 1$), the Cauchy problem is well-posed in the L^∞ class. Actually, the estimate (1.2.16) shows that well-posedness holds true in $L^1(\mathbb{R})$ too. We thus have an alternative to the L^2-theory. In particular, we may expect to work directly at the non-linear level.

The well-posedness of the Cauchy problem is a consequence of the following properties:

Comparison: If $a \leq b$ almost everywhere, then the entropy solutions u and v evolve accordingly: $u \leq v$,

Conservation: If $b - a \in L^1(\mathbb{R})$, then $v(\cdot, t) - u(\cdot, t) \in L^1(\mathbb{R})$ and

$$\int_{\mathbb{R}} (v(x,t) - u(x,t))\, dx = \int_{\mathbb{R}} (b(x) - a(x))\, dx, \qquad \forall t > 0.$$

Exercise: Show that the comparison and conservation imply together the *contraction*: If $b - a \in L^1(\mathbb{R})$, then

$$\int_{\mathbb{R}} |v(x,t) - u(x,t)|\, dx \leq \int_{\mathbb{R}} |b(x) - a(x)|\, dx, \qquad \forall t > 0.$$

At the discrete level, we should like to preserve the above properties of comparison and conservation. They imply a form of nonlinear stability. Together with the consistency, we expect that they yield convergence results.

2.3.1 Monotone schemes

We say that a conservative difference scheme for a scalar equation is *monotone* if it satisfies the following comparison principle: Given two approximate solutions, if $u_k^m \le v_k^m$ for a given m and every k, then $u_j^{m+1} \le v_j^{m+1}$. This amounts to saying that

H_Δ is a non-decreasing function of each of its arguments.

Exercise. Show that, under the CFL condition $\lambda|f'| < 1$, the Lax-Friedrichs and the Godunov schemes are monotone.

Exercise. Show that the Lax-Wendroff scheme cannot be monotone.

The drawback of monotone schemes is that they may not be high order accurate:

Proposition 2.1 *In the scalar case, consider a monotone scheme. Let us assume the consistency and the CFL condition (2.1.19).*

Then the scheme has positive numerical viscosity, unless the flux F depends either only on u_0 or only on u_1, and the CFL condition is an equality.[2]

Proof.
Denoting $a_k := d_k F(u, \dots, u)$, the numerical viscosity is

$$b(u) = -\lambda f'(u)^2 - \sum_{1-p}^{q} (2k-1)a_k = -\lambda \left| \sum_{1-p}^{q} a_k \right|^2 - \sum_{1-p}^{q} (2k-1)a_k.$$

Because of monotonicity, one has

$$a_1 \le a_2 \le \cdots \le a_q \le 0 \le a_{1-p} \le \cdots \le a_0, \qquad 1 + \lambda(a_1 - a_0) \ge 0.$$

Finally, the consistency and the CFL condition write

$$\lambda \left| \sum_{1-p}^{q} a_k \right| \le 1.$$

We thus have

$$b(u) \ge - \left| \sum_{1-p}^{q} a_k \right| - \sum_{1-p}^{q} (2k-1)a_k.$$

There remains to show that both

$$\pm \sum_{1-p}^{q} a_k - \sum_{1-p}^{q} (2k-1)a_k$$

[2] This combination of borderline properties is very much unlikely.

are positive, unless the equality case. This follows from the expressions

$$\sum_{1-p}^{q} a_k - \sum_{1-p}^{q}(2k-1)a_k = 2\sum_{1-p}^{q}(1-k)a_k$$

and

$$-\sum_{1-p}^{q} a_k - \sum_{1-p}^{q}(2k-1)a_k = -2\sum_{1-p}^{q} k a_k,$$

where the right-hand sides are sums of non-negative terms.

The equality case is left to the reader.

□

2.3.2 Kutznetsov's error estimate

The monotone schemes have been widely used in the 80's because of the following convergence result:

Theorem 2.2 *Let $a \in \bar{u} + L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ be a given initial data for a scalar conservation law (1.0.1). Let $u^{\Delta x}$ be an approximate solution, provided by a monotone difference scheme with a fixed grid ratio λ. We assume that the scheme is consistent with (1.0.1).*

Then $u^{\Delta x}$ converges boundedly almost everywhere towards the Kružkov solution u of the Cauchy problem.

Suppose in addition that the total variation of a is finite. Then one has the estimate

$$\|u^{\Delta x}(\cdot,t) - u(\cdot,t)\|_{L^1} \leq C\sqrt{t\Delta x}\, TV(a). \tag{2.3.21}$$

Comments.

