A beginner's course in finite volume approximation of scalar conservation laws

Pamplona – Pau – Toulouse – Zaragoza summer school on nonlinear conservation laws

Jaca 11-13/09/2008

Jérôme Droniou[1]

Version: October 10th, 2008

This is a first version, with probably typos errors (but hopefully no mathematical mistake...); feel free to contact me (see footnote) if you happen to notice some.

[1] Département de Mathématiques, UMR CNRS 5149, CC 051, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France. email: `droniou@math.univ-montp2.fr`

Contents

Summary

In this lecture, we present and study some methods to discretize scalar conservation laws $\partial_t u + \partial_x f(u) = 0$. Considering first the case of a linear equation ($f(u) = cu$) we try and understand a basic construction of numerical schemes (using finite volume techniques) and the issues related, mainly concerning the stability of the method. We then introduce the principle of monotone schemes for general non-linear equations, and give some classical examples (Lax-Friedrichs, Godunov); we try to explain the concept of numerical diffusion associated with such schemes (and its link with the discretization of parabolic equations), and we give some elements of study: stability, discrete entropy inequalities, convergence in the BV case. The numerical diffusion introduced by monotone fluxes allow to stabilize the scheme, but gives poor approximations of the qualitative properties of the continuous solution (shocks, rarefaction waves...); in the last chapter, we introduce some higher order methods (MUSCL techniques) which allow to obtain better approximations.

Chapter 1

Schemes for linear transport equations

1.1 Introduction, principle of the finite volume scheme

We first consider a simple linear transport (or convection) equation

$$\begin{cases} \partial_t u(t,x) + \partial_x(cu(t,x)) = 0 & t > 0\,,\ x \in \mathbb{R}\,, \\ u(0,x) = u_0(x) & x \in \mathbb{R} \end{cases} \tag{1.1.1}$$

where $c \in \mathbb{R}$ and $u_0 \in C^1_b(\mathbb{R})$ ([1]). The solution to this equation is quite obvious: $u(t,x) = u_0(x - ct)$, that is to say the initial data is transported with the velocity c (the solution at time t is the initial data translated by a factor ct). It is however interesting to study numerical approximations for this simple problem, since it allows to exhibit the main issues which arise for general non-linear conservation laws.

The principle to construct a finite volume scheme for a PDE is to decompose the domain into small parts (the control volumes) and to integrate the equation on these volumes. The domain of (1.1.1) is $[0,\infty[\times\mathbb{R}$ and the simplest way to decompose it is using small rectangles; let us thus take $\delta t > 0$ and $\delta x > 0$ some time and space lengths (also called "steps") and write $[0,\infty[\times\mathbb{R} = \cup_{n\geq 0}\cup_{i\in\mathbb{Z}}[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[$. Integrating the PDE in (1.1.1) on one small rectangle $[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[$, we obtain

$$\int_{i\delta x}^{(i+1)\delta x} u((n+1)\delta t, x)\,dx - \int_{i\delta x}^{(i+1)\delta x} u(n\delta t, x)\,dx \\ + \int_{n\delta t}^{(n+1)\delta t} cu(t,(i+1)\delta x)\,dt - \int_{n\delta t}^{(n+1)\delta t} cu(t, i\delta x)\,dt = 0. \tag{1.1.2}$$

Assume now that, for $n \geq 0$ and $i \in \mathbb{Z}$, u_i^n denotes an approximate value of u at time $t = n\delta t$ on the space mesh (or "control volume") $[i\delta x,(i+1)\delta x[$, and that f_i^n is an approximation of cu on $[n\delta t,(n+1)\delta t[$ at $x = i\delta x$ (see Figure 1.1). Then (1.1.2) divided by δt leads

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + f_{i+1}^n - f_i^n \approx 0. \tag{1.1.3}$$

If we could now express f_i^n in terms of $(u_j^n)_{j\in\mathbb{Z}}$, this equation (with $= 0$ instead of ≈ 0) would give a way to compute the values $(u_i^{n+1})_{i\in\mathbb{Z}}$ at time $t = (n+1)\delta t$ in function of the values $(u_i^n)_{i\in\mathbb{Z}}$ at time $t = n\delta t$;

[1] That is to say u_0 and u_0' exist and are continuous and bounded on $\mathbb{R}$; we take a regular initial data to avoid problems about the sense in which the solution is understood (strong, weak, entropy...), but we could as well consider $u_0 \in L^\infty(\mathbb{R})$.

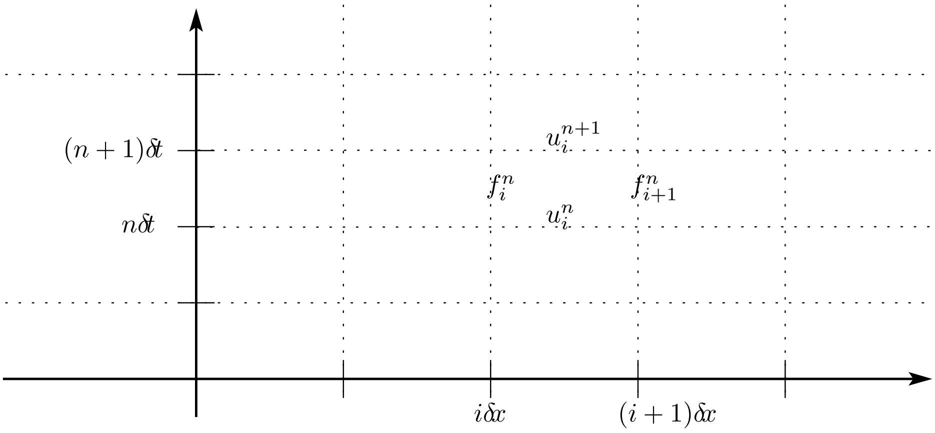


Figure 1.1: Discretization of $[0,\infty[\times\mathbb{R}$ and approximations of u and cu.

by induction, we would deduce all the $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$ from the values $(u_i^0)_{i\in\mathbb{Z}}$ of u at time $t=0$, i.e. the approximate values of u_0, for which there are several obvious simple choices, such as

$$\forall i\in\mathbb{Z}\ :\quad u_i^0=\frac{1}{\delta x}\int_{i\delta x}^{(i+1)\delta x}u_0(x)\,dx \tag{1.1.4}$$

or even, u_0 being regular, $u_i^0=u_0(i\delta x+\frac{\delta x}{2})$.

The remaining question is therefore to find appropriate expressions for f_i^n in function of $(u_j^n)_{j\in\mathbb{Z}}$ such that, if the time and space steps are small, these resulting computed values $(u_i^n)_{i\in\mathbb{Z}}$ are indeed close to the values of the solution to (1.1.1).

1.2 Centered scheme

The most obvious choice is to approximate of cu at $x=i\delta x$ using the mean value of the approximate values of u inside the control volumes on either side of $i\delta x$:

$$f_i^n=c\frac{u_{i-1}^n+u_i^n}{2}. \tag{1.2.1}$$

The scheme (1.1.3) is then written

$$\forall n\geq 0\,,\ \forall i\in\mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1}-u_i^n)+c\frac{u_{i+1}^n-u_{i-1}^n}{2}=0. \tag{1.2.2}$$

The resulting system [(1.2.2),(1.1.4)] (called the centered scheme) allows to compute all the $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$ and we can expect, from its construction, that it is a reasonable discretization of (1.1.1), that is to say that these $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$ are good approximations of the values of the solution u to (1.1.1).

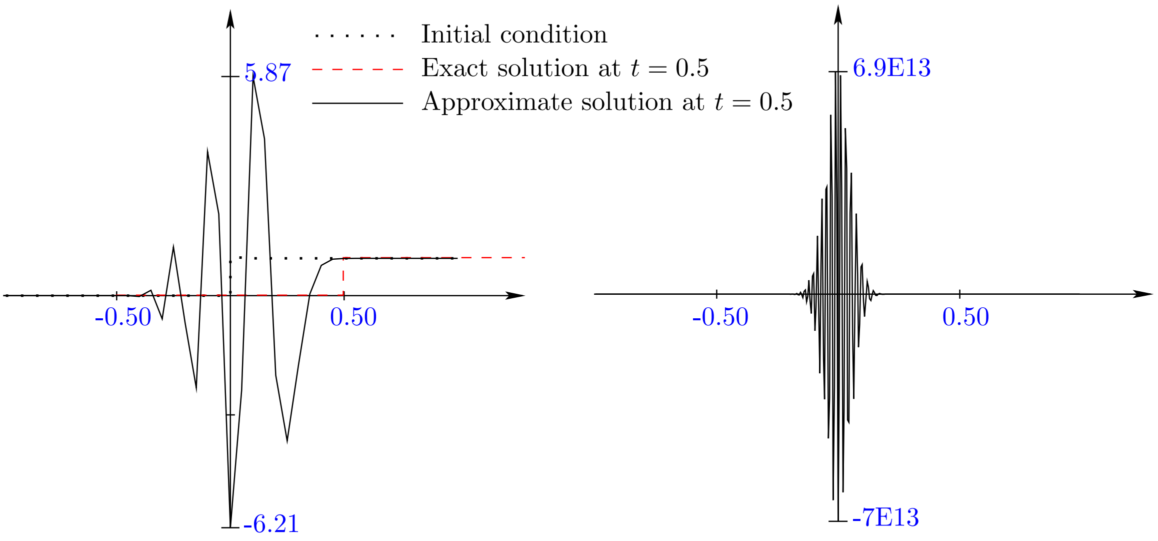
This is however not the case because, except in some trivial cases ($c=0$ or u_0 constant, for example), [(1.2.2),(1.1.4)] does not respect a fundamental feature of scalar conservation laws (linear or non-linear): *the maximum principle*, which states that the values of the solution to the PDE stays between the minimum and maximum values of the initial data (this is quite obvious in the linear case since the solution is $u(t,x)=u_0(x-ct)$).
Let us study this principle for the centered scheme. We have, for all $i\in\mathbb{Z}$, $u_i^1=u_i^0-\frac{\delta t}{\delta x}\frac{c}{2}u_{i+1}^0+\frac{\delta t}{\delta x}\frac{c}{2}u_{i-1}^0$.
If we take the Riemann initial condition $u_0(x)=0$ if $x<0$ and $u_0(x)=1$ if $x>0$, then (1.1.4) implies

$u_j^0 = 0$ for all $j < 0$ and $u_j^0 = 1$ for all $j \geq 0$. Thus,

$$u_{-1}^1 = -\frac{\delta t}{\delta x}\frac{c}{2} \quad \text{and} \quad u_0^1 = 1 - \frac{\delta t}{\delta x}\frac{c}{2}.$$

We therefore notice that, for any choice of δt and δx, if $c \neq 0$ then either $u_{-1}^1 < 0$ (for $c > 0$) or $u_0^1 > 1$ (for $c < 0$); at time $t = \delta t$, the approximate values are not between the minimum and maximum values of the initial approximate values: the centered scheme violates the maximum principle.

In fact, the situation is even worse than the simple preceding computation for $n = 1$ can make believe: when implementing the centered scheme, one can notice that the L^∞ norm of the approximate values explodes as the times increases ($\lim_{n\to\infty} \sup_{i\in\mathbb{Z}} |u_i^n| = +\infty$) and that these values have nothing to do with the exact solution, see Figure 1.2.



$\delta t = 0.05$, $\delta x = 0.05$

$\delta t = 0.005$, $\delta x = 0.005$

Figure 1.2: Centered scheme for $\partial_t u + \partial_x u = 0$ with two choices of the pair $(\delta t, \delta x)$: exact and approximate solutions at $t = 0.5$.

1.3 Upwind scheme

The choice, during the computation of f_i^n, of approximating u at $x = i\delta x$ by the mean value of u inside the two control volumes on either side of $i\delta x$ is therefore not appropriate. This could in fact have been predicted using the physical interpretation of the PDE: when obtaining the conservation law $\partial_t u + \partial_x(cu) = 0$, one first translates the conservation of u inside the control volume $[i\delta x, (i+1)\delta x[$ as (1.1.2), in which $\int_{n\delta t}^{(n+1)\delta t} cu(t, i\delta x)\,dt$ appears as the quantity of u transported through the interface $x = i\delta x$ by the velocity c. If $c > 0$ then this quantity comes from the control volume $[(i-1)\delta x, i\delta x[$ and enters the control volume $[i\delta x, (i+1)\delta x[$; in this case, it would therefore appear wiser to approximate $\int_{n\delta t}^{(n+1)\delta t} cu(t, i\delta x)\,dt$ using only the value of u inside the control volume $[(i-1)\delta x, i\delta x[$ form which originates the flux.

This mean that, if $c > 0$, we would rather take

$$f_i^n = cu_{i-1}^n. \tag{1.3.1}$$

Of course, if $c < 0$, the same reasoning would lead to choosing $f_i^n = cu_i^n$. From now on, let us consider only the case $c > 0$; with the choice (1.3.1), (1.1.3) is written

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + cu_i^n - cu_{i-1}^n = 0. \tag{1.3.2}$$

The scheme [(1.3.2),(1.1.4)] is called the *upwind* scheme, precisely because it is constructed using the upwind (with respect to the velocity c) choice (1.3.1) to compute the flux values.

1.3.1 L^∞ stability and convergence

On the contrary to the centered scheme, the upwind scheme respects the maximum principle and is therefore L^∞ stable: we can prove a L^∞ bound on the approximate values.

Proposition 1.3.1 (Stability of the upwind scheme for the linear transport equation) *Let $c > 0$ and assume that $\delta t > 0$ and $\delta x > 0$ are such that*

$$\frac{\delta t}{\delta x} \leq \frac{1}{c}. \tag{1.3.3}$$

If $(u_i^n)_{n\geq 0\,,\,i\in\mathbb{Z}}$ satisfies [(1.3.2),(1.1.4)], then

$$\forall n \geq 0\,, \forall i \in \mathbb{Z}\ :\quad \inf_{j\in\mathbb{Z}} u_j^n \leq u_i^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n. \tag{1.3.4}$$

In particular, all the values $(u_i^n)_{n\geq 0\,,\,i\in\mathbb{Z}}$ are between the infimum and supremum values of u^0 and

$$\sup_{n\geq 0\,,\,i\in\mathbb{Z}} |u_i^n| \leq ||u_0||_{L^\infty(\mathbb{R})}.$$

Remark 1.3.2 *The condition (1.3.3), linking the discretization steps with the velocity and called the CFL condition (for "Courant-Friedrichs-Levy"), states that a small space step imposes a small time step: if the space step is divided by 10, then the time step must also be divided by 10. We will further discuss this condition in Sections 1.3.2 and 2.2.3.*

Proof of Proposition 1.3.1
The proof is completely trivial if we rewrite (1.3.2) in the following way:

$$u_i^{n+1} = u_i^n - \frac{\delta t}{\delta x}cu_i^n + \frac{\delta t}{\delta x}cu_{i-1}^n = \left(1 - \frac{\delta t}{\delta x}c\right)u_i^n + \frac{\delta t}{\delta x}cu_{i-1}^n.$$

The sum of $1 - \frac{\delta t}{\delta x}c$ and $\frac{\delta t}{\delta x}c$ is equal to 1 and, under the condition (1.3.3), both these terms are non-negative; hence, u_i^{n+1} appears as a convex combination of u_i^n and u_{i-1}^n, and is therefore between the minimum and maximum value of these two real numbers. Relation (1.3.4) then follows, and the rest of the properties stated in the proposition are easy consequences of this relation. ■

We can now easily prove that the upwind scheme for (1.1.1) converges, i.e. that, for δt and δx small, the values $(u_i^n)_{n\geq 0\,,\,i\in\mathbb{Z}}$ computed by [(1.3.2),(1.1.4)] are indeed close (at least in a weak sense) to the values of the solution to (1.1.1)

Theorem 1.3.3 (Convergence of the upwind scheme for linear transport equations) *Let $c > 0$. For $\delta t > 0$ and $\delta x > 0$, let $u_{\delta t,\delta x} : [0,\infty[\times\mathbb{R} \to \mathbb{R}$ be the function equal to u_i^n on $[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[$ for all $n \geq 0$ and $i \in \mathbb{Z}$. Then, as δt and δx tend to 0 while satisfying the CFL condition (1.3.3), $u_{\delta t,\delta x}$ converges in $L^\infty(]0,\infty[\times\mathbb{R})$ weak-$*$ to the solution u of (1.1.1).*

Remark 1.3.4 *In fact, the convergence is much stronger than a weak-$*$ convergence, as we will see in Chapter 2. However, since (1.1.1) is linear, the weak convergence is enough to pass to the limit in the numerical scheme and we therefore only state this weak convergence in order to avoid unnecessary complexity at this stage.*

Proof of Theorem 1.3.3
Proposition 1.3.1 implies that, as δt and δx tend to 0 while satisfying (1.3.3), $u_{\delta t,\delta x}$ remains bounded in $L^\infty(]0,\infty[\times\mathbb{R})$. Up to a subsequence, we can therefore assume that it converges in L^∞ weak-$*$ to some u; if we prove that any such limit u of a subsequence of $u_{\delta t,\delta x}$ is the weak solution to (1.1.1) then, this weak solution being unique (the equation is linear), this will show that the whole sequence $u_{\delta t,\delta x}$ converges to this solution and will complete the proof.
Let us therefore consider that $u_{\delta t,\delta x}\to u$ weak-$*$ and prove that u is a weak solution to (1.1.1), i.e. that for all $\varphi\in C_c^\infty([0,\infty[\times\mathbb{R})$,

$$\int_0^\infty\int_{\mathbb{R}} u(t,x)\left(\partial_t\varphi(t,x)+c\partial_x\varphi(t,x)\right)dtdx+\int_{\mathbb{R}}u_0(x)\varphi(0,x)\,dx=0. \tag{1.3.5}$$

Take φ such a regular function and define $\varphi_i^n=\frac{1}{\delta t}\frac{1}{\delta x}\int_{n\delta t}^{(n+1)\delta t}\int_{i\delta x}^{(i+)1)\delta x}\varphi(t,x)\,dtdx$. Multiplying (1.3.2) by $\delta t\varphi_i^n$ and summing on $n\geq 0$ and $i\in\mathbb{Z}$ (notice that, since φ has a compact support, these sums in fact only involve a finite number of indices), we find

$$0=\sum_{i\in\mathbb{Z}}\delta x\sum_{n\geq 0}(u_i^{n+1}-u_i^n)\varphi_i^n+c\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}(u_i^n-u_{i-1}^n)\varphi_i^n. \tag{1.3.6}$$

But

$$\begin{aligned}\sum_{n\geq 0}(u_i^{n+1}-u_i^n)\varphi_i^n &= \sum_{n\geq 0}u_i^{n+1}\varphi_i^n-\sum_{n\geq 0}u_i^n\varphi_i^n\\ &= \sum_{n\geq 1}u_i^n\varphi_i^{n-1}-\sum_{n\geq 0}u_i^n\varphi_i^n\\ &= \sum_{n\geq 1}u_i^n(\varphi_i^{n-1}-\varphi_i^n)-u_i^0\varphi_i^0\end{aligned}$$

(this series of manipulation comes down to a discrete integration by parts: the "derivative" $u_i^{n+1}-u_i^n$ has been put on φ). Using a similar manipulation for the second sum on $i\in\mathbb{Z}$ in (1.3.6), we obtain

$$0=\sum_{i\in\mathbb{Z}}\delta x\sum_{n\geq 1}\delta t u_i^n\frac{\varphi_i^{n-1}-\varphi_i^n}{\delta t}-\sum_{i\in\mathbb{Z}}\delta x u_i^0\varphi_i^0+c\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}\delta x u_i^n\frac{\varphi_i^n-\varphi_{i+1}^n}{\delta x},$$

that is to say, owing to the definition of $u_{\delta t,\delta x}$ and of u_i^0,

$$0=-\int_{n\delta t}^\infty\int_{\mathbb{R}}u_{\delta t,\delta x}(t,x)\Gamma_{\delta t,\delta x}(t,x)\,dtdx-\int_{\mathbb{R}}u_0(x)\Theta_{\delta x}(x)\,dx-c\int_0^\infty\int_{\mathbb{R}}u_{\delta t,\delta x}(t,x)\Xi_{\delta t,\delta x}(t,x)\,dtdx \tag{1.3.7}$$

where $\Gamma_{\delta t,\delta x}$, $\Theta_{\delta x}$ and $\Xi_{\delta t,\delta x}$ are defined by

$$\begin{aligned}\Gamma_{\delta t,\delta x}&=\frac{\varphi_i^n-\varphi_i^{n-1}}{\delta t} && \text{on } [n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[,\\ \Theta_{\delta x}&=\varphi_i^0 && \text{on } [i\delta x,(i+1)\delta x[,\\ \Xi_{\delta t,\delta x}&=\frac{\varphi_{i+1}^n-\varphi_i^n}{\delta x} && \text{on } [n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[.\end{aligned}$$

By regularity of φ, we have $\Gamma_{\delta t,\delta x}\to\partial_t\varphi$, $\Theta_{\delta x}\to\varphi(0,\cdot)$ and $\Xi_{\delta t,\delta x}\to\partial_x\varphi$ uniformly as δt and δx tend to 0. The weak-$*$ convergence of $u_{\delta t,\delta x}$ then allows to pass to the limit in (1.3.7) and to see that u satisfies (1.3.5), thus concluding the proof. ■

Numerical implementations of the upwind scheme can be found in Figures 1.3 and 1.4, and confirm the preceding theoretical results: the upwind scheme gives respectable approximations of the solutions to (1.1.1).

