

Linear Foundations of Finite Volume Schemes
Preliminary version

Bruno Després

October 1, 2010

Introduction

This monograph is devoted to the numerical analysis of Finite Volume schemes (FV), as opposed to Finite Element Methods (FEM) or Finite Differences (FD). It is based on a M2 course given at the University of Paris VI, Lab. JLL. The discussion is restricted mainly to linear equations. A convenient model for expository purposes is the advection diffusion equation

$$\partial_t u + \mathbf{a} \cdot \nabla u - \Delta u = 0, \quad t > 0$$

in dimension two.

Most of material presented reflects also the own interests of the author about the **linear numerical analysis of Finite Volume schemes**.

This text is in construction and far to be error free. Remarks of any kind are welcomed at: despres@ann.jussieu.fr

Contents

1	Principles of Finite Volume	7
1.1	FV for the transport equation	7
1.1.1	Discretization in dimension $d = 1$	7
1.1.2	The transport equation in dimension $d \geq 2$	13
1.1.3	FV discretization in dimension $d > 1$	17
1.2	FV for the diffusion equation	20
1.2.1	Discretization in dimension $d = 1$	21
1.2.2	FV discretization in dimension $d = 1$	23
1.2.3	Relaxing hypothesis 1	25
1.3	Primal mesh-dual mesh	27
1.4	Elements of functional analysis	27
1.4.1	L^p spaces	28
1.4.2	Inequalities	29
1.4.3	Functions with bounded variation	29
1.4.4	Coarea formula	30
1.5	Exercices	31
2	Analysis of FV for advection	33
2.1	Maximum principle and CFL condition	33
2.1.1	A practical remark about the calculation of a_{jk}	34
2.1.2	The maximum principle	35
2.1.3	Geometrical analysis of the CFL condition	38
2.2	Default of consistency	40
2.2.1	What is consistency in the FD sense?	40
2.2.2	Consistency of FV	44
2.2.3	Strong consistency in 1D	45
2.2.4	Strong consistency in 2D	46
2.3	Weak consistency	50
2.4	An abstract formulation	52
2.4.1	Convergence of a non consistent iterative process	54
2.5	Convergence in L^2	56
2.5.1	Time estimate	57
2.5.2	Space estimate	58
2.6	Convergence in L^p for $p \neq 2$	59
2.6.1	$2 < p < \infty$	60
2.7	Approximation results	64

3	Non linear schemes for advection	71
3.1	Theory in 1D	71
3.1.1	The Muscl method	72
3.1.2	The construction of Lagoutière	75
3.1.3	Convergence	76
3.1.4	The repair method of Shashkov and Wendroff	79

Chapter 1

Principles of Finite Volume

Numerical discretization methods for linear partial differential equations may be grouped in three categories which are: Finite Differences (FD), the Finite Element Method (FEM) and Finite Volumes (FV). FD and FEM have been widely studied in the past. We refer to [12, 7] for FD and to [3, 6, 5] for FEM. See also references therein. An excellent presentation of the theory of FV schemes in the context of parabolic equations and non linear equations is [10], see also [8].

In this chapter we construct basic Finite Volume schemes for the transport equation and for the diffusion equation. We also compare these methods with standard FD and FEM schemes for the same equations. At the end of the chapter we introduce some elements of functional analysis which will be used in forthcoming chapters for the numerical analysis of Finite Volume schemes.

1.1 FV for the transport equation

The transport equation in free space in dimension d writes

$$\partial_t u + \mathbf{a} \cdot \nabla u = 0, \quad t > 0, \quad \mathbf{x} \in \mathbb{R}^d.$$

The function $(t, \mathbf{x}) \mapsto u(t, \mathbf{x})$ is the unknown: t is the time variable and $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ is the space variable. The gradient operator is

$$\nabla u = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_d} \right).$$

The vector field $\mathbf{x} \mapsto \mathbf{a}(\mathbf{x}) \in \mathbb{R}^d$ is given. It is called the velocity field for reasons that will appear evident in the following.

1.1.1 Discretization in dimension $d = 1$

For simplicity we consider the transport equation in dimension one $d = 1$ for a constant velocity $a \in \mathbb{R}$. The equation is now

$$\partial_t u + a \partial_x u = 0, \quad t > 0, \quad x \in \mathbb{R}. \quad (1.1)$$

We can assume for convenience that $a > 0$. The other case $a < 0$ is symmetrical to this one. This equation is supplemented with an initial condition at time $t = 0$

$$u(0, x) = u_0(x). \quad (1.2)$$

Lemma 1 *The unique solution of the transport equation (1.1) with initial condition (1.2) is*

$$u(t, x) = u_0(x - at). \quad (1.3)$$

With this formula it is evident that a is the velocity at which the initial data is advected.

Let assume for simplicity that the initial condition u_0 is a smooth C^1 function. Take the function defined in (1.3). One has $\partial_t u = -au'_0(x - at)$ and $\partial_x u = u'_0(x - at)$. So $\partial_t u + a\partial_x u = -au'_0 + au'_0 = 0$ which proves that u defined in (1.3) is indeed a solution.

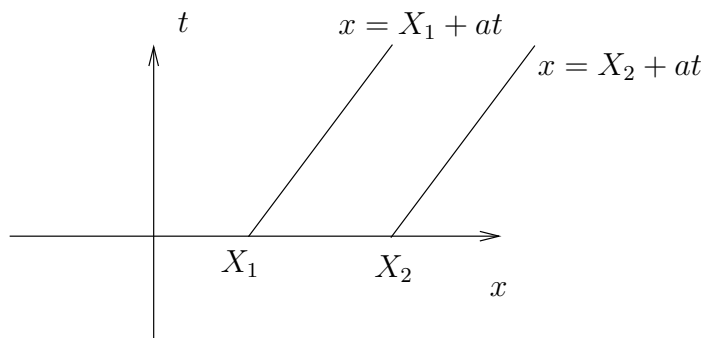


Figure 1.1: The solution of the advection equation is constant on the characteristic lines $x = X + at$

It remains to check the uniqueness. Let u_1 and u_2 be two a priori different smooth solutions of the advection equation with the same initial condition

$$u_1(0, x) = u_2(0, x) = u_0(x).$$

The function $u_1 - u_2$ is also a solution of the advection equation.

Let $x \mapsto \varphi_0(x)$ be a smooth non negative function $\varphi_0 \geq 0$ with compact support, that is $\varphi_0(x) = 0$ for $|x| \geq A$. Define $\varphi(t, x) = \varphi_0(x - at)$ which is a non negative solution of the advection equation with compact support. Define $v = (u_1 - u_2)^2 \varphi$. One has

$$\partial_t v + a\partial_x v = 2v\varphi(\partial_t(u_1 - u_2) + a\partial_x(u_1 - u_2)) + v^2(\partial_t \varphi + a\partial_x \varphi) = 0.$$

So v is also a solution of the advection equation. By construction v is non negative ($v \geq 0$) and has a compact support (it was not necessarily the case for $u_1 - u_2$). Therefore

$$0 = \int_{\mathbb{R}} (\partial_t v + a\partial_x v) dx = \int_{\mathbb{R}} \partial_t v dx.$$

By construction $v(0, x) = 0$. After integration in time we get

$$\int_{\mathbb{R}} v(T, x) dx = \int_{\mathbb{R}} v(0, x) dx + \int_0^T \left(\int_{\mathbb{R}} \partial_t v(t, x) dx \right) dt = 0.$$

This equality is true for all time $T > 0$ and all non negative function with compact support φ . Since $v \geq 0$ by construction, then v is zero everywhere. It shows that $u_1 = u_2$ and ends the proof.

Finite Difference approximation

Consider a grid in the upper plane $t > 0$. The time step is $\Delta t > 0$. The spatial step is $\Delta x > 0$. Set

$$v_j^n = u(j\Delta x, n\Delta t)$$

where u is a smooth solution of the advection equation. The time derivative is approximated by

$$\partial_t u(j\Delta x, n\Delta t) = \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(\Delta t).$$

Explicit FD approximations of the operator ∂_x can be constructed as follows

$$\begin{cases} \partial_x u(j\Delta x, n\Delta t) = \frac{v_j^n - v_{j-1}^n}{\Delta x} + O(\Delta x) & \text{(left-shifted),} \\ \partial_x u(j\Delta x, n\Delta t) = \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} + O(\Delta x^2) & \text{(centered),} \\ \partial_x u(j\Delta x, n\Delta t) = \frac{v_{j+1}^n - v_j^n}{\Delta x} + O(\Delta x) & \text{(right-shifted).} \end{cases}$$

In the case $a > 0$ it is known that the left-shifted approximation of the space derivative is better for stability reasons. It corresponds to the upwind method (also referred to as the donor cell choice). We write

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} + a \frac{v_j^n - v_{j-1}^n}{\Delta x} = O(\Delta x, \Delta t).$$

Dropping the residual in the right hand side, we get the standard (first order) upwind scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0. \quad (1.4)$$

Now u_j^n is the approximation of the exact solution at time $n\Delta t$ and position $j\Delta x$.

Finite Element approximation

Consider first the equation $\frac{d}{dx}u = f$ where the right hand side f is a given function. The Finite Element approximation of the operator $\frac{d}{dx}$ is based on a weak formulation. We consider the simplest one

$$\int_{\mathbb{R}} \frac{d}{dx}u(x)v(x)dx = \int_{\mathbb{R}} f(x)v(x)dx,$$

this is true for all test function v in an ad-hoc space of functions with compact support. One defines the standard P^1 hat function $x \mapsto \varphi_j(x)$

$$\begin{cases} \varphi_j(x) = 0 & \text{for } x \leq (j-1)\Delta x \text{ or } x \geq (j+1)\Delta x, \\ \varphi_j(x) = \frac{x - (j-1)\Delta x}{\Delta x} & \text{for } (j-1)\Delta x \leq x \leq j\Delta x, \\ \varphi_j(x) = \frac{(j+1)\Delta x - x}{\Delta x} & \text{for } j\Delta x \leq x \leq (j+1)\Delta x. \end{cases} \quad (1.5)$$

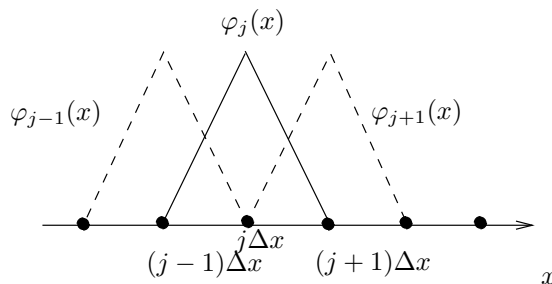


Figure 1.2: The hat function φ_j and its two neighbors φ_{j-1} and φ_{j+1}

The discrete Finite Element formulation writes

$$\int_{\mathbb{R}} \frac{d}{dx} u_{\Delta x}(x) \varphi_j(x) dx = \int_{\mathbb{R}} f(x) \varphi_j(x) dx, \quad \forall j.$$

Let assume that u may be approximated by $u_{\Delta x}$

$$u_{\Delta x} = \sum_i u_i \varphi_i.$$

Plugging into the discrete formulation, one gets

$$\sum_i \left(\int_{\mathbb{R}} \varphi_i'(x) \varphi_j(x) dx \right) u_i = \int_{\mathbb{R}} f(x) \varphi_j(x) dx, \quad \forall j.$$

We set $a_{i,j} = \int_{\mathbb{R}} \varphi_i'(x) \varphi_j(x) dx$. Elementary calculations show that

$$\begin{cases} a_{i,j} = 0 & i \leq j-2, \\ a_{i,j} = 0 & i \geq j+2, \\ a_{j+1,j} = \int_{j\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x} \times \frac{(j+1)\Delta x - x}{\Delta x} dx = \frac{1}{2}, \\ a_{j-1,j} = \int_{(j-1)\Delta x}^{j\Delta x} \frac{-1}{\Delta x} \times \frac{x - j\Delta x}{\Delta x} dx = -\frac{1}{2}, \\ a_{j,j} = \int_{\mathbb{R}} \frac{d}{dx} \left(\frac{\varphi_j^2}{2} \right) dx = 0. \end{cases}$$

Therefore the discrete equation takes the form

$$\frac{u_{j+1} - u_{j-1}}{2} = \int_{\mathbb{R}} f \varphi_j, \quad \forall j.$$

Define for convenience $f_j = \frac{1}{\Delta x} \int_{\mathbb{R}} f \varphi_j$. One gets the scheme

$$\frac{u_{j+1} - u_{j-1}}{2\Delta x} = f_j, \quad \forall j.$$

Lemma 2 *The discrete FEM approximation of $\frac{d}{dx}$ is centered.*

This result extends naturally to the space-time FEM approximation of $\partial_t u + a\partial_x u = 0$. Define the hat functions in the time variable

$$\begin{cases} \psi_n(x) = 0 & \text{for } t \leq (n-1)\Delta t \text{ or } t \geq (n+1)\Delta t, \\ \psi_n(x) = \frac{t - (n-1)\Delta t}{\Delta t} & \text{for } (n-1)\Delta t \leq t \leq n\Delta t, \\ \psi_n(x) = \frac{(n+1)\Delta t - t}{\Delta t} & \text{for } n\Delta t \leq t \leq (n+1)\Delta t. \end{cases}$$

We consider the space-time weak formulation

$$\int_{\mathbb{R}} \int_{\mathbb{R}} (\partial_t u + a\partial_x u) v(x, t) dx dt = 0, \quad \forall v \text{ in a suitable space.}$$

We replace the exact solution u by a discrete function

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

The discrete FEM writes

$$\int_{\mathbb{R}} \int_{\mathbb{R}} (\partial_t u_{\Delta x, \Delta t} + a\partial_x u_{\Delta x, \Delta t}) \varphi_j(x) \psi_n(t) dx dt = 0, \quad \forall j, n..$$

Plugging the expansion in the previous equality, one gets

$$\sum_{j, n} \left(\int_{\mathbb{R}} \int_{\mathbb{R}} (\varphi_j'(x) \psi_n(t) + a\varphi_j(x) \psi_n'(t)) \varphi_j(x) \psi_n(t) dx dt \right) u_j^m = 0, \quad \forall j, n,$$

that is

$$\sum_{j, n} \left(\int_{\mathbb{R}} a_{i, j} b_{m, n} + a b_{i, j} a_{m, n} \right) u_i^m = 0, \quad \forall j, n.$$

By definition

$$b_{i, j} = \int_{\mathbb{R}} \varphi_i(x) \varphi_j(x) dx = \begin{cases} \frac{2}{3} & \text{for } i = j, \\ \frac{1}{6} & \text{for } i = j \pm 1, \\ 0 & \text{for } i \neq j - 1, j, j + 1. \end{cases}$$

We obtain the scheme

$$\begin{aligned} & \frac{\frac{1}{6}u_{j-1}^{n+1} + \frac{2}{3}u_j^{n+1} + \frac{1}{6}u_{j+1}^{n+1} - \frac{1}{6}u_{j-1}^{n-1} - \frac{2}{3}u_j^{n-1} - \frac{1}{6}u_{j+1}^{n-1}}{\Delta t} \\ & + a \frac{\frac{1}{6}u_{j+1}^{n-1} + \frac{2}{3}u_j^n + \frac{1}{6}u_{j+1}^{n+1} - \frac{1}{6}u_{j-1}^{n-1} - \frac{2}{3}u_j^n - \frac{1}{6}u_{j-1}^{n+1}}{\Delta t} = 0. \end{aligned} \quad (1.6)$$

This scheme is centered in time and space. It is also non explicit (one cannot compute u_j^{n+1} directly).

An intermediate procedure could be to use the FEM for the space derivative, and a explicit FD procedure for the time derivative. The result is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \quad (1.7)$$

In any cases the spatial derivative is approximated with a centered formula.

Finite Volume approximation

The design principle of Finite Volume discretization is to integrate the equation in a cell. In dimension $d = 3$ this cell is a Volume, and this is the reason of the denomination Finite Volume methods. It is possible to integrate in a space-time volume, but for the sake of simplicity, we apply this methodology only to the integration in space volume. This is sufficient to present the basic idea. In dimension $d = 1$, we introduce the boundaries of cell j which are two points $x_{j-\frac{1}{2}}$ and $x_{j+\frac{1}{2}}$

$$\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}} \text{ is the length of cell } j.$$

We begin with the formula

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (\partial_t u + a \partial_x u) dx = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_t u dx + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a \partial_x u dx = 0. \quad (1.8)$$

The first integral is $\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_t u dx = \frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t, x) dx$. The quantity of fundamental interest in the FV method is $\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t, x) dx$. It is the mass (that is the total amount) of u in the cell. The next step consists in defining the mean value of u in cell j at time $n\Delta t$

$$v_j^n = \frac{\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(n\Delta t, x) dx}{\Delta x_j}.$$

Notice that no approximation has been made at this stage. Using a FD approximation of the operator $\frac{d}{dt}$ one gets immediately the approximate formula

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t, x) dx = \Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(\Delta t). \quad (1.9)$$

This is true for a smooth u . So somehow there is no real difficulty for the time derivative.

Let us turn to the approximation of

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a \partial_x u(n\Delta t, x) dx$$

which is the core of the method. The idea is to integrate in the cell (i.e. the volume)

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a \partial_x u(n\Delta t, x) dx = au(n\Delta t, x_{j+\frac{1}{2}}) - au(n\Delta t, x_{j-\frac{1}{2}}).$$

The boundary quantity $au(n\Delta t, x_{j+\frac{1}{2}})$ is called the flux. The fundamental question is now to approximate $u(n\Delta t, x_{j+\frac{1}{2}})$ by some combination of the mean values u_j^n . The usual choice is to upwind this choice accordingly to the sign of the

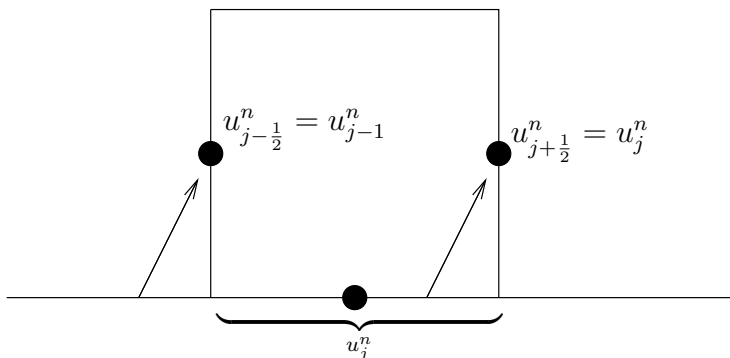


Figure 1.3: The value at $x_{j+\frac{1}{2}}$ is upwinded following the sign of the velocity ($a > 0$) and the slope of the characteristic lines

velocity. In the case $a > 0$, then we take

$$u(n\Delta t, x_{j+\frac{1}{2}}) = v_j^n + O(\Delta x), \quad \forall j.$$

So we get

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a \partial_x u(n\Delta t, x) dx = a(v_j^n - v_{j-1}^n) + O(\Delta x). \quad (1.10)$$

Inserting (1.9) and (1.10) in (1.8), one gets

$$\Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + a(v_j^n - v_{j-1}^n) = O(\Delta x) + O(\Delta t).$$

Dropping the right hand side, we get the definition of the Finite Volume scheme

$$\Delta x_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + a(u_j^n - u_{j-1}^n) = 0. \quad (1.11)$$

Lemma 3 Consider a regular grid: $\Delta x_j = \Delta x$ for all cell j . Then the FV scheme (1.11) is equal to the upwind FD scheme (1.4), and is different from the centered FEM schemes (1.6) and (1.7).