- In the estimate above, the constant C depends only on the Lipschitz constant of f and on that of the numerical flux F.
- When a is not of bounded variations, one still has an estimate, but the right-hand side involves the modulus

  $$\omega(h) := \|a(\cdot + h) - a\|_{L^1},$$

  and the dependence in t and Δx depends on the behaviour of ω near the origin.
- If $f'' > 0$ (genuine nonlinearity, then $\sqrt{t\Delta x}$ can be replaced by $t^{1/4}\sqrt{\Delta x}$. Then the estimates is valuable whenever $t\Delta x \ll 1$.

Chapter 3

Discrete shock profiles

So far, our analysis of the consistency of conservative difference schemes has been limited to the case where the underlying solution is smooth, thus slowly varying. This limit manifests itself in (2.1.4), where we understate that $u^n_{j-p}, \dots, u^n_{j+q}$ are close to a constant state v. This is by no means correct in presence of a jump. When the solution of the Cauchy problem admits a discontinuity from u_- to u_+ across a line $x = X(t)$, we expect that $u^n_{j-p} \sim u_-$ and $u^n_{j+q} \sim u_+$ when the mesh $(n\Delta t, j\Delta x)$ is somewhere close to the shock front. In this situation, we need another notion of consistency, which tells us how the approximate solution varies in a region of width $O(\Delta x)$ around the front. The appropriate concept is that of *Discrete shock profile*.

3.1 DSPs and conservation

We wish to mimic the notion of viscous shock profile. Let $(u_-, u_+; s)$ be a triplet satisfying (1.2.10). A discrete shock profile would be a travelling wave

$$U\left(\frac{x - st}{\epsilon}\right), \qquad U(\pm\infty) = u_\pm,$$

defined at every grid point $(t, x) = (n\Delta t, \Delta x)$. Therefore the argument $(x - st)/\epsilon$ runs over the additive subgroup $\mathbb{Z} + \eta\mathbb{Z}$, where we have defined

$$\eta := s\frac{\Delta t}{\Delta x} = s\lambda.$$

In addition, the wave has to be an exact solution of the numerical scheme (as a viscous profile was an exact solution of the parabolic conservation law (1.2.15)). Whence the definition:

Definition 3.1 *A* Discrete shock profile *(DSP) is a function*

$$U : \mathbb{Z} + \eta\mathbb{Z} \to \mathscr{U},$$

satisfying

$$\lim_{y\to\pm\infty} U(y) = u_\pm$$

and the profile equation

$$\begin{aligned} U(y-\eta) &= U(y) + \lambda\{F(U(y-p),\dots,U(y+q-1)) \\ &\quad -F(U(y-p+1),\dots,U(y+q))\}, \qquad \forall y \in \mathbb{Z}+\eta\mathbb{Z}. \end{aligned} \tag{3.1.1}$$

We immediately see a new feature in this beginning theory. The subgroup $\mathbb{Z}+\eta\mathbb{Z}$ can be discrete (when η is rational) or dense (when η is irrational). Thus one speaks of the *rational case* and of the *irrational case*.

Rational case. When $\eta \in \mathbb{Q}$, we write $\eta = \frac{m}{\ell}$ as an irreducible fraction. Then the domain $\mathbb{Z}+\eta\mathbb{Z}$ equals $\ell^{-1}\mathbb{Z}$. We ask that the profile equation (3.1.1) be satisfied for every $y \in \ell^{-1}\mathbb{Z}$. We shall often treat this equation as a dynamical system for a diffeomorphism.

Irrational case. When $\eta \notin \mathbb{Q}$, the domain is dense in $\mathbb{R}$. It becomes natural to look for a *continuous* travelling wave U, defined over the whole line $\mathbb{R}$. Then the equation (3.1.1) has to be satisfied for every $y \in \mathbb{R}$. This is a much more difficult situation from the analytical point of view.

We notice that we pass continuously from one case to the other one. We may either keep the triplet $(u_-, u_+; s)$ fixed and let vary the aspect ratio λ of the grid, or keep the grid fixed and let the triplet vary, as it would happen generically along a curved shock. We therefore expect that the qualitative results be the same in both cases. Sadly, we shall see that this is not true at all.