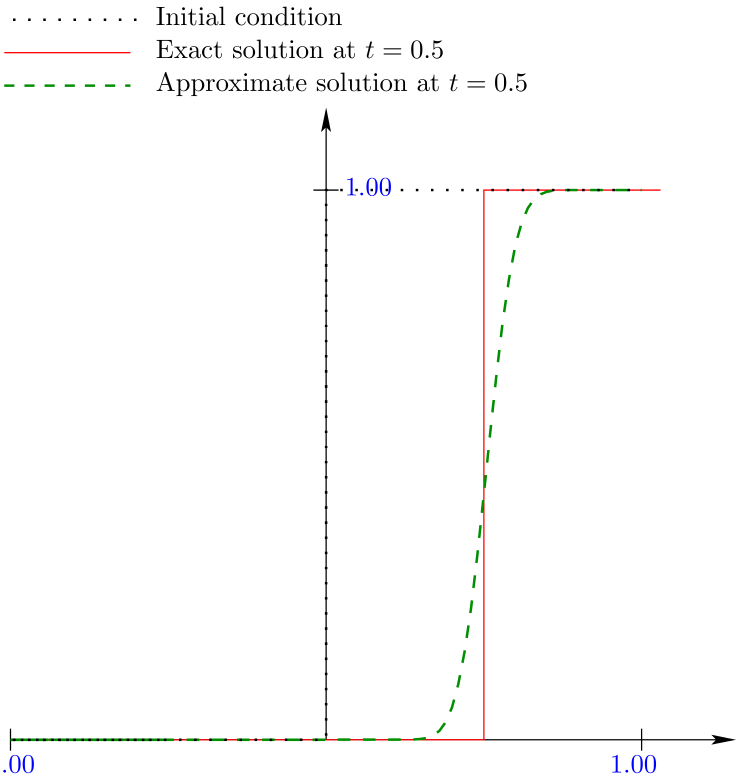


Figure 1.3: Upwind scheme for $\partial_t u + \partial_x u = 0$ and a Riemann initial data, with $\delta t = 0.01$ and $\delta x = 0.02$: exact and approximate solutions at $t = 0.5$.

Remark 1.3.5 *Note that if $\delta t = \frac{\delta x}{c}$ (limiting case in the CFL condition), then the scheme is* exact*; or, more precisely, its only error comes from the discretization of the initial condition: if $\widetilde{u}_0$ is the function equal to u_i^0 on $[i\delta x, (i+1)\delta x[$, then the solution to the scheme satisfies, for all $n \geq 0$ and all $x \in \mathbb{R}$, $u_{\delta t, \delta x}(n\delta t, x) = \widetilde{u}_0(x - cn\delta t)$ (i.e. the scheme transports with velocity c the discretized initial data).*

Remark 1.3.6 *The construction of the centered and upwind schemes lead to an interesting comparison of different intuitions coming from different scientific fields. The centered choice (1.2.1) is completely natural for the mathematician: f_i^n involves an approximate value of u at the interface between two cells, and a linear interpolation using the values inside each neighboring cell is therefore the best "simple" mathematical way to compute this interface value; however, it proves to be completely flawed, leading to a very bad scheme. The upwind choice (1.3.1) is, on the other side, the only reasonable choice from a physical point of view (because of the interpretation of the equation as a transport process); though it may seem not very efficient or natural to the mathematician, it proves to be a good way to compute the approximate fluxes. The morale of the story is that it is near to impossible to study certain equations without remembering their physical meaning...*

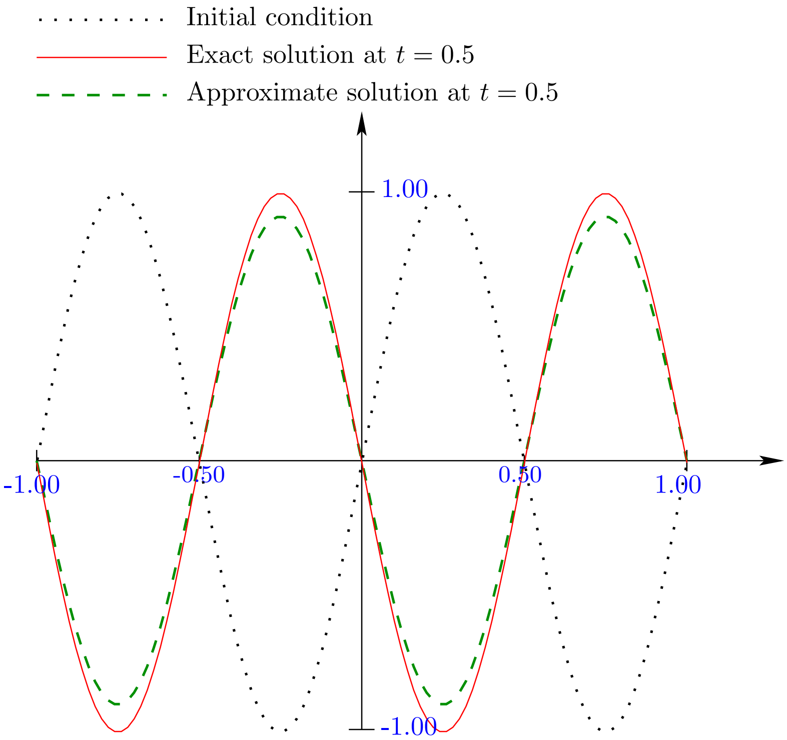


Figure 1.4: Upwind scheme for $\partial_t u + \partial_x u = 0$ and a sinus initial data, with $\delta t = 0.01$ and $\delta x = 0.02$: exact and approximate solutions at $t = 0.5$.

1.3.2 Relation with the discretization of convection-diffusion equations

Let us pause a moment to consider the case where the convection equation (1.1.1) is changed into a convection-diffusion equation

$$\left\{ \begin{array}{ll} \partial_t u(t,x) + \partial_x(cu(t,x)) - \nu \partial_x^2 u(t,x) = 0 & t > 0\,,\ x \in \mathbb{R}\,, \\ u(0,x) = u_0(x) & x \in \mathbb{R} \end{array} \right. \tag{1.3.8}$$

for some $\nu > 0$ and $c \in \mathbb{R}$ (no matter its sign). Using the same process as before, we can try and find a discretization of this equation by integrating it on time-space rectangles; noticing that

$$\int_{n\delta t}^{(n+1)\delta t} \int_{i\delta x}^{(i+1)\delta x} \partial_x^2 u(t,x)\, dt dx = \int_{n\delta t}^{(n+1)\delta t} \partial_x u(t,(i+1)\delta x)\, dt - \int_{n\delta t}^{(n+1)\delta t} \partial_x u(t, i\delta x)\, dt,$$

we are lead to the discretization

$$\frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + f_{i+1}^n - f_i^n - \nu(\partial_x u)_{i+1}^n + \nu(\partial_x u)_i^n = 0$$

where $f_i^n \approx cu$ and $(\partial_x u)_i^n \approx \partial_x u$ on $[n\delta t, (n+1)\delta t[\times \{i\delta x\}$. A natural approximation of the derivative of u at $x = i\delta x$ can be easily computed using the values of u in the control volumes on both sides of this interface (and considering for example that these values approximate u at the center of each control volume: $u_i^n \approx u(n\delta t, i\delta x + \frac{\delta x}{2})$):

$$(\partial_x u)_i^n = \frac{u_{i+1}^n - u_i^n}{\delta x}.$$

Taking then a *centered* discretization for f_i^n, we obtain the following discretization of the PDE in (1.3.8):

$$\frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + c\frac{u_{i+1}^n - u_{i-1}^n}{2} - \nu\frac{u_{i+1}^n - u_i^n}{\delta x} + \nu\frac{u_i^n - u_{i-1}^n}{\delta x} = 0. \tag{1.3.9}$$

What would it take for (1.3.9) to satisfy the maximum principle (which is valid for the PDE (1.3.8) it discretizes)? Let us make the same reasoning as in the proof of Proposition 1.3.1: from (1.3.9) we write

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

The sum of the coefficients of u_{i-1}^n, u_i^n and u_{i+1}^n is equal to 1 and u_i^{n+1} is thus a convex combination of u_{i-1}^n, u_i^n and u_{i+1}^n provided that all these coefficients are non-negative, that is to say:

$$2\nu\frac{\delta t}{\delta x^2} \leq 1 \tag{1.3.10}$$

and

$$\frac{|c|}{2}\delta x \leq \nu. \tag{1.3.11}$$

Remark 1.3.7 *Condition (1.3.10) plays the role of the CFL condition, albeit involving a relation between the discretization steps and the coefficient ν of the higher order term in (1.3.8) (i.e. the diffusion term); one can notice that this condition is much more restrictive than the CFL for the upwind discretization of the transport equation: the time step must here be of order the square of the space step (hence a small space step imposes a much smaller time step: if the space step is divided by 10, the time step must be divided by 100).*

Remark 1.3.8 *Although we have chosen a centered discretization for the convection term $\partial_x(cu)$, the presence of the diffusion term allows to bound the approximate solution provided that (1.3.11) is satisfied; this condition (called the Peclet condition) states that, if the convection and diffusion coefficients are fixed, then a small enough space step is enough to control the convection term using the diffusion term. However, in many practical situations, ν can be quite small with respect to c and (1.3.11) then imposes a strong condition on δx.*

Let us now come back to the upwind discretization (1.3.2) of (1.1.1) (we thus take $c > 0$), and let us rewrite it the following way:

$$\frac{\delta t}{\delta x}(u_i^{n+1} - u_i^n) + c\frac{u_{i+1}^n - u_{i-1}^n}{2} - \frac{c\delta x}{2}\frac{u_{i+1}^n - u_i^n}{\delta x} + \frac{c\delta x}{2}\frac{u_i^n - u_{i-1}^n}{\delta x} = 0. \tag{1.3.12}$$

Comparing this writing with (1.3.9), we notice that the upwind scheme for the linear pure convection equation is identical to the discretization of a convection-diffusion equation using the centered scheme for the convective part and the diffusion coefficient

$$\nu = \frac{c\delta x}{2}$$

(of course, this comparison is formal and only holds at the discrete level since the diffusion coefficient in a true convection-diffusion equation cannot depend on some kind of “time step”).
In other words, we can also say that the upwind discretization of a convection equation consists in adding a small discrete diffusion term (namely $-\frac{c\delta x}{2}\frac{u_{i+1}^n - u_i^n}{\delta x} + \frac{c\delta x}{2}\frac{u_i^n - u_{i-1}^n}{\delta x} \approx -\frac{c\delta x}{2}\int_{i\delta x}^{(i+1)\delta x} \partial_x^2 u$) to the centered discretization of the same convection equation (compare (1.2.2) and (1.3.12)).
One can also notice that, with this diffusion coefficient $\nu = \frac{c\delta x}{2}$ (recall that $c > 0$ here), the Peclet condition (1.3.11) of this formal discretization of a convection-diffusion equation is satisfied, and that the related CFL condition (1.3.10) is equivalent to the CFL condition (1.3.3) for the pure convection equation.

1.3.3 About the time-implicit discretization

In (1.1.3), f_i^n plays the role of an approximation of cu on $[n\delta t,(n+1)\delta t[\times\{i\delta x\}$. We discussed the use of approximate values of u in the control volumes on either side of $x = i\delta x$ to compute the numerical flux f_i^n; however, we did not discuss the time at which these approximate values should be taken: $t = n\delta t$ or $t = (n+1)\delta t$?

In fact, right from the beginning we only considered that f_i^n should be computed using $(u_j^n)_{j\in\mathbb{Z}}$ (i.e. only the values at time $t = n\delta t$); the advantage of this *a priori* choice is that it ensures that (1.1.3) is an equation allowing to simply and directly compute the approximate values at time $t = (n+1)\delta t$ from the approximate values at time $t = n\delta t$. This is a quite natural expectation for *evolution* equations such as (1.1.1) (which precisely tells how u evolves from a given initial state), but since f_i^n is supposed to approximate cu at $x = i\delta x$ on the whole time interval $[n\delta t,(n+1)\delta t[$, a similar natural expectation would also be to use the values at time $t = (n+1)\delta t$ to compute f_i^n.

Still considering the case of an upwind scheme with $c > 0$, we could in particular choose, instead of (1.3.1),

$$f_i^n = cu_{i-1}^{n+1}. \tag{1.3.13}$$

The resulting scheme is

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + cu_i^{n+1} - cu_{i-1}^{n+1} = 0, \tag{1.3.14}$$

to be completed with the discretization (1.1.4) of the initial condition. On the contrary to (1.3.2), (1.3.14) does not give a straightforward way to compute $(u_i^{n+1})_{i\in\mathbb{Z}}$ from $(u^n)_{i\in\mathbb{Z}}$; in fact, nothing ensures that, given $(u^n)_{i\in\mathbb{Z}}$, there exists $(u_i^{n+1})_{i\in\mathbb{Z}}$ which satisfies (1.3.14); put in another way: once (1.1.4) has given $(u_i^0)_{i\in\mathbb{Z}}$, will we even be able to find (let alone compute) $(u_i^1)_{i\in\mathbb{Z}}$ from (1.3.14) ?

This scheme [(1.3.14),(1.1.4)] is called "implicit" because it only gives (if it exists) $(u_i^{n+1})_{i\in\mathbb{Z}}$ from $(u_i^n)_{i\in\mathbb{Z}}$ in an implicit way: solving (1.3.14) to compute $(u_i^{n+1})_{i\in\mathbb{Z}}$ is not obvious. Implicit schemes are less easy to use in practical, but they sometimes have very interesting properties. The reader familiar with the Euler discretizations of ODE can for example remember that, for equations of the kind $y'(t) = -y(t)$, the explicit Euler scheme does not always preserve the positivity of the initial data (which is however preserved by the ODE), whereas the implicit scheme ensures that the approximate solution stays positive if the initial data is positive. For the conservation law (1.1.1), the implicit upwind scheme also has such an interesting property: it satisfies the maximum principle *whatever the values of the time and space steps* (recall that the explicit scheme only satisfies this principle if the steps satisfy the CFL condition (1.3.3)).

Proposition 1.3.9 (Stability of the implicit upwind scheme) *Let $c > 0$. For any $\delta t > 0$ and $\delta x > 0$, if $(u^n)_{n\geq 0\,,\,i\in\mathbb{Z}}$ satisfies [(1.3.14),(1.1.4)] and, for all $n \geq 0$, $\sup_{j\in\mathbb{Z}} |u_j^n| < +\infty$, then*

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \inf_{j\in\mathbb{Z}} u_j^n \leq u_i^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n. \tag{1.3.15}$$

Proof of Proposition 1.3.9
Let $n \geq 0$ and assume first that there exists $i_0 \in \mathbb{Z}$ such that $u_{i_0}^{n+1} = \sup_{j\in\mathbb{Z}} u_j^{n+1}$. The scheme (1.3.14) gives

$$u_{i_0}^{n+1} - u_{i_0}^n = \frac{\delta t}{\delta x} c(u_{i_0-1}^{n+1} - u_{i_0}^{n+1}). \tag{1.3.16}$$

Now, by definition of i_0 we have $u_{i_0-1}^{n+1} - u_{i_0}^{n+1} \leq 0$ and thus $\sup_{j\in\mathbb{Z}} u_j^{n+1} = u_{i_0}^{n+1} \leq u_{i_0}^n \leq \sup_{j\in\mathbb{Z}} u_j^n$, which proves the second inequality in (1.3.15).

If such an i_0 does not exist, then one can choose $(i_k)_{k\geq 1}$ such that $u_{i_k}^{n+1} \to \sup_{j\in\mathbb{Z}} u_j^{n+1}$ as $k \to \infty$ (recall that $(u_j^{n+1})_{j\in\mathbb{Z}}$ is bounded by assumption) and deduce from (1.3.16) with i_k instead of i_0:

$$u_{i_k}^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n + \frac{\delta t}{\delta x} c(\sup_{j\in\mathbb{Z}} u_j^{n+1} - u_{i_k}^{n+1}).$$

Passing to the limit $k \to \infty$ gives back $\sup_{j\in\mathbb{Z}} u_j^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n$.
The first inequality in (1.3.15) can be deduced with a similar reasoning using either i_0 such that $u_{i_0}^{n+1} = \inf_{j\in\mathbb{Z}} u_j^{n+1}$ or a sequence $(i_k)_{k\geq 1}$ such that $u_{i_k}^{n+1} \to \inf_{j\in\mathbb{Z}} u_j^{n+1}$ as $k \to \infty$. ■

It remains to prove that there exists a bounded solution to [(1.3.2),(1.1.4)]. This solution indeed exists, and is unique (see e.g. the end of Section 2.5 for an idea of the proof, or the comments on the bibliography at the end of the document).

