Numerical solution of hyperbolic equations

Christophe Besse and Pauline Lafitte

Laboratoire Paul Painlevé UMR CNRS 8524
U.F.R de Mathématiques Pures et Appliquées
Université Lille 1 Sciences et Technologies
Cité Scientifique
59655 Villeneuve d'Ascq – France

1 Introduction

The aim of this chapter is the study of *scalar* conservation laws in *one dimension.*

1.1 Why is it interesting to simulate numerically a solution ?

In general, explicit the expression of solution(s) $u :]0,T[\times\mathbb{R} \longrightarrow \mathbb{R}$ for $T \in \mathbb{R}^+$ of Cauchy problem of the following type

$$\begin{cases} \partial_t u + \partial_x f(u) = 0, \ t \in]0,T[, \ x \in \mathbb{R}, \\ u(0,x) = u^0(x), \ x \in \mathbb{R}, \end{cases} \tag{1}$$

with $f : \mathbb{R} \longrightarrow \mathbb{R}$ of C^∞ class, are not available.

The motivations of numerical simulations are the following:

- to obtain the shape of the functions and to have an idea of their properties, but also the properties of the studied conservations laws;
- to prepare or to replace the laboratory experiments by computations made with computers. Similarly, the experiments remain essential to derive the model equations, to give their parameters,... which therefore allow to validate the numerical solutions obtained in simplified cases.

1.2 General idea of the method

We replace the continuous space-time $[0,\infty[\times\mathbb{R}$ by a discrete set of points/intervals: the space line $\mathbb{R}$ becomes the union of the intervals $[x_{i-1/2}, x_{i+1/2})$, $i \in \mathbb{Z}$ and the time half-line $\mathbb{R}^+$ becomes the union of $[t^n, t^{n+1})$, $n \in \mathbb{N}$. The nodes of the mesh have coordinates $(t^n, x_i)_{n\in\mathbb{N}, i\in\mathbb{Z}}$.

Remarks

1. If we define an increasing sequence of points in space with $\Delta x_j = x_{j+1} - x_j$ and $\Delta t^n = t^n - t^{n-1}$, for all $j \in \mathbb{Z}$ and $n \in \mathbb{N}^*$, we speak about non uniform mesh. If Δt and Δx are constants, then the mesh is said to be uniform.
2. In practice, the number of points in space is obviously finite.
3. Since the involved scheme will be defined by a time recursion, we begin by fixing the space steps $(\Delta x_i)_i$, and then the (possibly recursively defined) time steps (Δt^n).

Let us now introduce the notion of numerically approximated solution. Instead of seeking the solution u of (1) which depends of $(t,x) \in [0,\infty) \times \mathbb{R}$, we compute by recursion the elements of a sequence with two indices $(v_i^n)_{n\in\mathbb{N},i\in\mathbb{Z}}$ which will describe an approximation of u in the following way:

- either v_i^n be an approximated value of u at point (t^n, x_i) – finite difference approximation;
- or v_i^n be an approximated value of the average of u on the domain $[t^n, t^{n+1}] \times [x_i, x_{i+1}]$ – finite volume approximation.

We obviously hope to well approximate the continuous solution of problem (1) when Δt and Δx are small!
We will discover the fundamental notions linked to these requirements: stability, consistency, order, and, of course, convergence.

Remark As we will see it in the next sections, it is important to think about the compromise between the expected precision and the number of points/computing time.

2 Linear equations

2.1 Influence of parameters

We consider the equation

$$u_t + au_x = \varepsilon u_{xx} + \eta u_{xxx},$$

and we are interested in the understanding of the influence of the parameters a, ε and η. Each of them characterizes the following properties:

- a: transport,
- ε: dissipation / damping,
- η: dispersion.

In order to understand their mutual influence, let us study the behavior of a typical plane wave solution

$$u(t,x) = e^{i(\omega t - kx)} e^{-t\alpha}.$$

The term $\omega t - kx$ denotes the phase, ω is the frequency, $T = 2\pi/\omega$ is the period, $\lambda = 2\pi/k$ represents the wave length and $v_p = \omega/k$ the phase velocity.

If u is a solution, then the different parameters have to satisfy

$$(i\omega - \alpha) - aik = -\varepsilon k^2 + i\eta k^3,$$

or again

$$\begin{cases} \alpha = \varepsilon k^2, \\ \omega - ak = \eta k^3. \end{cases}$$

The last relation is the so-called *dispersion relation*.

- *Influence of the dissipation*: The value of the modulus of the initial datum is $|u(0,x)| = 1$, while the modulus of the time evolution is $|u(t,x)| = e^{-t\alpha} = e^{-\varepsilon k^2 t}$. Three cases can be described
  - $\varepsilon = 0$: the modulus remains constant, equal to the value 1.
  - $\varepsilon > 0$: the modulus decays. Therefore, the parameter ε plays the role of a damping coefficient and the equations is **dissipative**. The damping is more important for bigger k.
  - $\varepsilon < 0$: the equation is **anti-dissipative**. It gives rise to unstable phenomena.

- *Influence of dispersion*: the dispersion relation allows to give an expression to the phase velocity
$$\frac{\omega}{k} = a + \eta k^2 := \beta.$$
Then, it is possible to write the wave as $u(t,x) = e^{-ik(x-\beta t)}e^{-\alpha t}$. Let us assume that $\varepsilon = 0$, which means that $\alpha = 0$. Therefore,
  - if $\eta = 0$: the phase velocity is constant, independent of k,
  - if $\eta \neq 0$: the phase velocity is varying.

- *Influence of the transport*: we take $\varepsilon = \eta = 0$. Then, the wave formulation becomes
$$u(t,x) = e^{i(x-at)},$$
which characterizes a translation a of the initial datum $u_0(x) = e^{ix}$ at time t. If $\varepsilon = 0$ and $\eta \neq 0$, the translation also depends on k.

These three phenomena appear with the numerical treatment of hyperbolic equations. It is necessary to limit the dissipation and the dispersion, the anti-dissipation being forbidden.

2.2 Reminder about characteristics

For $f \in C^\infty(\mathbb{R})$, we consider the equation
$$\begin{cases} u_t(t,x) + f(u)_x(t,x) = 0, \ t > 0, \ x \in \mathbb{R}, \\ u(0,x) = u^0(x), \ x \in \mathbb{R}. \end{cases} \tag{2}$$
We look at the trajectory of a particle whose movement is subject to this equation. Let us denote $x(t)$ the position of the particle, $v(t) = \dot{x}(t) = f'(u(x(t)))$ its velocity and $x(0) = x^0$ its initial position. What is the value of $U : t \mapsto u(t,x(t))$ along these curves? Let use derive this solution with respect to t. We have
$$\frac{d}{dt}u(t,x(t)) = \partial_t u(t,x(t)) + \frac{dx}{dt}\partial_x u(t,x(t)) = \partial_t u(t,x(t)) + f'(u(x(t))\partial_x u(t,x(t)) = 0.$$
Therefore, $U(t) = u(0,x^0)$: u is constant along the characteristic curves, which in our specific case are lines: this is a fundamental property of hyperbolic equations, that is **the propagation of finite velocity**, which needs to be preserved numerically.

Immediately, we are led to the problem of crossing characteristics: which meaning can we give to a solution of (2) ?

Indeed, in this case, the uniqueness is no longer ensured, and discontinuous solutions can arise after a certain amount of time even if the initial datum is smooth: these solutions are called shocks.

Then, the solution has a weak meaning: u is a solution to (2) if $u \in L^\infty_{t,x}$, that is
$\forall \phi \in C^1(\mathbb{R}^+ \times \mathbb{R})$, supp $\phi \subset K$ compact,

$$\int_t \int_K [u\phi_t + f(u)\phi_x] dx\, dt + \int_K [u^0(x)\phi(0,x)] dx \; = 0.$$

We call ***Riemann problem*** a problem for which the initial datum is of the following type

$$u(0,x) = \begin{cases} u_l \text{ if } x < 0 \\ u_r \text{ if } x > 0. \end{cases}$$

Its study is fundamental since it is clearly the problem that arises in numerical approximation.

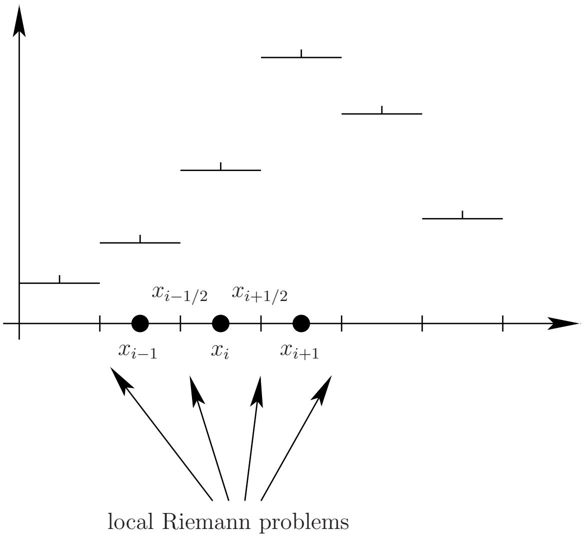


Figure 1: Local Riemann problems

For $i \in \mathbb{Z}$, we define I_i the interval $(x_{i-1/2}, x_{i+1/2})$ also called i$^{\text{th}}$ cell. The length of the interval $h_i = x_{i+1/2} - x_{i-1/2}$ is called mesh size. For $T > 0$, we define $t^0 = 0 < t^1 < \ldots < t^n < \ldots < T$ the time nodes sequence and $\Delta t^n := t^n - t^{n-1}$, $n \geq 1$.

2.3 Finite volume and finite difference schemes

We focus on the linear scalar case $f = au$, $a \in \mathbb{R}$:

$$\begin{cases} u_t(t,x) + au_x(t,x) = 0, \; t > 0, \; x \in \mathbb{R}, \\ u(0,x) = u^0(x), \; x \in \mathbb{R}. \end{cases} \tag{3}$$

By convention, the functions with continuous variables designate the solutions of continuous problems whereas the sequences designate the numerical solutions.