The profile equation can be integrated once, as in the viscous case, because of conservativeness. For instance, in the irrational case, we have

$$U(y) - U(y-\eta) = \frac{d}{dy}\int_{y-\eta}^{y} U(\xi)d\xi,$$

while

$$\begin{aligned} &F(U(y-p+1),\dots,U(y+q)) - F(U(y-p),\dots,U(y+q-1)) \\ &\qquad = \frac{d}{dy}\int_{y}^{y+1} F(U(\xi-p),\dots,U(\xi+q-1))d\xi. \end{aligned}$$

The profile equation is thus equivalent to the constency of

$$y \mapsto \int_{y-\eta}^{y} U(\xi)d\xi - \lambda\int_{y}^{y+1} F(U(\xi-p),\dots,U(\xi+q-1))d\xi.$$

From the condition at infinity, together with (2.1.4), we see that this constant must equal both of $\eta u_\pm - \lambda f(u_\pm)$. This in turn implies the Rankine-Hugoniot condition. In conclusion, the Rankine-Hugoniot condition is a necessary condition for the existence of

a DSP (as it was for the existence of a viscous profile). Finally, we have the integrated equation

$$\int_{y-\eta}^{y} U(\xi)d\xi - \lambda \int_{y}^{y+1} F(U(\xi-p),\dots,U(\xi+q-1))d\xi = \lambda(su_- - f(u_-)).$$

Exercise. Assume that the scheme is entropy-consistent. Prove that (1.2.13) is a necessary condition for the existence of a DSP.

In the rational case, the integrated profile equation involves finite sums instead of integrals. We leave it as an exercise.

3.1.1 The function Y

Let U be a DSP. In the rational case, we allow the domain of definition to be an arbitrary set $\mathscr{D}$ with the property that $\mathscr{D} + \ell^{-1}\mathbb{Z} = \mathscr{D}$. We sometimes write $(U;\mathscr{D})$ to recall what is the domain of definition of U. In this case, each of the restriction of U to translates $d + \ell^{-1}\mathbb{Z}$, where $d \in \mathscr{D}$ is given, is a discrete shock profile. Taking two base points d, d' that are not congruent modulo ℓ^{-1} amounts to comparing two distinct DSPs for the same shock. We shall see later on that this is relevant at least for small Lax shocks.

Let us define the following function

$$Y(x;h) := \sum_{y\in x+\mathbb{Z}} (U(y+h) - U(y)), \qquad \forall x, x+h \in \mathscr{D}.$$

Hereabove we assume that U has finite total variation in order to ensure the convergence of the series. This is true for instance in the scalar case, where one can prove the existence of a monotonous DSP with domain of definition $\mathscr{D} = \mathbb{R}$; see below.

Of course, the series

$$\sum_{y\in x+\mathbb{Z}} |U(y)|, \qquad \sum_{y\in x+\mathbb{Z}} |U(y+h)|$$

do not converge, especially because $u_+ \neq u_-$. Therefore one must take care with resummation. For instance,

$$\begin{aligned} Y(x;h+1) - Y(x;h) &= \sum_{y\in x+\mathbb{Z}} (U(y+h+1) - U(y+h)) \\ &= \sum_{y\in x+h+\mathbb{Z}} (U(y+1) - U(y)) = u_+ - u_-!! \end{aligned}$$

This identity does not involve the fact that U is a DSP. It only uses the limits of U at $\pm\infty$. On the contrary, the identity below, which at first glance looks very similar to the

former, is a consequence of (3.1.1):

$$\begin{aligned} Y(x;h-\eta)-Y(x;h) &= \sum_{y\in x+\mathbb{Z}} (U(y+h-\eta)-U(y+h)) \\ &= \sum_{y\in x+h+\mathbb{Z}} (G(y+h)-G(y+h+1)), \end{aligned}$$

where

$$G(y) := \lambda F(U(y-p),\dots,U(y+q-1)).$$

We therefore have

$$Y(x;h-\eta)-Y(x;h) = G(-\infty)-G(+\infty) = \lambda(f(u_-)-f(u_+)) = \eta(u_- - u_+),$$

where we have used (2.1.4) and (1.2.10).

Combining the above calculations, we deduce

Theorem 3.1 *Let us assume that the DSP $U : \mathscr{D} \to \mathscr{U}$ has bounded variations. Then the function Y satisfies*

$$Y(x;h)-Y(x;k) = (h-k)[u], \qquad \forall h,k \in \mathscr{D}.$$

In particular, if $\mathscr{D} = \mathbb{R}$ (for instance in the irrational case), we have

$$Y(x;h) = h[u].$$

3.1.2 Scalar case: monotone schemes

We consider in this paragraph the situation for a scalar equation, for which we employ a monotone scheme. For instance, it could be the Lax–Friedrichs scheme or the Godunov scheme, under the CFL condition. It is not difficult to show that a monotone scheme satisfies a comparison principle, as well as L^1-contraction: If (u_j^n) and (v_j^n) are approximate solutions governed by the scheme, we have