Let us conclude this discussion on the implicit discretization by the following remark: in the equation $\partial_t u + \partial_x(cu) = 0$, the same way the term $\partial_x(cu)$ models the convection of u at the velocity c along the space direction, one could interpret $\partial_t u$ as the convection of u along the time direction with a velocity 1 (evolution equations are nothing more than a transport toward the future). With such an interpretation in mind, our discussion on the discretization of convection terms would lead us to discretize $\partial_t u$ using an upwind choice; if we think of u_i^n as an approximation on $[n\delta t, (n+1)\delta t[\times[i\delta x, (i+1)\delta x[$ (and not only on $[i\delta x, (i+1)\delta x[$ at time $t = n\delta t$ as our in previous idea) — which is coherent with the definition of $u_{\delta t,\delta x}$ in Theorem 1.3.3 — one would thus replace the term

$$\frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n)$$

in (1.1.3) with

$$\frac{\delta x}{\delta t}(u_i^n - u_i^{n-1})$$

(approximating u at $t = (n+1)\delta t$ on $[i\delta x, (i+1)\delta x[$ with its time-upwind value, that is to say its value on $[n\delta t, (n+1)\delta t[\times[i\delta x, (i+1)\delta x[$). With such a change, the upwind scheme (1.3.2), with the same choice (1.3.1) for f_i^n as in the previous explicit discretization, becomes

$$\frac{\delta x}{\delta t}(u_i^n - u_i^{n-1}) + cu_i^n - cu_{i-1}^n = 0,$$

that is to say, up to a change of index $n \to n+1$, precisely the implicit upwind scheme (1.3.14).

Hence, the time-implicit discretization is nothing else than an upwind discretization of the temporal convection term in the equation...

Chapter 2

Schemes for non-linear conservation laws

2.1 Introduction

We now consider a non-linear scalar conservation law

$$\begin{cases} \partial_t u(t,x) + \partial_x(f(u(t,x))) = 0 & t>0\,,\ x\in\mathbb{R}\,,\\ u(0,x) = u_0(x) & x\in\mathbb{R} \end{cases} \tag{2.1.1}$$

with $f:\mathbb{R}\to\mathbb{R}$ locally Lipschitz-continuous and $u_0\in L^\infty(\mathbb{R})$. Our aim is, once again, to construct a stable converging scheme to approximate the solution to (2.1.1); since this equation is non-linear, its well-posedness requires the use of the notion of entropy solution (which we recall below) and we therefore have to be careful to construct approximations to the (unique) entropy solution, and not another weak solution.

Definition 2.1.1 (Entropy solution for (2.1.1)) *An entropy solution to (2.1.1) is* $u\in L^\infty(]0,\infty[\times\mathbb{R})$ *such that, for all* $\kappa\in\mathbb{R}$ *and all non-negative* $\varphi\in C_c^\infty([0,\infty[\times\mathbb{R})$*,*

$$\int_0^\infty\int_{\mathbb{R}} |u(t,x)-\kappa|\partial_t\varphi(t,x)+(f(u(t,x)\top\kappa)-f(u(t,x)\bot\kappa))\partial_x\varphi(t,x)\,dtdx \\ +\int_{\mathbb{R}}|u_0(x)-\kappa|\varphi(0,x)\,dx\geq 0\,, \tag{2.1.2}$$

where $a\top\kappa=\max(a,\kappa)$ *and* $a\bot\kappa=\min(a,\kappa)$*.*

Remark 2.1.2 *The usual definition of entropy solution makes use of general convex functions* $\eta:\mathbb{R}\to\mathbb{R}$ *and replaces* $|u(t,x)-\kappa|$ *with* $\eta(u(t,x))$ *and* $f(u(t,x)\top\kappa)-f(u(t,x)\bot\kappa)$ *with* $\phi(u(t,x))$*, where* $\phi(s)=\int_0^s\eta'(r)f'(r)\,dr$*. Definition 2.1.1, using only Krushkov's entropies* $\eta(s)=|s-\kappa|$*, is equivalent to the general definition.*

2.2 Monotone schemes

The principle of construction of a finite volume scheme for (2.1.1) is the same as in the linear case: we take time and space steps δt and δx, and we integrate the equation on a cell $[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[$, obtaining

$$\int_{i\delta x}^{(i+1)\delta x} u((n+1)\delta t,x)\,dx-\int_{i\delta x}^{(i+1)\delta x} u(n\delta t,x)\,dx$$

$$+\int_{n\delta t}^{(n+1)\delta t} f(u(t,(i+1)\delta x))\,dt - \int_{n\delta t}^{(n+1)\delta t} f(u(t,i\delta x))\,dt = 0.$$

If u_i^n is an approximate value of u at time $t = n\delta t$ on $[i\delta x, (i+1)\delta x[$ and f_i^n is an approximation of $f(u)$ on $[n\delta t, (n+1)\delta t[$ at $x = i\delta x$, dividing the preceding expression by δt, we are lead to consider the scheme

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + f_{i+1}^n - f_i^n = 0. \tag{2.2.1}$$

Assuming that f_i^n is expressed in terms of $(u_j^n)_{j\in\mathbb{Z}}$, completing (2.2.1) with the initial data

$$\forall i \in \mathbb{Z}\ :\quad u_i^0 = \frac{1}{\delta x}\int_{i\delta x}^{(i+1)\delta x} u_0(x)\,dx\,, \tag{2.2.2}$$

we obtain a scheme giving by induction the approximate values $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$. The hanging question is therefore to find an way to compute the approximation f_i^n of $f(u)$ in terms of the approximations $(u_j^n)_{j\in\mathbb{Z}}$ of u.

2.2.1 A first idea

Obviously, since it failed in the linear case (when $f(u) = cu$), there is little chance that the "natural" (for the mathematician) centered choice

$$f_i^n = \frac{f(u_{i-1}^n) + f(u_i^n)}{2} \tag{2.2.3}$$

works in general. However, following the reasoning in section 1.3.2 (linking, in the linear case, the upwind choice with the discretization of an additional diffusion term), we can try and modify this centered choice by adding some numerical diffusion.

The centered discretization (2.2.3) leads to the scheme

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + \frac{f(u_{i+1}^n) + f(u_i^n)}{2} - \frac{f(u_i^n) + f(u_{i-1}^n)}{2} = 0.$$

Adding some diffusion terms as in (1.3.12), we would write

$$\begin{aligned}\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\quad & \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + \frac{f(u_{i+1}^n) + f(u_i^n)}{2} - \frac{f(u_i^n) + f(u_{i-1}^n)}{2} \\ & -D(u_{i+1}^n - u_i^n) + D(u_i^n - u_{i-1}^n) = 0\end{aligned} \tag{2.2.4}$$

with some $D > 0$ to be chosen so as to ensure the stability of the scheme. This scheme corresponds to (2.2.1) with

$$f_i^n = \frac{f(u_{i-1}^n) + f(u_i^n)}{2} - D(u_i^n - u_{i-1}^n). \tag{2.2.5}$$

A proper choice of D can be made by trying, as in the linear case, to transform (2.2.4) in order to express u_i^{n+1} as a convex combination of u_{i-1}^n, u_i^n and u_{i+1}^n. We have

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\delta t}{\delta x}\frac{f(u_{i+1}^n) + f(u_i^n)}{2} + \frac{\delta t}{\delta x}\frac{f(u_i^n) + f(u_{i-1}^n)}{2} \\ &\quad + D\frac{\delta t}{\delta x}(u_{i+1}^n - u_i^n) - D\frac{\delta t}{\delta x}(u_i^n - u_{i-1}^n) \\ &= u_i^n - \frac{\delta t}{\delta x}\frac{f(u_{i+1}^n) - f(u_i^n)}{2} + \frac{\delta t}{\delta x}\frac{f(u_{i-1}^n) - f(u_i^n)}{2} \\ &\quad + D\frac{\delta t}{\delta x}(u_{i+1}^n - u_i^n) - D\frac{\delta t}{\delta x}(u_i^n - u_{i-1}^n) \\ &= u_i^n - \frac{\delta t}{\delta x}\alpha_i^n(u_{i+1}^n - u_i^n) + \frac{\delta t}{\delta x}\beta_i^n(u_i^n - u_{i-1}^n) \\ &\quad + D\frac{\delta t}{\delta x}(u_{i+1}^n - u_i^n) - D\frac{\delta t}{\delta x}(u_i^n - u_{i-1}^n) \end{aligned}$$

where $\alpha_i^n = \frac{f(u_{i+1}^n)-f(u_i^n)}{2(u_{i+1}^n-u_i^n)}$ and $\beta_i^n = \frac{f(u_i^n)-f(u_{i-1}^n)}{2(u_i^n-u_{i-1}^n)}$ (if one or the other of the denominators vanishes, we simply let $\alpha_i^n = 0$ or $\beta_i^n = 0$). This leads to

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\delta t}{\delta x}(\alpha_i^n - D)(u_{i+1}^n - u_i^n) + \frac{\delta t}{\delta x}(\beta_i^n - D)(u_i^n - u_{i-1}^n) \\ &= \left(1 + \frac{\delta t}{\delta x}(\beta_i^n + \alpha_i^n - 2D)\right) u_i^n + \frac{\delta t}{\delta x}(D - \alpha_i^n)u_{i+1}^n + \frac{\delta t}{\delta x}(D - \beta_i^n)u_{i-1}^n \end{aligned}$$

and u_i^{n+1} is therefore a convex combination of u_{i-1}^n, u_i^n and u_{i+1}^n if

$$|\alpha_i^n| \le D \quad \text{and} \quad |\beta_i^n| \le D \tag{2.2.6}$$

and

$$|\beta_i^n + \alpha_i^n - 2D|\frac{\delta t}{\delta x} \le 1. \tag{2.2.7}$$

By construction, if L^n is the Lipschitz constant of f on $[\inf_{i\in\mathbb{Z}} u_i^n, \sup_{i\in\mathbb{Z}} u_i^n]$ then $|\alpha_i^n| \le \frac{L^n}{2}$ and $|\beta_i^n| \le \frac{L^n}{2}$ and (2.2.6) and (2.2.7) are satisfied as soon as $\frac{L^n}{2} \le D$ and $(L^n + 2D)\frac{\delta t}{\delta x} \le 1$. It is easy to see by induction on n that if we choose D, δt and δx such that

$$\frac{\mathrm{Lip}_{u_0}(f)}{2} \le D \tag{2.2.8}$$

(with $\mathrm{Lip}_{u_0}(f)$ the Lipschitz constant of f on $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$) and

$$(\mathrm{Lip}_{u_0}(f) + 2D)\frac{\delta t}{\delta x} \le 1\,, \tag{2.2.9}$$

then (2.2.6) and (2.2.7) are satisfied for all $n \ge 0$ and the scheme [(2.2.1),(2.2.2),(2.2.5)] is stable: the computed approximate values all belong to $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$.
Equation (2.2.8) indicates how to take D in (2.2.5); we notice that D can be chosen only depending on f and u_0. Equation (2.2.9) is the CFL condition associated with the scheme: as in the linear case, the time step must not be to large with respect to the space step in order to obtain a stable approximation. This scheme is called the "modified" Lax-Friedrichs scheme ([1]).

2.2.2 General monotone schemes, L^∞ stability

We would like to find more general ways to write the f_i^n in (2.2.1) in terms of $(u_j^n)_{j\in\mathbb{Z}}$. In a first approach, it seems reasonable to decide that f_i^n will be computed using only u_{i-1}^n and u_i^n (this was the case in the linear upwind choice (1.3.1) and in the non-linear modified Lax-Friedrichs choice (2.2.5)), that is to say

$$f_i^n = F(u_{i-1}^n, u_i^n) \quad \text{with} \quad F : \mathbb{R}\times\mathbb{R} \to \mathbb{R}. \tag{2.2.10}$$

The question is to find properties on F which ensure that [(2.2.1),(2.2.2),(2.2.10)] is a "good" scheme.

Obviously, there must be some kind of relation between F and f, since f_i^n given by (2.2.10) is supposed to be an approximation of $f(u)$ on $[n\delta t, (n+1)\delta t[$ at $x = i\delta x$. A simple way to find such a relation is to consider the case where the approximate values u_{i-1}^n and u_i^n of u in the two cells neighboring $x = i\delta x$ are identical: there is then no reason to imagine that an approximate value of u at $x = i\delta x$ would differ from this common value and thus, in this situation, we would like to have $f_i^n = f(u_{i-1}^n) = f(u_i^n)$. This imposes

$$\forall a \in \mathbb{R} : \quad F(a,a) = f(a).$$

[1] The original Lax-Friedrichs scheme consists in taking $D = \frac{\delta x}{2\delta t}$; its convergence however imposes an "inverse" CFL condition $\frac{\delta t}{\delta x} \ge C$ (for some $C > 0$), therefore forcing δt and δx to have the same order (such a inverse CFL condition is not required to obtain the convergence of the modified Lax-Friedrichs scheme, as we show below).

This property is satisfied by $F(a,b) = \frac{f(a)+f(b)}{2} + D(a-b)$, the modified Lax-Friedrichs numerical flux.

As we already pointed out in the linear case and during the construction of the modified Lax-Friedrichs scheme, a crucial property of schemes is their stability. A way to ensure that F leads to a stable scheme is, once again, to check if (2.2.1) and (2.2.10) allow to see u_i^{n+1} as a convex combination of u_{i-1}^n, u_i^n and u_{i+1}^n. We begin with

$$u_i^{n+1} = u_i^n - \frac{\delta t}{\delta x} f_{i+1}^n + \frac{\delta t}{\delta x} f_i^n = u_i^n - \frac{\delta t}{\delta x} F(u_i^n, u_{i+1}^n) + \frac{\delta t}{\delta x} F(u_{i-1}^n, u_i^n) \tag{2.2.11}$$

and, subtracting and adding $F(u_i^n, u_i^n)$ to the last two terms,

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\delta t}{\delta x}(F(u_i^n, u_{i+1}^n) - F(u_i^n, u_i^n)) + \frac{\delta t}{\delta x}(F(u_{i-1}^n, u_i^n) - F(u_i^n, u_i^n)) \\ &= u_i^n + \frac{\delta t}{\delta x} a_i^n (u_{i+1}^n - u_i^n) + \frac{\delta t}{\delta x} b_i^n (u_{i-1}^n - u_i^n) \end{aligned} \tag{2.2.12}$$

where

$$a_i^n = -\frac{F(u_i^n, u_{i+1}^n) - F(u_i^n, u_i^n)}{u_{i+1}^n - u_i^n} \tag{2.2.13}$$

and

$$b_i^n = \frac{F(u_{i-1}^n, u_i^n) - F(u_i^n, u_i^n)}{u_{i-1}^n - u_i^n} \tag{2.2.14}$$

(as before, if one or the other denominator vanishes, so does the corresponding quantity a_i^n or b_i^n). From (2.2.12) we obtain

$$u_i^{n+1} = \left(1 - \frac{\delta t}{\delta x}(a_i^n + b_i^n)\right) u_i^n + \frac{\delta t}{\delta x} a_i^n u_{i+1}^n + \frac{\delta t}{\delta x} b_i^n u_{i-1}^n. \tag{2.2.15}$$

The convex combination is achieved if

$$a_i^n \geq 0\,, \quad b_i^n \geq 0 \quad \text{and} \quad \frac{\delta t}{\delta x}(a_i^n + b_i^n) \leq 1. \tag{2.2.16}$$

The non-negativity of a_i^n and b_i^n is ensured if we impose that F is non-decreasing with respect to its first argument, and non-increasing with respect to its second argument. If we assume that F is locally Lipschitz-continuous with respect to each of its arguments then, denoting L_1^n and L_2^n the Lipschitz constants with respect to its first and second variables on $[\inf_{i\in\mathbb{Z}} u_i^n, \sup_{i\in\mathbb{Z}} u_i^n]^2$, we have $|a_i^n| \leq L_1^n$ and $|b_i^n| \leq L_2^n$ and the last condition in (2.2.16) is satisfied if $\frac{\delta t}{\delta x}(L_1^n + L_2^n) \leq 1$.

The preceding reasoning leads us to the following general definition of a "good" way to compute the approximate fluxes f_i^n by a formula (2.2.10).

Definition 2.2.1 (Monotone numerical flux and monotone schemes) *A monotone (upwind) numerical flux for (2.1.1) is a function $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ which satisfies the following properties:*

$$\forall a \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0] \;:\quad F(a,a) = f(a)\,, \tag{2.2.17}$$

$$F \text{ is Lipschitz-continuous with respect to each of its variables on } [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^2\,, \tag{2.2.18}$$

$$\begin{gathered}\text{on } [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^2,\ F \text{ is non-decreasing with respect to its first variable} \\ \text{and non-increasing with respect to its second variable.}\end{gathered} \tag{2.2.19}$$

A monotone (upwind) scheme is [(2.2.1),(2.2.2),(2.2.10)] with F a monotone (upwind) numerical flux, that is to say

$$\begin{aligned} &\forall n \geq 0\,,\ \forall i \in \mathbb{Z} \;:\; \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n) = 0\,, \\ &\forall i \in \mathbb{Z} \;:\; u_i^0 = \frac{1}{\delta x}\int_{i\delta x}^{(i+1)\delta x} u_0(x)\, dx. \end{aligned} \tag{2.2.20}$$

Remark 2.2.2 *We ask for properties of F only on $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^2$ because, as stated in the following proposition, under a suitable CFL condition all the approximate values $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$ computed by the scheme stay in fact inside $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$.*

The following proposition, which establishes the stability of monotone schemes, is a simple consequence (by induction on n) of the above reasoning.

Proposition 2.2.3 (Stability of monotone schemes) *Assume that F is a monotone numerical flux for (2.1.1) and denote by $\mathrm{Lip}_{1,u_0}(F)$ and $\mathrm{Lip}_{2,u_0}(F)$ the Lipschitz constants of F with respect to its first and second variables on $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^2$. Then, under the CFL condition*

$$\frac{\delta t}{\delta x}(\mathrm{Lip}_{1,u_0}(F) + \mathrm{Lip}_{2,u_0}(F)) \leq 1, \tag{2.2.21}$$

if $(u_i^n)_{n\geq 0\,, i\in\mathbb{Z}}$ satisfies the scheme (2.2.20) we have

$$\forall n \geq 0\,, \forall i \in \mathbb{Z}\ :\quad \inf_{j\in\mathbb{Z}} u_j^n \leq u_i^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n.$$

In particular, all the values $(u_i^n)_{n\geq 0\,,\, i\in\mathbb{Z}}$ are between the infimum and supremum values of u^0 and

$$\sup_{n\geq 0\,,\, i\in\mathbb{Z}} |u_i^n| \leq ||u_0||_{L^\infty(\mathbb{R})}.$$

2.2.3 Examples of monotone fluxes, interpretation of the CFL condition

Here are three possible kinds of F satisfying (2.2.17)–(2.2.19).