2.3.1 Finite Difference Approach (FD)

We use Taylor expansions to give a direct formulation to differential operators. We treat in this part the approximations of u_t and u_x, assuming that the function u is sufficiently smooth. We have, for all $t > 0$, $x \in \mathbb{R}$,

- in time: forward FD $u_t(t,x) = \lim\limits_{\Delta t \to 0} \dfrac{u(t+\Delta t, x) - u(t,x)}{\Delta t}$
- in space:
  - forward FD $u_x(t,x) = \lim\limits_{\Delta x \to 0} \dfrac{u(t,x+\Delta x) - u(t,x)}{\Delta x}$,
  - backward FD $u_x(t,x) = \lim\limits_{\Delta x \to 0} \dfrac{u(t,x) - u(t,x-\Delta x)}{\Delta x}$,
  - centered FD $u_x(t,x) = \lim\limits_{\Delta t \to 0} \dfrac{u(t,x+\Delta x) - u(t,x-\Delta x)}{2\Delta x}$.

These differences lead to the three following schemes, defined by induction on n for all $i \in \mathbb{Z}$:

- Downwind scheme: $u_i^{n+1} = u_i^n - a\dfrac{\Delta t}{\Delta x}(u_{i+1}^n - u_i^n)$,
- Upwind scheme: $u_i^{n+1} = u_i^n - a\dfrac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n)$,
- Centered scheme: $u_i^{n+1} = u_i^n - a\dfrac{\Delta t}{2\Delta x}(u_{i+1}^n - u_{i-1}^n)$.

Another approach consists in carrying out a Taylor expansion of the solution u before discretizing with finite differences. Following this idea, we have

$$u(t+\Delta t, x) = u(t,x) + \Delta t u_t(t,x) + \frac{1}{2}\Delta t^2 u_{tt}(t,x) + \cdots$$

Although, thanks to the involved scalar hyperbolic equation, $u_t = -au_x$ and then $u_{tt} = -au_{tx} = -au_{xt} = a^2 u_{xx}$. This fact gives us

$$u(t+\Delta t, x) = u(t,x) - \Delta t a u_x(t,x) + \frac{1}{2}\Delta t^2 a^2 u_{xx}(t,x) + \cdots$$

Retaining the first three terms of this expansion and approximating u_x and u_{xx} with centered finite differences, we obtain the ***Lax-Wendroff*** scheme: for $(n,i) \in \mathbb{N}^* \times \mathbb{Z}$,

$$u_i^{n+1} = u_i^n - \frac{a}{2}\frac{\Delta t}{\Delta x}(u_{i+1}^n - u_{i-1}^n) + \frac{a^2}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 (u_{i+1}^n - 2u_i^n + u_{i-1}^n).$$

Exercise: implement these three schemes for $a = 1$ and $a = -1$ when $u^0 : x \mapsto e^{-10x^2}$.

2.3.2 Finite volumes

The idea consists in looking at the equation $u_t + (f(u))_x = 0$, $f(u) = au$ as a conservation law. Then, it is possible to write it as an integral formula

$$\forall t > 0,\ \forall\,(\alpha, \beta) \in \mathbb{R}^2, \quad \frac{d}{dt} \int_\alpha^\beta u(t,x)dx + [f(u(t,\beta)) - f(u(t,\alpha))] = 0.$$

Before going deeper in the details of the method, we have to define some notations. Let use denote $X = \left\{ f \in L^2,\ \forall i,\ f|_{I_i} = f_i = \mathrm{C^t} \right\}$ the set of piecewise functions, constant on each cell. Let g be a given function of L^2. Then, the projection operator on X is characterized by

$$P_X g(y) = \frac{1}{\Delta x_i} \int_{I_i} g(\xi)d\xi = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} g(\xi)d\xi, \quad \forall y \in I_i.$$

We now deal with the approximation of the linear problem $u_t + au_x = 0$, complemented with the initial datum $u(0,x) = u_0(x)$, using the finite volume idea. The approximation is led following three steps. In order to describe them, we assume that the mesh size is constant $\Delta x := \Delta x_i$, $\forall i$.

Step 1 *Projection* of u^0 on X: $u^{0,\Delta x} = u_i^0 = P_{I_i}(u^0)$.

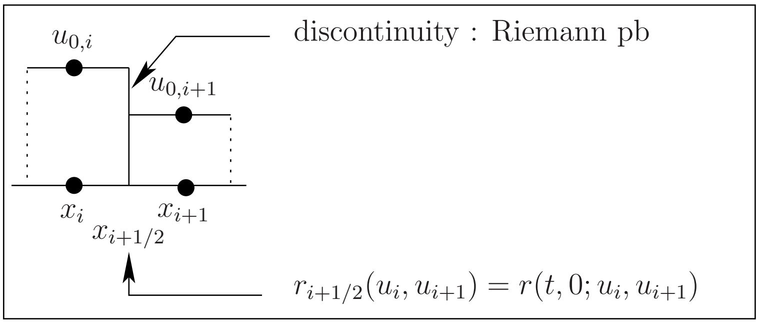

$u_{0,i}$
discontinuity : Riemann pb
$u_{0,i+1}$
x_i
x_{i+1}
$x_{i+1/2}$
$r_{i+1/2}(u_i, u_{i+1}) = r(t, 0; u_i, u_{i+1})$

Figure 2: Riemann problem

Step 2 *Resolution* of the Riemann problem. We look for an approximation of $u(t, x\cdot)$ in $t = \Delta t$. In the precise case of our problem, the propagation velocity of the information is constant, with the value a. This fact leads to a constraint. Indeed, the covered distance during time Δt is $a\Delta t$. Therefore, in order to avoid any crossing of characteristics, it is necessary that this length is less that the half mesh size $\Delta x/2$. This constraint, the so-called CFL condition, can therefore be written as $a\Delta t \leq \Delta x/2$.

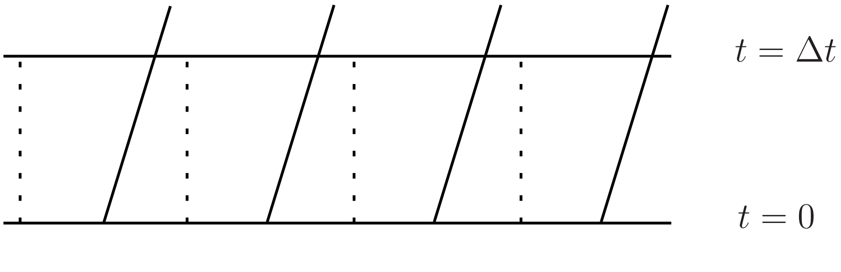


Figure 3: Characteristic lines

With this hypothesis, we have

$$\begin{aligned}
&\int_0^{\Delta t}\left[\frac{d}{dt}\int_{I_i} u(t,x)dx + [au^{\Delta x}(t,x_{i+1/2}) - au^{\Delta x}(t,x_{i-1/2})]\right]dt = 0\\
&\Leftrightarrow \int_{I_i} u^{\Delta x}(\Delta t,x)dx - \int_{I_i} u^{\Delta x}(0,x)dx + \int_0^{\Delta t}[au^{\Delta x}(t,x_{i+1/2}) - au^{\Delta x}(t,x_{i-1/2})]dt = 0\\
&\Leftrightarrow \int_{I_i} u^{\Delta x}(\Delta t,x)dx - \Delta x u_i^0 + \int_0^{\Delta t}[ar_{i+1/2} - ar_{i-1/2}]dt = 0.
\end{aligned}$$

Step 3 We *project again* the result on the constant-by-cell functions (in order to reboot the process to step 1, and to advance from $t=\Delta t$ to $t=2\Delta t$, and so on ...) defining a linear operator of reconstruction:

$$\mathcal{R} : (v_i)_i \in \mathbb{Z} \mapsto v : x \mapsto v_i \text{ if } x \in \ I_i,$$

so $u_i^1 = \dfrac{1}{\Delta x}\displaystyle\int_{I_i} u^{\Delta x}(\Delta t,x)dx$ which leads to

$$u_i^1 = u_i^0 - \frac{\Delta t}{\Delta x}(\quad \underbrace{ar_{i+1/2} - ar_{i-1/2}}_{\text{differences of numerical fluxes}}\quad) = 0$$

Exercise: solve the Riemann problem at point $x_{i+1/2}$ with respect to the sign of a and build the underlying numerical schemes. What are the similarities with the FD? What can you say about the CFL condition?

2.4 Analysis of the schemes

All the schemes that we built have only one time step and therefore can be recast as

$$u_i^{n+1} = G_{\Delta t}[u^n]_i,\ n\in\mathbb{N}^*,\ i\in\mathbb{Z}$$

or again $u^{n+1} = G[u^n],\ n\in\mathbb{N}^*$. For example, for the upwind scheme,

$$G_{\Delta t}[u^n]_i = u_i^n - a\frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n) = (1-a\sigma)u_i^n + a\sigma u_{i-1}^n,\ n\in\mathbb{N}^*,\ i\in\mathbb{Z}$$

with $\sigma = \Delta t/\Delta x$. We extend this notation to the functions by

$$v\in L^2 \mapsto x\in\mathbb{R}\mapsto G_{\Delta t}[v](x) = (1-a\sigma)v(x) + a\sigma v(x-\Delta x).$$

2.4.1 Notions of consistency and order

After n step of a one-step scheme, $n\geq 1$, for solving

$$\begin{cases} u_t + au_x = 0,\ t>0,\ x\in\mathbb{R},\\ u(t^0,x) = u^0(x),\ x\in\mathbb{R},\end{cases}$$

the numerical solution is defined by induction on n by

$$u^{n+1} = \underbrace{G_{\Delta t}\circ\cdots\circ G_{\Delta t}}_{n \text{ times}}[u^0].$$

Denoting $t^{n+1} = t^n + \Delta t$, we define the error **of truncation or of consistency** ε by

$$\varepsilon_i^n = \frac{u(t^{n+1}, x_i) - G_{\Delta t}[(u(t^n, x_j))_{j\in\mathbb{Z}}]_i}{\Delta t}, \ i \in \mathbb{Z}.$$

We will say that a one-step scheme is **consistent with the order p in time and q in space** if it exists $C > 0$ s.t.

$$\forall n \in \mathbb{N}, \ \|\varepsilon^n\| \leq C(\Delta t^p + \Delta x^q),$$

for a norm $\|\cdot\|$ on $\mathbb{R}^{\mathbb{Z}}$. In practice, we use the norms L^∞ $\|u^n\|_\infty = \max_i |u_i^n|$, L^1 $\|u^n\|_\infty = \Delta x \sum_i |u_i^n|$ and L^2 $\|u^n\|_2^2 = \Delta x \sum_i |u_i^n|^2$.
Let us take the example case of the upwind scheme. Then

$$\varepsilon_i^n = \frac{u(t^n + \Delta t, x_i) - (1 - a\sigma)u(t^n, x_i) - a\sigma u(t^n, x_i - \Delta x)}{\Delta t}, \ i \in \mathbb{Z}.$$

Let use assume that u is sufficiently smooth. Thanks to the two following Taylor expansions with respect to time and then with respect to space

$$u(t + \Delta t, x) = u(t, x) + \Delta t u_t(t, x) + \frac{\Delta t^2}{2} u_{tt}(t, x) + O(\Delta t^3)$$

and

$$u(t, x - \Delta x) = u(t, x) - \Delta x u_x(t, x) + \frac{\Delta x^2}{2} u_{xx}(t, x) + O(\Delta x^3),$$

we get

$$\varepsilon_i^n = \underbrace{u_t(t^n, x_i) + a u_x(t^n, x_i)}_{=0} + a(a\Delta t - \Delta x)u_{xx}(t^n, x_i) + O(\Delta t^2) + O(\Delta x^2).$$

We have $\|\varepsilon^n\| \leq C(\Delta x + \Delta t)$ and the scheme is therefore of order 1 in time and 1 in space.