1.1.2 The transport equation in dimension $d \geq 2$

Let $\Omega \subset \mathbb{R}^d$ be an open bounded subset of \mathbb{R}^d . The velocity field is $\mathbf{x} \mapsto \mathbf{a}(\mathbf{x})$. We assume that $\mathbf{a} \in C^1(\bar{\Omega})$ is a divergence free

$$\nabla \cdot \mathbf{a} = 0.$$

Then the transport equation may be rewritten in a conservative form

$$\partial_t u + \nabla \cdot (\mathbf{a}u) = \partial_t u + \mathbf{a} \cdot \nabla u + (\nabla \cdot \mathbf{a}) u = 0.$$

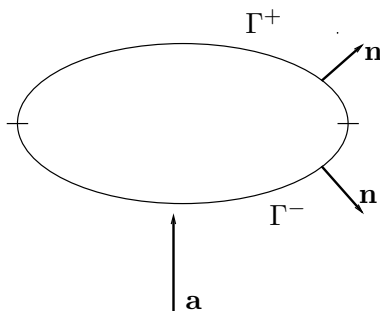


Figure 1.4: On this example Γ^+ is the upper part of the boundary. Γ^- is the lower part.

The boundary is split in two parts $\Gamma = \Gamma^- \cup \Gamma^+$

$$\Gamma^- = \{\mathbf{x} \in \Gamma, \mathbf{a} \cdot \mathbf{n} \leq 0\}, \quad \Gamma^+ = \{\mathbf{x} \in \Gamma, \mathbf{a} \cdot \mathbf{n} > 0\}.$$

We consider the problem with initial condition and boundary condition

$$\begin{cases} \partial_t u + \mathbf{a} \cdot \nabla u = 0, & \mathbf{x} \in \Omega, \quad t > 0, \\ u(0, \mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(t, \mathbf{x}) = u^-(t, \mathbf{x}), & \mathbf{x} \in \Gamma^-. \end{cases} \quad (1.12)$$

There is no boundary condition on Γ^+ . The difference between Γ^- and Γ^+ is evident in the following immediate result.

Lemma 4 Consider u_1 and u_2 two smooth solutions of (1.12). Assume u_1 and u_2 have the same initial condition and the same boundary condition on Γ^- . Then $u_1 = u_2$.

The difference $e = u_1 - u_2$ is solution of

$$\begin{cases} \partial_t e + \mathbf{a} \cdot \nabla e = 0, & \mathbf{x} \in \Omega, \quad t > 0, \\ e(0, \mathbf{x}) = 0, & \mathbf{x} \in \Omega, \\ e(t, \mathbf{x}) = 0, & \mathbf{x} \in \Gamma^-. \end{cases}$$

Define $E(t) = \frac{1}{2} \int_{\Omega} e^2(t, x) dx$. Then

$$\begin{aligned} E'(t) &= \int_{\Omega} e \partial_t e dx = - \int_{\Omega} e \mathbf{a} \cdot \nabla e dx = - \int_{\Omega} \nabla \cdot \left(\mathbf{a} \frac{e^2}{2} \right) dx \\ &= - \int_{\Gamma^-} (\mathbf{a}, \mathbf{n}) \frac{e^2}{2} d\sigma - \int_{\Gamma^+} (\mathbf{a}, \mathbf{n}) \frac{e^2}{2} d\sigma = - \int_{\Gamma^+} (\mathbf{a}, \mathbf{n}) \frac{e^2}{2} d\sigma \leq 0. \end{aligned}$$

We have used $e = 0$ on Γ^- . Since $E(0) = 0$ then $E(t) = 0$ for all $t > 0$. Therefore $u_1 = u_2$.

Characteristic forward in time

A general method to construct solutions is based on the characteristic curves $t \mapsto \mathbf{y}(t, \mathbf{X})$ defined by

$$\frac{d}{dt}\mathbf{y}(t, \mathbf{X}) = \mathbf{a}(\mathbf{X}) \text{ with } \mathbf{y}(0, \mathbf{x}) = \mathbf{X}.$$

These characteristic curves are correctly defined by application of the Cauchy-Lipschitz theorem (\mathbf{a} is C^1). A function u constant along the characteristic

$$u(t, \mathbf{y}(t, \mathbf{X})) = u_0(\mathbf{X})$$

is by construction such that

$$\partial_t|_{\mathbf{x}}u = \partial_tu + \frac{d}{dt}\mathbf{y}(t, \mathbf{X}) \cdot \nabla u = \partial_tu + \mathbf{a} \cdot \nabla u = 0.$$

So u is transported along the characteristic curves. The value of $u(t, \mathbf{x})$ is equal to $u(\mathbf{X})$ where

$$\mathbf{y}(t, \mathbf{X}) = \mathbf{x}. \quad (1.13)$$

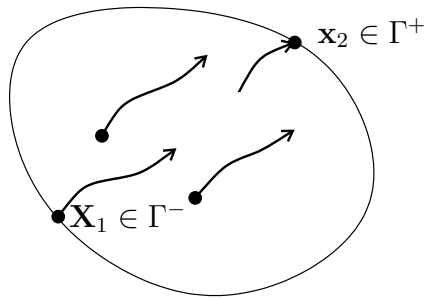


Figure 1.5: The function u is constant along the characteristic curves issued from the black circles

So to construct a solution one has to invert the equation (1.13) to find the starting point of characteristics. This is where the boundary conditions are involved. To clarify this issue we prefer to develop the same approach, but backward in time.

Characteristic backward in time

A second possibility to construct the solution is to construct the characteristic backward in time

$$\frac{d}{dt}\mathbf{X}(t, \mathbf{x}) = -\mathbf{a}(\mathbf{X}) \text{ with } \mathbf{X}(0, \mathbf{x}) = \mathbf{x} \in \Omega.$$

So $\mathbf{X}(t, \mathbf{x})$ is the starting point of the forward-in-time characteristic defined previously. This method is of course equivalent to the first one, but the discussion of the boundary source term u^- is easier.

We need to consider two cases: either \mathbf{X} is inside Ω , or \mathbf{X} has reached the boundary Γ . So we define the outgoing (backward) time

$$T(\mathbf{x}) = \inf(t) \text{ such that } \mathbf{X}(t, \mathbf{x}) \in \partial\Omega.$$

If $\mathbf{X}(t, \mathbf{x}) \in \Omega$ for all $t > 0$, we set $T(\mathbf{x}) = +\infty$. Then two cases occur.

First case: $t < T(\mathbf{x})$. We set

$$u(t, \mathbf{x}) = u_0(\mathbf{X}(t, \mathbf{x})). \quad (1.14)$$

Second case: $T(\mathbf{x}) \leq t$. At time $t = T(\mathbf{x})$ the characteristic line reaches the boundary, necessarily in Γ^- . So for all time $T(\mathbf{x}) \leq t$ we set

$$u(t, \mathbf{x}) = u^-(t - T(\mathbf{x}), \mathbf{X}(T(\mathbf{x}), \mathbf{x})). \quad (1.15)$$

By construction the function u defined by (1.14-1.15) satisfies the initial condition

$$u(0, \mathbf{x}) = u_0(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \quad (\text{this is (1.14) at } t = 0),$$

and satisfies the boundary condition

$$u(t, \mathbf{x}) = u^-(t, \mathbf{x}), \quad \forall \mathbf{x} \in \Gamma^-, \quad (\text{this is (1.15) for } \mathbf{x} \in \Gamma^-).$$

Since by construction

$$\mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x})) = \mathbf{X}(t, \mathbf{x}) \text{ for small } h > 0,$$

then one has the relation

$$u(t - h, \mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x}))) = u(t, \mathbf{x}) \text{ for small } h > 0. \quad (1.16)$$

One has the immediate result.

Lemma 5 *Assume the regularity $u_0 \in C^1(\Omega)$. Then the function u defined by (1.16) is locally C^1 and is solution of*

$$\partial_t u + \mathbf{a} \cdot \nabla u = 0 \quad \forall \mathbf{x} \in \Omega \quad \forall t < T(\mathbf{x}).$$

By construction the mapping $(t, \mathbf{x}) \mapsto \mathbf{X}(t, \mathbf{x})$ is C^1 around (t, \mathbf{x}) in the case $t < T(\mathbf{x})$. Then we differentiate (1.16) and get $\frac{d}{dh} u(t - h, \mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x}))) = 0$, that is

$$-\partial_t u - \frac{d}{dh} \mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x})) \cdot \nabla u = 0.$$

Since $\frac{d}{dh} \mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x})) = \mathbf{a}(\mathbf{X}(t - h, \mathbf{X}(h, \mathbf{x})))$, then for $h = 0$ it gives $-\partial_t u - \mathbf{a} \cdot \nabla u = 0$. It ends the proof.

The difficulty comes for $t \geq T(\mathbf{x})$ since the outgoing time $T(\mathbf{x})$ can be non continuous as well. In this case it is not clear how to differentiate (1.16) with respect to h since the function u may be discontinuous in the \mathbf{x} variable for example.

But if $T(\mathbf{x})$ and $\mathbf{X}(T(\mathbf{x}), \mathbf{x})$ are continuous and smooth functions, it is enough to insure that

$$\partial_t u + \mathbf{a} \cdot \nabla u = 0 \quad \forall \mathbf{x} \in \Omega \quad \forall t > T(\mathbf{x}).$$

If moreover the initial condition has some continuity with the boundary source term, then it is possible for the solution to be globally smooth across $t = T(\mathbf{x})$. However such conditions are non natural. We do not pursue in this direction. Instead it is possible to interpret (1.16) has an alternative formulation of the transport equation.

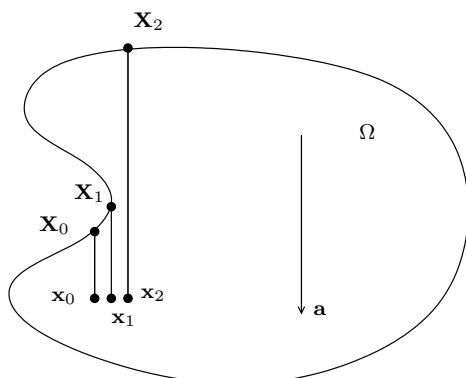


Figure 1.6: The velocity \mathbf{a} is vertical and uniform. The function $\mathbf{x} \mapsto \mathbf{X}(T(\mathbf{x}), \mathbf{x})$ is non continuous at \mathbf{x}_1 . The (backward) outgoing time $T(\mathbf{x})$ is also discontinuous at \mathbf{x}_1

1.1.3 FV discretization in dimension $d > 1$

We simplify the notations for expository purposes. So we take \mathbf{a} uniform in space. The domain Ω is polygonal in dimension $d = 2$. We consider a mesh of Ω . The mesh is a collection of **polygons** Ω_j such that

$$\overline{\Omega} = \cup_j \overline{\Omega_j}.$$

Denoting s_j the area of Ω_j we have

$$\text{mes}(\Omega) = \sum_j s_j.$$

The outgoing normal of Ω_j is \mathbf{n}_j . The interface between Ω_j and Ω_k is $\Sigma_{jk} = \Sigma_{kj}$. On the interface \mathbf{n}_j is also referred to as \mathbf{n}_{jk} . The length of the interface is $l_{jk} = l_{kj}$. It may be equal to zero if $\Sigma_{jk} = \emptyset$. By construction

$$\mathbf{n}_{jk} + \mathbf{n}_{kj} = 0 \text{ for } l_{jk} > 0.$$

The boundary of Ω_j is a collection of segments

$$\partial\Omega_j = \cup_k \Sigma_{jk} \cup \Gamma_j^- \cup \Gamma_j^+$$

where

$$\Gamma_j^- = \partial\Omega_j \cap \Gamma_{\text{int}}, \quad \text{that is } \mathbf{a} \cdot \mathbf{n}_j \leq 0 \text{ on } \Gamma_j^-,$$

and

$$\Gamma_j^+ = \partial\Omega_j \cap \Gamma^+, \quad \text{that is } \mathbf{a} \cdot \mathbf{n}_j > 0 \text{ on } \Gamma_j^+.$$

The length of Γ_j^- is l_j^- . The length of Γ_j^+ is l_j^+ . We need to define a characteristic length h of the mesh. We decide arbitrarily to consider the maximal value of all boundaries

$$h = \max \left(\max_{jk} l_{jk}, \max_j l_j^-, \max_j l_j^+ \right). \quad (1.17)$$

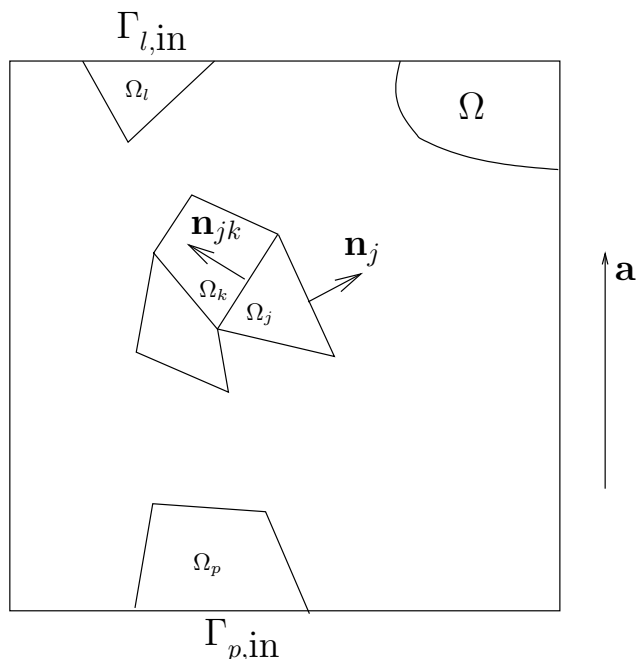


Figure 1.7: A mesh

To construct the FV scheme we integrate the equation in the cell Ω_j

$$\int_{\Omega_j} (\partial_t u + \nabla \cdot (\mathbf{a}u)) dx = 0.$$

We separate the time derivative and the space derivatives

$$\frac{d}{dt} \int_{\Omega_j} u dx + \int_{\Omega_j} \nabla \cdot (\mathbf{a}u) dx = 0. \quad (1.18)$$

For simplicity the solution u is as smooth as necessary, so that all expansions and approximations below can be justified with Taylor expansions. We rewrite the time contribution as

$$\left(\frac{d}{dt} \int_{\Omega_j} u dx \right) (n\Delta t) = s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(h^2 \Delta t). \quad (1.19)$$

Here v_j^n is the mean value of u in the cell at time $t_n = n\Delta t$

$$v_j^n = \frac{\int_{\Omega_j} u(n\Delta t, \mathbf{x}) dx}{s_j}.$$

Denote \mathbf{G}_j the center of mass of the cell. Then

$$v_j^n = u(n\Delta t, \mathbf{G}_j) + O(h^2). \quad (1.20)$$

Definition 1 We say that the natural degree of freedom in cell Ω_j is the value of the function at the center of mass \mathbf{G}_j .

Then we turn to the second contribution in (1.18). Let us integrate it directly in the cell. Assuming that Ω_j is in the interior of the domain, we obtain

$$\int_{\Omega_j} \nabla \cdot (\mathbf{a}u(n\Delta t, \mathbf{x})) dx = \int_{\partial\Omega_j} \mathbf{a} \cdot \mathbf{n}_j u(n\Delta t, \mathbf{x}) d\sigma = \sum_k (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) v_{jk}^n. \quad (1.21)$$

Here v_{jk}^n is the mean value on Σ_{jk} of u at time $t_n = n\Delta t$

$$v_{jk}^n = \frac{\int_{\Sigma_{jk}} u(n\Delta t, \mathbf{x}) d\sigma}{l_{jk}}.$$

The central idea of FV methods consists in the elimination of the flux v_{jk}^n in term of the cell centered values v_k^n for all k . For this task, we go back to the notion of characteristic lines. Two cases occur. For instance if the velocity direction \mathbf{a} goes from Ω_j into Ω_k then we shall take $v_{jk}^n \approx v_j^n$.

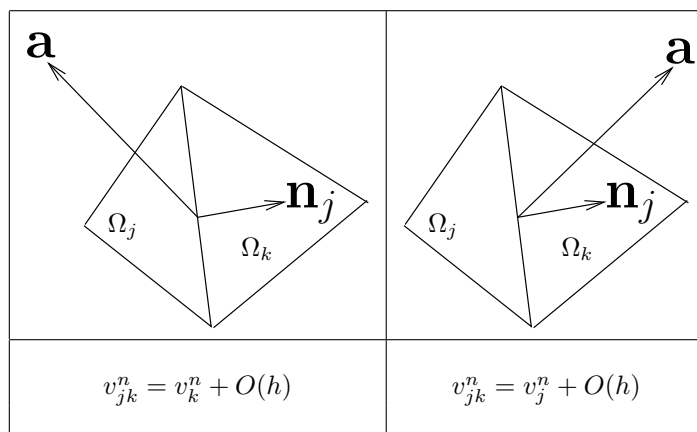


Figure 1.8: Upwinding at the boundaries

The same idea is used at boundaries. The small difference is for the incoming boundary Γ_j^- , because there is no neighbor. Instead the boundary source is used. In this case we use the mean value u_{in} on the boundary

$$u_j^{-,n} = \frac{\int_{n\Delta t}^{(n+1)\Delta t} \int_{\Gamma_j^-} u^-(s, \mathbf{x}) d\sigma ds}{\Delta t l_j^-}. \quad (1.22)$$

Another case is when $\mathbf{a} \cdot \mathbf{n}_{jk} = 0$. The value of v_{jk}^n does not matter, since it is multiplied by the weight $l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk} = 0$ in (1.21). Plugging all this (1.19-1.21) we obtain

$$\begin{aligned} & s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(h^2 \Delta t) \\ & + \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) (v_j^n + O(h)) + \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} < 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) (v_k^n + O(h)) \\ & + (l_j^- \mathbf{a} \cdot \mathbf{n}_j) v_j^{-,n} + (l_j^+ \mathbf{a} \cdot \mathbf{n}_j) (v_j^n + O(h)) = 0. \end{aligned}$$

The Finite Volume is defined by dropping all $O(\cdot)$ terms. It writes

$$\begin{aligned} & s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} \\ & + \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_j^n + \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} < 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_k^n \\ & + (l_j^- \mathbf{a} \cdot \mathbf{n}_j) u_j^{-,n} + (l_j^+ \mathbf{a} \cdot \mathbf{n}_j) u_j^{+,n} = 0. \end{aligned} \quad (1.23)$$

An important property is the following.