1. $F(a,b) = \frac{f(a)+f(b)}{2} + D(a-b)$ with $D \geq \frac{\mathrm{Lip}_{u_0}(f)}{2}$ (recall that $\mathrm{Lip}_{u_0}(f)$ is the Lipschitz constant of f on $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$). This is the modified Lax-Friedrichs numerical flux.
2. $F(a,b) = f_1(a) + f_2(b)$, where f is written $f(s) = f_1(s) + f_2(s)$ with f_1 non-decreasing and f_2 non-increasing (both functions being locally Lipschitz-continuous). This is the splitting numerical flux ([2]). The modified Lax-Friedrichs flux is a special case of splitting flux with $f_1(s) = \frac{f(s)}{2} + Ds$ and $f_2(s) = \frac{f(s)}{2} - Ds$.
3. $$F(a,b) = \begin{cases} \min_{[a,b]} f & \text{if } a \leq b\,, \\ \max_{[b,a]} f & \text{if } a > b. \end{cases} \tag{2.2.22}$$

   This is the Godunov numerical flux, which we now discuss in more depth.

Remark 2.2.4 *For a linear flux $f(s) = cs$, the modified Lax-Friedrichs numerical flux with $D = \frac{\mathrm{Lip}_{u_0}(f)}{2} = \frac{c}{2}$ and the Godunov numerical flux give back the upwind choice (1.3.1).*

Expression (2.2.22) is a simple and straightforward definition of the Godunov numerical flux, but this is not the original way it has been constructed. The idea of the Godunov flux is the following: in order to construct an approximation $f_i^n = F(a,b)$ of $f(u)$ on $[n\delta t, (n+1)\delta t]$ at $x = i\delta x$ when $u_{i-1}^n = a$ and $u_i^n = b$, consider the Riemann problem

$$\begin{cases} \partial_t v + \partial_x(f(v)) = 0 & t > 0\,,\ x \in \mathbb{R}\,, \\ v(0,x) = \begin{cases} a = u_{i-1}^n & \text{if } x < 0\,, \\ b = u_i^n & \text{if } x > 0. \end{cases} \end{cases} \tag{2.2.23}$$

[2] Not to be confused with the "splitting method" sometimes used in numerical analysis, see Remark 2.5.1.

and let $f_i^n = F(a,b) = f(v(\delta t, 0))$ (the solution v might not be pointwise well-defined at $x = 0$, but the flux $f(v)$ always is). In other words, we let the scalar conservation law evolve from the Riemann initial data defined by u_{i-1}^n and u_i^n and take the value of the resulting flux at time $t = \delta t$ at the interface $x = i\delta x$. It is known that the solution to (2.2.23) is self-similar and can therefore be written $v(t,x) = v^*(x/t)$; the Godunov flux also can be defined by $F(a,b) = f(v^*(0))$.
If f is convex (or concave), then there are simple expressions for v (depending on whether $a \leq b$ — v is then a rarefaction wave — or $a > b$ — v is then a shock) and it is easy to check that the definition of $F(a,b)$ through (2.2.23) is identical to (2.2.22). This is also true for more general f, but less easy to verify.

Obviously, in practical (scalar!) situations, only (2.2.22) is used, but the original definition using the Riemann problem is quite interesting to shed a new light on the CFL condition. We have represented in Figure 2.1 a possible solution of $\partial_t w + \partial_x f(w) = 0$ with the initial data corresponding to the approximate values of u at $t = n\delta t$ (in this figure, the solution is a shock at $x = i\delta x$ and a rarefaction wave at $x = (i+1)\delta x$).

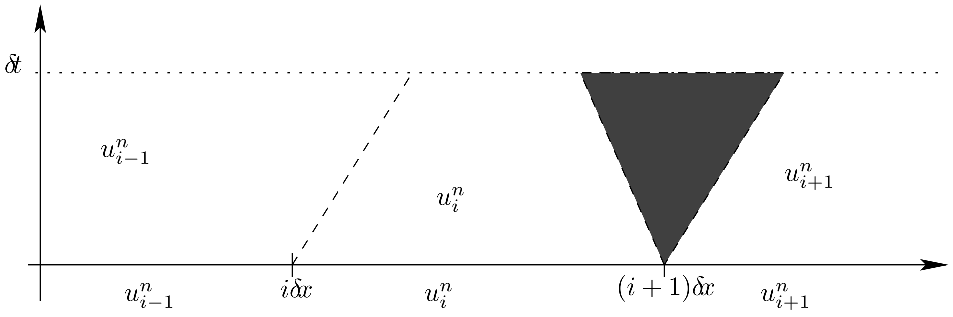


Figure 2.1: Example of solutions to the scalar conservation law with the initial condition made of the approximate values of u at $t = n\delta t$.

The principle of finite speed propagation for $\partial_t w + \partial_x(f(w)) = 0$ states that the value of the initial data at some point x can only influence, at time δt, the solution on $[x - L\delta t, x + L\delta t]$ with L the Lipschitz constant of f on an interval containing the range of the initial data. In the situation described by Figure 2.1 and under the condition

$$2\delta t \mathrm{Lip}_{u_0}(f) \leq \delta x \,, \tag{2.2.24}$$

this means that the waves coming from each interface $x = i\delta x$ cannot interact with each other on the time span $[0, \delta t]$. Hence, in this case, the values $(w(\delta t, i\delta x))_{i\in\mathbb{Z}}$ can be computed by solving each Riemann problem (2.2.23) separately, and the Godunov numerical fluxes $(F(u_{i-1}^n, u_i^n))_{i\in\mathbb{Z}}$ simply correspond to $(f(w(\delta t, i\delta x)))_{i\in\mathbb{Z}}$.
Noticing that, for the Godunov numerical flux, $\mathrm{Lip}_{1,u_0}(F) = \mathrm{Lip}_{2,u_0}(F) = \mathrm{Lip}_{u_0}(f)$, (2.2.21) is exactly (2.2.24): the CFL condition is therefore simply a way to ensure that the waves originating from each interface do not interact during the time δt. Under this condition, the Godunov scheme has a global interpretation as an evolution-projection process: it can be shown that it consists in letting $\partial_t w + \partial_x(f(w)) = 0$ evolve from the initial data defined by the values $(u_i^n)_{i\in\mathbb{Z}}$ and in defining $(u_i^{n+1})_{i\in\mathbb{Z}}$ as the L^2 projection of $w(\delta t, \cdot)$ on the piecewise constant functions, i.e. $u_i^{n+1} = \frac{1}{\delta x}\int_{i\delta x}^{(i+1)\delta x} w(\delta t, x)\, dx$.

Let us conclude that the Godunov scheme is one of the most favored schemes for (2.1.1), because it is simple to implement (thanks to (2.2.22)), simple to generalize to systems (thanks to its definition by (2.2.23)) and introduces in general less numerical diffusion than the modified Lax-Friedrichs or splitting fluxes.

2.2.4 Numerical diffusion

The modified Lax-Friedrichs scheme has been constructing by adding some numerical diffusion to the centered flux and, in fact, all monotone schemes can be seen as such.
If F is a monotone numerical flux, (2.2.10) gives

$$\begin{aligned} f_i^n &= \frac{F(u_{i-1}^n,u_{i-1}^n)+F(u_i^n,u_i^n)}{2}+\frac{1}{2}\left(F(u_{i-1}^n,u_i^n)-F(u_{i-1}^n,u_{i-1}^n)\right)+\frac{1}{2}(F(u_{i-1}^n,u_i^n)-F(u_i^n,u_i^n)) \\ &= \frac{f(u_{i-1}^n)+f(u_i^n)}{2}-g(u_{i-1}^n,u_i^n)(u_i^n-u_{i-1}^n) \qquad (2.2.25) \end{aligned}$$

with

$$g(u_{i-1}^n,u_i^n)=-\frac{F(u_{i-1}^n,u_i^n)-F(u_{i-1}^n,u_{i-1}^n)}{2(u_i^n-u_{i-1}^n)}+\frac{F(u_{i-1}^n,u_i^n)-F(u_i^n,u_i^n)}{2(u_{i-1}^n-u_i^n)}.$$

The monotony assumption (2.2.19) on F ensures that $g(u_{i-1}^n,u_i^n)\geq 0$ and (2.2.25) therefore shows that f_i^n indeed corresponds to a centered flux with the addition of a numerical diffusion term (compare with (2.2.5)), the coefficient of which depends on the unknowns at time step n.

2.3 Study of monotone schemes

We shall now prove that monotone schemes converge, as the time and space steps tend to 0 while satisfying the CFL condition, to (2.1.1). We have already proved in Proposition 2.2.3 an L^∞ estimate on the approximate solution, which implies (up to a subsequence) the weak convergence of this solution; however, because of the non-linearity in (2.1.1), this weak convergence is not enough to conclude: we have to prove stronger compactness properties on this solution, as well as some inequalities which ensure that its possible limits are not only weak solutions to (2.1.1), but also entropy solutions.

2.3.1 *BV* estimates

The strong compactness property of the approximate solution is a consequence of the following BV estimates.

Proposition 2.3.1 (Space BV estimates) *Assume that F is a monotone numerical flux for (2.1.1) and that (2.2.21) holds. If $(u_i^n)_{n\geq 0,i\in\mathbb{Z}}$ satisfies the scheme (2.2.20), then*

$$\forall n\geq 0\ :\quad \sum_{i\in\mathbb{Z}}|u_i^{n+1}-u_{i-1}^{n+1}|\leq\sum_{i\in\mathbb{Z}}|u_i^n-u_{i-1}^n|. \qquad (2.3.1)$$

In particular, if $u_0\in BV(\mathbb{R})$ then

$$\forall n\geq 0\ :\quad \sum_{i\in\mathbb{Z}}|u_i^n-u_{i-1}^n|\leq|u_0|_{BV(\mathbb{R})} \qquad (2.3.2)$$

where $|u_0|_{BV(\mathbb{R})}$ is the usual BV semi-norm of u_0.

Remark 2.3.2 *One can easily verify that $\sum_{i\in\mathbb{Z}}|u_i^n-u_{i-1}^n|$ is the total variation (the BV semi-norm) on $\mathbb{R}$ of the piecewise constant function equal to u_i^n on $[i\delta x,(i+1)\delta x[$ for all $i\in\mathbb{Z}$; (2.3.1) therefore states that the space BV semi-norm of the approximate solution is non-increasing with respect to the time. Schemes satisfying this property are called* Total Variation Decreasing *(TVD).*

Proof of Proposition 2.3.1

We use the expression (2.2.12) of the scheme to write

$$\begin{aligned}
u_i^{n+1}-u_{i-1}^{n+1} &= u_i^n+\frac{\delta t}{\delta x}a_i^n(u_{i+1}^n-u_i^n)+\frac{\delta t}{\delta x}b_i^n(u_{i-1}^n-u_i^n)\\
&\quad -u_{i-1}^n-\frac{\delta t}{\delta x}a_{i-1}^n(u_i^n-u_{i-1}^n)-\frac{\delta t}{\delta x}b_{i-1}^n(u_{i-2}^n-u_{i-1}^n)\\
&= \left(1-\frac{\delta t}{\delta x}(b_i^n+a_{i-1}^n)\right)(u_i^n-u_{i-1}^n)+\frac{\delta t}{\delta x}a_i^n(u_{i+1}^n-u_i^n)-\frac{\delta t}{\delta x}b_{i-1}^n(u_{i-2}^n-u_{i-1}^n).
\end{aligned}$$

Under (2.2.17)–(2.2.19) and (2.2.21), since all the values $(u_i^n)_{i\in\mathbb{Z}}$ belong to $[\inf_{\mathbb{R}} u_0,\sup_{\mathbb{R}} u_0]$ by Proposition 2.2.3, we have $1-\frac{\delta t}{\delta x}(b_i^n+a_{i-1}^n)\geq 0$, $a_i^n\geq 0$ and $b_{i-1}^n\geq 0$ (see (2.2.13) and (2.2.14)) and therefore

$$|u_i^{n+1}-u_{i-1}^{n+1}|\leq\left(1-\frac{\delta t}{\delta x}(b_i^n+a_{i-1}^n)\right)|u_i^n-u_{i-1}^n|+\frac{\delta t}{\delta x}a_i^n|u_{i+1}^n-u_i^n|+\frac{\delta t}{\delta x}b_{i-1}^n|u_{i-1}^n-u_{i-2}^n|.$$

Summing this inequality on $i\in\mathbb{Z}$ and re-indexing the sums, we find

$$\begin{aligned}
\sum_{i\in\mathbb{Z}}|u_i^{n+1}-u_{i-1}^{n+1}| &\leq \sum_{i\in\mathbb{Z}}\left(1-\frac{\delta t}{\delta x}(b_i^n+a_{i-1}^n)\right)|u_i^n-u_{i-1}^n|+\sum_{i\in\mathbb{Z}}\frac{\delta t}{\delta x}a_i^n|u_{i+1}^n-u_i^n|\\
&\quad +\sum_{i\in\mathbb{Z}}\frac{\delta t}{\delta x}b_{i-1}^n|u_{i-1}^n-u_{i-2}^n|\\
&\leq \sum_{i\in\mathbb{Z}}\left(1-\frac{\delta t}{\delta x}(b_i^n+a_{i-1}^n)\right)|u_i^n-u_{i-1}^n|+\sum_{i\in\mathbb{Z}}\frac{\delta t}{\delta x}a_{i-1}^n|u_i^n-u_{i-1}^n|\\
&\quad +\sum_{i\in\mathbb{Z}}\frac{\delta t}{\delta x}b_i^n|u_i^n-u_{i-1}^n|\\
&= \sum_{i\in\mathbb{Z}}|u_i^n-u_{i-1}^n|
\end{aligned}$$

and (2.3.1) is proved.
If $u_0\in BV(\mathbb{R})$, one can check ([3]) that

$$\sum_{i\in\mathbb{Z}}|u_i^0-u_{i-1}^0|\leq|u_0|_{BV(\mathbb{R})}$$

and the proof is therefore complete by induction on n. ■

Proposition 2.3.3 (Time BV estimates) *Assume that F is a monotone numerical flux for (2.1.1) and that (2.2.21) holds. If $(u_i^n)_{n\geq 0,i\in\mathbb{Z}}$ satisfies the scheme (2.2.20) and $u_0\in BV(\mathbb{R})$, then, for all $T\geq 0$,*

$$\sum_{i\in\mathbb{Z}}\delta x\sum_{n=0}^{[T/\delta t]}|u_i^{n+1}-u_i^n|\leq(T+\delta t)(\mathrm{Lip}_{1,u_0}(F)+\mathrm{Lip}_{2,u_0}(F))|u_0|_{BV(\mathbb{R})}$$

where $[T/\delta t]$ is the integer part of $T/\delta t$.

Remark 2.3.4 *One can check that $\sum_{n=0}^{[T/\delta t]}|u_i^{n+1}-u_i^n|$ is the BV semi-norm on $[0,T]$ of the piecewise constant function equal to u_i^n on $[n\delta t,(n+1)\delta t[$ for all $n\geq 0$. Proposition 2.3.3 therefore states a time BV estimate on the approximate solution, which is quite expected once we have the space BV estimate of Proposition 2.3.1: indeed, for the continuous equation $\partial_t u+\partial_x(f(u))=0$, a space BV estimate on u gives an estimate on $\partial_x(f(u))$, which therefore translates into a similar estimate on $\partial_t u$, i.e. a time BV estimate on u.*

[3] Assume first that $u_0\in C^1$ and write $u_i^0-u_{i-1}^0=\frac{1}{\delta x}\int_{i\delta x}^{(i+1)\delta x}(u_0(x)-u_0(x-\delta x))\,dx=\int_{i\delta x}^{(i+1)\delta x}\int_0^1 u_0'(x-s\delta x)\,dsdx$, so that $\sum_i|u_i^0-u_{i-1}^0|\leq\int_0^1\sum_i\int_{i\delta x}^{(i+1)\delta x}|u_0'(x-s\delta x)|\,dx\,ds=\int_0^1\int_{\mathbb{R}}|u_0'(x-s\delta x)|\,dx\,ds=||u_0'||_{L^1(\mathbb{R})}$; the general case $u_0\in BV(\mathbb{R})$ is obtained by a density argument.

Proof of Proposition 2.3.3
We have, from (2.2.12)–(2.2.14), (2.2.18) and the L^∞ estimate in Proposition 2.2.3,

$$\delta x|u_i^{n+1}-u_i^n| \leq \delta t \mathrm{Lip}_{2,u_0}(F)|u_{i+1}^n-u_i^n| + \delta t \mathrm{Lip}_{1,u_0}(F)|u_{i-1}^n-u_i^n|.$$

Summing on $i\in\mathbb{Z}$ and $n=0,\dots,[T/\delta t]$ (there is at most $\frac{T}{\delta t}+1$ such n) and using Proposition 2.3.1 we find

$$\sum_{i\in\mathbb{Z}}\delta x\sum_{n=0}^{[T/\delta t]}|u_i^{n+1}-u_i^n| \leq \delta t\left(\frac{T}{\delta t}+1\right)(\mathrm{Lip}_{1,u_0}(F)+\mathrm{Lip}_{2,u_0}(F))|u_0|_{BV(\mathbb{R})}$$

and the proof is complete. ■

Corollary 2.3.5 *Assume that F is a monotone numerical flux for (2.1.1) and that $u_0\in BV(\mathbb{R})$. Then, for all $T>0$, there exists $C=C(T,F,u_0)$ such that, if $\delta x>0$ and $\delta t\in]0,1[$ satisfy (2.2.21) and $(u_i^n)_{n\geq 0\,,i\in\mathbb{Z}}$ satisfies the scheme (2.2.20),*

$$|u_{\delta t,\delta x}|_{BV([0,T]\times\mathbb{R})}\leq C\,,$$

where $u_{\delta t,\delta x}:\mathbb{R}^+\times\mathbb{R}\to\mathbb{R}$ is the piecewise constant function equal to u_i^n on $[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[$ for all $n\geq 0$ and $i\in\mathbb{Z}$.

Proof of Corollary 2.3.5
We have $|u_{\delta t,\delta x}|_{BV([0,T]\times\mathbb{R})}\leq\int_0^T|u_{\delta t,\delta x}(t,\cdot)|_{BV(\mathbb{R})}\,dt+\int_{\mathbb{R}}|u_{\delta t,\delta x}(\cdot,x)|_{BV([0,T])}\,dx$ and the corollary follows from Propositions 2.3.1, 2.3.3 and Remarks 2.3.2, 2.3.4. ■

2.3.2 Discrete entropy inequalities

The previous BV estimates and Helly's theorem ensure the strong convergence of a subsequence of approximate solutions. However, in order to prove that the limit is not only a weak solution but the unique entropy solution to (2.1.1), we need some kind of discrete entropy inequalities (which, passing to the limit, will give the strong entropy inequalities for the limit of the approximations).