Remark 1. *If $a\sigma = 1$, then $\varepsilon = 0$ and the scheme is of infinite order. In fact, we have $u(t + \Delta t, x) = u(t, x - a\Delta t)$ which corresponds to an exact writing along the characteristic.*

We have to notice that the equation

$$u_t + a u_x = \frac{\Delta x}{2} a(1 - a\sigma)u_{xx}$$

at the order 2. We say that is it the ***equivalent equation*** associated to the numerical scheme at the order 2. We show that the scheme can introduce numerical diffusion:

- if $a\sigma < 1$: this is an equation of diffusion;
- if $a\sigma = 1$: there is no diffusion;
- if $a\sigma > 1$: this is an equation of anti-diffusion.

Exercise: Which is the equivalent equation to the upwind scheme at the order 3? Which notions are then involved?

2.4.2 Notion of stability

We have seen that the truncation error ε plays an important part in the relation

$$u(t^n + \Delta t, x_i) = G_{\Delta t}[(u(t^n, x_j))_j]_i + \varepsilon_i^n, \; i \in \mathbb{Z}, \; n \in \mathbb{N}.$$

Iterating this with time, the local errors are accumulated to make the global error

$$\begin{aligned} e_i^{n+1} &= u_i^{n+1} - u(t^{n+1}, i\Delta x) \\ &= G_{\Delta t}[u^n]_i - G_{\Delta t}[(u(t^n, x_j))_j]_i - \varepsilon_i^n. \end{aligned}$$

If the operator $G_{\Delta t}$ is linear with respect to v (which is the case by now since the schemes are linear), we have

$$e_i^{n+1} = G_{\Delta t}[e^n]_i - \varepsilon_i^n.$$

Therefore,

$$e^n = G_{\Delta t}(e^0) - \sum_{p=1}^{n} G_{\Delta t}^{n-p}[\varepsilon^{p-1}].$$

Thus, it is necessary to control the powers of $G_{\Delta t}$.

A scheme is said to be **stable** for the norm $\| \cdot \|$ if $\forall T$, $\exists C$ and $\Delta t^0 > 0$ such that

$$\forall \Delta t \leq \Delta t^0, \; 0 \leq n\Delta t \leq T, \quad \|G_{\Delta t}^n\| \leq C.$$

In order to better understand the notion of stability, we will study these two norms for the upwind scheme $u_i^{n+1} = (1 - a\sigma)u_i^n + a\sigma u_{i-1}^n$ with $a > 0$.

L^∞ **stability** $\|G_{\Delta t}^n\| \leq C$ is equivalent to $\displaystyle\max_{v \neq 0} \frac{\|G_{\Delta t}^n[v]\|}{\|v\|} \leq C$ or again $\|G_{\Delta t}^n[v]\| \leq \|v\| \; \forall v$. The scheme leads to

$$|u_i^{n+1}| \leq |1 - a\sigma| \, |u_i^n| + |a\sigma| \, |u_{i-1}^n|.$$

- If $0 < a\sigma \leq 1$, we have $\|G_{\Delta t}[u^n]\|_\infty = \|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ and so $\|G_{\Delta t}\|_\infty \leq 1$.
- If $a\sigma > 1$, we can show that the scheme is not L^∞ stable.

L^2 **stability** In order to proceed to the L^2 stability analysis, we are led to use the Fourier transform. We recall here the definition as an operator on $L^1(\mathbb{R})$, then extended to $L^2(\mathbb{R})$ by density of $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$,

$$\xi \in \mathbb{R} \mapsto \mathscr{F}(u)(\xi) = \hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u(x) e^{-ix\xi} dx \quad \text{(defined in the } L^2 \text{ meaning)}$$

an the Plancherel equality $\|\hat{u}\|_2 = \|u\|_2$. Since the Fourier transform is applied to functions, we use the reconstruction operator $\mathcal{R}$

$$\mathcal{R}[(u_i^{n+1})_i] = (1 - a\sigma)\mathcal{R}[(u_i^n)_i] + a\sigma\mathcal{R}[(u_{i-1}^n)_i].$$

After the application of the Fourier transform, we have

$$\hat{u}^{n+1}(\xi) = (1 - a\sigma)\hat{u}^n(\xi) + a\sigma e^{-i\Delta x\xi}\hat{u}^n(\xi) := h(\xi)\hat{u}^n(\xi).$$

The quantity $h(\xi)$ defines the amplification coefficient of the scheme. Therefore, if for all ξ, $|h(\xi)| \leq 1$, then $\|G_{\Delta t}\|_2 \leq 1$ and we have stability. This condition is the so-called **Von Neumann condition**. If $0 < a\sigma \leq 1$, then $|h(\xi)| \leq |1 - a\sigma| + |a\sigma| \, |e^{-i\Delta x\xi}|$ or again $|h(\xi)| \leq 1 - a\sigma + a\sigma \leq 1$.

CFL condition In conclusion, with the condition $0 < a\sigma \leq 1$, the scheme is L^∞ and L^2 stable. This condition is the so-called **CFL condition** following the names of Courant, Friedrichs and Lewy. If we fix for example $a\sigma = 0.8$, we say that the CFL is of 0.8.

Since we know now how to analyze the consistency and the stability, we have a necessary condition of convergence thanks to the next theorem, the so-called Lax theorem: **a linear scheme is convergent if and only if it is stable and consistent**. Its **order of convergence** is the truncation order.

Exercise: implement the different schemes seen until now for

$$\begin{cases} u_t + au_x = 0, \ t > 0, \, x \in \mathbb{R}, \\ u(0,x) = \begin{cases} u_l = 1 & x < 0 \\ u_r = -1 & x > 0 \end{cases} \end{cases}$$

Proceed to the analysis ot these schemes (order, equivalent equation, stability).

After the numerical experiments, one has to be able to see that the centered scheme leads to problems. In order to correct this bad behavior, we add diffusion. One obtains the following scheme:

$$\text{Lax-Friedrichs scheme} \quad u_i^{n+1} = \frac{1}{2}(u_{i+1}^n + u_{i-1}^n) - a\frac{\sigma}{2}(u_{i+1}^n - u_{i-1}^n).$$

Exercise: What are its order, its equivalent equation, its stability?

If we perform the same analysis for the Lax-Wendroff method, we see the formation of problem of under- and over-shooting. The equivalent equation is given by $u_t + au_x = (\Delta x^2/6)a(a^2\sigma - 1)u_{xxx}$. Thus, a dispersion term appears which is synonym to the propagation of harmonic waves with a velocity which depends on the wave numbers.

We can make a first report

Order 1 not very precise, very diffusive because of the numerical viscosity which makes the scheme **robust** to the discontinuities,

Order 2 very precise in the regular (smooth) regions but suffers of problems at discontinuities.

It is difficult to obtain these two behaviors simultaneously.

2.5 Monotone schemes, TVD

The continuous solution of the Cauchy problem (2) has to fundamental properties:

MAXIMUM PRINCIPLE : $\forall(t,x)$, $\min\limits_{R} u^0 \leq \min\limits_{\mathbb{R}^+\times\mathbb{R}} u \leq u(t,x) \leq \max\limits_{\mathbb{R}^+\times\mathbb{R}} u \leq \max\limits_{\mathbb{R}} u^0$,

DECAY OF THE TOTAL VARIATION : $\forall t$, $\int_{\mathbb{R}} |\partial_x u(t,x)|dx = \int_{\mathbb{R}} |(u^0)'(x)|dx$.

In order to succeed in the understanding of the both previous properties, we have to tackle the notions of monotony and of decay of total variation (TVD) with an application to three points schemes. Before to formulate these methods, we have to introduce some notations:

- explicit notation: $u_i^{n+1} = G(u_{i-1}^n, u_i^n, u_{i+1}^n)$;
- conservative notation: $u_i^{n+1} = u_i^n - \sigma(h_{i+1/2}^n - h_{i-1/2}^n)$, where $h_{i+1/2}^n$ designates the numerical flux;
- incremental notation: $u_i^{n+1} = u_i^n + C_{i+1/2}\Delta u_{i+1/2}^n - D_{i-1/2}\Delta u_{i-1/2}^n$ with $\Delta u_{i+1/2}^n = u_{i+1}^n - u_i^n$ and C and D some functions.

Exercise: write the previous schemes in the above proposed formulations.

A numerical scheme is said

MONOTONE if, in its explicit version, G is a nondecreasing function with respect to each of its arguments,

MONOTONY PRESERVING if, for n given, u^n is such that $u_{i+1}^n \geq u_i^n$, $\forall i$, then, $u_{i+1}^{n+1} \geq u_i^{n+1}$, $\forall i$ (and respectively for $\leq$).

The interest of these notations is that a monotone scheme on one hand does not produce oscillations (it is therefore robust), and on the other hand is monotony preserving. Thus, we have the following property: ”*if a scheme is linear, with a three points stencil and monotone, then it is of order 1 and dissipative*“.

However, we look for precise and robust schemes. This last proposition seems to indicate that it is impossible to build a scheme which satisfies these both properties. We introduce the notion of **total variation**

$$VT(u^n) = \sum_i |u_i^n - u_{i-1}^n|.$$

If on considers $u^{\Delta x}$, that is the piecewise constant function s.t. $u^{\Delta x}(x) = u_i^n$ for $x \in I_i$, then

$$VT(u^{\Delta x}) = \int_{\mathbb{R}} |u_x|\, dx.$$

Thus, a numerical scheme has to preserve this property. We will say that a scheme is **TVD (total variation diminishing)** if

$$\forall n \geq 0, \quad VT(u^{n+1}) \leq VT(u^n).$$

A TVD scheme can not generate oscillations at the discontinuities and therefore preserves the monotony. We focus on the construction of TVD schemes which will be almost of order 2, that is to say of order 2 in the continuous regions and of order 1 at discontinuity points. We will be interesting in this course only on ***Slope limiters TVD schemes***.