Lemma 6 *The scheme is conservative. That is the variation of the total mass is due to what flows in and out at boundaries*

$$\frac{\sum_j s_j u_j^{n+1} - \sum_j s_j u_j^n}{\Delta t} + \left(\sum_j (l_j^- \mathbf{a} \cdot \mathbf{n}_j) u_j^{-,n} \right) + \left(\sum_j (l_j^+ \mathbf{a} \cdot \mathbf{n}_j) u_j^{+,n} \right) = 0.$$

The proof is a consequence of definition of the flux. Consider an interface Σ_{jk} : What flows out of Ω_j is what flows in Ω_k . With mathematical notations

$$\begin{aligned} & \sum_j \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_j^n + \sum_j \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} < 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_k^n \\ & = \sum_j \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_j^n + \sum_k \sum_{j, \mathbf{a} \cdot \mathbf{n}_{kj} < 0} (l_{kj} \mathbf{a} \cdot \mathbf{n}_{kj}) u_j^n \\ & = \sum_j \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}) u_j^n + \sum_j \sum_{k, \mathbf{a} \cdot \mathbf{n}_{jk} > 0} (l_{jk} \mathbf{a} \cdot (-\mathbf{n}_{jk})) u_j^n = 0. \end{aligned}$$

It ends the proof.

1.2 FV for the diffusion equation

Let us consider the non stationary homogeneous diffusion in dimension $d = 2$ with Neumann boundary conditions

$$\begin{cases} \partial_t u - \Delta u = 0, & t > 0, \quad \mathbf{x} \in \Omega, \\ \nabla u \cdot \mathbf{n} = 0, & t > 0, \quad \mathbf{x} \in \Gamma, \\ u(0, \mathbf{x}) = u_0(\mathbf{x}) & \mathbf{x} \in \Omega. \end{cases} \quad (1.24)$$

This problem has a unique solution, see [6, 11]. The uniqueness is evident for smooth solutions.

Lemma 7 *Consider two solutions u_1 and u_2 with the same initial condition u_0 . Then $u_1 = u_2$.*

Let u be a solution of (1.24). Set $E(t) = \frac{1}{2} \int_{\Omega} u^2 dx$. One has $E'(t) = \int_{\Omega} u \partial_t u dx = \int_{\Omega} u \Delta u dx$. Then we integrate by parts $E'(t) = - \int_{\Omega} \nabla u \cdot \nabla u dx + \int_{\Gamma} u \nabla u \cdot \mathbf{n} dx = - \int_{\Omega} |\nabla u|^2 dx$. After integration in time, it yields the identity

$$\int_{\Omega} u(T, \mathbf{x})^2 dx + \int_0^T \int_{\Omega} |\nabla u(t, \mathbf{x})|^2 dx dt = \int_{\Omega} u_0(\mathbf{x})^2 dx.$$

Consider two solutions u_1 and u_2 with the same initial condition u_0 . So $u = u_1 - u_2$ is a solution of the same equation with a zero initial condition. Using the above identity, we obtain that $u \equiv 0$. It shows the uniqueness of the solution of (1.24).

1.2.1 Discretization in dimension $d = 1$

In order to compare FV, FEM and FD on a simple problem, we consider the diffusion equation in dimension $d = 1$ on the entire real line

$$\partial_t u - \partial_{xx} u = 0, \quad x \in \mathbb{R}, t > 0.$$

Finite Difference

The explicit Finite Difference approximation is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = 0, \quad \forall j \in \mathbb{Z}. \quad (1.25)$$

As before $\Delta t > 0$ is the time step, and $\Delta x > 0$ is the uniform mesh size.

Finite Element Method

The Finite Element Method is based on weak formulations. We first consider the equation $-u''(x) = f$. The weak formulation writes formally

$$\int_{\mathbb{R}} u'(x)v'(x)dx = \int_{\mathbb{R}} f(x)v(x)dx, \quad \text{for all } v \text{ in a convenient space.}$$

Take the P^1 hat function (1.5). The discrete weak formulation is

$$\int_{\mathbb{R}} u'_{\Delta x} \varphi'_j dx = \int_{\mathbb{R}} f \varphi_j dx.$$

Assume as before that $u_{\Delta x} = \sum_i u_i \varphi_i$. The discrete system writes

$$\sum_i \left(\int_{\mathbb{R}} \varphi'_i(x) \varphi'_j(x) dx \right) u_i = \int_{\mathbb{R}} f \varphi_j dx, \quad \forall j.$$

Set $c_{i,j} = \int_{\mathbb{R}} \varphi'_i(x) \varphi'_j(x) dx$ so that

$$\begin{cases} c_{i,j} = 0 & i \leq j - 2, \\ c_{i,j} = 0 & i \geq j + 2, \\ c_{j+1,j} = \int_{j\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x} \times \frac{-1}{\Delta x} dx = -\frac{1}{\Delta x}, \\ c_{j-1,j} = \int_{(j-1)\Delta x}^{j\Delta x} \frac{-1}{\Delta x} \times \frac{1}{\Delta x} dx = -\frac{1}{\Delta x}, \\ c_{j,j} = \int_{(j-1)\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x^2} dx = \frac{2}{\Delta x}. \end{cases}$$

Define for convenience $f_j = \frac{1}{\Delta x} \int_{\mathbb{R}} f \varphi_j$. One gets the scheme

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x} = f_j, \quad \forall j.$$

For simplicity we use an explicit discretization of $\partial_t u$. We obtain our FEM scheme for the diffusion equation in dimension $d = 1$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x} = 0, \quad \forall j. \quad (1.26)$$

This scheme is equal to the FD scheme (1.25).

Finite Volume

Finally we turn to the Finite Volume discretization. The principle is identical to what has been done for the transport equation in dimension one. So we integrate the equation in a cell $]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[$ and get

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t, x) dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_{xx} u(t, x) dx = 0.$$

That is

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t, x) dx - \partial_x u(t, x_{j+\frac{1}{2}}) + \partial_x u(t, x_{j-\frac{1}{2}}) = 0. \quad (1.27)$$

As before the mean value of u in the cell is noted

$$v_j^n = \frac{\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(n\Delta t, x) dx}{\Delta x_j} = u(n\Delta t, G_j) + O(\Delta x_j^2), \quad G_j = \frac{x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}}{2}, \quad (1.28)$$

and time derivative is approximated by the explicit forward difference (1.9). The natural approximation of the space derivative $\partial_x u(t, x_{j+\frac{1}{2}})$ is

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{u(n\Delta t, G_{j+1}) - u(n\Delta t, G_j)}{G_{j+1} - G_j} + O(G_{j+1} - G_j). \quad (1.29)$$

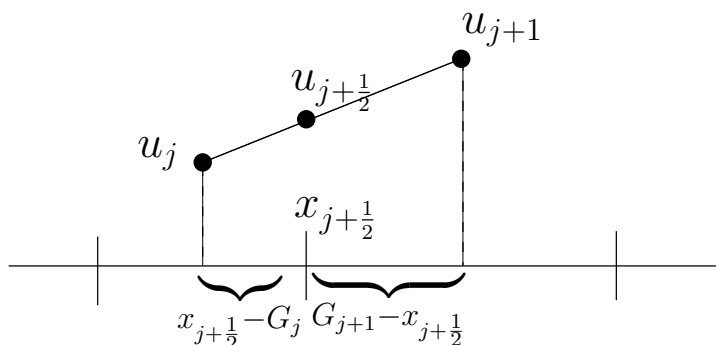


Figure 1.9: Interpolation at $x_{j+\frac{1}{2}}$ of the space derivative

It must be noted that if the mesh is uniform in the sense that

$$G_{j+1} - x_{j+\frac{1}{2}} = x_{j+\frac{1}{2}} - G_j \iff x_{j+\frac{3}{2}} - x_{j+\frac{1}{2}} = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}},$$

then the interpolation error is second order

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{u(n\Delta t, G_{j+1}) - u(n\Delta t, G_j)}{G_{j+1} - G_j} + O((G_{j+1} - G_j)^2).$$

But on the general case only (1.29) is true.

Replacing the point-wise value by the mean value with (1.28), we obtain

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{v_{j+1}^n - v_j^n}{G_{j+1} - G_j} + O(\max(\Delta x_{j+1}, \Delta x_j))$$

After elimination of the time contribution and replacement of the spatial derivatives in (1.27), one finds out

$$\Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} - \frac{v_{j+1}^n - v_j^n}{G_{j+1} - G_j} + \frac{v_j^n - v_{j-1}^n}{G_j - G_{j-1}} = O(\max(\Delta x_{j+1}, \Delta x_j, \Delta t)).$$

Finally we drop the right hand side and obtain the Finite Volume scheme for the diffusion equation

$$\Delta x_j \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - u_j^n}{G_{j+1} - G_j} + \frac{u_j^n - u_{j-1}^n}{G_j - G_{j-1}} = 0. \quad (1.30)$$

Proposition 1 *Consider a regular grid: $\Delta x_j = \Delta x$ for all cell j . The the FV scheme (1.30), the FEM scheme (1.26) and the FD scheme (1.25) are the same.*

1.2.2 FV discretization in dimension $d = 1$

Let us come back to the diffusion equation (1.24) in dimension $d = 2$ with a Neumann boundary condition. In dimension $d \geq 3$ the principles are the same. We use the notations of section 1.1.3. Let h be a characteristic length of the mesh, for example one can take the definition (1.17). The method of integration is the same, but it will appear that the construction of the FV scheme requires a compatibility assumption on the mesh. It makes the situation different compared with the FV discretization of the transport equation.

After integration of the diffusion equation in the cell Ω_j , explicit discretization of the time contribution and integration of the space derivatives in the cell, we obtain

$$s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} - \sum_k l_{jk} w_{jk}^n = O(h^2 \Delta t) \quad (1.31)$$

where by definition the mean value of the normal derivative is

$$w_{jk}^n = \frac{\int_{\Sigma_{jk}} \nabla u(n\Delta t, x) d\sigma}{l_{jk}} \cdot \mathbf{n}_{jk} = \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot \mathbf{n}_{jk} + O(h)$$

where the point \mathbf{x}_{jk} is the middle of the edge. **At this stage one adopts the philosophy of Finite Volume methods which is to construct an approximate value of the flux $l_{jk} w_{jk}^n$.** Unfortunately it is more difficult for a second order equation than for a first order equation. This will become evident in the following.

In formula (1.31) the interpolation error is second order in the definition of the mean normal derivative w_{jk}^n . One has

$$u(n\Delta t, \mathbf{G}_k) = u(n\Delta t, \mathbf{x}_{jk}) + \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{G}_k - \mathbf{x}_{jk}) + O(h^2),$$

and

$$u(n\Delta t, \mathbf{G}_j) = u(n\Delta t, \mathbf{x}_{jk}) + \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{G}_j - \mathbf{x}_{jk}) + O(h^2).$$

By subtraction

$$u(n\Delta t, \mathbf{G}_k) - u(n\Delta t, \mathbf{G}_j) = \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{G}_k - \mathbf{G}_j) + O(h^2).$$

So from the point-wise values at \mathbf{G}_k and \mathbf{G}_j , one can reconstruct the vector

$$\nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot \mathbf{f}_{jk} = \frac{u(n\Delta t, \mathbf{G}_k) - u(n\Delta t, \mathbf{G}_j)}{d_{jk}} + O(h). \quad (1.32)$$

Here

$$d_{jk} = |\mathbf{G}_k - \mathbf{G}_j| \text{ and } \mathbf{f}_{jk} = \frac{\mathbf{G}_k - \mathbf{G}_j}{d_{jk}} \text{ with } |\mathbf{f}_{jk}| = 1.$$

If the middle of the edge is also the middle of the centers of mass

$$\mathbf{x}_{jk} = \frac{\mathbf{G}_j + \mathbf{G}_k}{2}$$

then it is possible to show that the error in (1.32) is $O(h^2)$ which is better than $O(h)$. **The main problem is that we need to reconstruct the normal derivative in the direction \mathbf{n}_{jk} which can be completely different from the vector \mathbf{f}_{jk} . To continue the construction we are forced to make some hypotheses.**

Hypothesis 1 *We assume that*

$$\mathbf{f}_{jk} = \mathbf{n}_{jk}, \quad \forall j, k,$$

which means that the line joining the centers of mass is orthogonal to the edge.

This is a very strong hypothesis on the structure of the mesh. We shall relax it later on.

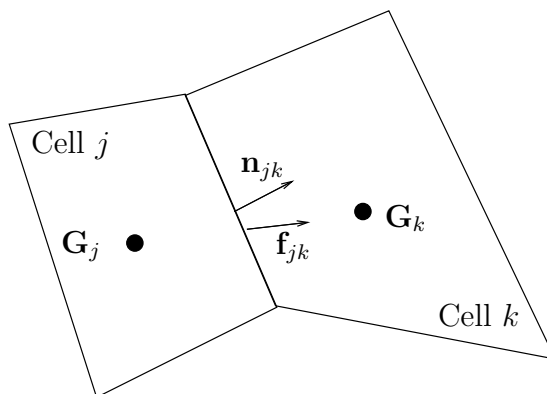


Figure 1.10: A mesh that does not satisfy hypothesis 1

So let us assume the mesh is such that the hypothesis 1 is true. Then

$$w_{jk}^n = \frac{u(n\Delta t, \mathbf{G}_k) - u(n\Delta t, \mathbf{G}_j)}{d_{jk}} + O(h)$$

and also

$$w_{jk}^n = \frac{v_k^n - v_j^n}{d_{jk}} + O(h).$$

It is now possible to eliminate w_{jk}^n in (1.31). After dropping all the $O(\cdot)$ terms, we get the definition of the Finite Volume scheme

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} - \sum_k l_{jk} \frac{u_k^n - u_j^n}{d_{jk}} = 0, \quad \forall j. \quad (1.33)$$

One can notice that the Neumann boundary condition is automatically taken into account in the scheme. This is because the sum is over all cells k around the cell j .

1.2.3 Relaxing hypothesis 1

A simple method exist to relax this hypothesis. The idea is to remark that v_j^n , which is the mean value in cell j , is a second order approximation of the point value of the function $u(n\Delta t, \mathbf{G}_j)$. But second order approximation is not necessary. Which means that it is possible to replace v_j^n by $u(n\Delta t, \mathbf{x}_j)$ where \mathbf{x}_j is a point in the cell (or at least $O(h)$ close to cell). We set

$$w_j^n = u(n\Delta t, \mathbf{x}_j)$$

where \mathbf{x}_j is not precisely defined for the moment.

What are the consequences? First the (1.31) is replaced by

$$s_j \frac{w_j^{n+1} - w_j^n}{\Delta t} - \sum_k l_{jk} w_{jk}^n = O(h^2 \Delta t) + O(h^3). \quad (1.34)$$

for smooth solutions u . The $O(h^3)$ term comes from the approximation

$$s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} = s_j \frac{w_j^{n+1} - w_j^n}{\Delta t} + O(h^3)$$

which can be proved for smooth functions. So now we shall try to approximate w_{jk}^n by a combination of w_j^n and w_k^n . One has

$$u(n\Delta t, \mathbf{x}_k) - u(n\Delta t, \mathbf{x}_j) = \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{x}_k - \mathbf{x}_j) + O(h^2),$$

so

$$\nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot \mathbf{f}_{jk} = \frac{u(n\Delta t, \mathbf{x}_k) - u(n\Delta t, \mathbf{x}_j)}{d_{jk}} + O(h)$$

where now

$$d_{jk} = |\mathbf{x}_k - \mathbf{x}_j| \quad \text{and} \quad \mathbf{f}_{jk} = \frac{\mathbf{x}_k - \mathbf{x}_j}{d_{jk}} \quad \text{with} \quad |\mathbf{f}_{jk}| = 1. \quad (1.35)$$

The rest of the construction is the same. So the FV scheme is formally the same, except that the definition of d_{jk} is based on (1.35). And of course hypothesis 1 needs to be replaced by hypothesis 2.

Hypothesis 2 *We assume that the vectors \mathbf{f}_{jk} (1.35) are such that*

$$\mathbf{f}_{jk} = \mathbf{n}_{jk}, \quad \forall j, k.$$

It means that the line joining the points \mathbf{x}_j and \mathbf{x}_k is orthogonal to the edge.

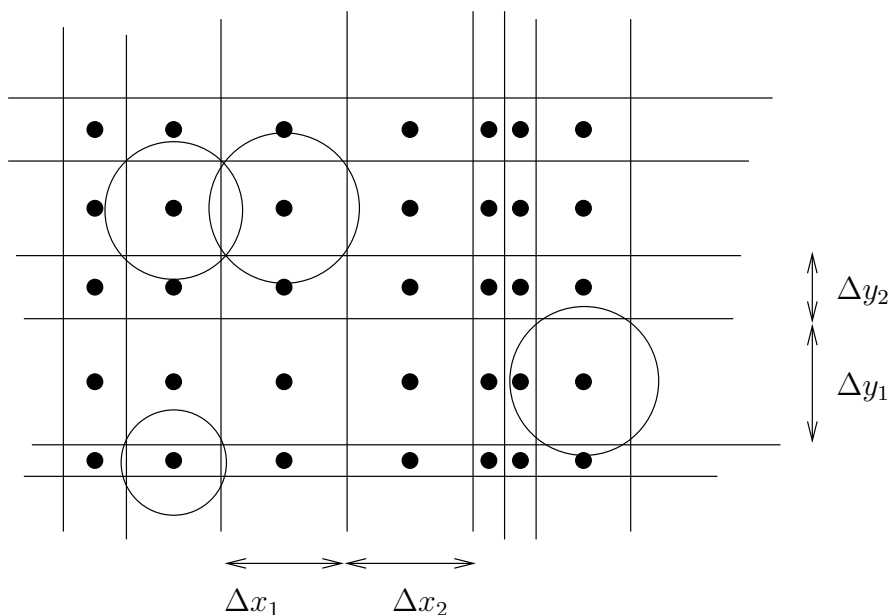


Figure 1.11: Example of quadrangular mesh which satisfies the hypothesis 2. For such a mesh, hypothesis 2 may result in a severe global structural constraint.

This hypothesis is more an equation about the points \mathbf{x}_j s that we need to satisfy in order to construct our FV scheme.

What is absolutely remarkable is that a simple local solution exists for a triangular mesh in the case all angles are small that $\frac{\pi}{2}$. This is what we describe below.