- If $u_j^n \le v_j^n$ for all $j \in \mathbb{Z}$, then $u_j^{n+1} \le v_j^{n+1}$,
- If $u_\cdot^n - v_\cdot^n \in \ell^1$, then

$$\sum_{j\in\mathbb{Z}} |v_j^{n+1}-u_j^{n+1}| \le \sum_{j\in\mathbb{Z}} |v_j^n - u_j^n|.$$

These properties have the consequences that whenever U is a discrete shock profile (with $\mathscr{D}$ minimal), it is monotonous. Using then the function Y, we deduce

Theorem 3.2 *Consider a monotone conservative scheme for a scalar conservation law. Then every DSP (with minimal domain) is monotonous and Lipschitz continuous:*

$$|U(y)-U(z)| \le |(y-z)[u]|.$$

The existence of a DSP for a given shock can be obtained in two steps. When dealing with a strictly monotone scheme in an interval that contains $u_\pm$, ordering arguments were employed by G. Jennings [12] to prove that every Lax shock with a rational η admits a one-parameter family of DSPs, whatever the strength $|u^r - u^l|$ of the shock. For every u^* taken in (u^l, u^r) (or (u^r, u^l)), there exists a unique DSP with $u_0 = u^*$. As mentioned above, this DSP is itself strictly monotone.

Jennings claimed that this result could be extended to irrational values of η. However, his density argument did not contain any detail. The question has been therefore considered as open for a long time. The gap was filled recently in [24], using the function Y described above. The Lipschitz estimate in Theorem 3.2 provides a compactness argument. This method allows to relax also the strict monotonicity. In particular, it handles the case of the Godunov scheme, for which Jennings' proof was powerless even in the rational case. We now have an as general as possible existence and uniqueness theorem, since only the monotonicity of the scheme over (u^-, u^+) is required. The final result is

Theorem 3.3 *Let us consider a scalar conservation law and assume that the conservative difference scheme is monotone (not necessarily strictly) in some interval I. Then every shock $(u^-, u^+; s)$ satisfying the Oleinik condition with strict inequalities admits a continuous DSP, defined on the whole line $\mathbb{R}$.*

In the rational case, this result tells us that the shock admits a continuum of DSPs. For this reason, we often speak of *continuous* DSPs ; even though this terminology is a bit paradoxical.

An explicit DSP. In general, it is rather difficult to provide DSPs in closed form. However there is a special case where this is possible. Let us consider the scalar equation

$$\partial_t u + \partial_x f(u) = 0, \qquad f(u) := -\frac{2}{\lambda} \log\cosh \frac{u}{2}$$

which we approximate through the Lax–Friedrichs scheme. Using the Hopf–Cole transformation, P. Lax found the following formula for the DSP:

$$U(y) = \log \frac{a^{y+1} + 1}{a^{y-1} + 1},$$

where a is the unique root of

$$2a^{-\eta} = a^{-1} + a, \quad a \neq 1.$$

This travelling wave connects the states $u = 0$ and $u = 2 \log a$, in an order that depends of the position of a w.r.to 1, that is of the sign of s (still, $\eta := s\lambda$). With the addition of a constant to U, this formula provides a DSP for every Oleinik shock of our conservation law.

We point out that the domain of definition of U is the whole line, as expected from Theorem 3.3. This justifies the fact that, even in the rational case, we look for continuous DSPs, at least in the case of Lax shocks.

3.2 Existence theory for rational η

We discuss in this section the tools for the existence of DSPs in the rational case. They follow the ideas used for VSPs, borrowed from dynamical systems theory. The main modification is that instead of working with vector fields, we work with diffeomorphisms. Thus the relevant theory is that of discrete dynamical systems.

For this procedure to apply, we need that the numerical flux be an invertible function of its extreme arguments. This works for instance for the Lax–Friedrichs scheme, but not for the Godunov scheme.

For a short account of the Center Manifold Theorem, taylored for its application to bifurcation analysis, we refer to [4].