Proposition 2.3.6 (Discrete entropy inequalities) *Assume that F is a monotone numerical flux for (2.1.1) and that (2.2.21) holds. If $(u_i^n)_{n\geq 0\,,i\in\mathbb{Z}}$ satisfies the scheme (2.2.20) then, for all $\kappa\in\mathbb{R}$,*

$$\begin{aligned}\forall n\geq 0\,,\ \forall i\in\mathbb{Z}\ :\quad &\frac{\delta x}{\delta t}\left[|u_i^{n+1}-\kappa|-|u_i^n-\kappa|\right]\\ &+\left[F(u_i^n\top\kappa,u_{i+1}^n\top\kappa)-F(u_i^n\bot\kappa,u_{i+1}^n\bot\kappa)\right]\\ &-\left[F(u_{i-1}^n\top\kappa,u_i^n\top\kappa)-F(u_{i-1}^n\bot\kappa,u_i^n\bot\kappa)\right]\leq 0.\end{aligned}\tag{2.3.3}$$

Remark 2.3.7 *Noticing that $F(\cdot\top\kappa,\cdot\top\kappa)-F(\cdot\bot\kappa,\cdot\bot\kappa)$ satisfies (2.2.17) with $f(\cdot\top\kappa)-f(\cdot\bot\kappa)$ instead of f, inequality (2.3.3) clearly is a discretization of*

$$\partial_t|u-\kappa|+\partial_x\left(f(u\top\kappa)-f(u\bot\kappa)\right)\leq 0\,,$$

which is just another way of writing the entropy inequalities in Definition 2.1.1.

Proof of Proposition 2.3.6
Let $H(a,b,c)=b+\frac{\delta t}{\delta x}F(a,b)-\frac{\delta t}{\delta x}F(b,c)$. The assumptions on F show that, under (2.2.21), H is non-decreasing on $[\inf_{\mathbb{R}}u_0,\sup_{\mathbb{R}}u_0]^3$ with respect to each of its variables, and we clearly have $H(a,a,a)=a$. By (2.2.11) we have $u_i^{n+1}=H(u_{i-1}^n,u_i^n,u_{i+1}^n)$ and the monotony properties of H therefore show that, for all $\kappa\in[\inf_{\mathbb{R}}u_0,\sup_{\mathbb{R}}u_0]$, since $\cdot\leq\cdot\top\kappa$,

$$u_i^{n+1}\leq H(u_{i-1}^n\top\kappa,u_i^n\top\kappa,u_{i+1}^n\top\kappa).\tag{2.3.4}$$

On the other hand, $\kappa = H(\kappa,\kappa,\kappa)$ and therefore, still using the monotony of H,

$$\kappa \le H(u_{i-1}^n \top \kappa, u_i^n \top \kappa, u_{i+1}^n \top \kappa). \tag{2.3.5}$$

We deduce from (2.3.4) and (2.3.5) that

$$u_i^{n+1} \top \kappa \le H(u_{i-1}^n \top \kappa, u_i^n \top \kappa, u_{i+1}^n \top \kappa). \tag{2.3.6}$$

Similarly, it is easy to show that

$$u_i^{n+1} \bot \kappa \ge H(u_{i-1}^n \bot \kappa, u_i^n \bot \kappa, u_{i+1}^n \bot \kappa). \tag{2.3.7}$$

Subtracting (2.3.7) from (2.3.6), and using $\cdot \top \kappa - \cdot \bot \kappa = |\cdot - \kappa|$ and the definition of H gives (2.3.3) when $\kappa \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$. If κ does not belong to this interval, then it is either greater or lower than all the values $(u_i^n)_{n\ge 0, i\in\mathbb{Z}}$ and one can check that the left-hand side of (2.3.3) then vanishes, thanks to (2.2.11). ■

Remark 2.3.8 *This representation of the scheme as*

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n) \tag{2.3.8}$$

(with H non-decreasing with respect to each of its variables and $H(a,a,a) = a$) has other applications; for example, one can prove Proposition 2.2.3 using this expression, by simply remarking that, if $m^n = \inf_{i\in\mathbb{Z}} u_i^n$ and $M^n = \sup_{i\in\mathbb{Z}} u_i^n$,

$$m^n = H(m^n, m^n, m^n) \le H(u_{i-1}^n, u_i^n, u_{i+1}^n) \le H(M^n, M^n, M^n) = M^n.$$

2.3.3 Convergence of the scheme

It is now quite easy to prove the convergence of the scheme.

Theorem 2.3.9 (Convergence of monotone schemes) *Assume that F is a monotone numerical flux for (2.1.1) and that $u_0 \in BV(\mathbb{R})$. For $\delta t > 0$ and $\delta x > 0$, denote by $u_{\delta t,\delta x}$ the piecewise constant function equal to u_i^n on $[n\delta t, (n+1)\delta t[\times[i\delta x, (i+1)\delta x[$, where $(u_i^n)_{n\ge 0\,,\,i\in\mathbb{Z}}$ is the solution to the scheme (2.2.20). Then, as δt and δx tend to 0 while satisfying (2.2.21), $u_{\delta t,\delta x}$ converges to the entropy solution of (2.1.1) weakly-$*$ in $L^\infty(]0,\infty[\times\mathbb{R})$ and strongly in $L^p_{\rm loc}([0,\infty[\times\mathbb{R})$ for all $p < \infty$.*

Remark 2.3.10 *The condition $u_0 \in BV(\mathbb{R})$ is not mandatory, see Section 2.6.2.*

Proof of Theorem 2.3.9
By Proposition 2.2.3, Corollary 2.3.5 and Helly's lemma (compact embedding of $L^1_{\rm loc} \cap BV_{\rm loc}$ into $L^1_{\rm loc}$), we can extract a subsequence, still denoted $u_{\delta t,\delta x}$, which converges to some u weakly-$*$ in $L^\infty(]0,\infty[\times\mathbb{R})$ and strongly in $L^p_{\rm loc}([0,\infty[\times\mathbb{R})$ for all $p < \infty$; if we prove that u is the entropy solution to (2.1.1), then its uniqueness ensures that the whole sequence converges, which proves the theorem.

Since u takes its values (as $u_{\delta t,\delta x}$) between the infimum and supremum of u_0, it is enough to prove (2.1.2) for κ also between these values; let such a κ, take $\varphi \in C_c^\infty([0,\infty[\times\mathbb{R})$ non-negative and multiply each equation of (2.3.3) by $\delta t \varphi_i^n$ with $\varphi_i^n = \frac{1}{\delta t\delta x}\int_{n\delta t}^{(n+1)\delta t}\int_{i\delta x}^{(i+1)\delta x} \varphi(t,x)\,dtdx$. Summing on i and n (notice that since φ has a compact support, these sums are in fact finite) and re-indexing some parts of these sums, we obtain

$$\begin{aligned} 0 \ \ \ge\ \ & \sum_{n\ge 0}\sum_{i\in\mathbb{Z}} \delta x \left[|u_i^{n+1} - \kappa| - |u_i^n - \kappa|\right] \varphi_i^n \\ & + \sum_{n\ge 0} \delta t \sum_{i\in\mathbb{Z}} \left[F(u_i^n \top \kappa, u_{i+1}^n \top \kappa) - F(u_i^n \bot \kappa, u_{i+1}^n \bot \kappa)\right] \varphi_i^n \end{aligned}$$

$$
\begin{aligned}
&-\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}\left[F(u_{i-1}^n\top\kappa,u_i^n\top\kappa)-F(u_{i-1}^n\bot\kappa,u_i^n\bot\kappa)\right]\varphi_i^n\\
\geq\ &\sum_{n\geq 1}\delta t\sum_{i\in\mathbb{Z}}\delta x|u_i^n-\kappa|\frac{\varphi_i^{n-1}-\varphi_i^n}{\delta t}-\sum_{i\in\mathbb{Z}}\delta x|u_i^0-\kappa|\varphi_i^0\\
&+\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}\delta x\left[F(u_{i-1}^n\top\kappa,u_i^n\top\kappa)-F(u_{i-1}^n\bot\kappa,u_i^n\bot\kappa)\right]\frac{\varphi_{i-1}^n-\varphi_i^n}{\delta x}. \qquad (2.3.9)
\end{aligned}
$$

We write

$$
\begin{aligned}
F(u_{i-1}^n\top\kappa,u_i^n\top\kappa)-F(u_{i-1}^n\bot\kappa,u_i^n\bot\kappa) \ =\ & F(u_{i-1}^n\top\kappa,u_i^n\top\kappa)-F(u_i^n\top\kappa,u_i^n\top\kappa)\\
&+F(u_i^n\top\kappa,u_i^n\top\kappa)-F(u_i^n\bot\kappa,u_i^n\bot\kappa)\\
&+F(u_i^n\bot\kappa,u_i^n\bot\kappa)-F(u_{i-1}^n\bot\kappa,u_i^n\bot\kappa).
\end{aligned}
$$

The consistency property (2.2.17) of F and its Lipschitz-continuity (2.2.18) (note that $(u_i^n)_{n\geq 0\,,\,i\in\mathbb{Z}}$ and κ stay in $[\inf_{\mathbb{R}} u_0,\sup_{\mathbb{R}} u_0]$) imply, from (2.3.9) and using the regularity of φ and Proposition 2.3.1,

$$
\begin{aligned}
0\ \geq\ &\sum_{n\geq 1}\delta t\sum_{i\in\mathbb{Z}}\delta x|u_i^n-\kappa|\frac{\varphi_i^{n-1}-\varphi_i^n}{\delta t}-\sum_{i\in\mathbb{Z}}\delta x|u_i^0-\kappa|\varphi_i^0\\
&+\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}\delta x(f(u_i^n\top\kappa)-f(u_i^n\bot\kappa))\frac{\varphi_i^n-\varphi_{i+1}^n}{\delta x}+\mathcal{O}\left(\delta x\sum_{n=0}^{[T/\delta t]}\delta t\sum_{i\in\mathbb{Z}}|u_i^n-u_{i-1}^n|\right)\\
0\ \geq\ &\sum_{n\geq 1}\delta t\sum_{i\in\mathbb{Z}}\delta x|u_i^n-\kappa|\frac{\varphi_i^{n-1}-\varphi_i^n}{\delta t}-\sum_{i\in\mathbb{Z}}\delta x|u_i^0-\kappa|\varphi_i^0\\
&+\sum_{n\geq 0}\delta t\sum_{i\in\mathbb{Z}}\delta x(f(u_i^n\top\kappa)-f(u_i^n\bot\kappa))\frac{\varphi_i^n-\varphi_{i+1}^n}{\delta x}+\mathcal{O}(\delta x)\,, \qquad (2.3.10)
\end{aligned}
$$

where T is some real number such that $\mathrm{supp}(\varphi)\subset[0,T]\times\mathbb{R}$.
Define $\Phi_{\delta t,\delta x}:\mathbb{R}^+\times\mathbb{R}\to\mathbb{R}$, $\Psi_{\delta t,\delta x}:\mathbb{R}^+\times\mathbb{R}\to\mathbb{R}$, $\Theta_{\delta x}:\mathbb{R}\to\mathbb{R}$ and $U_{\delta x}^{0,\kappa}:\mathbb{R}\to\mathbb{R}$ as the piecewise constant functions

$$
\begin{aligned}
&\Phi_{\delta t,\delta x}=\frac{\varphi_i^{n-1}-\varphi_i^n}{\delta t}\quad\text{ on }[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[\,,\\
&\Psi_{\delta t,\delta x}=\frac{\varphi_i^n-\varphi_{i+1}^n}{\delta t}\quad\text{ on }[n\delta t,(n+1)\delta t[\times[i\delta x,(i+1)\delta x[\,,\\
&\Theta_{\delta x}=\varphi_i^0\quad\text{ on }[i\delta x,(i+1)\delta x[\,,\\
&U_{\delta x}^{0,\kappa}=|u_i^0-\kappa|\quad\text{ on }[i\delta x,(i+1)\delta x[\,.
\end{aligned}
$$

The regularity of φ and (2.2.2) show that, as δt and δx tend to 0,

$$
\begin{aligned}
&\Phi_{\delta t,\delta x}\to-\partial_t\varphi\text{ and }\Psi_{\delta t,\delta x}\to-\partial_x\varphi\text{ uniformly on }[0,\infty[\times\mathbb{R}\,,\\
&\Theta_{\delta x}\to\varphi(0,\cdot)\text{ uniformly on }\mathbb{R}\text{ and }U_{\delta x}^{0,\kappa}\to|u_0-\kappa|\text{ in }L^1_{\mathrm{loc}}(\mathbb{R}).
\end{aligned} \qquad (2.3.11)
$$

Note also that $\Phi_{\delta t,\delta x}$, $\Psi_{\delta t,\delta x}$ and $\Theta_{\delta x}$ vanish outside some compact sets not depending on δt or δx. Equation (2.3.10) can then be written

$$
\begin{aligned}
0\ \geq\ &\int_0^\infty\int_{\mathbb{R}}|u_{\delta t,\delta x}-\kappa|\Phi_{\delta t,\delta x}(t,x)\,dtdx-\int_{\mathbb{R}}U_{\delta x}^{0,\kappa}(x)\Theta_{\delta x}(x)\,dx\\
&+\int_0^\infty\int_{\mathbb{R}}(f(u_{\delta t,\delta x}(t,x)\top\kappa)-f(u_{\delta t,\delta x}(t,x)\bot\kappa))\Psi_{\delta t,\delta x}(t,x)\,dtdx+\mathcal{O}(\delta x).
\end{aligned}
$$

The strong convergence in $L^1_{\mathrm{loc}}([0,\infty[\times\mathbb{R})$ of $u_{\delta t,\delta x}$ and (2.3.11) allow to pass to the limit in this expression and to conclude that u indeed satisfies (2.1.2), which completes the proof. ■

Remark 2.3.11 *As it is usual in finite volume methods, the proof of convergence of the scheme relies on compactness estimates and does not require to have established the existence of a solution to the PDE; in fact, this study of convergence of the scheme gives, as a by-product, the existence of a solution to the continuous problem.*

2.4 Some numerical results

Before presenting some numerical results, let us say a few words on the practical implementation. Each time step of the scheme (2.2.20) theoretically requires to compute an infinite number of values $(u_i^{n+1})_{i\in\mathbb{Z}}$, which a computer obviously cannot do; one therefore has to compute only some of the unknowns, and this can be achieved without loss thanks to the structure of the scheme.

Equation (2.2.11) shows that u_i^{n+1} only depends on u_{i-1}^n, u_i^n and u_{i+1}^n; hence, if we are interested in the approximate solution at time $t = N\delta t$ (for some $N \geq 0$) on the space interval $[-I\delta x, I\delta x]$ (for some $I \geq 0$), i.e. by $(u_i^N)_{|i|\leq I}$, we only need $(u_i^{N-1})_{|i|\leq I+1}$, which in turn only requires $(u_i^{N-2})_{|i|\leq I+2}$, etc., down to $(u_i^0)_{|i|\leq I+N}$.

Therefore, in order to find the approximate solution at $t = N\delta t$ on $[-I\delta x, I\delta x]$, the only values we need to compute are u_i^n for $n = 0, \dots, N$ and $|i| \leq I + N - n$; in particular, we only need to know $(u_i^0)_{|i|\leq I+N}$, i.e. u_0 on $[-(I+N)\delta x, (I+N)\delta x]$, in order to compute $u_{\delta t,\delta x}(N\delta t, \cdot)$ on $[-I\delta x, I\delta x]$. This phenomenon is the discrete equivalent of the well-known "finite speed propagation property" of scalar conservation laws: this property states that the solution to (2.1.1) at time $t = T$ on $[-R, R]$ only depends on the initial data on $[-R - \mathrm{Lip}_{u_0}(f)T, R + \mathrm{Lip}_{u_0}(f)T]$ (dependency cone). In the discrete setting, if $T = N\delta t$ and $R = I\delta x$, the approximate solution at $t = T$ on $[-R, R]$ depends on the initial data on $[-R - \frac{\delta x}{\delta t}T, R + \frac{\delta x}{\delta t}T]$; owing to the CFL condition (2.2.21), in the best case scenario this means that we need u_0 on $[-R - (\mathrm{Lip}_{1,u_0}(F) + \mathrm{Lip}_{2,u_0}(F))T, R + (\mathrm{Lip}_{1,u_0}(F) + \mathrm{Lip}_{2,u_0}(F))T]$; in general, one has $\mathrm{Lip}_{i,u_0}(F) \geq \mathrm{Lip}_{u_0}(f)$ (with equality for the Godunov flux, for example): this means that the discrete dependency cone is larger than the continuous dependency cone (its slope is, at best, twice as large).

We now illustrate the behavior of the modified Lax-Friedrichs scheme (with the smallest possible D) and the Godunov scheme on the Burgers problem with a Riemann initial data

$$\left\{\begin{array}{ll} \partial_t u(t,x) + \partial_x\left(\frac{u(t,x)^2}{2}\right) = 0 & t > 0\,,\ x \in \mathbb{R}\,, \\ u_0(x) = \left\{\begin{array}{l} u_l\,, \\ u_r \end{array}\right. & \begin{array}{l} \text{if } x < 0\,, \\ \text{if } x > 0. \end{array} \end{array}\right. \tag{2.4.1}$$

We consider both the cases where the solution is a rarefaction wave (taking $u_l = -1$, $u_r = 1$) and a shock (with $u_l = 1$, $u_r = -1$), and we are interested in the solution at $t = 0.5$ on $[-1, 1]$; we have plotted the results, for some δt and δx satisfying (2.2.21) and indicated in the captions, in Figures 2.2 and 2.3.

These figures clearly show that the Godunov scheme is a slightly better than the modified Lax-Friedrichs scheme; this concurs with what we wrote at the end of Section 2.2.3, namely that the Godunov scheme is less diffusive than the modified Lax-Friedrichs scheme (and thus provides a better approximation of the solution).

2.5 Semi-linear parabolic equations

As we clearly showed during the construction of monotone schemes, the discretization of scalar conservation laws has strong links with the discretization of convective-diffusive equations. It might therefore be interesting to say a few words on the discretization of (2.1.1) when a diffusion term is added:

$$\left\{\begin{array}{ll} \partial_t u(t,x) + \partial_x(f(u(t,x))) - \nu\partial_{xx}u(t,x) = 0 & t > 0\,,\ x \in \mathbb{R}\,, \\ u(0,x) = u_0(x) & x \in \mathbb{R} \end{array}\right. \tag{2.5.1}$$

(with $\nu > 0$ and, as before, $f : \mathbb{R} \to \mathbb{R}$ locally Lipschitz continuous and $u_0 \in L^\infty(\mathbb{R})$).