These schemes are non linear even if the equation is linear.

Affine by cells approximation: process

1. $\{u_i^n\}$ average values by cells → construction $\tilde{u}^n(t^n, \cdot) = R(\cdot, u^n)$ affine by cells reconstruction
2. Exact resolution → u^{n+1}.
3. Projection on the constant-by-cell functions.

Example for $u_t + au_x = 0$ with $a > 0$: $\tilde{u}^n(t^n, x) = u_i^n + \gamma_i^n(x - x_i)$, with γ_i^n the slope of the line $\gamma_i^n = (u_{i+1}^n - u_i^n)/\Delta x$. Then, the reconstruction is

$$R(x, u^n) = u_i^n + (x - x_i)\frac{u_{i+1}^n - u_i^n}{\Delta x}.$$

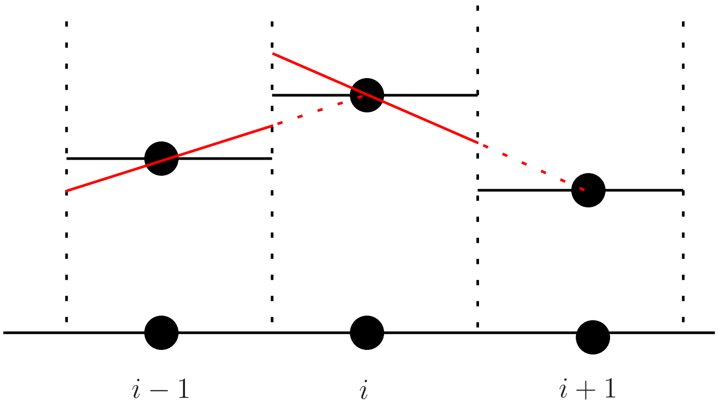


Figure 4: Reconstruction by slope

With this choice, we recover the Lax-Wendroff scheme. Problem: it is not TVD.

Exercise: show it numerically.

However, it is of order 2. One can show that steps 2 and 3 cannot make the scheme TV diminishing. The loss of the monotonicity property comes from the reconstruction step. We see on figure 2.5 in cell I_i that the slope γ_i^n of the line is forced to increase. Therefore, we have to limit the slope. It would be necessary to ensure the TVD property that extremities of this affine line belongs to the interval $[u_i^n, u_{i+1}^n]$. Thus, $\gamma_i^n = (u_{i+1}^n - u_i^n)/\Delta x$ if one does not increase the TV, but has to be chosen smaller otherwise.

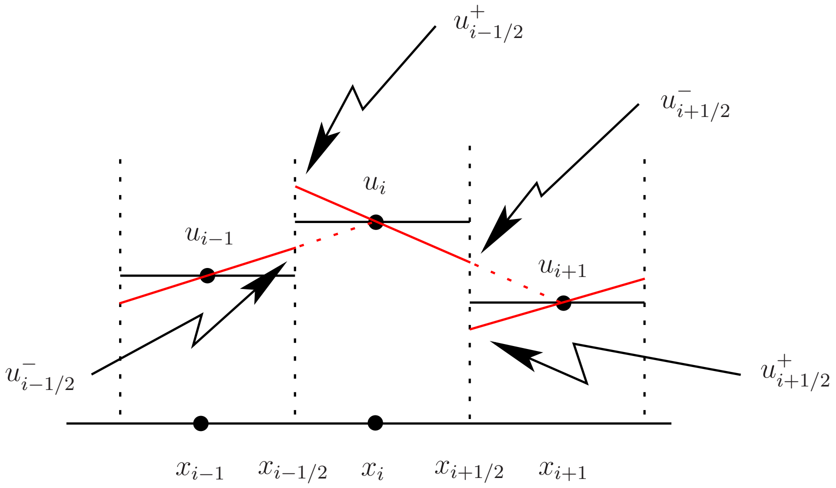


Figure 5: Reconstruction by limited slope

We impose to the modified slope γ_i^{mod} that $u^\pm_{i-1/2} \in [u_{i-1}, u_i]$. With the choice $\gamma_{i-1} = (u_i - u_{i-1})/\Delta x$, we have $u^-_{i-1/2} = (u^n_i + u^n_{i-1})/2 \in [u_{i-1}, u_i]$. Concerning $u^+_{i-1/2}$, three cases may appear

1. $u_{i-1} \geq u_i \geq u_{i+1}$ which expresses a local decay. We want that $u_{i-1} \geq u^+_{i-1/2} \geq u_i$ which is equivalent to $u_{i-1} \geq u_i + \gamma_i(x_{i-1/2} - x_i) \geq u_i$. The second inequality is always true since $\gamma_i < 0$ and $x_{i-1/2} - x_i < 0$. Only the first inequality is therefore a true constraint.

$$\begin{aligned} u_{i-1} \geq u_i - \frac{\Delta x}{2}\gamma_i \quad &\Leftrightarrow \quad -\frac{\gamma_i}{2}\Delta x \leq u_{i-1} - u_i \\ &\Leftrightarrow \quad \frac{|\gamma_i|}{2}\Delta x \leq u_{i-1} - u_i = |u_i - u_{i-1}| \\ &\Leftrightarrow \quad \Delta x|\gamma_i| \leq 2|u_i - u_{i-1}|. \end{aligned}$$

   If no constraint is necessary, then the slope is defined through $\gamma_i = (u^n_{i+1} - u^n_i)/\Delta x$. Thus, the modified slope is defined by

$$\Delta x \gamma_i^{mod} = -\min(2|u_i - u_{i-1}|, |u_{i+1} - u_i|).$$

2. $u_{i-1} \leq u_i \leq u_{i+1}$, which expresses a local growth. We want $u_{i-1} \leq u_i - \gamma_i \Delta x/2 \leq u_i$. Only the first inequality is a constraint. By a reasoning similar to the previous one, we get

$$\Delta x \gamma_i^{mod} = \min(2|u_i - u_{i-1}|, |u_{i+1} - u_i|).$$

3. $u_i \leq u_{i+1}$ and $u_i \leq u_{i-1}$, which is characteristic of a local minimum, then $\gamma_i^{mod} = 0$.

In conclusion, we have

$$\Delta x \gamma_i^{mod} = \text{minmod}(2(u_i - u_{i-1}), (u_{i+1} - u_i)),$$

where

$$\text{minmod}(x_1, x_2, \cdots, x_m) = \begin{cases} 0 \text{ if } x_i \text{ are not of the same sign} \\ \text{sgn}(x_i)\min_i |x_i| \text{ otherwise.} \end{cases}$$

The numerical scheme is therefore written

$$\begin{aligned} u_i^{n+1} &= u_i^n - a\sigma(u_i^n - u_{i-1}^n) + \frac{1}{2}a\sigma(a\sigma-1)\Delta x(\gamma_i^{mod,n} - \gamma_{i-1}^{mod,n}) \\ &= u_i^n - a\sigma(u_i^n - u_{i-1}^n) + \frac{\Delta x}{2}a\sigma(a\sigma-1)(\phi_i^n\gamma_i^n - \phi_{i-1}^n\gamma_{i-1}^n) \end{aligned}$$

where $\phi_i = \text{minmod}(2\frac{u_i-u_{i-1}}{u_{i+1}-u_i}, 1)$ is the slope limiter.
All the previous approaches can be repeated with an other slope choice, for example $\gamma_i^n = (u_{i+1}^n - u_{i-1}^n)/2\Delta x$.

There exist other approaches. We present here Harten's with a scheme written in an incremental formulation.

$$u_i^{n+1} = u_i^n + C_{i+1/2}^n \Delta u_{i+1/2}^n - D_{i-1/2}^n \Delta u_{i-1/2}^n.$$

The sufficient condition to get a TVD scheme written with an incremental formulation is

$$\forall i, \quad C_{i+1/2}^n \geq 0, \quad D_{i-1/2}^n \geq 0, \quad C_{i+1/2}^n + D_{i+1/2}^n \leq 1.$$

The difficulty is that the incremental formulation corresponding to a given numerical scheme is not unique. If one considers the following scheme

$$u_i^{n+1} = u_i^n - a\sigma(u_i^n - u_{i-1}^n) + \frac{\Delta x}{2}a\sigma(a\sigma-1)(\phi_i^n\gamma_i^n - \phi_{i-1}^n\gamma_{i-1}^n)$$

then, we can choose $C_{i+1/2}^n = 0$ and

$$D_{i-1/2}^n = a\sigma - \frac{1}{2}a\sigma(a\sigma-1)\frac{\phi_i^n(u_{i+1}^n - u_i^n) - \phi_{i-1}^n(u_i^n - u_{i-1}^n)}{u_i^n - u_{i-1}^n}.$$

Let use denote $\theta_i = (u_i - u_{i-1})/(u_{i+1} - u_i)$, $\nu = a\sigma$, and define a function ϕ such that $\phi(\theta_i) = \phi_i$. Thus, in order to have a TVD scheme, it is sufficient that $0 \leq D_{i-1/2}^n \leq 1$ which can be expressed by

$$0 \leq D_{i-1/2}^n = \nu\left(1 + \frac{(1-\nu)}{2}\left(\frac{\phi(\theta_i^n)}{\theta_i^n} - \phi(\theta_{i-1}^n)\right)\right) \leq 1.$$

It is possible to show that this is satisfied if the function ϕ satisfies $|\frac{\phi(\theta)}{\theta} - \phi(\theta)| \leq 2$, which can be interpreted as $0 \leq \phi(\theta)/\theta \leq 2$ and $0 \leq \phi(\theta) \leq 2$.