3.2.1 DSPs for small steady Lax shocks

For the sake of simplicity, we assume a three-point scheme. As mentionned above, our flux $F(a,b)$ is invertible with respect to both a and b. The profile equation is integrated once. If $\eta = m/\ell$, this yields

$$\sum_{j=0}^{m-1} U\left(y-\frac{j}{\ell}\right) - \lambda \sum_{k=1}^{\ell} F\left(U\left(y-1+\frac{k}{\ell}\right), U\left(y+\frac{k}{\ell}\right)\right) = m(u_- - sf(u_-)).$$

By the CFL condition, we know that $|m| < \ell$, and therefore this equation is equivalent to an induction of the form

$$U(y+1) = \phi\left(U(y-1), \dots, U\left(y+\frac{\ell-1}{\ell}\right)\right).$$

Expanding our unknow as

$$V(y) := \left(U(y-1), \dots, U\left(y+\frac{\ell-1}{\ell}\right)\right),$$

our problem can be recast as finding a heteroclinic orbit from $V_- := (u_-, \dots, u_-)$ to $V_+ := (u_+, \dots, u_+)$ of a dynamical system

$$V\left(y+\frac{1}{\ell}\right) = \Phi(V(y)). \tag{3.2.2}$$

As in Section 1.5, we have to let the triplet $(u_-, u_+; s)$ vary. However, we need that the integrated profile equation keep a fixed form. In particular, we want to keep a same size ℓn of the unknown. For this reason, we ask that η remains constant. This can be achieved in two ways: Either one keeps a fixed grid and let vary u_- and u_+ *simultaneously*, in such a way that the shock velocity s remains constant. Or one keeps u_- fixed, let vary u_+ and choose the grid ratio according to $\lambda := m/s\ell$. In both cases, we start from a triplet $(u_-, u_-; \lambda_k(u_-))$. When applying a center manifold theorem, we need that the differential

of the diffeomorphism (extended to a larger unknown, say (V, u_+)) have a well-identified eigenspace associated to the eigenvalue $\mu = 1$, and no over eigenvalue on the unit circle (this last part is called *non-resonance*). Then the dynamics reduces to a simpler dynamics over the center manifold.

For a Lax shock, this manifold is of dimension $n+1$. Each line of equation $u_+ = \text{cst}$ is invariant under the dynamics. We can therefore reduce the analysis to such lines, on which we find two fixed points, corresponding to the Hugoniot triplets $(u_-, u_+; s)$ and $(u_+, u_+; \lambda_k(u_+))$. The restriction of the diffeomorphism over such a line $\gamma(u_+)$ preserves the orientation. Therefore every point $V_0 \in \gamma(u_+)$ between the fixed points serves to build a DSP, through $V(j) := \Phi^{(j)}(V_0)$. We obtain in this way a continuum of DSPs. In other words, we are in presence of a 'continuous' DSP. Mind that such a continuous DSP is far from unique, even up to a shift. For if $\rho = \mathbb{R} \to \mathbb{R}$ is strictly increasing and such that $\rho(y + 1/\ell) = \rho(y) + 1/\ell$, and if U is a continuous DSP, then $U \circ \rho$ is an other one. In practice, we do not distinguish both.

In the construction above, the Lax shock inequalities (1.3.17,1.3.18) guarantee that the profile goes from u_- to u_+, as in the viscous case. For a complete proof of the following result, we refer to ([25]. The result itself is due to Majda and Ralston [18]. See also [19].

Theorem 3.4 *Assume that the numerical flux be invertible to its extreme arguments. Let $u_- \in \mathscr{U}$ and λ_0 be given, in such a way that the k-th characteristic field be GNL at u_-, and $\eta := \lambda_k(u_-)\lambda$ be rational, $\eta = m/\ell$. We also assume some linear stability and a non-resonance condition for the scheme (details are omitted).*

Then there exists two neighbourhoods $\mathscr{V} \subset \mathscr{V}_1$ of u_- such that, for every entropy-admissible shock $(u_l, u_r; s)$ with $u_{l,r} \in \mathscr{V}$, and every grid ratio λ such that $s\lambda = m/\ell$, there exists a continuous DSP with values in $\mathscr{V}_1$. In addition this DSP is unique, up to transformations ρ as described above.

Remarks.

- The smallness assumption, represented by the neighbourhood $\mathscr{V}$, depends dramatically upon the denominator ℓ. This set shrinks to $\{u_-\}$ as $\ell \to +\infty$. Therefore this theorem cannot be used to prove an existence result for irrational η's by passing to the limit from rationals to irrationals.

- Amazingly, the non-resonance condition is not satisfied by the Lax–Friedrichs scheme, for an obvious reason: this scheme acts on the grid points with $j+n$ even on the one hand, and on the grid points with $j+n$ odd on the other hand, independently. To get rid of this decoupling, we can iterate twice the scheme and then restrict to the coarser grid of points with j, n even. This new grid has time and space lengths $2\Delta t$ and $2\Delta x$, respectively (λ is thus inchanged). The scheme then rewrites

$$u_j^{n+2} = u_j^n + \lambda(F(u_{j-2}^n, u_j^n) - F(u_j^n, u_{j+2}^n))$$

with numerical flux

$$F_{LF2}(a,b) = \frac{1}{4\lambda}(a-b) + \frac{1}{4}(f(a)+f(b)) + \frac{1}{2}f\left(\frac{a+b}{2} + \frac{\lambda}{2}(f(a)-f(b))\right).$$

For reasonnable fluxes, this new scheme is non-resonant.