Lemma 8 *Consider a triangular mesh. Assume that the angles of all triangles are less than $\frac{\pi}{2}$. Define \mathbf{x}_j the center of the circumscribed circle for all cell j . Then $\mathbf{x}_j \in \Omega_j$ for all cell j and $\mathbf{x}_k - \mathbf{x}_j$ is parallel to \mathbf{n}_{jk} for all interface (j, k) , so that $\mathbf{f}_{jk} = \mathbf{n}_{jk}$ everywhere.*

By construction the line that joins \mathbf{x}_j and \mathbf{x}_k cuts the edge in the middle. So $\mathbf{x}_k - \mathbf{x}_j$ is always parallel to \mathbf{n}_{jk} . It remains to prove that the orientation is the same, that is the normal \mathbf{n}_{jk} is oriented from \mathbf{x}_j towards \mathbf{x}_k .

Let us define \mathcal{A} the set of all triangles with angles $< \frac{\pi}{2}$ and \mathcal{B} the set of triangles with one angle equal to $\frac{\pi}{2}$.

Note that \mathcal{B} is also the set of all triangle-rectangles, and that for such triangles the center of the circumscribed circle is the middle of the hypotenuse, thus is on the boundary of the triangle.

Consider a general open triangle $\mathcal{T} \in \mathcal{A}$. It is possible to design a continuous family of open triangles $t \mapsto \mathcal{T}(t)$ such that $\mathcal{T}(t) = \mathcal{T}$, $\mathcal{T}(t) = \mathcal{T}_E$ is equilateral triangle and $\mathcal{T}(t) \in \mathcal{A}$ for all $t \in [0, 1]$. Let $\mathbf{x}(t)$ be the center of the circumscribed circle of $\mathcal{T}(t)$. By construction $\mathbf{x}(t)$ is never on the boundary (if so $\mathcal{T}(t) \in \mathcal{B}$ for some $0 < t < 1$, which is not possible). Since $\mathbf{x}(1)$ is inside \mathcal{T}_E , then by continuity $\mathbf{x}(t) \in \mathcal{T}$ for all $0 \leq t \leq 1$. Then $\mathbf{x}_{\mathcal{T}} \in \mathcal{T}$ for all triangle in \mathcal{A} . It ends the proof of the lemma.

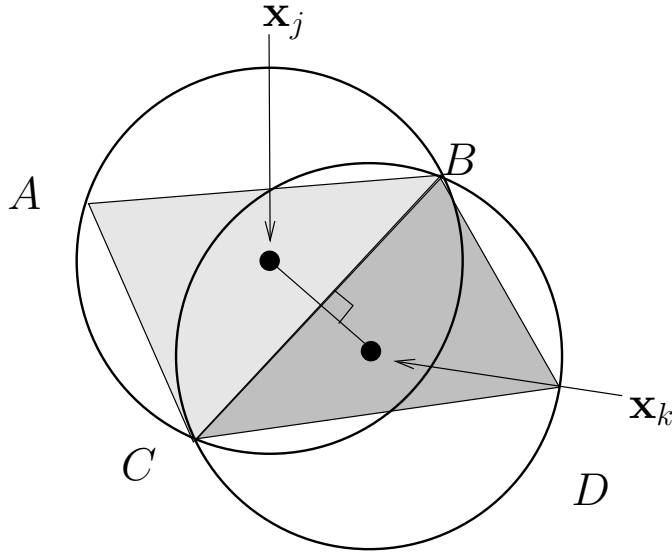


Figure 1.12: Centers of the circumscribed circles

It is possible to construct triangular meshes with all angles strictly less than $\frac{\pi}{2}$. This is not completely evident at the practical level. We shall come back to this issue later one. But nevertheless it insures that the construction of the FV scheme (1.33) with the definition d_{jk} (1.35) is correct.

1.3 Primal mesh-dual mesh

FV schemes can be used on various meshes. A important variant, with respect to what we have presented so far, is to consider control volume around vertices of the initial mesh. The initial mesh is the primal mesh. The idea is to define a dual mesh from this first one. On Figure 1.13 is an example of such a construction. In the example the grey cell around vertex r is a dual cell. By definition the dual mesh covers the whole domain. Some extra definitions are needed near the boundary Γ nevertheless.

We leave to the reader the generalization on such meshes of the FV schemes previously discussed.

1.4 Elements of functional analysis

For any numerical method for the discretization of a partial differential equation, a fundamental question is to prove that the error between the exact solution and the numerical solution is small. For this task we define norms in convenient Banach spaces.

Definition 2 *A real Banach space V is a real vectorial space, associated with a norm $u \mapsto \|u\|$ defined for all $u \in V$. The properties of a norm are*

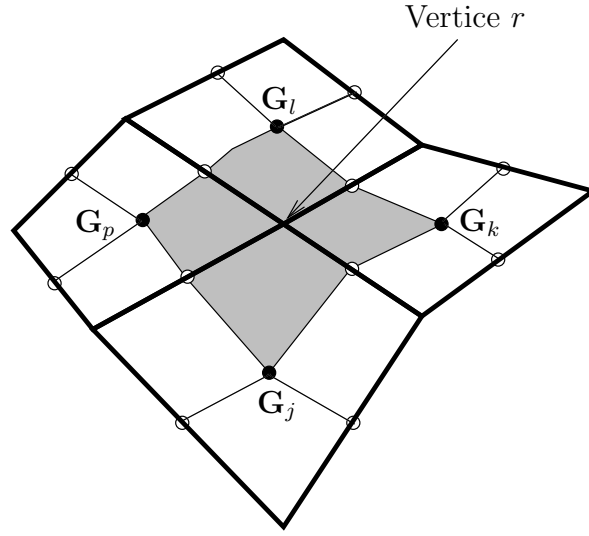


Figure 1.13: A dual mesh. The grey region Θ_r around the vertice r is a dual cell. In this example the boundary of the dual cell joins the centers of mass and the middle of the edges

- $\|u\| \geq 0$ for all $u \in V$,
- $\|u\| = 0$ if and only if $u = 0$,
- $\|\lambda u\| = |\lambda| \|u\|$ for all $\lambda \in \mathbb{R}$,
- $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in V$.

We refer to [2] for a comprehensive presentation of functional analysis. In particular the definition we use of the partial derivative of a measurable function is the one of [2].

1.4.1 L^p spaces

Let Ω be an open subset of \mathbb{R}^d .

Definition 3 Let $p \in [1, \infty]$.

- For $1 \leq p < \infty$, the space L^p is the set of all measurable functions such that $\int_{\Omega} |u(\mathbf{x})| dx < \infty$. The L^p norm is

$$\|u\|_p = \left(\int_{\Omega} |u(\mathbf{x})| dx \right)^{\frac{1}{p}}.$$

- For $p = \infty$, L^∞ is the set of all measurable and bounded functions. The L^∞ norm is

$$\|u\|_\infty = \sup \{ \lambda; \text{mes}(|u(\mathbf{x})| > \lambda) \neq 0 \}.$$

The partial derivatives [2] of u are noted

$$u^{(k_1, \dots, k_d)} = \frac{\partial^{k_1 + \dots + k_d}}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} u.$$

Definition 4 *The set of all measurable functions with all derivatives (the total order of derivation less than q) in L^p is referred to as $W^{q,p}$. For $1 \leq p \leq \infty$ a norm in $W^{p,q}$ is*

$$\|u\|_{W^{q,p}} = \sum_{k_1 + \dots + k_d \leq q, 0 \leq k_i} \|u^{(k_1, \dots, k_d)}\|_p.$$

1.4.2 Inequalities

Let $p \in [1, \infty]$ and $q \in [1, \infty]$ be two positive real numbers (possibly infinite) such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

We say p and q are **conjugate numbers**.

Lemma 9 (Hölder inequality) *Let $u \in L^p$ and $v \in L^q$ with p and q conjugate numbers. Then*

$$\left| \int_{\Omega} u(\mathbf{x})v(\mathbf{x})dx \right| \leq \|u\|_p \times \|v\|_q.$$

In case $p = q = 2$, the Hölder inequality is called the Cauchy-Schwarz inequality.

1.4.3 Functions with bounded variation

Discontinuous functions have naturally bounded variations. For a vector $\varphi = (\varphi_1, \dots, \varphi_d)$ we note

$$|\varphi| = \sqrt{\varphi_1^2 + \dots + \varphi_d^2}.$$

We define

$$C_0^{1,b} = \left\{ \varphi \in (C_0^1(\Omega))^d, |\varphi(\mathbf{x})| \leq 1 \forall \mathbf{x} \right\}$$

the set of all C^1 vectorial functions with compact support, which are also bounded by 1.

Definition 5 *Consider $u \in L^1$. The (possibly infinite) number*

$$|u|_{\text{BV}} = \sup_{\varphi \in C_0^{1,b}} \left(- \int_{\Omega} u(\mathbf{x}) \nabla \cdot \varphi(\mathbf{x}) dx \right)$$

is called the total variation of u .

Example 1 *Consider the case, in dimension one, where $\Omega = \mathbb{R}$ is the entire line. Let $u_1 \in W^{1,1}$. Then $|u_1|_{\text{BV}} = \|u_1^{(1)}\|_1 = \int_{\mathbb{R}} |u_1'(x)| dx$.*

It comes from the integration by parts formula

$$-\int_{\mathbb{R}} u_1(x)\varphi'(x)dx = \int_{\mathbb{R}} u_1'(x)\varphi(x)dx.$$

Then the supremum over all φ such that $|\varphi| \leq 1$ everywhere yields

$$\sup_{|\varphi| \leq 1} \left(\int_{\mathbb{R}} u_1'(x)\varphi(x)dx \right) = \int_{\mathbb{R}} |u_1'(x)| dx = \|u_1^{(1)}\|_1.$$

Example 2 *The unit square in \mathbb{R}^2 is*

$$S = \{\mathbf{x} = (x_1, x_2), 0 < x_1, x_2 < 1, \}.$$

Consider the indicatrix function of the unit square: $u_2(\mathbf{x}) = 1$ if $\mathbf{x} \in S$; $u_2(\mathbf{x}) = 0$ elsewhere. Then $|u_2|_{\text{BV}} = 4$.

One has

$$-\int_{\Omega} u(\mathbf{x})\nabla \cdot \varphi(\mathbf{x})dx = -\int_{\mathbf{x} \in S} \nabla \cdot \varphi(\mathbf{x})dx = \int_{\mathbf{x} \in \partial S} \varphi(\mathbf{x}) \cdot \mathbf{n}_S dx \leq 4.$$

The bound 4 is reached using a convenient sequence of functions φ_n . It shows that $|u_2|_{\text{BV}} = 4$, which is the length of the perimeter of S .

Lemma 10 *Define BV, the set of all functions in L^1 with a bounded total variation. The norm in BV is*

$$\|u\|_{\text{BV}} = |u|_{\text{BV}} + \|u\|_1.$$

One has the strict inclusion

$$W^{1,1} \subset \text{BV}.$$

It is an immediate consequence of the definition and examples.

1.4.4 Coarea formula

Let $u \geq 0$ be a measurable non negative function. Let us define

$$E_\lambda = \{\mathbf{x} \in \Omega, u(\mathbf{x}) > \lambda\} \subset \Omega.$$

The perimeter of E_λ is

$$\|E_\lambda\| = \sup_{\varphi \in C_0^{1,b}} \left(-\int_{E_\lambda} \nabla \cdot \varphi(\mathbf{x})dx \right).$$

It is also equal to $|1_E|_{\text{BV}}$, where 1_E is the indicatrix function of E . Note that for any non negative function bounded,

$$u(\mathbf{x}) = \int_0^\infty 1_{E_\lambda}(\mathbf{x})d\lambda.$$

Lemma 11 *Let $u \in \text{BV}$ be a non negative function, $u \geq 0$. Then*

$$|u|_{\text{BV}} = \int_0^\infty \|E_\lambda\|d\lambda.$$

1.5 Exercices

Ex. 1 Take the transport equation $\partial_t u + \mathbf{a} \cdot \nabla u = 0$. Consider that the vector field $\mathbf{a} \in C^1(\bar{\Omega})$ is not divergence free. Show that a possible FV scheme writes

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_{k^+} m_{jk} u_j^n + \sum_{k^-} m_{jk} u_k^n = \sum_k m_{jk} u_j^n.$$

where $m_{jk} = |(l_{jk} \mathbf{a}_{jk} \cdot \mathbf{n}_{jk})| \geq 0$.

Ex. 2 Show the previous scheme satisfies the maximum principle under CFL.

Ex. 3 Consider the system

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \partial_t (\rho c) + \nabla \cdot (\rho \mathbf{u} c) = 0. \end{cases}$$

The physical meaning is: $\rho > 0$ is a density, $0 \leq c \leq 1$ a mass fraction and \mathbf{u} a velocity. We take $\mathbf{u} \in C^1(\bar{\Omega})$ for simplicity.

Show that c satisfy the transport equation $\partial_t c + \mathbf{u} \cdot \nabla c = 0$. Discretize the both equations with a conservative FV scheme. Show that c^n satisfies the maximum principle under CFL.

Chapter 2

Analysis of FV for advection

FV schemes for advection and transport satisfy the maximum principle under a condition on the time step. This time step restriction is the famous CFL condition, coined after the seminal work of Courant, Friedrichs and Levy [9]. The maximum principle is a very strong property which insures the robustness of the method. It explains the success of such methods for complicated problems with non linear physics involved. In what follows we detail this issue for the transport equation. Then we show that a consequence is the loss of consistency in the FD sense, wich means that it is not possible to prove the convergence by using the simple FD strategy. But we show that weak consistency is the rule (that is consistency in the sense of distributions). Finally we give a proof of convergence in L^p for all $1 \leq p \leq \infty$ with an order of convergence $\frac{1}{2}$.

2.1 Maximum principle and CFL condition

We consider the problem

$$\begin{cases} \partial_t u + \mathbf{a} \cdot \nabla u = 0, & \mathbf{x} \in \Omega, & t > 0, \\ u(0, \mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(t, \mathbf{x}) = u_{\text{in}}(t, \mathbf{x}), & \mathbf{x} \in \Gamma_{\text{in}}. \end{cases} \quad (2.1)$$

We do not assume that the velocity field $\mathbf{a} \in C^1(\overline{\Omega})$ is constant: however is is divergence free

$$\nabla \cdot \mathbf{a} = 0.$$

We use the notations of section 1.1.3. Let us consider

$$a_{jk} = \frac{1}{l_{jk}} \int_{\Sigma_{jk}} \mathbf{a}(\mathbf{x}) \cdot \mathbf{n}_{jk}(\mathbf{x}) d\sigma.$$

This quantity is the mean value of the scalar product of \mathbf{a} against the outgoing normal. Assuming that \mathbf{a} is the direction of some beam of light, then $l_{jk} a_{jk}$ is the total amount of light that impinges on the edge Σ_{jk} . Let us define for convenience

$$I^+(j) = \{k \text{ such that } a_{jk} > 0\} \text{ and } I^-(j) = \{k \text{ such that } a_{jk} < 0\}$$

We shall use two convention of notations. The **first convention** is to shorten the notations using

$$k^\pm \text{ instead of } k \in I^\pm(j).$$

The **second convention** is to incorporate the boundary conditions on Γ^+ in the indices $k^+ \iff k \in I^+(j)$.

We set

$$m_{jk} = l_{jk} |a_{jk}| \quad \forall k, \quad \text{and } m_j^- = l_j^- |a_j^-|.$$

With all these notations the standard FV scheme which generalizes (1.23) to non constant velocity fields writes

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_{k^+} m_{jk} u_j^n - \sum_{k^-} m_{jk} u_k^n - m_j^- u_j^{-,n} = 0. \quad (2.2)$$

2.1.1 A practical remark about the calculation of a_{jk}

In dimension $d = 2$ a simple method makes possible the calculation of a_{jk} with little cost.

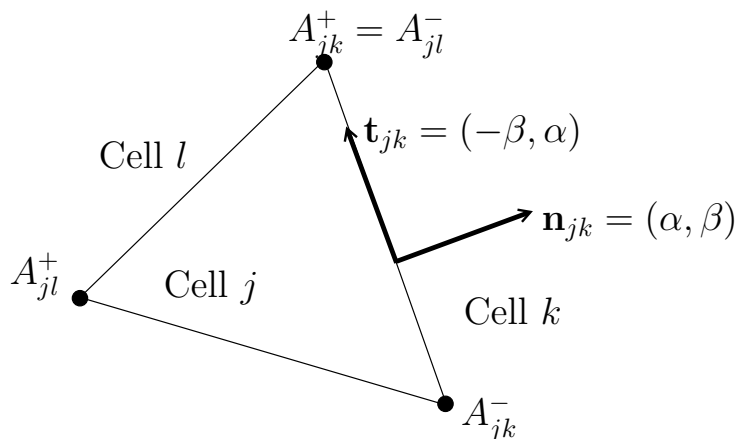


Figure 2.1: Orientation of the edge

Assume that \mathbf{a} is the rotational of a scalar potential $q \in C^2(\bar{\Omega})$

$$\mathbf{a} = \nabla \wedge q = (-\partial_{x_2} q, \partial_{x_1} q).$$

By construction $\nabla \cdot \mathbf{a} = \partial_{x_1}(-\partial_{x_2} q) + \partial_{x_2}(\partial_{x_1} q) = 0$. Then

$$\begin{aligned} a_{jk} &= \frac{1}{l_{jk}} \int_{\Sigma_{jk}} \nabla \wedge q \cdot \mathbf{n}_{jk} d\sigma \quad (\mathbf{n}_{jk} = (\alpha, \beta)) \\ &= \frac{1}{l_{jk}} \int_{\Sigma_{jk}} (-\partial_{x_2} q \alpha + \partial_{x_1} q \beta) d\sigma = -\frac{1}{l_{jk}} \int_{\Sigma_{jk}} \nabla q \cdot \mathbf{t} d\sigma = -\frac{1}{l_{jk}} \int_{\Sigma_{jk}} \frac{\partial q}{\partial \mathbf{t}} d\sigma, \end{aligned}$$

that is

$$a_{jk} = \frac{q(A_{jk}^-) - q(A_{jk}^+)}{l_{jk}}.$$

By convention (A_{jk}^-, A_{jk}^+) is oriented clockwise on the edge Σ_{jk} , and $A_{jk}^- = A_{kj}^+$. Note also the $A_{jl}^- = A_{jk}^+$ if the edge Σ_{jl} is after Σ_{jk} (clockwise).

Lemma 12 *By construction one has $\sum_k l_{jk} a_{jk} + l_j^- a_j^- = 0$.*

It is sufficient to consider the case where $l_j^- = 0$. One has

$$\sum_k l_{jk} a_{jk} = \sum_k l_{jk} \left(\frac{q(A_{jk^-}) - q(A_{jk^+})}{l_{jk}} \right) = \sum_k (A_{jk^-} - A_{jk^+}) = 0$$

because the contour is closed. Another proof comes from the free-divergence condition

$$\sum_k l_{jk} a_{jk} = \int_{\partial\Omega_j} \mathbf{a} \cdot \mathbf{n}_j \, d\sigma = \int_{\Omega_j} \nabla \cdot \mathbf{a} \, dx = 0.$$

Lemma 13 *One has $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk} + m_j^-$.*

This is immediate from the definition of m_{jk} and lemma 12.