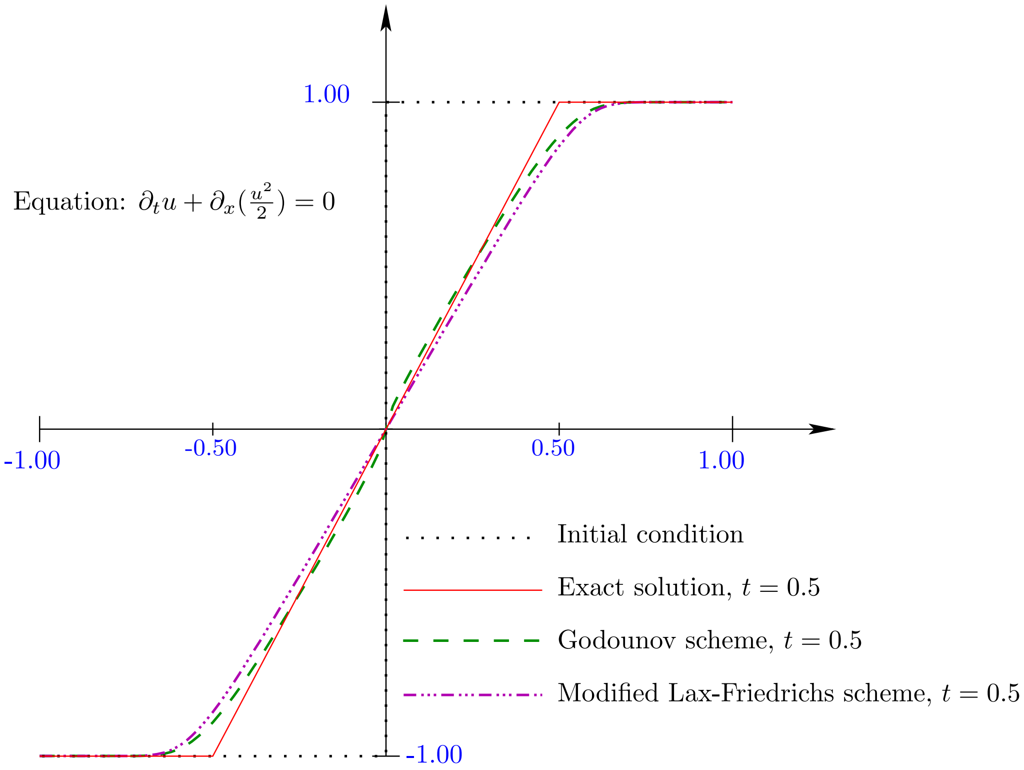


Figure 2.2: Comparison between the Godunov and the modified Lax-Friedrichs scheme for the Burgers equation $\partial_t u + \partial_x(\frac{u^2}{2}) = 0$ in the case of a rarefaction wave: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

Discretizing the diffusion term as in Section 1.3.2 and using a monotone flux for the hyperbolic term, a scheme for (2.5.1) can be written:

$$\forall n \geq 0\,, \forall i \in \mathbb{Z}\ :\ \frac{\delta x}{\delta t}(u_i^{n+1} - u_i^n) + F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n) - \nu \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\delta x} = 0\,, \qquad (2.5.2)$$

completed with the discretization (2.2.2) of the initial data.
As the pure scalar conservation law, the semi-linear parabolic equation (2.5.1) satisfies a maximum principle: the solution is bounded from below and from above by the infimum and supremum values of the initial data; we thus expect a scheme for this equation to behave the same way. Using the writing (2.2.15) for the hyperbolic terms of the equation, we easily see that (2.5.2) is

$$\begin{aligned} u_i^{n+1} &= \left(1 - \frac{\delta t}{\delta x}(a_i^n + b_i^n)\right) u_i^n + \frac{\delta t}{\delta x} a_i^n u_{i+1}^n + \frac{\delta t}{\delta x} b_i^n u_{i-1}^n + \nu \frac{\delta t}{(\delta x)^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) \\ &= \left(1 - \frac{\delta t}{\delta x}(a_i^n + b_i^n) - 2\nu \frac{\delta t}{(\delta x)^2}\right) u_i^n + \left(\frac{\delta t}{\delta x} a_i^n + \nu \frac{\delta t}{(\delta x)^2}\right) u_{i+1}^n + \left(\frac{\delta t}{\delta x} b_i^n + \nu \frac{\delta t}{(\delta x)^2}\right) u_{i-1}^n \end{aligned}$$

with a_i^n and b_i^n defined by (2.2.13) and (2.2.14). A stability condition for (2.5.2) is therefore

$$\frac{\delta t}{\delta x}(a_i^n + b_i^n) + 2\nu \frac{\delta t}{(\delta x)^2} \leq 1\,, \qquad (2.5.3)$$

$$a_i^n + \frac{\nu}{\delta x} \geq 0 \quad \text{and} \quad b_i^n + \frac{\nu}{\delta x} \geq 0. \qquad (2.5.4)$$

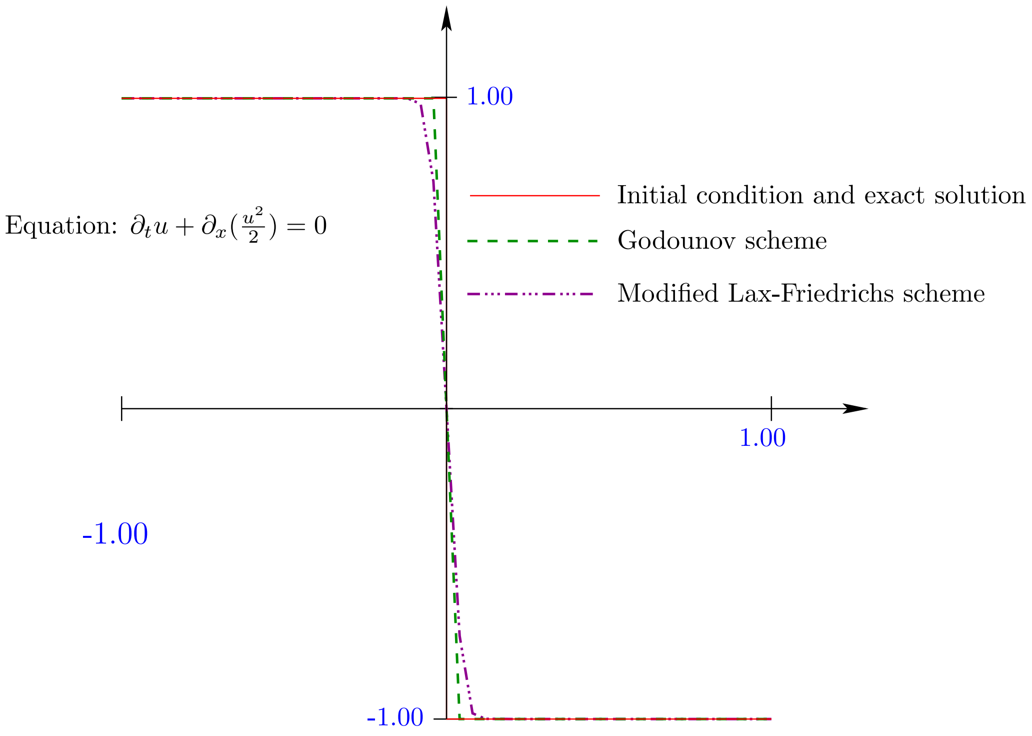


Figure 2.3: Comparison between the Godunov and the modified Lax-Friedrichs scheme for the Burgers equation $\partial_t u + \partial_x(\frac{u^2}{2}) = 0$ in the case of a shock: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

Equations (2.5.4) are always satisfied if F is a monotone flux (because $a_i^n \geq 0$ and $b_i^n \geq 0$ if $(u_i^n)_{i\in\mathbb{Z}}$ all belong to $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$), but they show that, in presence of a diffusion term, the monotony assumptions on F can be relaxed: a_i^n and b_i^n do not necessarily need to be nonnegative, since they can be compensated by the diffusion term. In particular, if

$$\nu \geq \max(\mathrm{Lip}_{1,u_0}(F), \mathrm{Lip}_{2,u_0}(F))\delta x \tag{2.5.5}$$

then no monotony is required on F in order that (2.5.4) is satisfied; we already noticed this in the linear case, and if F is the centered flux $F(a,b) = \frac{1}{2}(f(a) + f(b))$, (2.5.5) is the equivalent of the linear Peclet condition (1.3.11). In fact, one can understand from (2.5.4) exactly how to relax the monotony assumptions on F. Define the "lower and upper Lipschitz constants" of F by

$$\mathrm{Lip}_{1,u_0}^- = \sup\left\{\left(\frac{F(a,c) - F(b,c)}{a-b}\right)^- \; ; \; (a,b,c) \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^3\right\}$$

and

$$\mathrm{Lip}_{2,u_0}^+ = \sup\left\{\left(\frac{F(c,a) - F(c,b)}{a-b}\right)^+ \; ; \; (a,b,c) \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^3\right\}.$$

Since that $a_i^n \geq -\mathrm{Lip}_{2,u_0}^+(F)$ and $b_i^n \geq -\mathrm{Lip}_{1,u_0}^-(F)$ (if $(u_i^n)_{i\in\mathbb{Z}}$ all belong to $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$), we see that (2.5.4) is satisfied if we impose (2.5.5) with $\mathrm{Lip}_{1,u_0}^-(F)$ and $\mathrm{Lip}_{2,u_0}^+(F)$ instead of $\mathrm{Lip}_{1,u_0}(F)$ and $\mathrm{Lip}_{2,u_0}(F)$ (this gives a less restrictive condition, which is for example always satisfied in the case of a monotone flux since we have then $\mathrm{Lip}_{1,u_0}^-(F) = \mathrm{Lip}_{2,u_0}^+(F) = 0$).

In order to ensure (2.5.3), one basically has to impose

$$\frac{\delta t}{\delta x}(\mathrm{Lip}_{1,u_0}(F)+\mathrm{Lip}_{2,u_0}(F))+2\nu\frac{\delta t}{(\delta x)^2}\le 1 \tag{2.5.6}$$

This condition, non-linear equivalent of (1.3.10) ([4]), shows that the diffusion term imposes a more restrictive condition on the time and space steps than the hyperbolic term since it leads to a bound of the kind $\delta t\le C(\delta x)^2$ (this has already been noticed in the linear case, see Remark 1.3.7). There is however a way to avoid such a restrictive CFL condition: it consists in discretizing the diffusion term in a *implicit* way rather than an explicit one.

We write, instead of (2.5.2),

$$\forall n\ge 0\,,\forall i\in\mathbb{Z}\;:\;\frac{\delta x}{\delta t}(u_i^{n+1}-u_i^n)+F(u_i^n,u_{i+1}^n)-F(u_{i-1}^n,u_i^n)-\nu\frac{u_{i+1}^{n+1}-2u_i^{n+1}+u_{i-1}^{n+1}}{\delta x}=0. \tag{2.5.7}$$

In order to study the stability of this semi-implicit scheme (the hyperbolic term is discretized in a explicit way, the diffusive term in an implicit way), we define

$$v_i^n=u_i^n-\frac{\delta t}{\delta x}F(u_i^n,u_{i+1}^n)+\frac{\delta t}{\delta x}F(u_{i-1}^n,u_i^n) \tag{2.5.8}$$

and we notice that, under the usual hyperbolic CFL condition (2.2.21) (not involving $(\delta x)^2$), if $(u_i^n)_{i\in\mathbb{Z}}$ all belong to $[\inf_{\mathbb{R}}u_0,\sup_{\mathbb{R}}u_0]$ then the values $(v_i^n)_{i\in\mathbb{Z}}$ also belong to this interval, since these are simply the values computed by the monotone scheme defined by F for the pure hyperbolic scalar conservation law (see (2.2.11)). Assume now that $(u_i^n)_{i\in\mathbb{Z}}$ are given real numbers in $[\inf_{\mathbb{R}}u_0,\sup_{\mathbb{R}}u_0]$ and that there exists a bounded sequence $(u_i^{n+1})_{i\in\mathbb{Z}}$ satisfying (2.5.7). Then

$$u_i^{n+1}+\nu\frac{\delta t}{(\delta x)^2}(u_i^{n+1}-u_{i+1}^{n+1})+\nu\frac{\delta t}{(\delta x)^2}(u_i^{n+1}-u_{i-1}^{n+1})=v_i^n \tag{2.5.9}$$

and, taking $(i_k)_{k\ge 0}$ a sequence such that $u_{i_k}^{n+1}\to\sup_{j\in\mathbb{Z}}u_j^{n+1}$ and applying (2.5.9) to $i=i_k$, we find

$$u_{i_k}^{n+1}+\nu\frac{\delta t}{(\delta x)^2}(u_{i_k}^{n+1}-\sup_{j\in\mathbb{Z}}(u_j^{n+1}))+\nu\frac{\delta t}{(\delta x)^2}(u_{i_k}^{n+1}-\sup_{j\in\mathbb{Z}}(u_j^{n+1}))\le\sup_{j\in\mathbb{Z}}v_j^n\le\sup_{\mathbb{R}}u_0.$$

Passing then to the limit $k\to\infty$, we infer $\sup_{j\in\mathbb{Z}}u_j^{n+1}\le\sup_{\mathbb{R}}u_0$; similarly, we could show that $\inf_{j\in\mathbb{Z}}u_j^{n+1}\ge\inf_{\mathbb{R}}u_0$. This shows that the semi-implicit scheme (2.5.7) satisfies the maximum principle (and is therefore stable) under the same CFL (2.2.21) as in the absence of a diffusion term; this CFL is always less restrictive than (2.5.6), and much more so in the case of small space steps.
There however remains the question of the existence, given $(u_i^n)_{i\in\mathbb{Z}}$, of $(u_i^{n+1})_{i\in\mathbb{Z}}$ satisfying (2.5.7); on the contrary to the case of the fully explicit scheme (2.5.2), this existence is not obvious. However, once *a priori* estimates on the possible solution $(u_i^{n+1})_{i\in\mathbb{Z}}$ have been obtained, one can apply classical techniques which ensure the existence of this solution. Here, this would for example consist in cutting (2.5.7) in order to consider only a finite number of equations, say for $|i|\le I$, to notice that the preceding *a priori* estimates still holds for this finite-dimensional system and therefore ensure the existence and uniqueness of its solution, and to pass to the limit $I\to\infty$ (still using the estimates on the solution) to obtain a solution to the full system (2.5.7); it is also possible to prove that this solution is unique.

Remark 2.5.1 *As we have noticed, $(v_i^n)_{i\in\mathbb{Z}}$ defined by (2.5.8) is computed, from $(u_i^n)_{i\in\mathbb{Z}}$, by applying one time iteration of the monotone scheme for the pure hyperbolic conservation law (2.1.1). One can also notice that (2.5.9) consists in computing $(u_i^{n+1})_{i\in\mathbb{Z}}$ from $(v_i^n)_{i\in\mathbb{Z}}$ by applying one time iteration of the*

[4] The term $\frac{\delta t}{\delta x}(\mathrm{Lip}_{1,u_0}(F)+\mathrm{Lip}_{2,u_0})$ does not appear in (1.3.10) because, in the linear case, one has $a_i^n+b_i^n=0$.

scheme for the pure diffusion equation (i.e. (2.5.1) with $f = 0$). Each time iteration of the semi-implicit scheme (2.5.7) for

$$\partial_t u + \partial_x(f(u)) - \nu \partial_{xx} u = 0 \tag{2.5.10}$$

therefore appears as the successive application of one time iteration of a scheme for

$$\partial_t u + \partial_x(f(u)) = 0$$

and one time iteration of a scheme for

$$\partial_t u - \nu \partial_{xx} u = 0.$$

This technique, which consists in cutting the evolution of (2.5.10) in two equations, is known in numerical analysis as the splitting method.

2.6 Two concluding remarks

2.6.1 Implicit discretization of the fluxes

As in the linear case, another natural choice of flux discretization in (2.2.1) is to use an implicit form, replacing (2.2.10) with

$$f_i^n = F(u_{i-1}^{n+1}, u_i^{n+1}). \tag{2.6.1}$$

One can then prove that, if F is a monotone numerical flux, the resulting scheme [(2.2.1),(2.2.2),(2.6.1)] is L^∞ stable without any CFL assumption. Indeed, it leads to

$$\forall n \geq 0\,,\ \forall i \in \mathbb{Z}\ :\ u_i^n = u_i^{n+1} + \frac{\delta t}{\delta x} F(u_i^{n+1}, u_{i+1}^{n+1}) - \frac{\delta t}{\delta x} F(u_{i-1}^{n+1}, u_i^{n+1}) =: G(u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^{n+1}) \tag{2.6.2}$$

with, on $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^3$, G non-increasing with respect to its first and third variables, and non-decreasing with respect to its second variable and $G(a,a,a) = a$; assuming that there exists a solution $(u_i^{n+1})_{i\in\mathbb{Z}} \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^{\mathbb{Z}}$ to (2.6.2) and taking i such that $u_i^{n+1} = \sup_{j\in\mathbb{Z}} u_j^{n+1}$ (or, if such an i does not exist, a sequence $(i_k)_{k\geq 0}$ such that $u_{i_k}^{n+1} \to \sup_{j\in\mathbb{Z}} u_j^{n+1}$ as in Section 2.5), we have

$$\sup_{j\in\mathbb{Z}} u_j^n \geq G(u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^{n+1}) \geq G(u_i^{n+1}, u_i^{n+1}, u_i^{n+1}) = u_i^{n+1} = \sup_{j\in\mathbb{Z}} u_j^{n+1}.$$

Similarly, we would show that $\inf_{j\in\mathbb{Z}} u_j^n \leq \inf_{j\in\mathbb{Z}} u_j^{n+1}$. These *a priori* estimates allow to prove the existence of a solution $(u_i^{n+1})_{i\in\mathbb{Z}} \in [\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]^{\mathbb{Z}}$ to the scheme (2.6.2).

Implicit schemes for scalar conservation laws (2.1.1) are however not as used as for diffusion equations (2.5.1), because the resulting system (2.6.2) to solve is non-linear, and the study of MUSCL methods (see Chapter 3) for the implicit discretization is not obvious.

2.6.2 Convergence without BV estimates

We proved the convergence of the monotone scheme for (2.1.1) under the assumption that the initial data has a bounded variation (see Theorem 2.3.9); the BV assumption has been used to obtain the compactness of the approximate solution, and to prove that the error term in (2.3.10) tends to 0 with the space step.

It is however possible to prove the convergence of the scheme without any assumption on u_0 besides the fact that it belongs to $L^\infty(\mathbb{R})$. An idea is, instead of trying to prove an *a priori* compactness property on the approximate solution, to use Young's measure theory in order to pass to the limit (in a "non-linear weak-$*$" sense) in the non-linear terms of the entropy inequalities (2.3.3); the resulting limit is no longer a function of (t,x), but a Young measure (roughly speaking, a family of probability measures $(\nu_{t,x})_{t,x}$ on $\mathbb{R}$), which can be represented by its repartition function (a function of (t,x,α), where α is an additional variable) called “entropy process solution”. One then shows a strong uniqueness property for this entropy process solution, which proves that it does not depend on α and is therefore a classical entropy solution (a function of (t, x)); this gives, as an *a posteriori* by-product, the strong convergence of the approximate solutions toward this entropy solution.

Notice however that, even if one does not need a BV estimate on the approximate solution to obtain a non-linear weak-$*$ compactness property on it, some kind of “weak BV estimate” is required in order to control the error term in (2.3.10). This estimate is written

$$\sum_{i\in\mathbb{Z}} |u_i^n - u_{i-1}^n| \leq \frac{C}{\sqrt{\delta x}}$$

(compare with (2.3.2)); the error term in (2.3.10) is then not $\mathcal{O}(\delta x)$ but $\mathcal{O}(\sqrt{\delta x})$ and therefore still vanishes as $\delta x \to 0$.