Other shocks. For shocks with larger amplitude, we cannot write such a general result. What we can do is to figure out the shape of the intersection of the stable manifold of (3.2.2) at V_+, and its unstable manifold at V_-. While in the case of a Lax shock the sum of their dimensions exceeds by one the dimension $N = \ell n$ of the ambient space, it equals N for undercompressive shock, as defined by (1.5.32). Since these manifolds do not need to have a common tangent vector (contrary to the viscous case), the generic picture is that they intersect transversally along a discrete set. Therefore the situation for under-compressive shocks is completely different from the viscous case: – on the one hand the existence of DSPs is generic instead of exceptional of codimension one, – on the other hand the DSP is not 'continuous', but genuinely discrete. It was shown actually in [21] that the number of 'distinct' DSP for a shock is even. In particular, it is not unique.

3.2.2 DSPs for steady Lax shocks: the Godunov scheme

The procedure described above does not work for the Godunov scheme because its numerical flux is not invertible with respect to either of its arguments. For non-stationary shocks, the existence of DSPs is an open problem, except in the scalar case where we have Theorem 3.3. However, the case of steady shocks can be treated explicitly. In particular, there is no need of a smallness assumption. Given a steady shock $(u_-, u_+; s = 0)$, we shall make two natural assumptions:

- The Riemann problem with a constant initial data $u(t = 0, x) \equiv a$ admits only the constant solution $u \equiv a$. We point out that this assumption is a direct consequence of the entropy inequality (1.2.13) if there is a convex entropy.
- The equation $f(v) = f(u_-)$ has only the two solutions u_- and u_+.

The profile equation reduces to $F_{God}(u_j, u_{j+1}) = f(u_-)$, that is $f(R(u_j, u_{j+1}; 0)) = f(u_-)$. By assumption, this means that

$$u_{j+1/2} := R(u_j, u_{j+1}; 0) \in \{u_-, u_+\}, \qquad \forall j \in \mathbb{Z}.$$

Lemma 3.1 *If $u_{j-1/2} = u_+$, then $u_{j+1/2} = u_+$. Equivalently, if $u_{j+1/2} = u_-$, then $u_{j-1/2} = u_-$.*

Proof.

We proceed *ad absurdum*. Let us assume that $u_{j-1/2} = u_+$ and $u_{j+1/2} = u_-$. This means on the one hand that one can pass from u_+ to u_j with only forward waves (*i.e.* waves with positive velocities) ; we denote by W this self-similar solution. On the other hand, we pass from u_j to u_- by backward waves ; we denote by Z this self-similar solution. Then we construct a solution of the Riemann problem between u_j and itself, using first W, then the steady shock $u_- \mapsto u_+$, and finally Z. This contradicts our uniqueness assumption for $a = u_j$.

□

Lemma 3.1 amounts to saying that there exists an index j_0 such that if $j \leq j_0$, then $u_{j-1/2} = u_-$, while if $j \geq j_0$, then $u_{j+1/2} = u_+$.

Lemma 3.2 *If $j < j_0$, then $u_j = u_-$, while if $j > j_0$, then $u_j = u_+$.*

Proof.

If $j < j_0$, then $u_{j\pm 1/2} = u_-$. This means that we can pass from u_- to u_j by forward waves, and from u_j to u_- by backward waves. By the uniqueness assumption, we deduce that that the Riemann problem between u_j and itself passes through u_-, and therefore $u_j = u_-$, by uniqueness.

□

There remains to identify u_{j_0}. Since $u_{j_0-1/2} = u_-$ and $u_{j_0+1/2} = u_+$, we can pass from u_- to u_{j_0} by forward waves, and from u_{j_0} to u_+ by backward waves. The first property is written $u_{j_0} \in \mathscr{W}_+^f(u_-)$, where f means *forward*, and the subscript + means that u_- is at right (!!) of u_{j_0} in this Riemann problem. Likewise, the second property is written $u_+ \in \mathscr{W}_+^b(u_{j_0})$, or equivalently $u_{j_0} \in \mathscr{W}_-^b(u_+)$. The set of DSPs for the steady shock $(u_-, u_+; 0)$ is thus parametrized by a pair (j_0, a) where $j_0 \in \mathbb{Z}$ and a is any point of the intersection

$$\Lambda := \mathscr{W}_+^f(u_-) \cap \mathscr{W}_-^b(u_+).$$

For a Lax shock $(u_-, u_+; 0)$, this intersection is usually a curve with end points $u_\pm$. We warn the reader that the pairs (j_0, u_-) and $(j_0 + 1, u_+)$ define the same DSP. Therefore the set of DSPs is again a one-parameter set, an infinite 'periodic' curve, smooth away from the points (j_0, u_-).