2.1.2 The maximum principle

The exact solution of the transport equation satisfies a maximum principle. This is an immediate consequence of the representation formulas (1.14-1.15). For the discrete solution (2.2) one has

$$u_j^{n+1} = \left(1 - \frac{\Delta t}{s_j} \sum_{k^+} m_{jk} \right) u_j^n + \frac{\Delta t}{s_j} \sum_{k^+} m_{jk} u_k^n + \frac{\Delta t}{s_j} m_j^- u_j^{-,n}. \quad (2.3)$$

Lemma 14 *Assume the CFL condition*

$$\frac{\Delta t}{s_j} \sum_{k^+} m_{jk} \leq 1, \quad \forall j. \quad (2.4)$$

Then the FV scheme (2.2) satisfies the maximum principle under the form

$$\min \left(\min_k u_k^n, \min_l u_l^{-,n} \right) \leq u_j^{n+1} \leq \max \left(\max_k u_k^n, \max_l u_l^{-,n} \right). \quad (2.5)$$

Set $m = \min(\min_k u_k^n, \min_l u_l^{-,n})$ the minimum over all cells and incoming boundary conditions at the previous time step. We want to prove that $m \leq u_j^{n+1}$ for all cell j . Let us assume for simplicity that the cell is in the interior of the domain, that is $m_j^- = 0$. One has

$$u_j^{n+1} - m = \left(1 - \frac{\Delta t}{s_j} \sum_{k^+} m_{jk} \right) (u_j^n - m) + \frac{\Delta t}{s_j} \sum_{k^+} m_{jk} (u_k^n - m).$$

All coefficients, that is $1 - \frac{\Delta t}{s_j} \sum_{k^+} m_{jk}$ and the $\frac{\Delta t}{s_j} \sum_{k^+} m_{jk}$, are non negative and their sum is equal to one. So $u_j^{n+1} - m$ is a mean of the previous values of the discrete solutions and of the incoming boundary condition. Therefore $m \leq u_j^{n+1}$ for all cell j . The proof generalizes without difficulty to boundary cells such that $m_j^- > 0$. A similar inequality is proved for the upper bound. It ends the proof.

More generally let us consider a continuous and convex function $u \mapsto \varphi(u)$ such that

$$\varphi(\theta u_1 + (1 - \theta)u_2) \leq \theta \varphi(u_1) + (1 - \theta)\varphi(u_2)$$

for all u_1 and u_2 and for all $\theta \in [0, 1]$.

Lemma 15 *Assume the CFL condition (2.4). Then one has the inequality*

$$\sum_j s_j \varphi(u_j^{n+1}) \leq \sum_j s_j \varphi(u_j^n) + \Delta t \sum_j a m_j^- \varphi(u_j^{-,n}) - \Delta t \sum_j m_j^+ \varphi(u_j^n), \quad (2.6)$$

where $m_j^+ = m_{jk}$ for $k \in I^+(j)$ (that is for an outgoing edge).

Any convex function satisfies the inequality $\varphi(\sum \theta_i u_i) \leq \sum \theta_i \varphi(u_i)$ provided $\sum \theta_i u_i$ is a convex combination, that is $\theta_i \geq 0$ for all i and $\sum \theta_i = 1$. For any cell in the domain the equality (2.3) is a convex combination (all coefficients are non negative and their sum is equal to 1). Therefore

$$\varphi(u_j^{n+1}) \leq \left(1 - \frac{\Delta t}{s_j} \sum_{k^+} m_{jk}\right) \varphi(u_j^n) + \frac{\Delta t}{s_j} \sum_{k^+} m_{jk} \varphi(u_k^n).$$

Then we sum over all cells

$$\sum_j \varphi(u_j^{n+1}) \leq \sum_j \varphi(u_j^n) - \sum_j \sum_{k^+} \frac{\Delta t m_{jk}}{s_j} \varphi(u_j^n) + \sum_j \sum_{k^+} \frac{\Delta t m_{jk}}{s_j} \varphi(u_k^n).$$

One has

$$\sum_j \sum_{k^+} \frac{\Delta t m_{jk}}{s_j} \varphi(u_j^n) = \sum_j \sum_{k, a_{jk} > 0} \frac{\Delta t m_{jk}}{s_j} \varphi(u_j^n) + \sum_j \frac{\Delta t m_j^+}{s_j} \varphi(u_j^n)$$

and

$$\sum_j \sum_{k^-} \frac{\Delta t m_{jk}}{s_j} \varphi(u_k^n) = \sum_j \sum_{k, a_{jk} < 0} \frac{\Delta t m_{jk}}{s_j} \varphi(u_k^n) + \sum_j \frac{\Delta t m_j^-}{s_j} \varphi(u_j^{-,n}).$$

One has the equality

$$\sum_j \sum_{k, a_{jk} > 0} \frac{\Delta t m_{jk}}{s_j} \varphi(u_j^n) = \sum_j \sum_{k, a_{jk} < 0} \frac{\Delta t m_{jk}}{s_j} \varphi(u_k^n)$$

since these quantities concern the edges only in the interior of the domain. The rest of the proof is evident.

Lemma 16 *The FV scheme (2.2) is stable in all L^p under the same CFL (2.4). That is*

$$\|u^{n+1}\|_{L^p(\Omega)}^p \leq \|u^n\|_{L^p(\Omega)}^p + \Delta t \|u^{-,n}\|_{L^p(\Gamma^-)}^p, \quad 1 \leq p < \infty, \quad (2.7)$$

and

$$\|u^{n+1}\|_{L^\infty(\Omega)} \leq \max(\|u^n\|_{L^\infty(\Omega)}, \|u^{-,n}\|_{L^\infty(\Gamma^-)}). \quad (2.8)$$

First we consider the case $1 \leq p < \infty$. We apply the previous inequality with $\varphi(u) = |u|^p$. It yields (2.7). For the case $p = \infty$ we apply the maximum principle (2.5).

The inequality of lemma 15 offers the possibility of a different proof of the maximum principle. Assume the numerical solution is known at time step n . We also know the boundary condition of course. We set

$$m = \min \left(\min_k u_k^n, \min_l u_l^{-,n} \right)$$

and

$$M = \max \left(\max_k u_k^n, \max_l u_l^{-,n} \right).$$

We want to prove that, under CFL,

$$m \leq u_j^{n+1} \leq M, \quad \forall j.$$

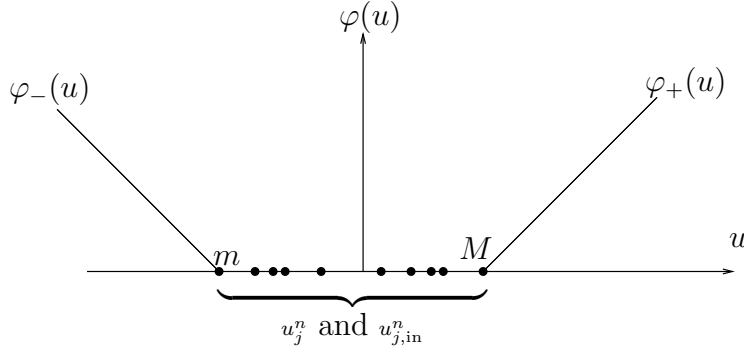


Figure 2.2: φ_- and φ_+

We first design a function

$$\varphi_-(u) = \begin{cases} m - u & \text{for } u \leq m, \\ 0 & \text{for } m \leq u. \end{cases}$$

Then φ_- is continuous and convex. So we apply the inequality (15). By construction the right hand side vanishes. Therefore

$$\sum_j s_j \varphi_-(u_j^{n+1}) \leq 0.$$

But φ_- is a non negative function. So $\varphi_-(u_j^{n+1}) = 0$ for all j , which turns into $m \leq u_j^{n+1}$ for all j . For the other inequality, namely $u_j^{n+1} \leq M$, we design a second continuous, non negative and convex function

$$\varphi_+(u) = \begin{cases} 0 & \text{for } u \leq M, \\ u - M & \text{for } M \leq u. \end{cases}$$

Using a similar argument, we obtain the claim: the maximum principle holds under CFL.

2.1.3 Geometrical analysis of the CFL condition

In this section, we analyze the geometrical meaning of the CFL condition. This inequality is mainly a restriction on the time step Δt , that is Δt must be smaller than a certain given quantity

$$\Delta t \leq \frac{s_j}{\sum_{k^+} m_{jk}} \quad \forall j, \quad (2.9)$$

which depends on the geometry of the mesh. This inequality needs to be satisfied for all j to guaranty the stability in practice. In order to simplify the analysis we take \mathbf{a} constant, that is

$$\mathbf{a} \in \mathbb{R}^2 \text{ is independent of } \mathbf{x} \in \Omega. \quad (2.10)$$

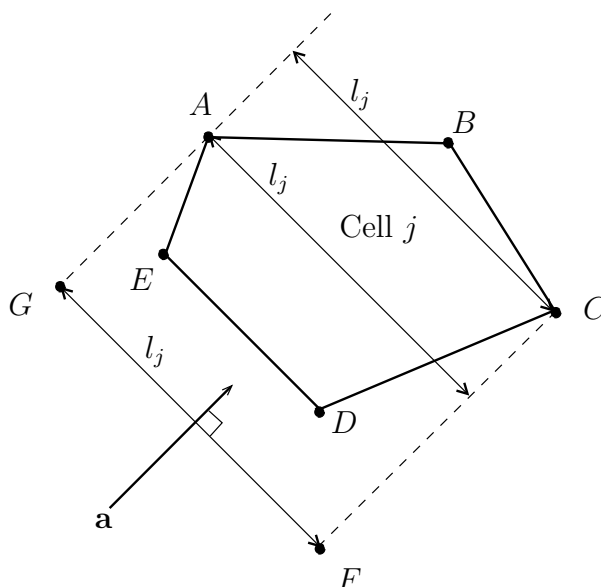


Figure 2.3: Apparent length of the cell

The apparent length l_j of a cell Ω_j is the transverse length of the cell, visible from an observatory in the direction of \mathbf{a} .

Lemma 17 *Assume the cell j is convex. Then the stability inequality (2.9) may be rewritten as*

$$\Delta t \leq \frac{s_j}{|\mathbf{a}|l_j}. \quad (2.11)$$

For example consider the pentagonal convex cell Ω_j with vertices $ABCDE$ in the figure 2.3. We construct a bigger cell $\Omega_{j'}$ with vertices $ABCFG$, where the lines AG and CF are parallel to the vector \mathbf{a} . Since Ω_j is convex and \mathbf{a} is constant, the outgoing edges $k \in I^+(j)$ (i.e. AB and BC on the figure) form a connected broken line. Similarly the incoming edges $k \in I^-(j)$ (i.e. CD , DE

and EA on the figure) form a connected broken line. For the bigger cell $\Omega_{j'}$, the outgoing edges $k \in I^+(j')$ are the same as for Ω , that is

$$I^+(j) = I^+(j').$$

So

$$\sum_{k \in I^+(j)} m_{jk} = \sum_{k \in I^+(j')} m_{jk} = \sum_{k \in I^-(j')} m_{jk} = |\mathbf{a}|l_j.$$

The edges AG and CF being parallel to \mathbf{a} , they do not contribute. Therefore (2.9) is equivalent to the claim.

Example 3 *If the cell is non convex, then*

$$\sum_{k \in I^+(j)} m_{jk} > |\mathbf{a}|l_j$$

and the time step restriction is more restrictive than in (2.11).

For a given convex cell Ω_j we define r_j^- the biggest radius of all interior circles and r_j^+ the smallest radius of all exterior circles. Notice that

$$\text{diam}(\Omega_j) = 2r_j^+.$$

Lemma 18 *Assume the cell j is convex. Then the stability inequality (2.9) is true as soon as*

$$\Delta t \leq \frac{\pi(r_j^-)^2}{2|\mathbf{a}|r_j^+}. \quad (2.12)$$

By definition $s_j \geq \pi(r_j^-)^2$ and $l_j \leq 2r_j^+$. Therefore (2.11) is a consequence of (2.12). It ends the proof.

Definition 6 *We define the quality factor (or aspect ratio) of the mesh*

$$Q = \max_j \left(\frac{r_j^+}{r_j^-} \right) \geq 1,$$

and the characteristic length of the mesh

$$h = \max_j (\text{diam}(\Omega_j)).$$

Note that the definition of h is an alternative to a similar definition already encountered in (1.17). With this definition, the time restriction (2.12) is true as soon as

$$|\mathbf{a}| \left(\frac{Q^2}{\pi} \right) \frac{\Delta t}{h} \leq 1.$$

In practice, our interest is to have the greatest time step as possible. Since h is proportional the maximal length in the mesh, it is related to the total number of cells in the mesh. We consider it is somehow a given quantity. On the contrary Q is not given, in the sense that it is related to the shape of cells which we consider is a freedom in the definition of the cells. This inequality

shows the interest to have the smallest Q as possible for time step requirements, that is to have the best aspect ratio.

At this point it is necessary to make the following remark whose importance will appear later: **the quality factor is also an important quantity to bound in order to establish a convergence result.** Consider for example a sequence of mesh $n = 0, 1, 2, \dots$. Each mesh has its own characteristic length h_n and its own quality factor Q_n . We make the assumption that the characteristic length tends to zero

$$h_n \rightarrow 0 \text{ as } n \rightarrow \infty$$

which means that the cells are smaller and smaller. So what about the quality factor?

Definition 7 *A sequence of mesh is regular if Q_n is bounded uniformly with respect to the characteristic length h*

$$1 \leq Q_n \leq C, \quad \forall n.$$

With a little abuse of language, we shall say that a given mesh is regular if its quality factor Q is bounded uniformly with respect to the important variables of the analysis.

Proves of convergence will be given for such regular meshes.

2.2 Default of consistency

It is well known that a linear scheme which is stable and consistent is convergent. Unfortunately FV schemes are not consistent in the Finite Difference sense. This is the consistency default of FV schemes. We will show that FV schemes are nevertheless convergent.

2.2.1 What is consistency in the FD sense?

Consider the FD scheme (1.4) for advection in dimension $d = 1$

$$\begin{cases} \frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0, \\ u_j^0 = u_0(j\Delta x). \end{cases} \quad (2.13)$$

The initial condition is $x \mapsto u_0(x)$. We can also write

$$u^{n+1} = (I + \Delta t A_{\Delta x}) u^n, \text{ where } u^n = (u_j^n)_{j \in \mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$$

is the vector of unknowns, I is the identity operator in $\mathbb{R}^{\mathbb{Z}}$ and $A_{\Delta x} : \mathbb{R}^{\mathbb{Z}} \rightarrow \mathbb{R}^{\mathbb{Z}}$ is the operator defined by

$$A_{\Delta x} u = (w_j) \text{ with } w_j = -a \frac{u_j^n - u_{j-1}^n}{\Delta x}.$$

To show the convergence we wish to compare the numerical solution u^n with the exact solution v^n at grid points

$$v_j^n = u(n\Delta t, j\Delta x) = u_0(j\Delta x - a n\Delta t).$$

Let us define the error

$$e^n = v^n - u^n.$$

Lemma 19 *The error satisfies the iterative process*

$$\begin{cases} \frac{e^{n+1} - e^n}{\Delta t} = A_{\Delta x} e^n + r^n, \\ e^0 = 0, \end{cases}$$

where $r^n = (r_j^n)$ is the truncation error: $r_j^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} + a \frac{v_j^n - v_{j-1}^n}{\Delta x}$.

This is evident. One makes the difference between (2.13) and the definition of the truncation error.

To go further in the analysis, let us assume that the initial data u_0 is sufficiently smooth, $u_0 \in W^{2,\infty}(\mathbb{R})$ for example. In this case the solution u is also twice differentiable, all second derivatives in time and space being bounded. Then one has the following Taylor expansions

$$v_j^{n+1} = v_j^n + \Delta t \partial_t u(n\Delta t, j\Delta x) + (\Delta t^2 \|\partial_t^2 u\|_\infty) \alpha_j^n, \quad |\alpha_j^n| \leq \frac{1}{2},$$

and

$$v_{j-1}^n = v_j^n - \Delta x \partial_x u(n\Delta t, j\Delta x) + (\Delta x^2 \|\partial_x^2 u\|_\infty) \beta_j^n, \quad |\beta_j^n| \leq \frac{1}{2}.$$

Plugging in r_j^n we obtain

$$r_j^n = \partial_t u(n\Delta t, j\Delta x) + (\Delta t \|\partial_t u\|_\infty) \alpha_j^n + a \partial_x u(n\Delta t, j\Delta x) + a (\Delta x \|\partial_x^2 u\|_\infty) \beta_j^n.$$

Noticing that $\|\partial_t^2 u\|_\infty = a^2 \|\partial_x^2 u_0\|_\infty$ and $\|\partial_x^2 u\|_\infty = \|\partial_x^2 u_0\|_\infty$, we obtain

$$r_j^n = a (\Delta x \alpha_j^n + a \Delta t \beta_j^n) \|\partial_x^2 u_0\|_\infty.$$

Making no restriction, we assume stability of the scheme which is insured for $\Delta t \leq \frac{\Delta x}{a}$. It implies

$$\|r^n\|_\infty \leq a \Delta x \|\partial_x^2 u_0\|_\infty. \quad (2.14)$$

Since the residual tends to zero with Δx , then we say the scheme is consistent in L^∞ .

Under CFL, one also has the stability estimate $\|I + \Delta t A_{\Delta x}\|_\infty \leq 1$ which is another way of writing the inequality (2.7) in L^∞

$$\|u^{n+1}\|_\infty = \|(I + \Delta t A_{\Delta x}) u^n\|_\infty \leq \|u^n\|_\infty, \quad \forall u^n.$$

Lemma 20 *Consider a smooth initial condition $u_0 \in W^{2,\infty}(\mathbb{R})$. Assume the CFL condition is true. Let $T > 0$ be some given time. Then for all time $t_n = n\Delta t \leq T$, one has the error estimate*

$$\|e^n\|_\infty \leq \|\partial_x^2 u_0\|_\infty (aT\Delta x). \quad (2.15)$$

One says the scheme is first order because of the first order dependence with respect to Δx .

The proof is immediate. One has $e^{n+1} = (I + \Delta t A_{\Delta x}) e^n + \Delta t r^n$. So

$$\|e^{n+1}\|_\infty \leq \|(I + \Delta t A_{\Delta x}) e^n\|_\infty + \Delta t \|r^n\|_\infty,$$

that is with the stability estimate $\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + \Delta t \|r^n\|_\infty$, and therefore, since $e^0 = 0$, $\|e^n\|_\infty \leq \Delta t \sum_{p=0}^{n-1} \|r^p\|_\infty$. Using (2.14) the claim is proved.