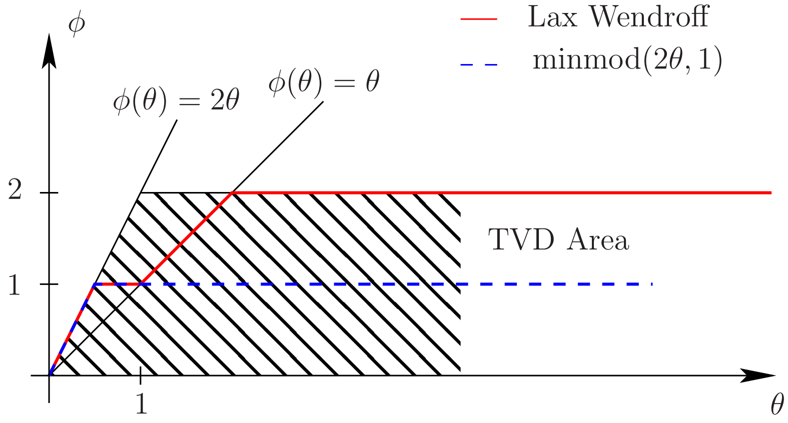


Figure 6: TVD Area

A limiter defined as $\phi(\theta) = \theta$ leads to the Beam-Warming scheme.
If we want to obtain a second order TVD scheme, the TVD area is slightly more forced.

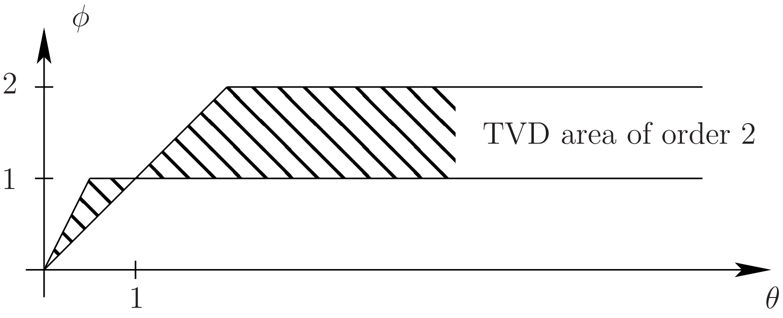


Figure 7: TVD Area

Rewriting of the schemes in conservative formulation

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}(F_{j+1/2}^n - F_{j-1/2}^n),$$

with $F_{j+1/2} = F(u_{j-l}, u_{j-k+1}, \cdots, u_j, \cdots, u_{j+r})$. To guarantee the consistency with the equation $u_t + (f(u))_x = 0$, it is necessary that

$$F(u, u, \cdots, u) = f(u).$$

Example for the equation $u_t + au_x = 0$, with a of any sign. Then, the upwind or downwind scheme is written as

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{\Delta x}\begin{cases} u_j^n - u_{j-1}^n & a > 0 \\ u_{j+1}^n - u_j^n & a < 0 \end{cases}$$

or again

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

Therefore, the conservative formulation is valid if one defines the numerical flux as

$$F_{j+1/2} = \frac{1}{2}a(u_{j+1} + u_j) - \frac{|a|}{2}(u_{j+1} - u_j).$$

For the Lax-Wendroff scheme, the numerical scheme is

$$F_{j+1/2}^{LW} = \frac{1}{2}a(u_{j+1} + u_j) - \frac{a^2}{2}\frac{\Delta t}{\Delta x}(u_{j+1} - u_j)$$

and for Beam-Warming

$$\begin{aligned} F_{j+1/2}^{BW} &= \tfrac{a}{4}(u_{j+2}(-1 - a\tfrac{\Delta x}{\Delta t}) + u_{j+1}(3 + a\tfrac{\Delta x}{\Delta t}) + u_j(3 - a\tfrac{\Delta x}{\Delta t}) + u_{j-1}(-1 + a\tfrac{\Delta x}{\Delta t})) \\ &\quad -\tfrac{|a|}{4}(u_{j+2}(-1 - a\tfrac{\Delta x}{\Delta t}) + u_{j+1}(3 + a\tfrac{\Delta x}{\Delta t}) - u_j(3 - a\tfrac{\Delta x}{\Delta t}) - u_{j-1}(-1 + a\tfrac{\Delta x}{\Delta t})). \end{aligned}$$

In a similar way as the one we chose to limit the slope in the numerical schemes written with explicit of incremental notation, we can again apply a limitation technique. This time, the idea is to average a flux of lower order f^{low} (for example up or downwind) and an higher order flux f^{high} (for example Lax Wendroff). The problem is the balance in the way of limiting the flux.

$$\begin{aligned} F_{j+1/2} &= f^{low}_{j+1/2} - \phi(\theta_j)(f^{low}_{j+1/2} - f^{high}_{j+1/2}) \\ F_{j-1/2} &= f^{low}_{j-1/2} - \phi(\theta_{j-1})(f^{low}_{j-1/2} - f^{high}_{j-1/2}) \end{aligned}$$

where as previously the slope $\theta_j = (u_j - u_{j-1})/(u_{j+1} - u_j)$. Thereby, $F_{j+1/2} = f^{high}_{j+1/2}$ if $\phi = 1$, and $F_{j+1/2} = f^{low}_{j+1/2}$ if $\phi = 0$.

Interpretation:

$$F_{j+1/2} = f^{low}_{j+1/2} \qquad \underbrace{-\phi(\theta_j)(f^{low}_{j+1/2} - f^{high}_{j+1/2})}_{\substack{\text{antidiffusion term to} \\ \text{make the scheme precise far} \\ \text{from the discontinuities } (\phi = 1)}}$$

$$F_{j+1/2} = f^{high}_{j+1/2} \qquad \underbrace{-(1-\phi(\theta_j))(f^{high}_{j+1/2} - f^{low}_{j+1/2})}_{\substack{\text{diffusion term when dealing } \phi < 1 \\ \text{with discontinuities areas} \\ \text{to make the scheme robust}}}$$

List of some second order limiters:

minmod	$\phi(\theta) = \max(0, \min(1, \theta))$
Osher	$\phi(\theta) = \max(0, \min(\theta, \beta))$, $1 \le \beta \le 2$
Superbee	$\phi(\theta) = \max(0, \min(2\theta, 1), \min(\theta, 2))$
Sweby	$\phi(\theta) = \max(0, \min(\beta\theta, 1), \min(\theta, \beta))$, $1 \le \beta \le 2$
van Leer	$\phi(\theta) = (\theta + \|\theta\|)/(1 + \theta)$

Exercise: Implement the schemes with slope limiters. A function allowing the choice of the velocity, the initial datum and the limiter will be written. Check numerically that the schemes are of higher order.

Exercise: What can be your suggestion for systems of linear equations?

3 Nonlinear scalar equations

3.1 Weak solution - entropy solution

Method of characteristics We are interested in this section in the numerical solution of nonlinear hyperbolic equations. They can be written on the one hand through conservative formulation or on the other hand through a characteristic formulation.

Conservative formulation $\quad u_t + (f(u))_x = 0.$

In this formulation, the equations can be integrated and written

$$\frac{d}{dt}\int_a^b u(t,x)dx + [f(u(t,b) - f(u(t,a))] = 0.$$

This is the approach that we have mentioned when we derived the Godunov scheme, which is of ” Finite volume“ type.

Characteristic formulation $\qquad u_t + f'(u)u_x = 0 = u_t + a(u)u_x = 0.$

As for linear equations, it is interesting to define the notion of characteristic curves. They are the trajectories $t \mapsto X(t)$ which are solutions to ODEs $\dot{X}(t) = a(u(t, X(t)))$ with initial datum $X(0) = x^0$. Along these trajectories, we have $(d/dt)u(\cdot, X(\cdot)) = 0$, thus $u(t, X(t)) = u(0, X(0))$. Since u is constant along some characteristics, the characteristics are lines since $\dot{X}(t) = a(u(t, X(t)) = a(u^0(X(0)))$.

Example of Burgers equation Let us take the Burgers equation, with flux $f(u) = u^2/2$. Thereby, $f'(u) = a(u) = u$. The characteristics are solutions to $\dot{X}(t) = u(t, X(t))$. Since u is constant along the characteristic, it is a linear ODE with solution $X(t) = x^0 + ut$ and so $u(t, x) = u^0(x - ut)$ which leads to an implicit definition for u. However, an other interpretation is still possible. Since u is constant along the characteristics, $u(t, X(t)) = u(0, X(0)) = u^0(x^0)$ and the ODE $\dot{X}(t) = u(t, X(t))$ can also be written as $\dot{X}(t) = u^0(x^0)$.

Let us take the initial datum

$$u^0(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

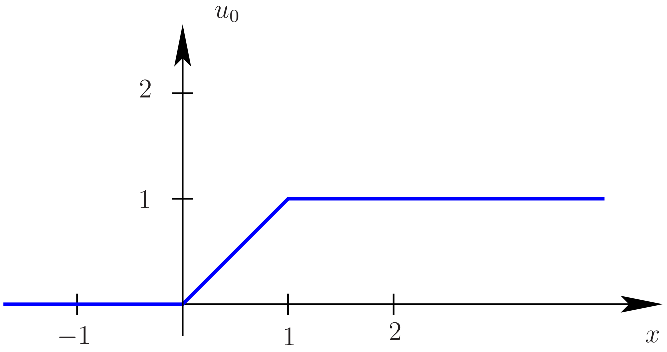


Figure 8: Initial datum

Three cases can be described

- Solution to the ODE $\begin{cases} \dot{X}(t) = u^0(x^0) \\ x^0 \leq 0 \end{cases}$. For the initial datum $x^0 \leq 0$, the ODE is reduced to $\dot{X}(t) = 0$ that is $X(t) = x^0$ and so $u(t, X(t)) = u^0(x^0) = 0$.
- Solution to the ODE $\begin{cases} \dot{X}(t) = u^0(x^0) \\ 0 < x^0 < 1 \end{cases}$. For the initial datum , the ODE is reduced to $\dot{X}(t) = u^0(0 < x^0 < 1) = x^0$ and so $X(t) = (1 + t)x^0$. Thereby, $u(t, X(t)) = u^0(x^0) = x^0 = x/(1 + t)$.
- Solution to the ODE $\begin{cases} \dot{X}(t) = u^0(x^0) \\ x^0 \geq 1 \end{cases}$. For the initial datum , the ODE is reduced to $\dot{X}(t) = u^0(x^0 > 1) = 1$ and so $X(t) = t + x^0$. We therefore have $u(t, X(t)) = u^0(x^0 > 1) = 1$.