For instance, let us consider an extreme shock, say an n-shock $(u_-, u_+; 0)$. Then every velocity $\lambda_k(u_+)$ are negative. Therefore $\mathscr{W}_-^b(u_+)$ is a neighbourhood of u_+ ; it is a 'half-space', bounded by the set of states a such that the Riemann problem between a and u_+ is such that $u \equiv a$ precisely on $x < 0$ and only there. In particular, u_- is a boundary point of $\mathscr{W}_-^b(u_+)$. On the other hand, all velocities $\lambda_k(u_-)$ but the last one are negative. Therefore $\mathscr{W}_+^f(u_-)$ is the part of the n-th wave curve of u_-, made of states for which the n-wave is entirely a forward wave. Typically, it is a 'half' of the n-th wave curve, bounded by u_+ since the wave between u_- and u_+ is precisely a steady wave. It is clear on this example that γ reduces to the segment of the n-th wave curve of u_-, of extremities u_- and u_+.

3.2.3 What can go wrong ?

If the realm of DSPs was a perfect world, then there would be some kind of well-posedness, with the following properties:

- For every small Lax shock of a GNL field, there would exists a unique (modulo a diffeomorphism ρ if η is rational) continuous DSP, no matter whether η is rational or not,

- This DSP would be absolutely continuous. In particular, it would have bounded variations,
- As the shock data $(u_-, u_+; s)$ and the grid ratio λ vary, the DSP would vary smoothly.

If all this is true, the function Y varies smoothly with the shock data and λ, thus with η. In particular, Theorem 3.1 would extend to the rational case since irrational numbers are dense. In passing, this would fix the DSP up to a translation. However, the identity $Y(x; h) = h[u]$ would mean that given two genuinely discrete shock profiles U and V over the domain $\ell^{-1}\mathbb{Z}$, the sum of the series

$$\sum_{j\in\mathbb{Z}} (U(j) - V(j)) \tag{3.2.3}$$

is parallel to $u_+ - u_-$. In practice, there is no reason why this should be true, and it is not too difficult to build counter-examples.

For instance, let us consider the Godunov scheme with a steady shock (thus $\ell = 1$). We have described the profiles in the previous paragraph. The profiles U and V are characterized respectively by the pairs (j_0, a) and (j_1, b) with $j_0, j_1 \in \mathbb{Z}$ and $a, b \in \Lambda$. Then the sum above equals $(j_1 - j_0)[u] + a - b$. If this was parallel to $[u]$ for every choice of U and V, then Λ would be the straight segment $[u_-, u_+]$. It is easy to see that in most cases, this is false. For instance, it fails in gas dynamics, where every shock is an extreme shock. As explained above, Λ is the segment between u_- and u_+ in a wave curve, and this wave curve is never a straight line.

In conclusion, something must go wrong in the theory of DSPs. Either there are some small shocks for which no DSP exist. Or DSPs exist but they lack regularity. Actually, Bressan & coll. constructed [2] an example where the tail of a DSP oscillates so much that the profile has unbounded variations. Or the DSPs do not depend smoothly enough on the shock data and the grid ratio.

Special cases. After this discussion, one may wander why everything can go as best as possible in the scalar case. Theorem 3.3 tells us that the continuous DSP always exists, that it is monotone and thus of bounded variation, and uniformly Lipschitz. This seems to contradict the analysis made here, but there is no contradiction at all. In the scalar case, any two numbers are parallel vectors!

Something similar might happen for systems under the following circumstances:

- Every component f_j of the flux, but one, is linear. This applies for instance to the so-called p-system

$$\partial_t u_1 + \partial_x u_2 = 0, \qquad \partial_t u_2 + \partial_x p(u_1) = 0,$$

  or to the full gas dynamics with the equation of state $p = 2\rho e$ (meaning $\gamma = 3$).
- The scheme is Lax–Friedrichs.

It turns out that under these assumptions, the sum (3.2.3) for two DSPs is automatically parallel to $[u]$. Thus there is no obstruction to a well-posedness theory for DSPs. This theory remains however fully open.