Definition 8 Consistency. *In dimension d , we say that a scheme is consistent in the Finite Difference setting if there exist points attached to the cells (denoted $\mathbf{x}_j \in \Omega$ in the following) such that the truncation error tends to zero in L^∞ uniformly with the mesh size (provided the exact solution is sufficiently differentiable).*

In practice it is sufficient to satisfy an inequality like

$$\|r^n\|_{L^\infty} \leq C(\dots) \Delta x^p$$

where $p > 0$ and $C(\dots)$ is independent of the mesh size $h = \Delta x$ for the scheme to be consistent. The constant $C(\dots)$ depends on the derivatives of u .

Fractional orders of convergence are common for an initial data for which only the first derivative is bounded. In this direction we quote the result of lemma 21. The method of the proof of lemma 21 has its own interest. It consists in the application of lemma 20 together with a regularization of the initial data.

Definition 9 Regularization. *Let $\varphi \in C_0^1(\mathbb{R})$ be a non negative function such that*

$$\int_{\mathbb{R}} \varphi(z) dz = 1.$$

For a given function $w \in W^{1,\infty}(\mathbb{R})$ we regularize w by convolution and define a new function w_ε

$$w_\varepsilon(x) = \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) w(y) dy. \quad (2.16)$$

Lemma 21 *One has the inequalities*

$$\|w_\varepsilon\|_\infty \leq \|w\|_\infty, \quad \|w'_\varepsilon\|_\infty \leq \|w'\|_\infty, \quad \|w''_\varepsilon\|_\infty \leq \frac{\int |\varphi'(z)| dz}{\varepsilon} \|w'\|_\infty, \quad (2.17)$$

and

$$\|w_\varepsilon - w\|_\infty \leq \varepsilon \int \varphi(z) |z| dz \|\partial_x w\|_\infty, \quad (2.18)$$

This is completely standard [2]. From the definition of w^ε one has the inequality

$$|w^\varepsilon(x)| \leq \left(\frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) dy \right) \|w\|_\infty.$$

Since $\frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) dy = \int_{\mathbb{R}} \varphi(z) dz = 1$, then it proves immediately $\|w^\varepsilon\|_\infty \leq \|w\|_\infty$.

One also has the equality

$$w'_\varepsilon(x) = -\frac{1}{\varepsilon^2} \int_{\mathbb{R}} \varphi' \left(\frac{x-y}{\varepsilon} \right) w(y) dy = \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi \left(\frac{x-y}{\varepsilon} \right) w'(y) dy$$

from which shows $\|w'_\varepsilon\|_\infty \leq \|w'\|_\infty$.

Then one considers the equality

$$w''_\varepsilon(x) = -\frac{1}{\varepsilon^2} \int_{\mathbb{R}} \varphi' \left(\frac{x-y}{\varepsilon} \right) w'(y) dy$$

which turns into

$$|w''_\varepsilon(x)| \leq \left(\frac{1}{\varepsilon^2} \int_{\mathbb{R}} \varphi' \left(\frac{x-y}{\varepsilon} \right) |w'(y)| dy \right) \|w'\|_\infty = \frac{\int |\varphi'(z)| dz}{\varepsilon} \|w'\|_\infty.$$

It remains to prove (2.18). One has by construction

$$w_\varepsilon(x) - w(x) = \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi \left(\frac{x-y}{\varepsilon} \right) (w(y) - w(x)) dy$$

therefore

$$|w_\varepsilon(x) - w(x)| \leq \left(\frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi \left(\frac{x-y}{\varepsilon} \right) |x-y| dy \right) \|w'\|_\infty = \varepsilon \int \varphi(z) |z| dz \|w'\|_\infty.$$

The proof is finished.

Lemma 22 *Consider a smooth initial condition $u_0 \in W^{1,\infty}(\mathbb{R})$. Assume the CFL condition is true. Let $T > 0$ be some given time. Then for all time $t_n = n\Delta t \leq T$, one has the error estimate*

$$\|e^n\|_\infty \leq \frac{4}{\sqrt{3}} \|\partial_x u_0\|_\infty \sqrt{aT\Delta x}.$$

Let us define the regularization of the initial data

$$u_{0,\varepsilon}(x) = \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi \left(\frac{x-y}{\varepsilon} \right) u_0(y) dy.$$

The numerical solution issued from this regularized initial data is u_ε^n

$$\begin{cases} \frac{u_\varepsilon^{n+1} - u_\varepsilon^n}{\Delta t} = A_{\Delta x} u_\varepsilon^n, \\ u_{\varepsilon,j}^n = u_{0,\varepsilon}(j\Delta x). \end{cases}$$

One has the triangular inequality

$$\|e^n\|_\infty = \|v^n - u^n\|_\infty \leq \|v^n - v_\varepsilon^n\|_\infty + \|v_\varepsilon^n - u_\varepsilon^n\|_\infty + \|u_\varepsilon^n - u^n\|_\infty. \quad (2.19)$$

Here $v_\varepsilon^n = (u_\varepsilon(n\Delta t, j\Delta x))_{j \in \mathbb{Z}}$. Our task is to bound each of the right hand side terms.

The stability of the scheme shows that $\|u_\varepsilon^n - u^n\|_\infty \leq \|u_\varepsilon^0 - u^0\|_\infty \leq \|u_{0,\varepsilon} - u_0\|_\infty$. By commutation of regularization and advection, one has that $\|v^n - v_\varepsilon^n\|_\infty \leq \|u_{0,\varepsilon} - u_0\|_\infty$. Using (2.17) one gets $\|u_{0,\varepsilon} - u_0\|_\infty \leq \varepsilon \int \varphi(z) |z| dz \|u'_0\|_\infty$.

Using (2.15) for the regularized solution for which the second derivative may be bounded using (2.17), one gets

$$\|v_\varepsilon^n - u_\varepsilon^n\|_\infty = \|e_\varepsilon^n\|_\infty \leq \frac{\int |\varphi'(z)| dz}{\varepsilon} \|u'_0\|_\infty (aT\Delta x).$$

Plugging these inequalities in (2.19) we find that

$$\|e^n\|_\infty \leq \left(2\varepsilon \int \varphi(z)|z| dz + \frac{\int |\varphi'(z)| dz}{\varepsilon} aT\Delta x \right) \|u'_0\|_\infty.$$

The optimal value of ε is the one that gives the smallest result, namely $\varepsilon = \left(\frac{aT\Delta x \int |\varphi'(z)| dz}{2 \int \varphi(z)|z| dz} \right)^{\frac{1}{2}}$. Finally

$$\|e^n\|_\infty \leq 2 \left(2aT\Delta x \int |\varphi'(z)| dz \int \varphi(z)|z| dz \right)^{\frac{1}{2}} \|u'_0\|_\infty.$$

Take for simplicity φ with compact support

$$\varphi(z) = 1 - |z| \text{ for } |z| \leq 1, \text{ and } \varphi(z) = 0 \text{ for } |z| \geq 1.$$

Then $\int |\varphi'(z)| dz \times \int \varphi(z)|z| dz = \frac{2}{3}$. The rest of the proof is evident.

2.2.2 Consistency of FV

Let us analyze the consistency of the general Finite Volume scheme for advection (2.2) or (2.3) for general meshes. It is sufficient to consider a cell interior to the domain of computation. The truncation error of the scheme is by definition

$$\begin{aligned} r_j^n &= \frac{v_j^{n+1} - v_j^n}{\Delta t} + \frac{1}{s_j} \sum_{k^+} m_{jk} v_j^n - \frac{1}{s_j} \sum_{k^-} m_{jk} v_k^n \\ &= \frac{v_j^{n+1} - v_j^n}{\Delta t} - \frac{1}{s_j} \sum_{k^-} m_{jk} (v_k^n - v_j^n). \end{aligned}$$

where v_j^n , v_j^{n+1} and the v_k^n 's are a priori some point-wise representation of the exact solution

$$v_j^n = u(n\Delta t, \mathbf{x}_j).$$

The point \mathbf{x}_j can be chosen to be equal to the center of mass \mathbf{G}_j , but it is much better to relax this restriction. Let us assume that u_0 is in $W^{2,\infty}$. Then one has the Taylor expansions

$$v_j^{n+1} = v_j^n + \partial_t u(n\Delta t, \mathbf{x}_j) \Delta t + O(\Delta t^2) = v_j^n - \mathbf{a} \cdot \nabla u(n\Delta t, \mathbf{x}_j) \Delta t + O(\Delta t^2)$$

and

$$v_k^n = v_j^n + \nabla u(n\Delta t, \mathbf{x}_j) \cdot (\mathbf{x}_k - \mathbf{x}_j) + O(h^2)$$

where the characteristics length of the mesh is such that

$$\sup_{k \in I(j)} |\mathbf{x}_j - \mathbf{x}_k| \leq Ch \text{ with } C \text{ independent of } h.$$

Here $k \in I(j) = I^+(j) \cap I^-(j)$ describes the set of all neighbors of cell Ω_j . So

$$r_j^n = -\mathbf{a} \cdot \nabla u(n\Delta t, \mathbf{x}_j) \Delta t - \frac{1}{s_j} \sum_{k^-} l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk} \nabla u(n\Delta t, \mathbf{x}_j) \cdot (\mathbf{x}_k - \mathbf{x}_j) + O(\Delta t) + O(h),$$

that is

$$r_j^n = (\mathbf{M}_{jk}^t \mathbf{a}) \cdot \nabla u(n\Delta t, \mathbf{x}_j) + O(\Delta t) + O(h). \quad (2.20)$$

where the matrix is $\mathbf{M}_{jk} = -\mathbf{I} + \frac{1}{s_j} \sum_{k^-} l_{jk} \mathbf{n}_{jk} \otimes (\mathbf{x}_k - \mathbf{x}_j)$. So in order for r_j^n to be small like $O(\Delta t) + O(h)$ then we need the vector $\mathbf{M}_{jk}^t \mathbf{a}$ to be equal to zero.

Definition 10 Strong consistency *We say that the FV scheme is strongly consistent iff there exists points (\mathbf{x}_j) which solve the equation $\mathbf{M}_{jk} = 0$, that is*

$$\sum_{k^-} l_{jk} \mathbf{n}_{jk} \otimes (\mathbf{x}_k - \mathbf{x}_j) = s_j \mathbf{I}, \quad \forall j. \quad (2.21)$$

If moreover $\mathbf{x}_j \in \overline{\Omega_j}$ then we say that it is a local solution.

Notice that the sum is taken over all k^- , which still shows a dependency with respect to \mathbf{a} . This definition of strong consistency will be weakened later one.

2.2.3 Strong consistency in 1D

Let us consider the advection equation $\partial_t u + a \partial_x u = 0$ in dimension $d = 1$. Let us assume that $a > 0$. The numbering of the cells is logical, from the left to the right. So by construction $I^-(j) = \{j - 1\}$. So the consistency equation (2.21) simplifies into

$$x_j - x_{j-1} = \Delta x_j.$$

If $a < 0$ the consistency equation (2.21) writes

$$x_{j+1} - x_j = \Delta x_j.$$

Lemma 23 *In dimension $d = 1$, there exists one and only one local solution of (2.21). The solution is*

$$\begin{cases} \text{if } a > 0 & \text{then } x_j = x_{j+\frac{1}{2}} \forall j, \\ \text{if } a < 0 & \text{then } x_j = x_{j-\frac{1}{2}} \forall j. \end{cases}$$

As a consequence the FV scheme for transport is strongly consistent in dimension one and is therefore convergent with optimal rates. Set

$$h = \sup_j \Delta x_j \infty.$$

The projection of the exact solution on the mesh is

$$v_j^n = \frac{1}{\Delta x_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(n\Delta t, x) dx. \quad (2.22)$$

Lemma 24 Consider the FV scheme (1.11) for advection in dimension $d = 1$, with the projection wise initial condition $u^0 = v^0$. Assume the initial condition is smooth $u_0 \in W^{2,\infty}$. Assume the CFL condition holds. Then

$$\|u^n - v^n\|_\infty \leq (2\|u'_0\|_\infty + aT\|u''_0\|_\infty)h, \quad n\Delta t \leq T. \quad (2.23)$$

If the initial condition is less regular $u_0 \in W^{1,\infty}$, then one has

$$\|u^n - v^n\|_\infty \leq 2\|u'_0\|_\infty h + \frac{4}{\sqrt{3}}\|u'_0\|_\infty \times \sqrt{aTh}, \quad n\Delta t \leq T. \quad (2.24)$$

The proof is fundamentally the same than the one of lemma 20. But we need to evaluate the interpolation error between the projection v^n of the exact solution and w^n which is the pointwise value of the exact solution. This is the reason of the extra term $2\|u'_0\|_\infty h$ in (2.23). Note that this term does not depend upon the time variable T .

Assuming $a > 0$ the pointwise value of the exact solution is defined by

$$w_j^n = u(n\Delta t, x_{j+\frac{1}{2}}). \quad (2.25)$$

We also define z^n which is the numerical solution issued from the pointwise initial condition. One has the triangular inequality

$$\|u^n - v^n\|_\infty \leq \|u^n - z^n\|_\infty + \|z^n - w^n\|_\infty + \|w^n - v^n\|_\infty. \quad (2.26)$$

Since the scheme is stable $\|u^n - z^n\|_\infty \leq \|u^n - z^0\|_\infty = \|v^0 - w^0\|_\infty$. It is evident using a Taylor expansion that the projection-wise and the pointwise initial condition are $O(h)$, that is $\|v^0 - w^0\|_\infty \leq h\|u'_0\|_\infty$. Similarly a Taylor expansion shows that $\|v^n - w^n\|_\infty \leq h\|u'(n\Delta t, \cdot)\|_\infty = h\|u'_0\|_\infty$.

Finally the last term $\|z^n - w^n\|_\infty$ is bounded using the consistency of the scheme that we have established for the pointwise definition v^n . Let us recall the main ingredient of the proof. Define the truncation error

$$r_j^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} + a \frac{v_j^n - v_{j-1}^n}{\Delta x_j}.$$

Then standard Taylor expansions based on (2.25) show that

$$\|r^n\|_\infty \leq ah\|u''_0\|_\infty.$$

This inequality is the generalization of (2.14). So $\|z^n - w^n\|_\infty \leq \|u''_0\|_\infty(aTh)$. Plugging in (2.26) it ends the proof of (2.23).

If the initial condition is less regular, namely $u_0 \in W^{1,\infty}$, it does not affect the interpolation error which is still bounded by $2\|u'_0\|_\infty h$. The term $\|z^n - w^n\|_\infty$ is bounded using the regularization technique (2.16) used in the proof of lemma 21. The proof is completed.

2.2.4 Strong consistency in 2D

Surprisingly the situation is much more complicated in dimension greater than because strong consistency does not hold in the general case. We begin with an example which shows that strong consistency is true for cartesian grids.

Example 4 Consider a cartesian mesh in dimension $d = 2$. Assume that $\mathbf{x}_j = \mathbf{G}_j$ for all cell j . Then the consistency requirement (2.21) is true for all \mathbf{a} .

It comes from $s_j = \Delta x^2$, $l_{jk} = \Delta x$ and $\mathbf{x}_k - \mathbf{x}_j = \Delta x \mathbf{n}_{jk}$. In the sum (2.21) only two edges contribute. The proof is ended.

However it is not possible to find, in the general case, some points (\mathbf{x}_j) such that (2.21) is true. Indeed let us consider that $\mathbf{x}_j \in \mathbb{R}^d$ is the unknown in (2.21). The system (2.21) is a linear system of d^2 scalar equations. So the number of independent unknowns per cell is d , while the number of equations d^2 . This is incompatible.

Example 5 Consider the triangular cells of figure 2.4. There is no solution for the strong consistency requirement (2.21).

Consider the figure 2.4. Only the cell k is in $I^-(j)$. Therefore the sum in (2.21) reduces to a single contribution $l_{jk} \mathbf{n}_{jk} \otimes (\mathbf{x}_k - \mathbf{x}_j)$. This matrix is of rank one whatever are \mathbf{x}_j and \mathbf{x}_k . Therefore it cannot be equal to \mathbf{I} .

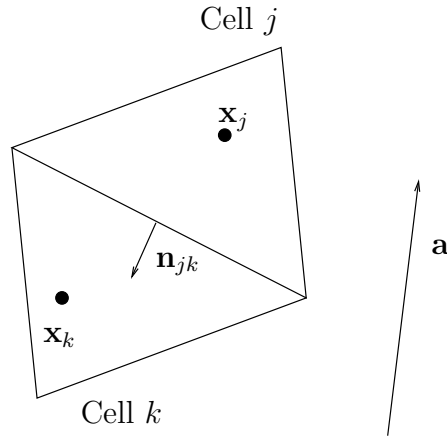


Figure 2.4: Exemple

It means that the strong form of the consistency equation (2.21) has no solution in the general case, that is for general meshes in dimension greater than one.

In order to make progress in the study of consistency, one can study the weaker equation $\mathbf{M}_{jk}^t \mathbf{a} = 0$, which writes also

$$\mathbf{x}_j = \sum_{k^-} \left(\frac{l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}}{\sum_{r^-} l_{jr} \mathbf{a} \cdot \mathbf{n}_{jr}} \right) \mathbf{x}_k - \frac{s_j}{\sum_{k^-} l_{jk} \mathbf{a} \cdot \mathbf{n}_{jk}} \mathbf{a}. \quad (2.27)$$

Indeed if (2.27) is true, then the truncation error (2.20) is $O(\Delta t) + O(h)$. To solve (2.27) a possibility is to consider the \mathbf{x}_k s are given, so it determines the point \mathbf{x}_j as an average value of the \mathbf{x}_k s plus a geometric correction. In some sense this equation propagates the definition of \mathbf{x}_j from cell to cell. It is quite difficult to study the solutions of this equation in the general case. Even the

fact that the solution is local $\mathbf{x}_j \in \Omega_j$ is not evident. However we refer to [4] for some results concerning the consistency analysis based (2.27). The example 6, inspired from [4], shows that the solution of (2.27) may diverge: the distance between \mathbf{x}_j and the cell may go to infinity.

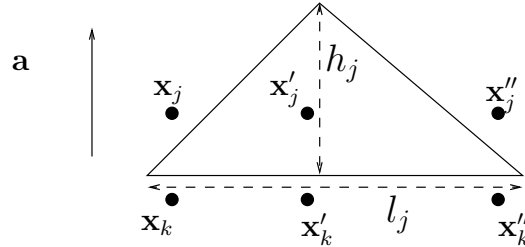


Figure 2.5: The equation (2.27) becomes $\mathbf{x}_j = \frac{h_j}{2|\mathbf{a}|}\mathbf{a} + \mathbf{x}_k$. The height of the triangle is $h_j = \frac{s_j}{l_j}$. So if the right hand side of (2.27) is \mathbf{x}_k , \mathbf{x}'_k or \mathbf{x}''_k , then $\mathbf{x}_j \in \Omega_j$ or not.