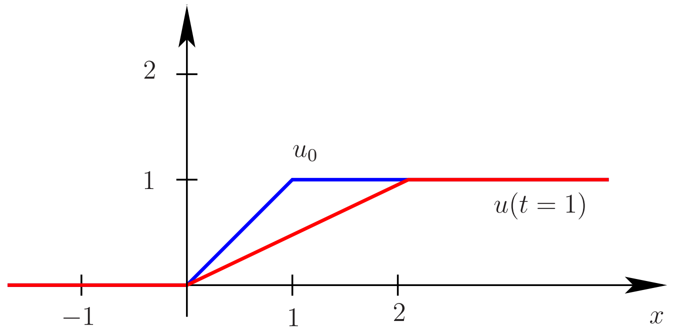


Figure 9: Solution at $t = 1$: the wave is rarefying

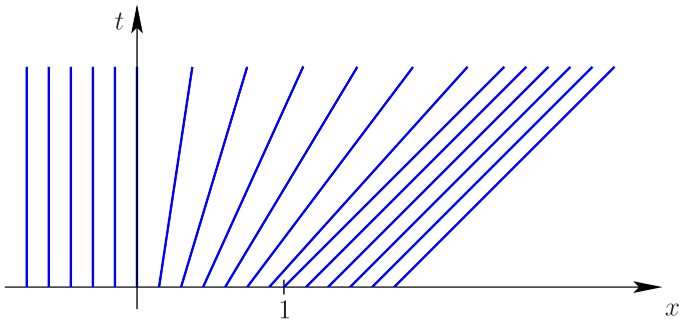


Figure 10: Evolution of t with respect to x

What happens if we change the initial datum ? For example, let us take

$$u^0(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 1 - x & \text{if } 0 < x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

Solution

- $x^0 \leq 0$: we have $\dot{X}(t) = 1$, that is $X(t) = t + x^0$ and $u(t,x) = 1$, for $x \leq t$.
- $0 < x^0 < 1$: we have $\dot{X}(t) = 1 - x^0$, that is $X(t) = (1 - x^0)t + x^0$, therefore $x^0 = (x - t)/(1 - t)$ and $u(t,x) = 1 - x^0 = (1 - x)/(1 - t)$ for $t < x < 1$.
- $x^0 \geq 1$: we have $\dot{X}(t) = 0$, therefore $X(t) = x^0$, and $u(t,x) = 0$ for $x > 1$.

We of course remark that after time $t = 1$ a problem appears for the solution: a singularity will be created since $u(t,x) = 1$ if $x < 1$ and 0 if $x > 1$. A crossing of the characteristics appears at $t = 1$. There is a formation of a **shock** giving rise to a compressive wave.

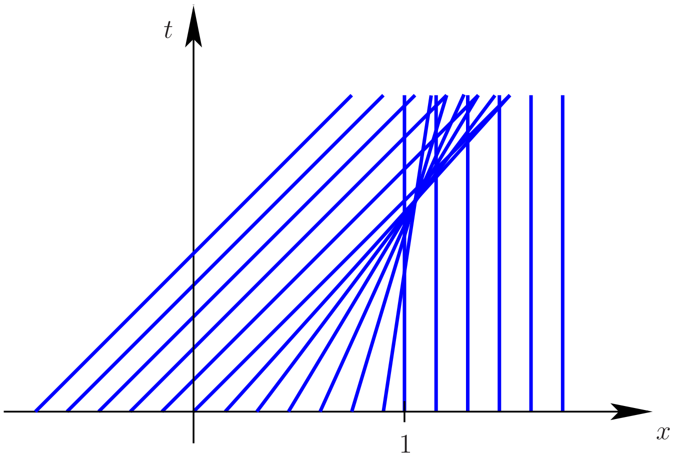


Figure 11: Evolution of t with respect to x: crossing of characteristics

Clearly, the method of characteristics allowing to reconstruct the solution with the help of the initial datum is inadequate. The solution involves discontinuities in finite time. Therefore, one needs to define a notion of weak solution.

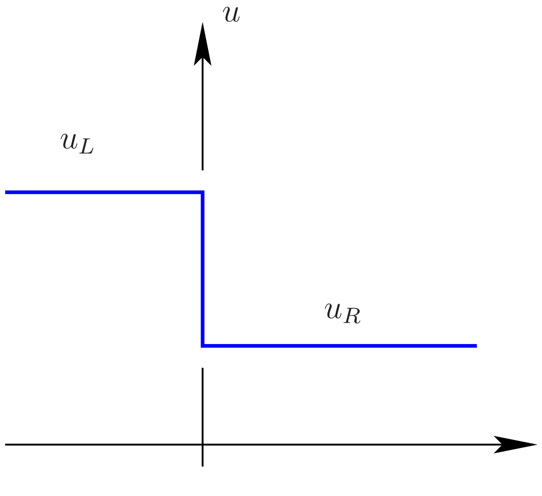
Weak solution We say that $u \in L^\infty(\mathbb{R}_t^+ \times \mathbb{R})$ is **a weak solution** if for all $\varphi \in C_0^1(\mathbb{R}_t^+ \times \mathbb{R})$, we have

$$\int_0^{+\infty} \int_{\mathbb{R}} u\varphi_t + f(u)\varphi_x dxdt = -\int_{\mathbb{R}} u(0,x)\varphi(0,x)dx.$$

A weak solution satisfies at discontinuity points the **Rankine-Hugoniot condition**

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

where $[\cdot]$ denotes the jump of a quantity.
Example:



$$s = \frac{f(u_R) - f(u_L)}{u_R - u_L}$$

Figure 12: Rankine Hugoniot condition

For the previous example, $u_L = 1$ and $u_R = 0$. Thereby, $s = 1/2$. Indeed, for the Burgers equation, we have $s = \frac{1}{2}\frac{u_R^2 - u_L^2}{u_R - u_L} = \frac{u_L + u_R}{2}$.

After the formation of the shock (that is the emergence of the crossing of characteristics), the velocity is s, thus $\dot{X}(t) = s$. In the previous example, this implies that $X(t) = (1+t)/2$: the shock moves forward with the velocity $s = 1/2$.

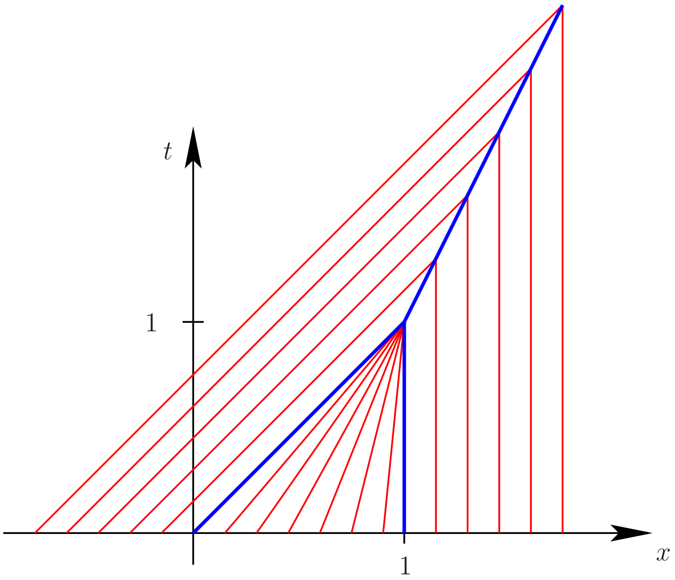


Figure 13: Evolution of t with respect to x

Example:

$$u^0(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } 0 < x < 1 \\ 0 & \text{if } x \ge 1 \end{cases}$$

Solution

- $x^0 \le 0$: we have $\dot{X}(t) = 0$, that is $X(t) = x^0$ and $u(t,x) = 0$.
- $0 < x^0 < 1$: we have $\dot{X}(t) = x^0$, that is $X(t) = x^0(1+t)$, thus $x^0 = x/(1+t)$ and $u(t,x) = x/(1+t)$.
- $x^0 \ge 1$: we have $\dot{X}(t) = 0$, thus $X(t) = x^0$, and $u(t,x) = 0$.

All of the above relations are valid only if one forgets the discontinuity. Now, let us apply the Rankine-Hugoniot relation at the discontinuity point. Then, we have

$$s = \frac{u_R + u_L}{2} = \frac{1}{2}\frac{x}{1+t}.$$

We are led to solve the ODE $\dot{X}(t) = s = X(t)/(2(1+t))$ with $X(0) = 1$, whose solution is $X(t) = \sqrt{1+t}$.

The notion of weak solution therefore allows to solve the problem of discontinuous wave generating a shock in finite time. Nevertheless, it does not allow to ensure the uniqueness. We can see this fact on the next example

$$u^0(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}$$

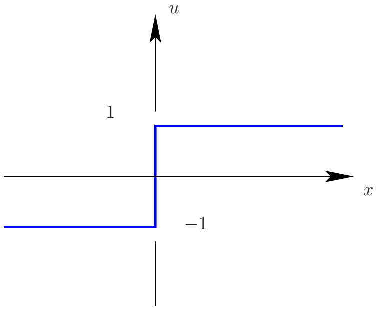


Figure 14: Initial datum

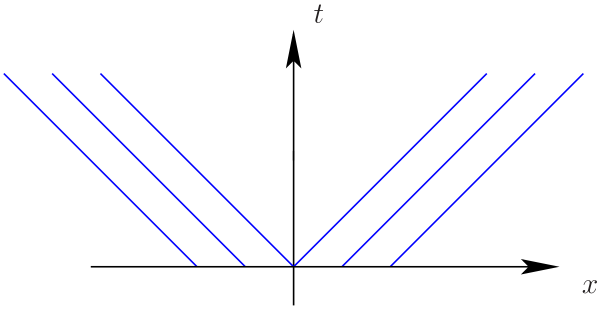


Figure 15: Evolution of t with respect to x

The Rankine Hugoniot relation gives the velocity of the propagation front at the discontinuity point which is $s = (u_L + u_R)/2 = 0$. We therefore have at least two solutions

1. Solution 1: $u(t,x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}$

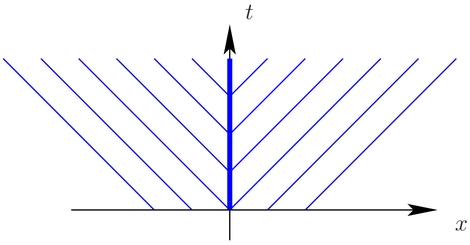


Figure 16: Evolution of t with respect to x

2. Solution 2: $u(t,x) = \begin{cases} -1 & \text{if } x < -t \\ x/t & \text{if } -t \leq x \leq t \\ 1 & \text{if } x > t \end{cases}$

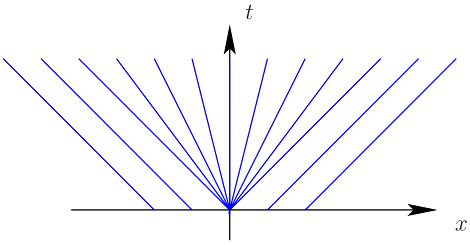


Figure 17: Evolution of t with respect to x

This last solution leads to a rarefaction wave.