Bibliography

[1] V. I. Arnold. *Geometrical methods in the theory of ordinary differential equations*. Grundlehren der mathematischen Wissenschaften **250**. Springer–Verlag, New York (1983).

[2] P. Baiti, A. Bressan, H.-K. Jenssen. Preprint (2003).

[3] S. Benzoni-Gavage, D. Serre. *Multi-dimensional hyperbolic partial differential equations. First-order systems and applications*. Oxford Univ. Press (2007). Oxford, UK.

[4] A. Bressan. Tutorial on the Center Manifold Theorem. *Hyperbolic systems of balance laws*. CIME cours (Cetraro 2003). Springer Lect. Notes in Maths **1911**, pp 327–344. Springer-Verlag, Heidelberg (2007).

[5] M. Bultelle, M. Grassin, D. Serre. Unstable Godunov discrete profiles for steady shock waves. *SIAM J. Numer. Anal.*, **35** (1998), pp 2272–2297.

[6] C. Dafermos. *Hyperbolic conservation laws in continuum physics*. Grundlehren der mathematischen Wissenschaften, **325**. Springer–Verlag (2000), Heidelberg.

[7] B. Fiedler, J. Scheurle. *Discretization of homoclinic orbits, rapid forcing and "invisible chaos"*. Memoirs of the Amer. Math. Soc., **119** (1996), no 570.

[8] E. Godlewski, P.-A. Raviart. *Numerical approximations of hyperbolic systems of conservation laws*. Springer–Verlag, New–York (1996).

[9] S. Godunov. A difference scheme for numerical calculations of discontinuous solutions of the equations of hydrodynamics. *Math. Sb.*, **47** (1959), pp 271–306.

[10] H. Fan. Existence and uniqueness of travelling waves and error estimates for Godunov schemes of conservation laws. *Math. Computation*, **70** (1998), pp 87–109.

[11] H. Fan. Existence of discrete shock profiles of a class of monotonicity preserving schemes for conservation laws. *Math. Computation*, **67** (2000), pp 1043–1069.

[12] G. Jennings. Discrete shocks. *Comm. Pure Appl. Math.*, **27** (1974), pp 25–37.

[13] P. D. Lax. Hyperbolic systems of conservation laws (II). *Comm. Pure Appl. Math.*, **10** (1957), pp 537–566.

[14] R. J. Leveque. *Numerical methods for conservation laws*. Birkhäuser, Basel (1990).

[15] J.-G. Liu, Z. Xin. L^1-stability of stationary discrete shocks. *Math. Comput.*, **60** (1993), pp 233–244.

[16] J.-G. Liu, Z. Xin. Nonlinear stability of discrete shocks for systems of conservation laws. *Arch. Rational Mech. Anal.*, **125** (1994), pp 217–256.

[17] T.-P. Liu, H.-S. Yu. Continuum shock profiles for discrete conservation laws. *Comm. Pure Appl. Math.*, **52** (1999), I. Construction, pp 85–127 & II. Stability, pp 1047–73.

[18] A. Majda, J. Ralston. Discrete shock profiles for systems of conservation laws. *Comm. Pure Appl. Math.*, **32** (1979), pp 445–482.

[19] D. Michelson. Discrete shocks for difference approximations to systems of conservation laws. *Adv. Appl. Math.*, **5** (1984), pp 433–469.

[20] D. Serre. Remarks about the discrete profiles of shock waves. *Matemática Contemporânea*, **11** (1996), pp 153–170.

[21] D. Serre. Discrete shock profiles and their stability. *Hyperbolic problems: Theory, Numerics and Applications*, Zurich 1998. M. Fey, R. Jeltsch eds. ISNM **130**, Birkäuser (1999), pp 843–854.

[22] D. Serre. *Systems of conservation laws, I.* Cambridge Univ. Press. Cambridge (1999).

[23] D. Serre. *Systems of conservation laws, II.* Cambridge Univ. Press. Cambridge (2000).

[24] D. Serre. L^1-stability of nonlinear waves in scalar conservation laws. *Handbook of Differential Equations*, C. Dafermos, E. Feireisl editors. North-Holland, Amsterdam, 2004.

[25] D. Serre. Discrete shock profiles: Existence and stability. *Hyperbolic systems of balance laws*. CIME cours (Cetraro 2003). Springer Lect. Notes in Maths **1911**, pp 79–158. Springer-Verlag, Heidelberg (2007).

[26] L. Ying. Asymptotic stability of discrete shock waves for the Lax–Friedrichs scheme to hyperbolic systems of conservation laws. *Japan J. Indus. Appl. Math.*, **14** (1997), pp 437–468.