Chapter 3

MUSCL methods

3.1 Position of the problem, principle of MUSCL schemes

The finite volume method presented in Chapter 2 approximates the solution with functions which are constant on each space cell; in particular, the numerical fluxes computed at an interface $x = i\delta x$ uses the two values u_{i-1}^n and u_i^n inside the neighboring cells: if one (quite naturally) considers these values as pointwise approximations of the solution at the center of each cell, this means that the interface values are computed using values at a distance $\delta x/2$ of the said interface. The order of the resulting scheme is therefore not very high (the consistency error on the fluxes is at best a $\mathcal{O}(\delta x)$) and, as can be seen in Figure 2.2 and 2.3, the resulting approximations, though correct, are not very good, especially near the points where the exact solution is not smooth. We would like to present here some methods which allow to increase the quality of monotone schemes.

The MUSCL methods (Monotone Upwind Scheme for Conservation Laws) consist, instead of considering piecewise constant approximation of the solution, in using piecewise linear (discontinuous) approximations. u_i^n is then considered as the approximate value at the center $\frac{i\delta x+(i+1)\delta x}{2}$ of the space cell $[i\delta x,(i+1)\delta x[$ and slopes $(p_i^n)_{i\in\mathbb{Z}}$ are computed inside each cell, which allows to obtain approximate values $(\widetilde{u}_i^n)^- = u_{i-1}^n + p_{i-1}^n\frac{\delta x}{2}$ and $(\widetilde{u}_i^n)^+ = u_i^n - p_i^n\frac{\delta x}{2}$ of the solution on the left and right of each interface $x = i\delta x$ (see Figure 3.1). These values are (hopefully) better approximations of u at $x = i\delta x$ than u_{i-1}^n and u_i^n, and can then be used in (2.2.10) to compute the approximate fluxes:

$$f_i^n = F((\widetilde{u}_i^n)^-,(\widetilde{u}_i^n)^+) = F\left(u_{i-1}^n + p_{i-1}^n\frac{\delta x}{2}, u_i^n - p_i^n\frac{\delta x}{2}\right). \tag{3.1.1}$$

The resulting scheme [(2.2.1),(3.1.1)] is

$$u_i^{n+1} = u_i^n + \frac{\delta t}{\delta x}F\left(u_{i-1}^n + p_{i-1}^n\frac{\delta x}{2}, u_i^n - p_i^n\frac{\delta x}{2}\right) - \frac{\delta t}{\delta x}F\left(u_i^n + p_i^n\frac{\delta x}{2}, u_{i+1}^n - p_{i+1}^n\frac{\delta x}{2}\right), \tag{3.1.2}$$

completed with the discretization of the initial condition (2.2.2). The only remaining task is to find a way to compute the fluxes p_i^n so that (3.1.2) gives rise to a stable scheme.

3.2 General stability and entropy lemmas

Let us first state two general lemmas for schemes written under a slightly more general form than (2.3.8). The first lemma gives a very general description of stable schemes, and the second lemma shows that such schemes satisfy the entropy inequalities.

Lemma 3.2.1 (General stability result) *Let $u = (u_i)_{i\in\mathbb{Z}}$ and $v = (v_i)_{i\in\mathbb{Z}}$ be two bounded sequences, $A = \inf_{i\in\mathbb{Z}} u_i$ and $B = \sup_{i\in\mathbb{Z}} u_i$. The following properties are equivalent:*

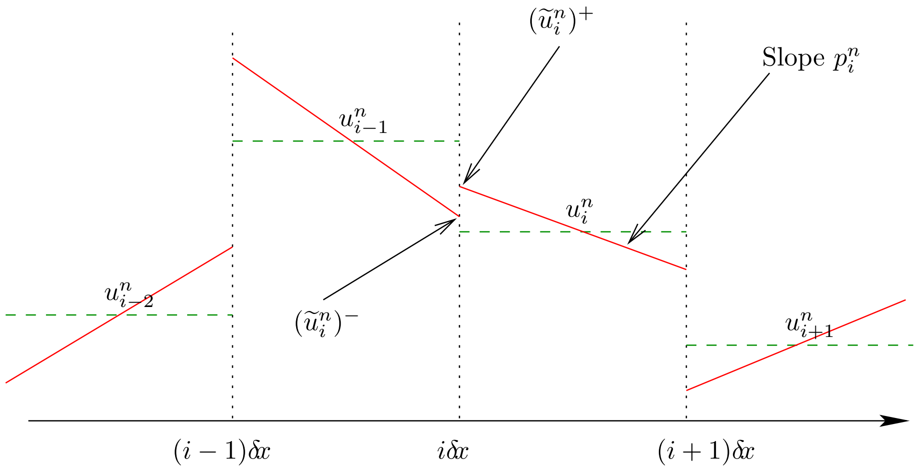


Figure 3.1: Piecewise constant and piecewise linear approximate solutions, slopes inside the cells and approximate values on either side of the interfaces.

1. *There exists* $H_{u,v} : [A,B]^3 \times \mathbb{Z} \to \mathbb{R}$ *satisfying*

$$\begin{array}{l} H_{u,v} \textit{ is non-decreasing with respect to each of its first three variables and} \\ \forall r \in [A,B]\,,\ \forall i \in \mathbb{Z}\,,\ H_{u,v}(r,r,r,i) = r \end{array} \tag{3.2.1}$$

and such that, for all $i \in \mathbb{Z}$, $v_i = H_{u,v}(u_{i-1}, u_i, u_{i+1}, i)$.

2. *For all* $i \in \mathbb{Z}$, $v_i \in [\min(u_{i-1}, u_i, u_{i+1}), \max(u_{i-1}, u_i, u_{i+1})]$.

3. *For all* $i \in \mathbb{Z}$, v_i *is a convex combination of* u_{i-1}, u_i *and* u_{i+1}.

Proof of Lemma 3.2.1
Assume that Item 1 holds, let $i \in \mathbb{Z}$ and denote $m_i = \min(u_{i-1}, u_i, u_{i+1})$ and $M_i = \max(u_{i-1}, u_i, u_{i+1})$. Since $H_{u,v}(\cdot,\cdot,\cdot,i)$ is non-decreasing with respect to each of its variables, we have

$$H_{u,v}(m_i, m_i, m_i, i) \leq H_{u,v}(u_{i-1}, u_i, u_{i+1}, i) = v_i \leq H_{u,v}(M_i, M_i, M_i, i)$$

and (3.2.1) gives $m_i \leq v_i \leq M_i$, that is to say Item 2.
If Item 2 is satisfied, then Item 3 is obvious: taking $(j,k) \in \{i-1, i, i+1\}$ such that $u_j = \min(u_{i-1}, u_i, u_{i+1})$ and $u_k = \max(u_{i-1}, u_i, u_{i+1})$, v_i is in fact a convex combination of u_j and u_k.
Finally, if Item 3 holds then, for all $i \in \mathbb{Z}$ there exists $a_i(u,v)$, $b_i(u,v)$ and $c_i(u,v)$ non-negative with sum equal to 1 such that $v_i = a_i(u,v)u_{i-1} + b_i(u,v)u_i + c_i(u,v)_{i+1}$, and we then let $H_{u,v}(\alpha, \beta, \gamma, i) = a_i(u,v)\alpha + b_i(u,v)\beta + c_i(u,v)\gamma$ to get Item 1. ■

Lemma 3.2.2 (Discrete entropy inequalities) *If* $u = (u_i)_{i\in\mathbb{Z}}$ *and* $v = (v_i)_{i\in\mathbb{Z}}$ *are bounded sequences which satisfy Item 1 in Lemma 3.2.1 then, for all* $\kappa \in \mathbb{R}$ *and all* $i \in \mathbb{Z}$,

$$v_i \top \kappa \leq H_{u,v}(u_{i-1} \top \kappa, u_i \top \kappa, u_{i+1} \top \kappa, i)$$

and

$$v_i \bot \kappa \geq H_{u,v}(u_{i-1} \bot \kappa, u_i \bot \kappa, u_{i+1} \bot \kappa, i).$$

Proof of Lemma 3.2.2
The proof is the same as the proof of Proposition 2.3.6: (3.2.1) gives

$$v_i = H_{u,v}(u_{i-1}, u_i, u_{i+1}, i) \le H_{u,v}(u_{i-1}\top\kappa, u_i\top\kappa, u_{i+1}\top\kappa, i)$$

and

$$\kappa = H_{u,v}(\kappa, \kappa, \kappa, i) \le H_{u,v}(u_{i-1}\top\kappa, u_i\top\kappa, u_{i+1}\top\kappa, i).$$

Taking the supremum of these two inequalities gives the first inequality in the lemma, and a similar reasoning allows to prove the second inequality. ■

3.3 Example of a MUSCL scheme

We have to find ways to compute slopes p_i^n, using $(u_j^n)_{j\in\mathbb{Z}}$, which ensure that the solution to (3.1.2) satisfies the maximum principle. It is quite simple to compute slopes inside the cells, taking for example $p_i^n = \frac{u_{i+1}^n - u_{i-1}^n}{2\delta x}$, but such simple choices usually do not lead to stable schemes; one can quite easily understands why by looking at the example in Figure 3.2 (in which the slope inside the cell is precisely computed using the two values on the neighboring cells).

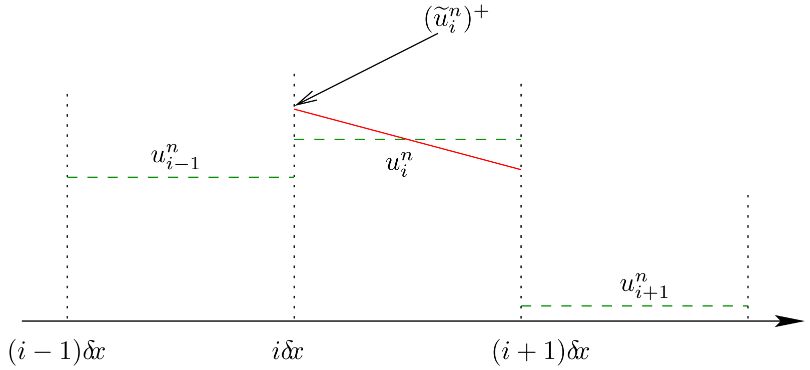


Figure 3.2: Example of “bad choices” of slopes in the MUSCL method.

We can see in this figure that the approximated value $(\widetilde{u}_i^n)^+$ computed on the right of $x = i\delta x$ is outside $[\min(u_{i-1}^n, u_i^n, u_{i+1}^n), \max(u_{i-1}^n, u_i^n, u_{i+1}^n)]$; this means that u_i^{n+1} from (3.1.2) also has chances to be out of this interval, which is clearly not a good omen for the stability of the scheme (and one can indeed confirm, implementing this scheme, that it is not stable and explodes, as the centered scheme explodes for linear transport equations).

In choosing the slopes, one therefore has to be cautious not to create interface values $(\widetilde{u}_i^n)^\pm$ outside the range of the values in the cells on either side of the interface. A proper choice of the slopes consists in defining $\widetilde{p}_i^n = \frac{u_{i+1}^n - u_{i-1}^n}{2\delta x}$ and in taking $\alpha_i^n \in [0,1]$ the largest possible number such that $p_i^n = \alpha_i^n \widetilde{p}_i^n$ is an acceptable slope in the sense that $(\widetilde{u}_i^n)^+ \in [\min(u_{i-1}^n, u_i^n), \max(u_{i-1}^n, u_i^n)]$ and $(\widetilde{u}_{i+1}^n)^- \in [\min(u_i^n, u_{i+1}^n), \max(u_i^n, u_{i+1}^n)]$. It is easy to see that this comes down to taking

$$p_i^n = \mathrm{minmod}\left(\frac{u_{i+1}^n - u_i^n}{\delta x}, \frac{u_i^n - u_{i-1}^n}{\delta x}\right) \tag{3.3.1}$$

where

$$\mathrm{minmod}(a,b) = \begin{cases} \mathrm{sgn}(a)\min(|a|,|b|) & \text{if } a \text{ and } b \text{ have the same sign,} \\ 0 & \text{otherwise,} \end{cases}$$

i.e. p_i^n vanishes if the two slopes computed with (u_{i-1}^n, u_i^n) and (u_i^n, u_{i+1}^n) have different signs, and is the smallest (in absolute value) possible slope between those two if they have the same sign.

Let us look at the stability and entropy inequalities for the scheme [(3.1.2),(3.3.1),(2.2.2)] thus obtained, beginning with a few simple properties.

Lemma 3.3.1 *Let $[a, b]$ be the set of numbers between a and b, whatever the order of a and b, and $\mathbf{1}_A$ the characteristic function of $A \subset \mathbb{R}$. The following properties hold:*

$$(a, b) \to \mathrm{minmod}(a, b) \textit{ is non-decreasing with respect to each of its variables,} \tag{3.3.2}$$

$$\textit{Defining } m_i^n = \tfrac{u_{i-1}^n + u_{i+1}^n}{2}, \textit{ we have } p_i^n \tfrac{\delta x}{2} = \mathbf{1}_{[u_{i-1}^n, m_i^n[}(u_i^n) \tfrac{u_i^n - u_{i-1}^n}{2} + \mathbf{1}_{[m_i^n, u_{i+1}^n]}(u_i^n) \tfrac{u_{i+1}^n - u_i^n}{2}, \tag{3.3.3}$$

$$\begin{array}{c} (u_{i-1}^n, u_i^n, u_{i+1}^n) \to u_i^n + p_i^n \frac{\delta x}{2} \textit{ is non-increasing with respect to } u_{i-1}^n, \\ \textit{non-decreasing with respect to } u_i^n \textit{ and } u_{i+1}^n \textit{ and:} \\ a) \textit{ Lipschitz-continuous with constant } 1.5 \textit{ with respect to } u_i^n \textit{ if } u_i^n \in [u_{i-1}^n, m_i^n], \\ b) \textit{ Lipschitz-continuous with constant } 0.5 \textit{ with respect to } u_i^n \textit{ if } u_i^n \in [m_i^n, u_{i+1}^n], \\ c) \textit{ Lipschitz-continuous with constant } 1 \textit{ with respect to } u_i^n \textit{ elsewhere,} \end{array} \tag{3.3.4}$$

$$\begin{array}{c} (u_{i-1}^n, u_i^n, u_{i+1}^n) \to u_i^n - p_i^n \frac{\delta x}{2} \textit{ is non-increasing with respect to } u_{i+1}^n, \\ \textit{non-decreasing with respect to } u_{i-1}^n \textit{ and } u_i^n \textit{ and:} \\ a) \textit{ Lipschitz-continuous with constant } 0.5 \textit{ with respect to } u_i^n \textit{ if } u_i^n \in [u_{i-1}^n, m_i^n], \\ b) \textit{ Lipschitz-continuous with constant } 1.5 \textit{ with respect to } u_i^n \textit{ if } u_i^n \in [m_i^n, u_{i+1}^n], \\ c) \textit{ Lipschitz-continuous with constant } 1 \textit{ with respect to } u_i^n \textit{ elsewhere.} \end{array} \tag{3.3.5}$$

Proof of Lemma 3.3.1
We deduce (3.3.2) from the symmetry of minmod and a study of two cases: if b is non-negative, $\mathrm{minmod}(a, b) = 0$ for $a \leq 0$, $= a$ for $0 \leq a \leq b$ and $= b$ for $a > b$ and, if b is non-positive, $\mathrm{minmod}(a, b) = 0$ for $a \geq 0$, $= a$ for $b \leq a \leq 0$ and $= b$ for $b < a$.
Property (3.3.3) can be checked by studying all possible situations: $u_i^n \notin [u_{i-1}^n, u_{i+1}^n]$ (the two slopes used to define p_i^n then have opposite signs, and $p_i^n = 0$), $u_i^n \in [u_{i-1}^n, m_i^n[$ (we have then $|u_i^n - u_{i-1}^n| \leq |u_{i+1}^n - u_i^n|$) and $u_i^n \in [m_i^n, u_{i+1}^n]$ (in which case $|u_i^n - u_{i-1}^n| \geq |u_{i+1}^n - u_i^n|$).
Both (3.3.4) and (3.3.5) are consequences of (3.3.1) and (3.3.2) (which show that p_i^n is non-decreasing with respect to u_{i+1}^n and non-increasing with respect to u_{i-1}^n) and of (3.3.3) (for the properties concerning u_i^n). ■

Equation (3.1.2) can be written

$$u_i^{n+1} = u_i^n + \frac{\delta t}{\delta x} F\left(c_1^i(u_{i-1}^n, u_i^n), d_1^i(u_i^n, u_{i+1}^n)\right) - \frac{\delta t}{\delta x} F\left(d_2^i(u_i^n, u_{i-1}^n), c_2^i(u_i^n, u_{i+1}^n)\right) \tag{3.3.6}$$

where

$$c_1^i(u_{i-1}^n, u_i^n) = u_{i-1}^n + p_{i-1}^n \frac{\delta x}{2}, \qquad c_2^i(u_i^n, u_{i+1}^n) = u_{i+1}^n - p_{i+1}^n \frac{\delta x}{2},$$
$$d_1^i(u_i^n, u_{i+1}^n) = u_i^n - p_i^n \frac{\delta x}{2} \quad \text{and} \quad d_2^i(u_i^n, u_{i-1}^n) = u_i^n + p_i^n \frac{\delta x}{2}$$

depend on other $(u_j^n)_{j \in \mathbb{Z}}$ than the ones explicitly stated, but satisfy the following properties (whatever the values of these other $(u_j^n)_{j \in \mathbb{Z}}$):

- c_1^i and c_2^i are non-decreasing with respect to their two variables,
- d_1^i and d_2^i are non-decreasing with respect to their first variable and non-increasing with respect to their second variable,

- for all $(u_{i-1}^n, u_i^n, u_{i+1}^n)$,

  a) d_1^i and d_2^i are Lipschitz-continuous with respective constants 0.5 and 1.5 with respect to u_i^n if $u_i^n \in [u_{i-1}^n, m_i^n]$,

  b) d_1^i and d_2^i are Lipschitz-continuous with respective constant 1.5 and 0.5 with respect to u_i^n if $u_i^n \in [m_i^n, u_{i+1}^n]$,

  c) d_1^i and d_2^i are Lipschitz-continuous with constant 1 with respect to u_i^n elsewhere,

- $c_1^i(s,s) = c_2^i(s,s) = d_1^i(s,s) = d_2^i(s,s) = s$ for all $s \in \mathbb{R}$

(this last property comes from $\mathrm{minmod}(0,r) = \mathrm{minmod}(r,0) = 0$ for all $r \in \mathbb{R}$). From these remarks and the monotony (2.2.19) of F, we see that the right-hand side of (3.3.6) is non-decreasing with respect to u_i^n, when all values $(u_j^n)_{j\in\mathbb{Z}}$ are inside $[\inf_{\mathbb{R}} u_0, \sup_{\mathbb{R}} u_0]$, as soon as

$$\frac{\delta t}{\delta x} \max \Big[0.5\mathrm{Lip}_{2,u_0}(F) + 1.5\mathrm{Lip}_{1,u_0}(F)\,;\, 1.5\mathrm{Lip}_{2,u_0}(F) + 0.5\mathrm{Lip}_{1,u_0}(F)\,; \\ \mathrm{Lip}_{2,u_0}(F) + \mathrm{Lip}_{1,u_0}(F) \Big] \quad \le \quad 1. \qquad (3.3.7)$$

Under this assumption, the monotony and consistency properties of F, c_1^i, c_2^i, d_1^i and d_2^i thus prove that (3.3.6) can be written

$$u_i^{n+1} = H_{u^n,u^{n+1}}(u_{i-1}^n, u_i^n, u_{i+1}^n, i)$$

with $u^{n+1} = (u_j^{n+1})_{j\in\mathbb{Z}}$, $u^n = (u_j^n)_{j\in\mathbb{Z}}$ and $H_{u^n,u^{n+1}}$ satisfying (3.2.1) (the first three variables of $H_{u^n,u^{n+1}}$ are the u_{i-1}^n, u_i^n, u_{i+1}^n explicitly appearing in the right-hand side of (3.3.6)). The L^∞ stability (in fact the maximum principle) and the entropy inequalities for the MUSCL scheme [(3.1.2),(3.3.1),(2.2.2)] follow then from Lemmas 3.2.1 and 3.2.2.