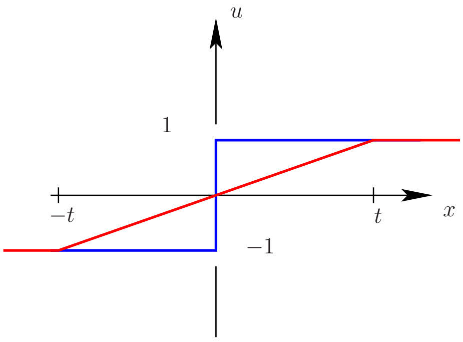


Figure 18: Rarefaction wave solution

We will attempt to give an explanation concerning the choice of the " good solution ". We can see the Burgers equation as the inviscid limit $\varepsilon = 0$ of the viscous Burgers equation

$$u_t + (f(u))_x = \varepsilon u_{xx}.$$

This equation is of parabolic type. Thus, there exists a unique smooth solution. In the smooth areas, u_{xx} is a small term. On the contrary, in the areas of strong gradients (discontinuity), u_{xx} is big and this is the dominant term. We will characterize the good weak solution as being the limit of the solution to the viscous Burgers equation.

Entropy solution We say that $E(u)$ is an entropy function if it is positive, convex and if there exists an entropy flux such that $F'(u) = E'(u)f'(u)$ (which obviously means that $f \in C^1$). Then

$$u_t + (f(u))_x = \varepsilon u_{xx} \Leftrightarrow u_t + f'(u)u_x = \varepsilon u_{xx}$$

One multiplies by $E'(u)$, which leads to

$$E'(u)u_t + F'(u)u_x = \varepsilon E'(u)u_{xx}.$$

that is

$$E_t + F_x = \varepsilon E' u_{xx} = \varepsilon (E_{xx} - E''(u_x)^2).$$

But, E is chosen to be convex, thus $E'' > 0$ and thereby

$$E_t + F_x \leq \varepsilon E_{xx}.$$

At the limit $\varepsilon \longrightarrow 0$, we thus ask an equivalent relation and we can define entropy solutions by the next property.

We say that a solution $u(t,x)$ is **an entropy solution** if for every entropy function E,

$$E(u)_t + F(u)_x \leq 0.$$

There exists a **unique** entropy solution.

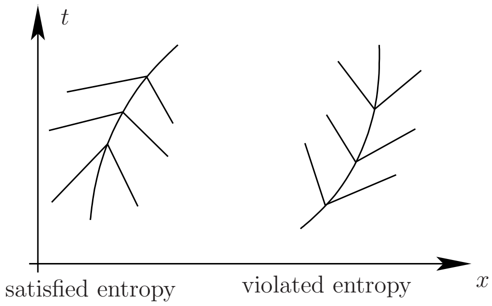
For the Burgers equation, we choose $E(u) = u^2$ and $F(u) = 2u^3/3$.

Characterization of an entropy solution

- First characterization: If f is convex (or concave), that is $f''(u) > 0$ (or $f''(u) < 0$) for all u, the entropy condition is

$$a(u_L) = f'(u_L) > s > f'(u_R) = a(u_R)$$

which means that the characteristics enter into the shock when t grows.



If we come back to the previous example, in the first solution, we have $a(u_L) = u_L = -1$, $s = 0$ and $a(u_R) = u_R = 1$ and we obviously do not have $-1 > s > 1$. In the second solution, we do not have shocks, thus no entropy condition has to be satisfied and we thus obtain an entropy solution.

- Second characterization: Oleinik condition. We say that u is an entropy solution if for every discontinuity satisfies

$$\frac{f(u) - f(u_L)}{u - u_L} \geq s \geq \frac{f(u) - f(u_R)}{u - u_R}, \quad \forall u \in [u_L, u_R].$$

3.2 Numerical schemes for scalar hyperbolic problems

Let us try to solve numerically the problem

$$\left\{ \begin{array}{l} u_t + u\, u_x = 0, \\ u(0,x) = \left\{ \begin{array}{ll} 1 & x < 0 \\ 0 & x \geq 0 \end{array} \right. \end{array} \right. .$$

The initial datum leads to a shock and, for every positive time, the solution remains positive or zero. In this form, the velocity of propagation is $a(u) = u > 0$ and we could be tempted in a first attempt to apply the upwind scheme. Let us define the parameter $\sigma = \Delta t/\Delta x$. The scheme is, for $n \geq 1,\ j \in \mathbb{Z}$

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

with $u_j^0 = 1$ if $j < 0$ and 0 if $j \geq 0$. Thus, if $j < 0$, $u_j^1 = 1$ and if $j \geq 0$, then $u_j^1 = 0$. Therefore, $u^1 = u^0$, and consequently $u^n = u^0$, $\forall n \geq 0$ for all σ. The numerical solution converges with $u(t,x) = \left\{ \begin{array}{l} 1 \text{ if } x < 0 \\ 0 \text{ if } x \geq 0 \end{array} \right.$, which is not a weak solution! Actually, the scheme can not be written as

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

with $F(u, \cdots, u) = f(u)$.

Examples of conservative schemes

Lax-Friedrichs

$$F^{LF}_{j+1/2} = \frac{1}{2}\left(f_j + f_{j+1} - \frac{\Delta x}{\Delta t}(u_{j+1} - u_j)\right) \text{ where } f_j = f(u_j).$$

Lax-Wendroff

$$F^{LW}_{j+1/2} = \frac{1}{2}\left(f_j + f_{j+1} - a^2_{j+1/2}\frac{\Delta t}{\Delta x}(u_{j+1} - u_j)\right) \text{ where } a_{j+1/2} = \begin{cases} \dfrac{f_{j+1} - f_j}{u_{j+1} - u_j} \text{ if } u_{j+1} \neq u_j, \\ \\ f'(u_j) \text{ if } u_{j+1} = u_j. \end{cases}$$

Beam-Warming

$$\begin{aligned} F^{BW}_{j+1/2} &= \frac{1}{4}(-f_{j+2} + 3f_{j+1} + 3f_j - f_{j-1}) - a^2_{j+1/2}\frac{\Delta x}{4\Delta t}(u_{j+2} - u_{j+1} + u_j - u_{j-1}) \\ &\quad - \frac{s_{j+1/2}}{4}(-f_{j+2} + 3f_{j+1} - 3f_j + f_{j-1}) + s_{j+1/2}a^2_{j+1/2}\frac{\Delta x}{4\Delta t}(u_{j+2} - u_{j+1} - u_j + u_{j-1}) \end{aligned}$$

with $s_{j+1/2} = a_{j+1/2}/|a_{j+1/2}|$.

Mac Cormack Let $\bar{u}^n_j = u^n_j - \dfrac{\Delta t}{\Delta x}(f(u^n_{j+1}) - f(u^n_j))$. The scheme reads

$$u^{n+1}_j = \frac{u^n_j + \bar{u}^n_j}{2} - \frac{\sigma}{2}(f(\bar{u}^n_j) - f(\bar{u}^n_{j-1})).$$

These schemes are adaptations of the schemes for linear scalar hyperbolic equations.

Finite volume schemes One can build some new schemes applying again the idea of the finite volume schemes to nonlinear equations.

1. averaged values for each cell,
2. exact solution on $[t^n, t^{n+1}]$,
3. projection on constant-by-cell functions to reboot the process.

As previously, the second step leads to the solution of Riemann problems at each interface and the finite volume schemes read

$$u^{n+1}_i = u^n_i - \frac{\Delta t}{\Delta x}\left[f(r_{i+1/2}(u^n_i, u^n_{i+1})) - f(r_{i-1/2}(u^n_{i-1}, u^n_i))\right].$$

Thanks to this base, we can build many schemes. The most famous is the **Godunov scheme** which consists in exactly solving the Riemann problem at each interfaces, but can be computationally costly and complicated.

$$u^{n+1}_i = u^n_i - \frac{\Delta t}{\Delta x}\left[f^{G,n}_{i+1/2} - f^{G,n}_{i-1/2}\right]$$

with $f^{G,n}_{i+1/2} = f(r_{i+1/2}(u^n_i, u^n_{i+1}))$. The CFL condition associated to this scheme is a little bit more complicated compared to the linear one since we have to guarantee that

$$\max_i |a(u^n_i)|\sigma \leq \frac{1}{2},$$

in order that the characteristics do not cross the cell (to avoid Riemann problems interferences). This time, the solution of the previous example leads to a shock moving with the velocity 1/2.

It can be sometimes subtle to exactly solve the Riemann problem. It can be preferable to solve an approximated Riemann problem. This is the idea of the **Murman-Roe** scheme. To do that, we solve the approximated equation

$$u_t + a_{j+1/2}u_x = 0, \quad x \in I_j$$

with

$$a_{j+1/2} = \begin{cases} \dfrac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} \text{ if } u_j \neq u_{j+1} \\ f'(u_j) \text{ if } u_j = u_{j+1}. \end{cases}$$

we therefore have a new Riemann problem but for a linear equation whose solution can be written as

$$r_{j+1/2}(u_j, u_{j+1}) = u_j \frac{a^+_{j+1/2}}{a_{j+1/2}} + u_{j+1}\frac{a^-_{j+1/2}}{a_{j+1/2}}$$

where $a^+ = (a + |a|)/2$ and $a^- = (a - |a|)/2$. Therefore, the Murman-Roe flux reads

$$F^{MR}_{j+1/2} = \frac{1}{2}(f_j + f_{j+1} - |a_{j+1/2}|(u_{j+1} - u_j)).$$

We can wonder if these schemes converge to the entropy solution. Let us try to apply the Murman-Roe scheme to the Burgers equation with the initial datum $u(x, 0) = 1$ if $x \geq 0$ and -1 otherwise. Then, the velocity is $a_{j+1/2} = 0$ and so $u_{j+1} = u_j$. Thus, we have $F^{MR}_{j+1/2} = 1$ for all j and the scheme leads to $u^{n+1}_j = u^n_j$. Therefore, we recover a non entropy solution. The scheme is not able to make the difference between a shock and a rarefaction wave. It is the same difficulty for the Lax-Wendroff and Beam-Warming schemes. On the other hand, the Godunov scheme is entropic by construction.
A possible approach is to compute an entropy function for the discrete schemes and to prove a cell entropy inequality

$$\frac{E(u^{n+1}_j) - E(u^n_j)}{\Delta t} + \frac{H^n_{j+1/2} - H^n_{j-1/2}}{\Delta x} \leq 0$$

where $H^n_{j+1/2}$ denotes the associated numerical flux. This is a difficult operation and one prefers to verify that a scheme is in the class of entropy schemes. Here is two classes of schemes that are of entropy type:

Monotonous schemes a scheme is said to be monotonous if $u^{n+1}_j = H(u^n_{j-l}, u^n_{j-l+1}, \cdots, u^n_j, \cdots, u^n_{j+r})$ with $\partial_{u_i} H \geq 0$, $\forall i$. It is of entropy type and moreover of first order.