Example 6 Consider the following mesh of the quarter of the plane. The mesh

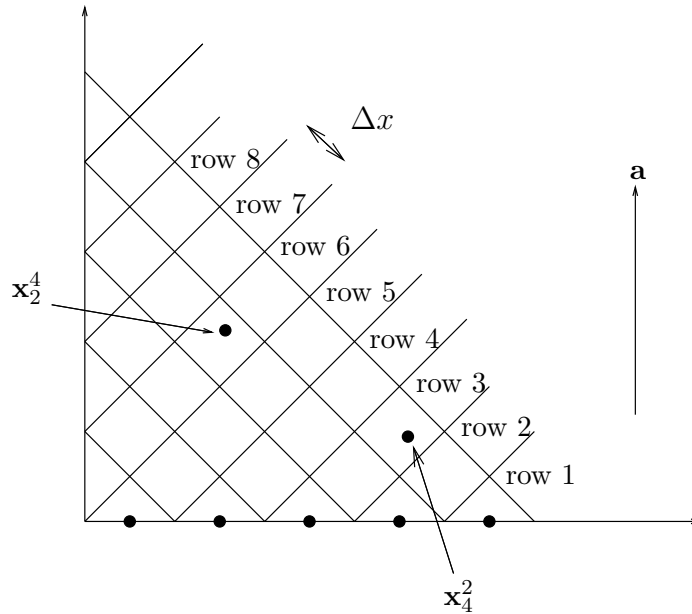


Figure 2.6: The mesh of example 6

is made with Cartesian cells except at the boundaries where the cells are triangles. The mesh is divided in rows. A new ordering is used for simplicity. The cell in row p at position j (counting from the left boundary) is index \cdot_j^p . Define the points (\mathbf{x}_j^p) which are the solutions of the equations (2.27). We assume that

all points in row 0 are centered $\mathbf{x}_j^0 = \frac{\Delta x}{\sqrt{2}}(j-1)$. Then

$$\sup_j |\mathbf{x}_1^p - \mathbf{G}_1^p| = \infty. \quad (2.28)$$

All angles are 45 degrees. Equation (2.27) writes

$$\mathbf{x}_1^{2q+1} = \mathbf{x}_1^{2q} + \frac{\Delta x}{\sqrt{2}}\mathbf{a}, \quad \forall q, \quad (2.29)$$

$$\mathbf{x}_j^{2q+1} = \frac{1}{2}\mathbf{x}_{j-1}^{2q} + \frac{1}{2}\mathbf{x}_j^{2q} + \frac{\Delta x}{\sqrt{2}}\mathbf{a}, \quad \forall q, \forall j \geq 2, \quad (2.30)$$

and

$$\mathbf{x}_j^{2q} = \frac{1}{2}\mathbf{x}_j^{2q-1} + \frac{1}{2}\mathbf{x}_{j+1}^{2q-1} + \frac{\Delta x}{\sqrt{2}}\mathbf{a}, \quad \forall q, \forall j \geq 1. \quad (2.31)$$

Define $\mathbf{y}_j^p = \mathbf{x}_j^p - \mathbf{G}_j^p$ which is the difference between \mathbf{x}_j^p and the center of mass. Note the the centers of mass \mathbf{G}_j^p satisfy the two last equations (2.30-2.31). So \mathbf{y}_j^p satisfies by construction homogeneous equations. Define $\mathbf{b} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$. One has the equations

$$\begin{cases} \mathbf{y}_0^{2q+1} = \mathbf{y}_0^{2q} + \frac{\Delta x}{\sqrt{2}}\mathbf{b}, & \forall q, \\ \mathbf{y}_j^{2q+1} = \frac{1}{2}\mathbf{y}_{j-1}^{2q} + \frac{1}{2}\mathbf{y}_j^{2q}, & \forall q, \forall j \geq 1, \\ \mathbf{y}_j^{2q} = \frac{1}{2}\mathbf{y}_j^{2q-1} + \frac{1}{2}\mathbf{y}_{j+1}^{2q-1}, & \forall q, \forall j \geq 0. \end{cases}$$

Notice that $\mathbf{y}_j^0 = 0$ for all j (in row 0). Therefore the vector $\mathbf{y}_j^p = \begin{pmatrix} \alpha_j^p \\ 0 \end{pmatrix}$ has two components, the second component being always 0 by recurrence, and the first one being non positive everywhere by recurrence: that is $\alpha_j^p \leq 0$. The recurrence relation is ($\gamma = -\frac{\Delta x}{\sqrt{2}} < 0$)

$$\begin{cases} \alpha_0^{2q+1} = \alpha_0^{2q} + \gamma, & \forall q, \\ \alpha_j^{2q+1} = \frac{1}{2}\alpha_{j-1}^{2q} + \frac{1}{2}\alpha_j^{2q}, & \forall q, \forall j \geq 1, \\ \alpha_j^{2q} = \frac{1}{2}\alpha_j^{2q-1} + \frac{1}{2}\alpha_{j+1}^{2q-1}, & \forall q, \forall j \geq 0, \end{cases}$$

that is after elimination of the even superscripts

$$\begin{cases} \alpha_0^{2q+1} = \frac{1}{2}\alpha_0^{2q-1} + \frac{1}{2}\alpha_1^{2q-1} + \gamma, & \forall q, \\ \alpha_j^{2q+1} = \frac{1}{4}\alpha_{j-1}^{2q-1} + \frac{1}{2}\alpha_j^{2q-1} + \frac{1}{4}\alpha_{j+1}^{2q-1}, & \forall q, \forall j \geq 1. \end{cases}$$

The initial condition is simply $\alpha_j^{-1} = 0$ for all j . To find the analytical solution we symetrize α : we set $\beta_j^{2q+1} = \alpha_{|j|}^{2q+1}$. The recurrence relation becomes

$$\begin{cases} \beta_0^{2q+1} = \frac{1}{4}\beta_{-1}^{2q-1} + \frac{1}{2}\beta_0^{2q-1} + \frac{1}{4}\beta_1^{2q-1} + \gamma, & \forall q, \\ \beta_j^{2q+1} = \frac{1}{4}\beta_{j-1}^{2q-1} + \frac{1}{2}\beta_j^{2q-1} + \frac{1}{4}\beta_{j+1}^{2q-1}, & \forall q, \forall j \neq 0. \end{cases}$$

Define δ_j^{2q+1} such that

$$\begin{cases} \delta_0^{2q+1} = \frac{1}{4}\delta_{-1}^{2q-1} + \frac{1}{2}\delta_0^{2q-1} + \frac{1}{4}\delta_1^{2q-1} + \gamma, & \forall q, \\ \delta_j^{2q+1} = \frac{1}{4}\delta_{j-1}^{2q-1} + \frac{1}{2}\delta_j^{2q-1} + \frac{1}{4}\delta_{j+1}^{2q-1}, & \forall q, \forall j \neq 0, \end{cases}$$

with the initial condition $\delta_j^{-1} = 0$ for all j , so that, by construction, $\beta_j^{2q+1} = \sum_{r=0}^q \delta_j^{2r+1}$. The analytical value of this last recurrence is easy the binomial coefficient $\delta_j^{2r+1} = \frac{1}{2^{2r}} \binom{2r}{r-j}$. Therefore

$$\alpha_0^{2q+1} = \gamma \sum_{r=0}^q \frac{1}{2^{2r}} \frac{(2r)!}{(r!)^2}.$$

The Stirling formula is $n! \approx \frac{1}{\sqrt{2\pi n}} \left(\frac{n}{e}\right)^n$. Then

$$\frac{1}{2^{2r}} \frac{(2r)!}{(r!)^2} \approx \frac{1}{2^{2r}} \frac{\frac{1}{\sqrt{2\pi 2r}} \left(\frac{2r}{e}\right)^{2r}}{\left(\frac{1}{\sqrt{2\pi r}} \left(\frac{r}{e}\right)^r\right)^2} = \gamma \sqrt{2\pi} \frac{1}{\sqrt{r}}$$

for large rs . We deduce that α_0^{2q+1} diverges

$$\alpha_0^{2q+1} \approx \gamma \sqrt{2\pi} \sum_{r \geq 1}^q \frac{1}{\sqrt{r}} \approx \left(\gamma \sqrt{\frac{\pi}{2}}\right) \sqrt{q}.$$

But α_0^{2q+1} is the horizontal difference between the \mathbf{x}_0^{2q+1} and G_j^{2q+1} . Therefore the difference tends to $-\infty$, which implies (2.28).

2.3 Weak consistency

Since the very basic and simple Finite Volume scheme fails to pass the strong consistency test (2.21) for general meshes, it is worthwhile to study a weaker notion.

Definition 11 *Let $\Omega \subset \mathbb{R}^d$ be a closed subset of \mathbb{R}^d . We say that $u \in L^\infty(0, T; \Omega)$ is a weak solution of the transport equation inside Ω iff the weak formulation*

$$\int_0^T \int_{\Omega} u (\partial_t \varphi + \mathbf{a} \cdot \nabla \varphi) dt dx = 0 \quad (2.32)$$

holds true for all test functions $\varphi \in C_0^1(0, T; \Omega)$.

Let us perform a change of variable $\tau = t$ and $\mathbf{y} = \mathbf{x} - \mathbf{a}\tau$. The chain rule gives $\partial_\tau = \partial_t + \mathbf{a} \cdot \nabla$ so the weak formulation rewrites $\int \int (u \partial_\tau \varphi) d\tau dy = 0$. Since the test function is arbitrary in C_0^1 , it means that $\partial_\tau u = 0$ in the weak sense.

Lemma 25 *If u , solution of the weak formulation, is regular enough, for example $u \in C^1(0, T; \Omega)$, then u is constant along the characteristics.*

From $\partial_\tau u = 0$, one deduces that $u(t, \mathbf{y}) = u(t, \mathbf{x} - \mathbf{a}t)$ is independent of \mathbf{y} . It ends the proof.

The question that we adress now is to determine if this notion of weak consistency is enough to recover the consistency of the Finite Volume Scheme on general meshes.

In this direction we consider the following condition.

Definition 12 Weak consistency *Let us consider a mesh in any dimension. Each cell is assumed to have planar faces without exterior normal \mathbf{n}_{jk} : j is the index of the current cell and k is the index of a neighboring cell. The measure in \mathbb{R}^{d-1} of the interface is $\sigma_{jk} > 0$. The volume of the current cell is V_j . We say the FV scheme is weakly consistent if and only there exists points (\mathbf{x}_{jk}) associated to the interfaces such that*

$$\sum_k \sigma_{jk} \mathbf{n}_{jk} \otimes \mathbf{x}_{jk} = V_j \mathbf{I} \quad (2.33)$$

The interest of this condition is evident on the following example. Let u be some given bounded function $u \in W^\infty(\Omega)$. The mean value in the cell is

$$u_j = \frac{1}{V_j} \int_{\Omega_j} u dx. \quad (2.34)$$

We consider the cell-wise functions

$$\begin{cases} u_h = \sum_j u_j \mathbf{I}_{x \in \Omega_j}, \\ \mathbf{p}_h = \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) \mathbf{I}_{x \in \Omega_j}, \end{cases} \quad (2.35)$$

The quantity u_{jk} is associated to the interfaces, it is a constant per interface.

Lemma 26 *Assume that $u \in BV(\Omega) \cap L^\infty(\Omega)$. Let h be the mesh size of a regular mesh (definition 7). Assume that the difference between the cell centered values and the edge values is BV bounded in the sense*

$$\sum_j \sum_{k \in V(j)} h^{d-1} |u_j - u_{jk}| \leq C. \quad (2.36)$$

Assume the weak consistency condition (2.33) holds. Then $\mathbf{p}_h = (\nabla u)_h + O(h)$ in the weak sense.

One has

$$\int \mathbf{p}_h \varphi dx = \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) \frac{\int_{\Omega_j} \varphi dx}{V_j}.$$

Consider that \mathbf{x}_j is the center of mass of the cell. Then $\frac{\int_{\Omega_j} \varphi dx}{V_j} = \varphi(\mathbf{x}_j) + O(h^2)$, so

$$\begin{aligned} \int \mathbf{p}_h \varphi dx &= \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) (\varphi(\mathbf{x}_j) + O(h^2)) \\ &= \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) \varphi(\mathbf{x}_j) + O(h) \\ &= \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) (\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{jk})) + O(h) \end{aligned}$$

where \mathbf{x}_{jk} is a point a priori on the interface. The Taylor expansion holds $\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{jk}) = (\nabla \varphi(\mathbf{x}_j) \cdot (\mathbf{x}_j - \mathbf{x}_{jk})) + O(h^2)$. So

$$\int \mathbf{p}_h \varphi dx = \sum_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} u_{jk} \right) (\nabla \varphi(\mathbf{x}_j) \cdot (\mathbf{x}_j - \mathbf{x}_{jk})) + O(h).$$

Next we use the hypothesis (2.36) together with elementary bounds to obtain

$$\begin{aligned} \int \mathbf{p}_h \varphi dx &= \sum_j u_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} \otimes (\mathbf{x}_j - \mathbf{x}_{jk}) \right) \nabla \varphi(\mathbf{x}_j) + O(h) \\ &= - \sum_j u_j \left(\sum_k \sigma_{jk} \mathbf{n}_{jk} \otimes \mathbf{x}_{jk} \right) \nabla \varphi(\mathbf{x}_j) + O(h) \end{aligned}$$

since $\sum_k \sigma_{jk} \mathbf{n}_{jk} = 0$. The hypothesis (2.33) is here used to show that

$$\int \mathbf{p}_h \varphi dx = - \sum_j u_j \nabla \varphi(\mathbf{x}_j) V_j + O(h).$$

It is then a classroom exercise to show that

$$\int \mathbf{p}_h \varphi dx = - \int u \nabla \varphi dx + O(h)$$

provided that u is a regular function.

The interpretation is that under very weak conditions on the regularity of u and on the flux (u_{jk}) then the gradient operator is weakly consistent.

Lemma 27 *Assume that the cells have planar faces. There always exists a solution of the weak consistency condition (2.33). If the scheme is consistent in the strong sense (2.21), then there exists two solutions to the weak consistency condition.*

Define

$$\mathbf{x}_{jk} = \frac{\int_{\partial\Omega_j \cap \partial\Omega_k} \mathbf{x} d\sigma}{\text{meas}(\partial\Omega_j \cap \partial\Omega_k)} \in \partial\Omega_j \cap \partial\Omega_k$$

the center of mass of the face between cell j and cell k . One has

$$V_j \delta_p^q = \int_{\Omega_j} \partial_{x_p} x_q dx = \int_{\partial\Omega_j} n_q x_p d\sigma.$$

It shows that $(\mathbf{x}_{jk})_k$ is a solution to the weak consistency condition.

Assume now that there exists a solution of the strong consistency condition. Then we set $\mathbf{x}_{jk} = \mathbf{x}_j$, $k \in I^+(j)$ and $\mathbf{x}_{jk} = \mathbf{x}_k$, $k \in I^-(j)$. The rest of the verifications are evident.

This analysis shows that the default of strong consistency may be viewed as an artefact of the analysis.

2.4 An abstract formulation

In what follows we use an abstract formalism which will be an help to study the strong convergence in L^p spaces. We will also use this approach in the next chapters.

We consider an abstract linear evolution problem in a Banach space V

$$\begin{cases} u'(t) = Au(t), & t > 0, \\ u(0) = u_0. \end{cases} \quad (2.37)$$

The problem under consideration (2.1) can be recast in this form provided the boundary condition is vanishing $u_{\text{in}} \equiv 0$. We will see this is enough to analyze the nature of what is a non consistent scheme. This is why we concentrate in this section on the abstract evolution problem (2.37).

We make natural assumptions. We assume that the solution $t \mapsto u(t) \in V$ always exist for any initial data $u_0 \in V$. It means that even if the operator A may be non continuous in A , the evolution operator is nevertheless bounded

$$\forall T > 0, \quad \exists C(T) > 0, \quad \|e^{At}\| \leq C(T). \quad (2.38)$$

Therefore one has the inequality

$$\|u(t)\| = \|e^{At}u_0\| \leq C(t)\|u_0\|. \quad (2.39)$$

A discrete approximation is constructed as follows. Let us consider a discrete subspace V_h

$$V_h \subset V. \quad (2.40)$$

The subscript \cdot_h means by convention that V_h is “smaller” than V . For example the dimension of V_h can be finite. The time step is $\Delta t > 0$. The discrete approximation at time $t_n = n\Delta t$ is $u_h^n \in V_h$. This discrete solution satisfies the following forward Euler iteration

$$\begin{cases} \frac{u_h^{n+1} - u_h^n}{\Delta t} = A_h u_h^n, \\ u_h^0 = P_h u_0. \end{cases} \quad (2.41)$$

Two operators appear. The first operator is

$$A_h : V_h \rightarrow V_h.$$

It is an “approximation” of A in a sense to be precised. The other operator

$$P_h : V \rightarrow V_h$$

is used to project the initial condition in V_h . Most of the time P_h is also a projection operator: $P_h^2 = P_h$.

Example 7 For the general FV scheme (2.2), the space V can be any of the L^p . The discrete subspace $V_h \subset V$ is the space of functions of V constant by cell. The operator A_h is defined by

$$A_h u_h|_j = -\frac{1}{s_j} \left(\sum_{k^+} m_{jk} u_j - \sum_{k^-} m_{jk} u_k \right), \quad u_h = (u_j) \in V_h.$$

The projection operator P_h is defined by $P_h u|_j = \frac{1}{s_j} \int_{\Omega_j} u(x) dx$, $u \in V$.

A stability property is attached to this A_h under an abstract CFL condition.

Definition 13 Abstract CFL condition There exists a function

$$h \mapsto \tau(h) > 0. \quad (2.42)$$

Assume the abstract CFL condition $\Delta t \leq \tau(h)$ is true. Then for all n such that $n\Delta t \leq T$, one has the inequality

$$\|(I_h + \Delta t A_h)^n\| \leq C(T), \quad n\Delta t \leq T, \quad (2.43)$$

where $C(T)$ is a constant which depends only on the time T and I_h is the identity operator in V_h .