Let us conclude by noting that (3.3.7) is the CFL condition associated with the 5-points ([1]) MUSCL method [(3.1.2),(3.3.1),(2.2.2)], and that it is equivalent to

$$\frac{\delta t}{\delta x} \max \Big[1.5\mathrm{Lip}_{1,u_0}(F) + 0.5\mathrm{Lip}_{2,u_0}(F)\,;\, 0.5\mathrm{Lip}_{1,u_0}(F) + 1.5\mathrm{Lip}_{2,u_0}(F) \Big] \le 1. \qquad (3.3.8)$$

This condition for the MUSCL implementation of the chosen monotone scheme is slightly more demanding than the initial CFL (2.2.21), but not very much (it is the same if $\mathrm{Lip}_{1,u_0}(F) = \mathrm{Lip}_{2,u_0}(F)$).

Remark 3.3.2 *One can also increase the number of unknowns used to compute each slope, taking for example*

$$p_i^n = \mathrm{minmod}\left(\frac{u_{i+2}^n - u_i^n}{2\delta x}, \frac{u_{i+1}^n - u_i^n}{\delta x}, \frac{u_i^n - u_{i-1}^n}{\delta x}, \frac{u_i^n - u_{i-2}^n}{2\delta x} \right)$$

(we define $\mathrm{minmod}(a,b,c,d) = \mathrm{sgn}(a)\min(|a|,|b|,|c|,|d|)$ *if a, b, c and d all have the same sign, and* $\mathrm{minmod}(a,b,c,d) = 0$ *otherwise). This choice leads to a 7-points scheme, and one can check that its stability condition is the same as for the 5-points scheme, that is to say (3.3.8). However, in many practical situations, the 7-points scheme appears more diffusive than the 5-points scheme (see Figure 3.3 in the next section) and is therefore less interesting.*

3.4 Numerical results

We noticed in Section 2.4 that, to obtain the approximate solution at $t = N\delta t$ on $[-I\delta x, I\delta x]$ using a classical (3-points) monotone scheme, one has to compute all the approximate values u_i^n for $n = 0, \ldots, N$ and $|i| \le I + N - n$; this was due to the fact that the computation of u_i^{n+1} required u_{i-1}^n, u_i^n and u_{i+1}^n.

[1] The term "5-points" refers to the fact that [(3.1.2),(3.3.1),(2.2.2)] gives a relation involving u_i^{n+1} and five values at time $t = n\delta t$: u_{i-2}^n, u_{i-1}^n, u_i^n, u_{i+1}^n and u_{i+2}^n (the values corresponding to $i-2$ and $i+2$ are needed to compute the slopes p_{i-1}^n and p_{i+1}^n).

The 5-points MUSCL scheme requires two additional unknowns, namely u_{i-2}^n and u_{i+2}^n, for the same computation; hence, in order to obtain the approximate solution at $t = N\delta t$ on $[-I\delta x, I\delta x]$ using the 5-points scheme, we have to compute u_i^n for $n = 0, \ldots, N$ and $|i| \leq I + 2N - 2n$: the discrete dependency cone is therefore larger than for the 3-points scheme (this was expected...).

Let us now give some numerical results involving MUSCL schemes. The first result, presented in Figure 3.3, compares the 5-points and 7-points MUSCL schemes (using a Godunov numerical flux) in the simple case of a linear transport of an initial discontinuity; this result confirms what we said in Remark 3.3.2: the 7-points scheme is slightly more diffusive than the 5-points scheme. Hence, hereafter, all the numerical MUSCL results we present are obtained using the 5-points scheme.

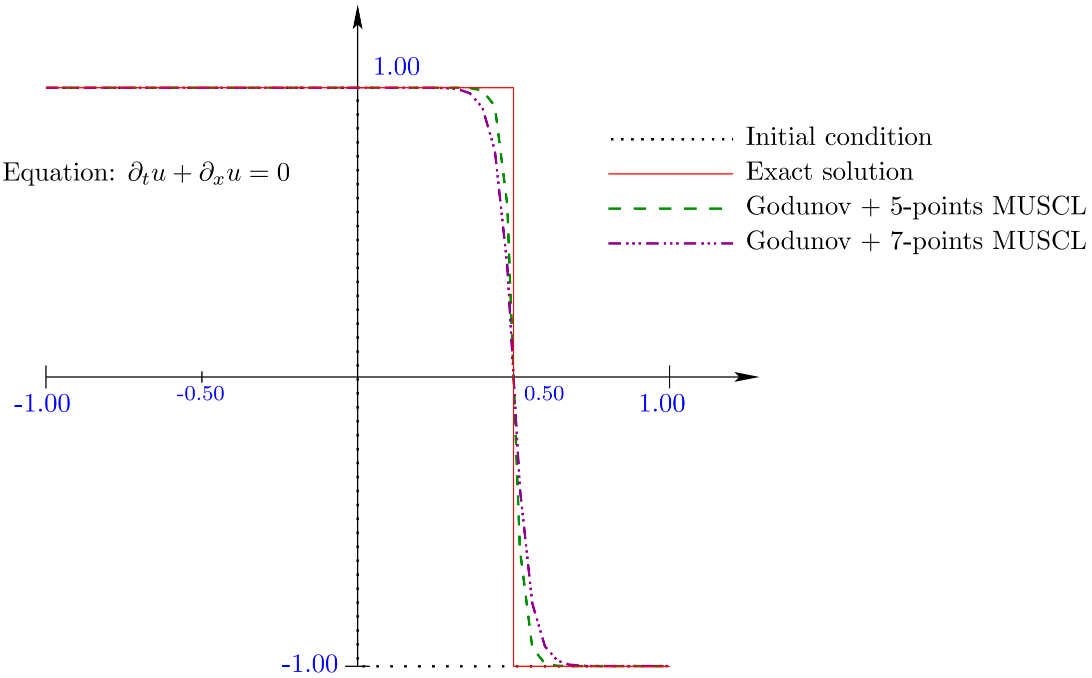


Figure 3.3: Comparison between the 5-points and 7-points MUSCL methods on the Godunov scheme for a linear transport equation $\partial_t u + \partial_x u = 0$ and a discontinuous initial data: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

In Figures 3.4 and 3.5, we compare the pure Godunov scheme and its MUSCL modification in two situations (linear transport of a discontinuous initial data and a rarefaction wave solution to the Burgers equation); these results show the remarkable efficiency of the MUSCL method in reducing the numerical diffusion.

The case of a shock wave is illustrated in Figures 3.6 and 3.7. For the equation and initial condition considered in Figure 3.6, the Godunov numerical flux is already exact (up to the discretization of the initial data, see Figure 2.3), so we applied the modified Lax-Friedrichs numerical flux; the Godunov scheme is no longer “exact” in the situation presented in Figure 3.7, which allows for a comparison with its MUSCL modification. In both these situations, the reduction of numerical diffusion by the MUSCL technique is perhaps a bit less astonishing than in the previous tests, but it is nevertheless perceptible.

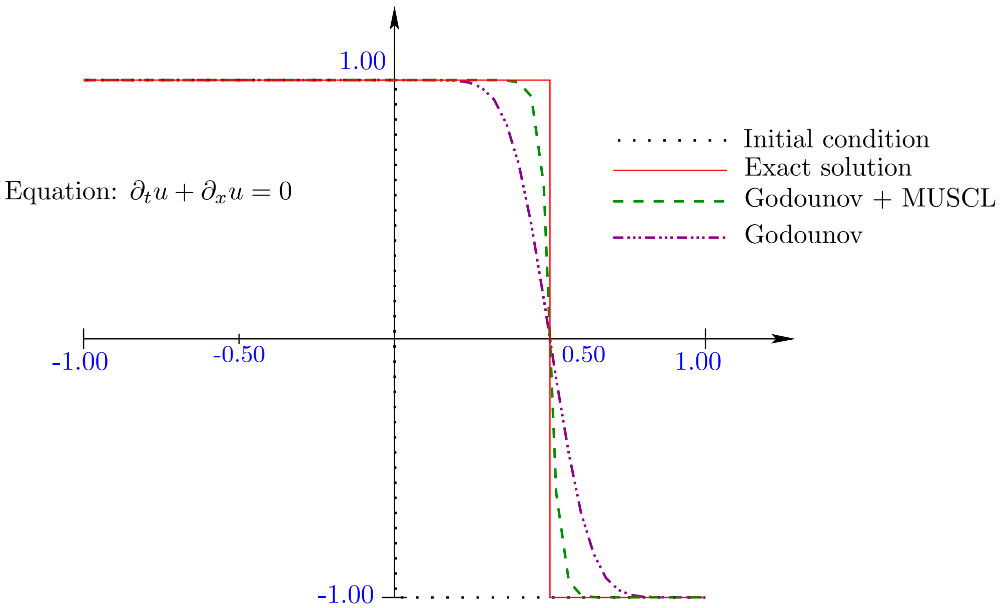


Figure 3.4: Effect of the MUSCL method on the Godunov scheme for the linear transport equation $\partial_t u + \partial_x u = 0$ and a discontinuous initial data: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

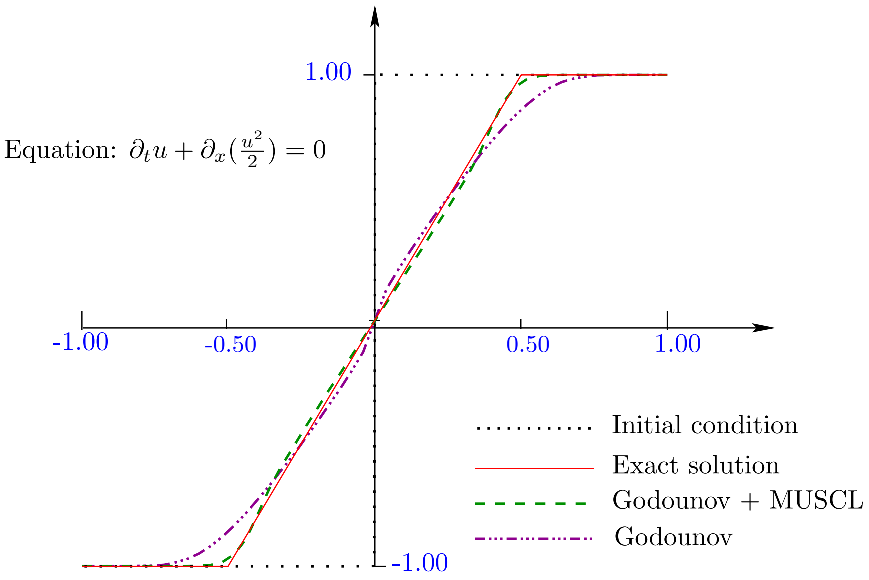


Figure 3.5: Effect of the MUSCL method on the Godunov scheme for the Burgers equation $\partial_t u + \partial_x (\frac{u^2}{2}) = 0$ in the case of a rarefaction wave: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

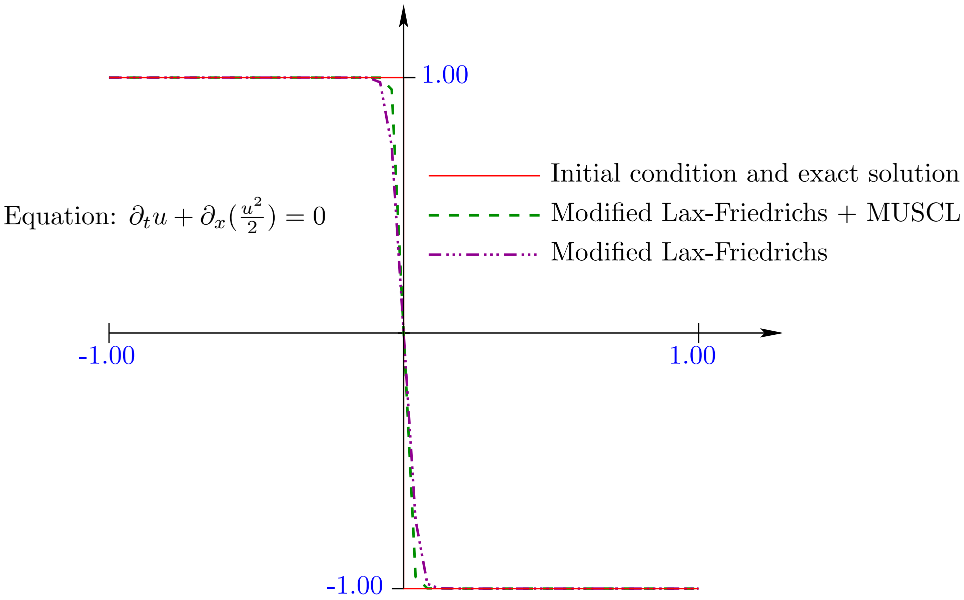


Figure 3.6: Effect of the MUSCL method on the modified Lax-Friedrichs scheme for the Burgers equation $\partial_t u + \partial_x(\frac{u^2}{2}) = 0$ in the case of a shock: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

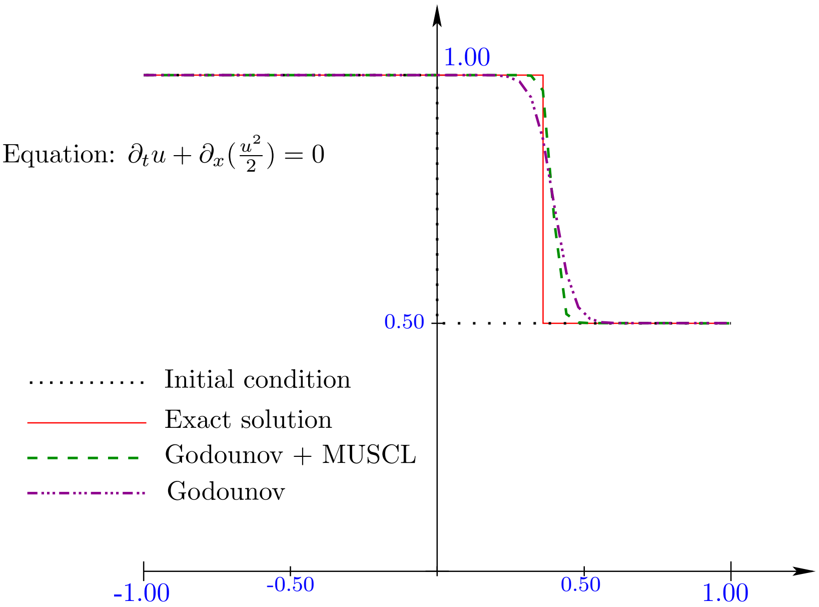


Figure 3.7: Effect of the MUSCL method on the Godunov scheme for the Burgers equation $\partial_t u + \partial_x(\frac{u^2}{2}) = 0$ in the case of a shock: exact and approximate solutions at $t = 0.5$ with $\delta t = 0.02$ and $\delta x = 0.04$.

Bibliography

[1] Chainais-Hillairet C., PhD Thesis, ENS Lyon, 1998.

[2] Clain S. and Clauzon V., *The Multi-Slope MUSCL Method*, Finite volumes for complex applications, V (Aussois, 2008), 297–304, Hermes Sci. Publ., Paris, 2008.

[3] Droniou J., *A numerical method for fractal conservation laws*, submitted for publication.

[4] Depres B. and Dubois F., Systèmes hyperboliques de lois de conservation, Editions de l'Ecole Polytechnique, 2005.

[5] Eymard R., Gallouët T., Herbin R., Finite volume methods, *Handbook of numerical analysis, Vol. VII, North-Holland, Amsterdam* (2000), 713–1020.

[6] Godlewski E., Raviart P.-A., Hyperbolic systems of conservation laws, *Mathématiques & Applications, 3/4, Ellipses, Paris* (1991), 252 pp.

[7] Vovelle J., *Convergence of finite volume monotone schemes for scalar conservation laws on bounded domains*, Numer. Math. **90** (2002), no. 3, 563–596.

Comments

Most of Chapters 1 and 2 of this document are a simplified version of the theory developed in [5]; we tried and make a more detailed and lengthy presentation, illustrated with several numerical results, in order that the reader with little or no background in numerical analysis can grasp both the theoretical basis on finite volume schemes for scalar conservation laws, as well as some understanding of their qualitative behaviour. The reader interested in delving into more complex details (for example the proof of existence of solutions to implicit schemes, the proof of convergence without BV estimates mentioned in Section 2.6.2 or the multidimensional versions of the schemes) should look into this reference.

As we showed in Section 2.5, the discretization of hyperbolic terms in PDE can be mixed with the discretization of other terms, such as diffusive terms. A more detailed presentation of this section can be found in [3], in which the Laplacian operator $-\nu\partial_{xx}u$ is replaced by a Lévy opérator (a fraction of the Laplacian); the situation is therefore a little bit more complex, but in many ways similar to the case quickly described in Section 2.5 (which can therefore serve as an introduction to [3]).

We only considered scalar conservation laws on the whole space, but two natural situations can then be considered: bounded domains, and systems of conservation laws. The theory on the corresponding PDEs is much more complex than in the scalar case, and this complexity finds some echo in the numerical approximation of these problems. A presentation and study of finite volume schemes for scalar conservation laws on bounded (multidimensional) domains can be found in [7], and [4] gives a good basis on the numerical approximation of systems of conservation laws. Of course, both these situations re-use and adapt the basic ideas presented here for scalar equations on the whole domain; as this is very often the case in Mathematics, the understanding of complex situations comes from the understanding of simpler cases (even if they appear only academic).

The basic study of generic multi-dimensional MUSCL methods can be found in [6] or [1], and specific recent examples of such schemes are presented in [2].