E-SCHEMES If $F_{j+1/2}$ satisfies

$$\operatorname{sign}(u^n_{j+1}-u^n_j)(F^n_{j+1/2}-f(u))\leq 0, \quad \forall u\in[u_j,u_{j+1}],$$

then the scheme is an E-scheme which is of entropy type and at most of first order.

Cure to make Murman-Roe of entropy type: the problem comes from the viscous term in $F^{MR}_{j+1/2}$ which can cancel. To modify this bad behavior, one modifies the term $|a_{j+1/2}|$ by $\psi_\varepsilon(a_{j+1/2})$ where the function ψ_ε is defined by

$$\psi_\varepsilon(v)=\begin{cases} v \text{ if } |v|\geq\varepsilon \\ \dfrac{v^2+\varepsilon^2}{2\varepsilon} \text{ if } |v|<\varepsilon \end{cases}$$

We can take ε defined by

$$\varepsilon=\sup_{u\in(u_l,u_r)}\max(0,s(u_l,u_r)-s(u_l,u),s(u,u_r)-s(u_l,u_r))$$

where $s(u_l,u_r)$ is the velocity given by the Rankine-Hugoniot relation.
For Burgers equation, in the case of " stationary " rarefaction wave, $u_l=-1$ and $u_r=1$, then

$$\varepsilon=\sup_{u\in[-1,1]}\max(0,\frac{1-u}{2},\frac{1+u}{2})=1$$

3.3 Incremental formulation and numerical viscosity

One says that a scheme can be written in an incremental formulation if there exist some functions C and D which belong to $\mathbb{R}^{2k}$ and $\mathbb{R}$ such that, if we define the incremental coefficients

$$\begin{aligned} C^n_{j+1/2} &= C(v^n_{j-k+1},\dots,v^n_{j+k}), \\ D^n_{j+1/2} &= D(v^n_{j-k+1},\dots,v^n_{j+k}), \end{aligned}$$

the scheme reads

$$\forall j,\ n,\ v^{n+1}_j=v^n_j+C^n_{j+1/2}\Delta v^n_{j+1/2}-D^n_{j-1/2}\Delta v^n_{j-1/2},$$

with $\Delta v_{j+1/2}=v_{j+1}-v_j$.

Remark 2. *What can we say in the linear case?*

Every scheme with a three point stencil and consistent with (1.1) with numerical flux g locally lipschitz can be written in incremental formulation with

$$C(u,v)=\lambda\frac{f(u)-g(u,v)}{v-u};D(u,v)=\lambda\frac{g(u,v)-f(v)}{v-u}.$$

Indeed, let $n\in\mathbb{N}$, $j\in\mathbb{Z}$, (v_j) a real sequence.
We have for $j\in\mathbb{Z}$

$$C^n_{j+1/2}\Delta v^n_{j+1/2}-D^n_{j-1/2}\Delta v^n_{j-1/2}=-\lambda(g^n_{j+1/2}-g^n_{j-1/2}.$$

Taking $v_j = v_{j-1}$, we find $C^n_{j+1/2}\Delta v^n_{j+1/2} = -\lambda(g^n_{j+1/2} - f(v_j))$.
Taking $v_j = v_{j+1}$, we find $D^n_{j-1/2}\Delta v^n_{j-1/2} = -\lambda(f(v_j) - g^n_{j-1/2})$.

Similarly, every incremental three points stencil scheme satisfying

$$D_{j+1/2} - C_{j+1/2} = \lambda\frac{\Delta f_{j+1/2}}{\Delta v_{j+1/2}}$$

is conservative and consistent with (1.1).
We can show that setting $g(u,v) = f(u) - C(u,v)/\lambda$.

Theorem (Harten's criterion 1)**.** *If $v^{n+1} = H_\Delta(v^n)$ can be written in incremental formulation with $C \geq 0$, $D \geq 0$ and $C + D \leq 1$, then the scheme is TVD.*

Proof

We try to show that $\forall n \in \mathbb{N}$, $\sum_j |\Delta v^{n+1}_{j+1/2}| \leq \sum_j |\Delta v^n_{j+1/2}|$.
Let $n \in \mathbb{N}$, $j \in \mathbb{Z}$. We have

$$\begin{aligned}\Delta v^{n+1}_{j-1/2} &= \Delta v^n_{j-1/2} + C^n_{j+1/2}\Delta v^n_{j+1/2} - D^n_{j-1/2}\Delta v^n_{j-1/2} - C^n_{j-1/2}\Delta v^n_{j-1/2} + D^n_{j-3/2}\Delta v^n_{j-3/2}\\ &= C^n_{j+1/2}\Delta v^n_{j+1/2} + (1 - C^n_{j-1/2} - D^n_{j-1/2})\Delta v^n_{j-1/2} + D^n_{j-3/2}\Delta v^n_{j-3/2}\end{aligned}$$

Therefore

$$|\Delta v^{n+1}_{j-1/2}| \leq C^n_{j+1/2}|\Delta v^n_{j+1/2}| + (1 - C^n_{j-1/2} - D^n_{j-1/2})|\Delta v^n_{j-1/2}| + D^n_{j-3/2}|\Delta v^n_{j-3/2}|.$$

Thus

$$\begin{aligned}\sum_{j\in\mathbb{Z}} |\Delta v^{n+1}_{j-1/2}| &\leq \sum_{j\in\mathbb{Z}} C^n_{j+1/2}|\Delta v^n_{j+1/2}| + \sum_{j\in\mathbb{Z}}(1 - C^n_{j+1/2} - D^n_{j+1/2})|\Delta v^n_{j+1/2}| + \sum_{j\in\mathbb{Z}} D^n_{j+1/2}|\Delta v^n_{j+1/2}|\\ &\leq \sum_j |\Delta v^n_{j+1/2}|.\end{aligned}$$

□

We say that a scheme $v^{n+1} = H_\Delta(v^n)$ is written in numerical viscosity formulation if there exists $Q : \mathbb{R}^{2k} \longrightarrow \mathbb{R}$ such that

$$\forall n, j,\ v^{n+1}_j = v^n_j - \frac{\lambda}{2}\Delta f_{j+1/2} + \frac{1}{2}(Q_{j+1/2}\Delta v_{j+1/2} - Q_{j-1/2}\Delta v_{j-1/2}),$$

with $\Delta_1 f_j = f(v_{j+1}) - f(v_{j-1})$.

For example, for the Lax-Friedrichs scheme, $Q \equiv 1$; for the Lax-Wendroff scheme, $Q = \lambda^2 a_{j+1/2}\Delta f_{j+1/2}/\Delta v_{j+1/2}$; for the Godunov scheme, $Q = \lambda(f(u) + f(v) - 2g(u,v))/(v-u)$.
Exercise: What is the link with the linear case ?

Every three point stencil scheme, conservative and consistent with (1.1) has a numerical viscosity formulation with $Q = C + D$.

Remark 3. *In general, for a three point stencil scheme, we can write $Q(u,v) = q(\lambda a(u,v))$.*

Theorem (Harten criterion 2)**.** *Let us assume that $v^{n+1} = H_\Delta(v^n)$ with three point stencil can be written in numerical viscosity formulation precisely with $Q(u,v) = q(\lambda a(u,v))$, q continuous. Then, there exists $\mu \in [0,1]$ such that $\forall \theta \in [-\mu,\mu]$, $|\theta| \leq q(\theta) \leq 1$, the scheme is TVD if the CFL condition $\lambda \sup |a(u)| \leq 1$ is satisfied.*

Proof

[Idea of the proof] Show that if $C(u,v) = (q(\lambda a(u,v)) - \lambda a(u,v))/2$ and $D(u,v) = (q(\lambda a(u,v)) + \lambda a(u,v))/2$ then the first Harten criterion is satisfied.

□

Remark 4. *Show that for three point stencil schemes the Harten criteria are also necessary.*

3.4 E-schemes

A conservative scheme is an E-scheme if

$$\forall c \in \ I(v_j, v_{j+1}),\ (v_{j+1} - v_j)(g_{j+1/2} - f(c)) \leq 0.$$

Example : Let us consider a conservative scheme consistent with (1), monotonous and with a three point stencil:

$$\begin{array}{ccccc} g(u,v) - f(c) & = & g(u,v) - g(u,c) & + & g(u,c) - g(c,c) \\ u \leq c \leq v & & \leq 0 & & \leq 0 \\ u \geq c \geq v & & \geq 0 & & \geq 0 \end{array}$$

A conservative scheme consistent with (1) is an E-scheme if and only if $Q_{j+1/2} \geq Q_{j+1/2}^{Godunov}$, $\forall j \in \mathbb{Z}$. Indeed, let $u \leq v$.

$$\begin{aligned} \forall c \in [u,v],\ g(u,v) - f(c) \leq 0 \ &\Leftrightarrow\ g(u,v) \leq \min_{u \leq c \leq v} f(c), \\ &\Leftrightarrow\ g(u,v) \leq g^G(u,v), \\ &\Leftrightarrow\ \lambda(f(u) + f(v) - 2g(u,v)) \geq \lambda(f(u) + f(v) - 2g^G(u,v)) \\ &\Leftrightarrow\ Q \geq Q^G. \end{aligned}$$

To finish, we introduce the following modified Lax-Friedrichs scheme:

$$\begin{aligned} \forall n,\ j,\ v_j^{n+1} &= v_j^n + \frac{1}{4}((v_{j+1}^n - v_j^n) - (v_j^n - v_{j-1}^n)) - \lambda \frac{f(v_{j+1}^n) - f(v_{j-1}^n)}{2} \\ &= \frac{v_{j+1}^n + v_{j-1}^n}{2} - \lambda \frac{f(v_{j+1}^n) - f(v_{j-1}^n)}{2} - \frac{1}{4}(v_{j+1}^n - 2v_j^n + v_{j-1}^n). \end{aligned}$$

we have $Q_{j+1/2}^{LFM} = 1/2$.

Theorem. *If an E-scheme is such that $Q^G \leq Q \leq 1/2$, then it is consistent with all entropy conditions.*

[]