It is natural to assume that P_h is more and more accurate as $h \rightarrow 0$. So we also assume that

$$\lim_{h \rightarrow 0} P_h u = u, \quad \forall u \in V. \quad (2.44)$$

Let us now turn to the approximation property of A_h . We desire to express this in a way which is close to the usual notion of consistency. So we consider a general solution (2.37) and we project it at each time step

$$v_h^n = P_h u(n\Delta t).$$

Definition 14 *The truncation error is by definition*

$$r_h^n = \frac{v_h^{n+1} - v_h^n}{\Delta t} - A_h v_h^n \in V_h. \quad (2.45)$$

The problem of non consistency of Finite Volume schemes admits the following abstract formulation. Assume the truncation error does not tend to 0 with h , that is $\|r_h^n\| = O(1)$. **What can we say about the convergence of the discrete solution u_h^n towards the projection of the exact solution v_h^n ?**

2.4.1 Convergence of a non consistent iterative process

In order to answer some of the questions addressed in the previous section, we study e_h^n solution of

$$\begin{cases} \frac{e_h^{n+1} - e_h^n}{\Delta t} = A_h e_h^n + r_h^n, \\ e_h^0 = 0. \end{cases} \quad (2.46)$$

One has the representation formula

$$e_h^n = \Delta t \sum_{p=0}^{n-1} (I + \Delta t A_h)^{n-1-p} r_h^p. \quad (2.47)$$

Definition 15 Compatibility condition *We say the truncation error r_h^n is compatible with A_h if there exists $s_h^n \in V_h$ such that*

$$r_h^n = \tau(h) A_h s_h^n \text{ and } \|s_h^n\| \leq C'(T), \quad n\Delta t \leq T. \quad (2.48)$$

Note that the only thing that we can deduce from (2.42-2.43) and (2.48) is $\|r_h^n\| \leq (1 + C(t))C'(T)$. That is r_h^n is $O(1)$ with respect to h .

Lemma 28 *Assume (2.42-2.43) and (2.48). Then one has the inequality*

$$\|e_h^n\| \leq \frac{C(T)}{\sqrt{\nu(1-\nu)}} \tau(h)^{\frac{1}{2}}, \quad \nu = \frac{\Delta t}{\tau(h)} \in]0, 1[\quad (2.49)$$

for all n such that $n\Delta t \leq T$.

One defines $T_h = I + \tau(h)A_h$, such that $\|T_h^p\| \leq K$ for all $p \in \mathbb{N}$. The idea is to prove various inequalities for T_h .

One has $I + \Delta t A_h = (1 - \nu)I + \nu T_h$ so $(I + \Delta t A_h)^q r_h^p = ((1 - \nu)I + \nu T_h)^q (T_h - I) s_h^q$ and

$$(I + \Delta t A_h)^q r_h^p = \sum_j \left(\nu^{j-1} (1 - \nu)^{q-j+1} \binom{j-1}{q} - \nu^j (1 - \nu)^{q-j} \binom{j}{q} \right) T_h^j s_h^q.$$

The binomial coefficient is $\binom{j}{q} = \frac{q!}{j!(q-j)!}$. If $j < 0$ or $j > q$ then $\binom{j}{q} = 0$.

The series $j \mapsto \nu^j (1 - \nu)^{q-j} \binom{q}{j}$ increases from $-\infty$ to some integer j_0 , and decreases from j_0 to $+\infty$. Thus the sum is a telescopic one

$$\begin{aligned} \sum_j \left| \nu^{j-1} (1 - \nu)^{q-j+1} \binom{q}{j-1} - \nu^j (1 - \nu)^{q-j} \binom{q}{j} \right| \\ \leq 2\nu^{j_0} (1 - \nu)^{q-j_0} \binom{j_0}{q}. \end{aligned} \quad (2.50)$$

One has the integral formula

$$\nu^{j_0} (1 - \nu)^{q-j_0} \binom{j_0}{q} = \frac{1}{2\pi} \int_0^{2\pi} e^{-ij_0\theta} ((1 - \nu) + \nu e^{i\theta})^q d\theta$$

from which we deduce the bound

$$\nu^{j_0} (1 - \nu)^{q-j_0} \binom{j_0}{q} \leq \frac{1}{2\pi} \int_0^{2\pi} |(1 - \nu) + \nu e^{i\theta}|^q d\theta.$$

Therefore

$$\|(I + \Delta t A_h)^q r_h^p\| \leq 2KC'(T) \frac{1}{2\pi} \int_0^{2\pi} |(1 - \nu) + \nu e^{i\theta}|^q d\theta. \quad (2.51)$$

Obtaining a bound for the integral is a classroom exercise. One has

$$\begin{aligned} \int_0^{2\pi} |(1 - \nu) + \nu e^{i\theta}|^q d\theta &= 4 \int_0^{\frac{\pi}{2}} (1 - 4\nu(1 - \nu) \sin^2 \mu)^{\frac{q}{2}} d\mu \\ &\leq 4 \int_0^{\frac{\pi}{2}} \left(1 - \frac{16}{\pi^2} \nu(1 - \nu) \mu^2 \right)^{\frac{q}{2}} d\mu \end{aligned}$$

since $\frac{2}{\pi} \mu \leq \sin \mu$ for $\mu \in [0, \frac{\pi}{2}]$. Next

$$\begin{aligned} \int_0^{\frac{\pi}{2}} \left(1 - \frac{16}{\pi^2} \nu(1 - \nu) \mu^2 \right)^{\frac{q}{2}} d\mu &\leq \int_0^{\frac{\pi}{2}} e^{(1 - \frac{16}{\pi^2} \nu(1 - \nu) \mu^2) \frac{q}{2}} d\mu \\ &\leq \int_0^{\infty} e^{-\frac{8}{\pi^2} \nu(1 - \nu) q \mu^2} d\mu \leq \frac{1}{\left(\frac{8}{\pi^2} \nu(1 - \nu) q \right)^{\frac{1}{2}}} \int_0^{\infty} e^{-\mu^2} d\mu \\ &\leq \frac{C}{(\nu(1 - \nu) q)^{\frac{1}{2}}} \leq \frac{\bar{C}}{(\nu(1 - \nu)(q + 1))^{\frac{1}{2}}}. \end{aligned}$$

By insertion in (2.51) one gets the bound

$$\|(I + \Delta t A_h)^q r_h^p\| \leq \frac{\tilde{C}}{(\nu(1-\nu)(q+1))^{\frac{1}{2}}}.$$

The last part of the proof is evident. We use this inequality to estimate the series in (2.47). One has

$$\begin{aligned} \|u_h^n - v_h^n\| &\leq \Delta t \sum_{p=0}^{n-1} \|(I + \Delta t A_h)^{n-1-p} r_h^p\| \\ &\leq \Delta t \frac{\tilde{C}}{\sqrt{\nu(1-\nu)}} \sum_{q=0}^{n-1} \frac{1}{\sqrt{q+1}} \leq \Delta t \frac{\tilde{C}}{\sqrt{\nu(1-\nu)}} \int_1^n \frac{dx}{\sqrt{x}} \\ &\leq \Delta t \frac{\tilde{C}}{2\sqrt{\nu(1-\nu)}} \sqrt{n} = \frac{2\tilde{C}}{\sqrt{(1-\nu)}} \sqrt{(n\Delta t)\tau(h)}. \end{aligned}$$

The proof is ended.

A proof for non constant time steps is in [1].

2.5 Convergence in L^2

In this section we prove the convergence in L^2 of the upwind scheme. We use the notations of section 2.1, except that consider the torus \mathcal{T} so that the domain of calculation has no boundaries. The velocity field $\mathbf{a} \in \mathbb{R}^2$ is assumed to be constant, that is independant of the time and of the space. The initial data will be assumed to be $u_0 \in H^1(\mathcal{T})$.

The mesh is regular. For simplicity we assume that the number of interfaces with neighboring cells is bounded.

Theorem 1 *Assume $u(t) \in H^1(\mathcal{T})$ uniformly for all $t > 0$, the CFL condition is true, the sequence of meshes is regular and the cells are convex. Then the upwind scheme is convergent with an optimal rate of convergence*

$$\|u_h^n - \Pi_h u(n\Delta t)\|_{L^2(\mathcal{T})} \leq C \|\nabla u\|_{L^2(\mathcal{T})} (Th)^{\frac{1}{2}}, \quad n\Delta t \leq T. \quad (2.52)$$

The difficulty comes from the fact that the scheme is not strongly consistent. So it is not possible to prove a stronger estimate $O(h)$ for a solution twice differentiable in L^2 , and then to weaken the estimate to obtain $O(h^{\frac{1}{2}})$ for a solution just once differentiable. We have to developp another method of analysis.

The strategy of the proof is to split it in two stages. In the first stage (time estimate) we bound the difference between the numerical solution and the semi-discrete solution. This is done using the previous material. In the second stage (space estimate) we focus on the discrete space operator.

Notice that the rate of convergence is optimal if we made no additional assumption on the regularity of the exact solution u .

2.5.1 Time estimate

We first use the abstract lemma 28 to estimate the difference between the fully discrete FV scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{1}{s_j} \left(\sum_{k^+} m_{jk} u_j^n - \sum_{k^-} m_{jk} u_k^n \right) = 0,$$

and the continuous in time FV scheme

$$w_j'(t) + \frac{1}{s_j} \left(\sum_{k^+} m_{jk} w_j(t) - \sum_{k^-} m_{jk} w_k(t) \right) = 0.$$

The continuous in time FV scheme is also called the semi-discrete scheme. The initial condition is the same for both schemes

$$u_j^0 = w_j(0) = \frac{1}{s_j} \int_{\Omega_j} u_0(\mathbf{x}) dx.$$

Lemma 29 *Assume the initial data is $u_0 \in H^1(\mathcal{T})$. Then*

$$\|u_h^n - w_h(n\Delta t)\|_{L^2(\mathcal{T})} \leq C \|\nabla u\|_{L^2(\mathcal{T})} (Th)^{\frac{1}{2}}, \quad n\Delta t \leq T, \quad (2.53)$$

where the constant C is function of the quality factor Q .

Let us define the mean value at time $t_n = n\Delta t$ of the semi-discrete solution $w_j^n = w_j(n\Delta t)$. By construction $\frac{w_h^{n+1} - w_h^n}{\Delta t} = A_h w_h^n + r_h^n$, where the residual is

$$\begin{aligned} r_h^n &= \frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} w'(s) ds - A_h w_h(n\Delta t) \\ &= A_h \left(\frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} (w(s) - w(n\Delta t)) ds \right). \end{aligned}$$

This problem has exactly the structure of definition 15, that is $r_h^n = \tau(h) A_h s_h^n$ with a uniform bound $\|s_h^n\|_2 \leq C$, provided we prove $t \mapsto w(t)$ is differentiable in L^2 . In this case we are able to apply the lemma 28 which gives the estimate we need. Define $y(t) = w'(t)$ which satisfies

$$y_j'(t) + \frac{1}{s_j} \left(\sum_{k^+} m_{jk} y_j(t) - \sum_{k^-} m_{jk} y_k(t) \right) = 0.$$

The initial condition is

$$y_j(0) = -\frac{1}{s_j} \left(\sum_{k^+} m_{jk} w_j(0) - \sum_{k^-} m_{jk} w_k(0) \right) = \frac{1}{s_j} \sum_{k^-} m_{jk} (w_k(0) - w_j(0)).$$

By the Cauchy-Schwarz inequality

$$s_j |y_j|^2 \leq \frac{1}{s_j} \left(\sum_{k^-} m_{jk} (w_k(0) - w_j(0))^2 \right) \left(\sum_{k^-} m_{jk} \right)$$

Using the technical inequality of lemma 40 for $p = q = 2$ and the various assumptions about the mesh and the initial data u_0 , it is immediate to prove that $\|y(0)\|_{L^2(\mathcal{T})}^2 = \sum_j s_j |y_j|^2 \leq C^2 \|u_0\|_{L^2(\mathcal{T})}^2$. The stability of the scheme implies that

$$\|w'(t)\|_{L^2(\mathcal{T})} = \|y(t)\|_{L^2(\mathcal{T})} \leq \|y(0)\|_{L^2(\mathcal{T})} \leq C \|u_0\|_{L^2(\mathcal{T})}.$$

Therefore w is differentiable in L^2 . It end the proof.

2.5.2 Space estimate

Now that we have bound the difference between the discrete in time and the continuous in time schemes, we study the difference between the solution of the continuous in time scheme and the exact solution. Let $u(t)$ be the solution of the advection equation in 2D and let $w_h(t)$ be the solution of the continuous in time linear finite volume scheme: $w_h(t)' = A_h w_h(t)$, $w_h(0) = u_h^0$. Recall that v_h^n is the solution of the discrete in time linear scheme.

Lemma 30 *Assume $u(t) \in H^1(\mathcal{T})$ uniformly for all $t > 0$, the CFL condition is true, and the sequence of meshes is uniformly regular. Then one has the convergence estimate*

$$\|u(n\Delta t) - w_h(n\Delta t)\|_{L^2} \leq (C(\nabla u) \left(h + (Th)^{\frac{1}{2}} \right)), \quad n\Delta t \leq T. \quad (2.54)$$

We study

$$E(t) = \frac{1}{2} \int_{\mathcal{T}} (u(t) - w_h(t))^2. \quad (2.55)$$

One has $E(0) \leq C(\nabla u_0)^2 h^2$ using lemma 36.

It remains to prove $E'(t) \leq C \|\nabla u_t\|_{L^2}^2 h$. This is a matter of elementary algebra. One has $E(t) = \frac{1}{2} \int_{\Omega} u(t)^2 + \frac{1}{2} \int_{\Omega} w_h(t)^2 - \int_{\Omega} w_h(t)u(t)$. So

$$\begin{aligned} E'(t) &= \frac{d}{dt} \left(\frac{1}{2} \int_{\mathcal{T}} u(t)^2 \right) + \left(-\frac{1}{2} \sum_j \sum_{k \in I^+(j)} m_{jk} (w_{h,j} - w_{h,k})^2 \right) \\ &\quad - \sum_j \left(\frac{-\sum_{k^+} m_{jk} w_{h,j} - \sum_{k^-} m_{jk} w_{h,k}}{s_j} \right) \int_{\Omega_j} u(t) \\ &\quad - \sum_j w_{h,j} \left(\sum_{k^+} m_{jk} u_{jk} - \sum_{k^-} m_{jk} u_{jk} \right). \end{aligned}$$

Here u_{jk} denotes the mean value of the exact solution on the edge $\partial\Omega_j \cap \partial\Omega_k$. One has the decomposition $E'(t) = A_1 + A_2 + A_3 + A_4$. The first contribution A_1 is non positive since the L^2 norm of the exact solution decreases

$$\frac{d}{dt} \int_{\Omega} \frac{u(t)^2}{2} = - \int_{\Omega} u(t) \mathbf{a} \cdot \nabla u = - \int_{\Omega} \nabla \cdot \left(\mathbf{a} \frac{u^2}{2} \right) = 0.$$

The second contribution $A_2 = \frac{d}{dt} \int_{\Omega} \frac{w_h(t)^2}{2}$ is non positive since the linear scheme is dissipative. The third and fourth contributions A_3 and A_4 are *a priori* non vanishing. But their sum is homogeneous to $\approx \int_{\Omega} \mathbf{a} \cdot \nabla (u w_h) = 0$. Let us

check that. Since $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk}$, then $A_3 = -\sum_j \sum_{k^-} m_{jk}(w_{h,j} - w_{h,k})u_{jk}$ ($= A_5$) + $\sum_j \sum_{k^-} m_{jk}(w_{h,j} - w_{h,k}) \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)$ ($= A_6$). A discrete integration by parts shows that $A_5 = -A_4$. On the other hand the Cauchy-Schwarz implies $A_6 \leq -A_2 + \frac{1}{2} \sum_j \sum_{k^-} m_{jk} \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)^2$. So finally one gets

$$E'(t) \leq \frac{1}{2} \sum_j \sum_{k^-} m_{jk} \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)^2.$$

Then use the estimate $\frac{1}{2} \sum_j \sum_{k^-} m_{jk} \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)^2 \leq Ch \|\nabla u_0\|_{L^2}^2$ which is nothing but lemma 40 for $p = q = 2$. It ends the proof.

Notice also that

$$A_6 \leq -\alpha A_2 + \frac{1}{2\alpha} \sum_j \sum_{k^-} m_{jk} \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)^2, \quad \forall \alpha > 0.$$

Using this inequality we obtain

$$E'(t) + \frac{1-\alpha}{2} \sum_j \sum_{k \in I^+(j)} m_{jk}(w_{h,j} - w_{h,k})^2 \leq \frac{1}{2\alpha} \sum_j \sum_{k^-} m_{jk} \left(u_{jk} - \frac{\int_{\Omega_j} u(t)}{s_j}\right)^2.$$

Taking $\alpha = \frac{1}{4}$ for example, we obtain the estimate

$$\sum_j \sum_{k \in I^+(j)} m_{jk}(w_{h,j} - w_{h,k})^2 \leq Ch \|\nabla u_0\|_{L^2}^2, \quad C > 0. \quad (2.56)$$

2.6 Convergence in L^p for $p \neq 2$

For the time estimate the strategy of the proof is the same as for $p = 2$ since all estimates in the proof of lemma 29 are independent of p , except the last one which is a bound for

$$y_j(0) = \frac{1}{s_j} \sum_{k^-} m_{jk}(w_k(0) - w_j(0)), \quad w_j(0) = \frac{1}{s_j} \int_{\Omega_j} u_0(\mathbf{x}) dx.$$

Theorem 2 *Assume that $u_0 \in W^{1,p}(\mathcal{T})$. Then $\|y(0)\|_{L^p(\mathcal{T})} \leq C \|\nabla u_0\|_{L^p(\mathcal{T})}$.*

We rewrite $y_j = \frac{1}{s_j} \sum_{k^-} \left(m_{jk}^{\frac{1}{p}}(w_k(0) - w_j(0))\right) m_{jk}^{\frac{1}{q}}$. As a consequence of the Hölder inequality one has

$$|y_j| \leq \frac{1}{s_j} \left(\sum_{k^-} m_{jk}(w_k(0) - w_j(0))^p\right)^{\frac{1}{p}} \left(\sum_{k^-} m_{jk}\right)^{\frac{1}{q}},$$

so

$$s_j |y_j|^p \leq \frac{1}{s_j^{p-1}} \left(\sum_{k^-} m_{jk}(w_k(0) - w_j(0))^p\right) \left(\sum_{k^-} m_{jk}\right)^{\frac{p}{q}}$$

$$\leq \frac{1}{s_j^{p-1}} h^p q \|u_0\|_{L^p(\Omega_j)}^p (Ch)^p q \leq \tilde{C}^p h^{2\frac{p}{q}-2p+2} \|u_0\|_{L^p(\Omega_j)}^p = \tilde{C}^p \|u_0\|_{L^p(\Omega_j)}^p.$$

After summation over all cells it ends the proof of the claim.

Therefore the time estimate is easy to prove for all p . We now turn to the space estimate.

2.6.1 $2 < p < \infty$

We adapt the method used in the case $p = 2$. We consider the power p of the L^p norm of the error, scaled by a factor p

$$E(t) = \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^p = \int_{\mathcal{T}} \varphi(u(t) - w_h(t)) dx \text{ with } \varphi(v) = \frac{|v|^p}{p}.$$

Lemma 31 *One has the formula $E'(t) = -A + B$ with*

$$A = \sum_{j,k^+} m_{jk} \left(\frac{1}{s_k} \int_{\Omega_k} (\varphi_u(w_j) - \varphi_u(w_k) - (w_j - w_k) \varphi'(w_k)) d\sigma \right) \quad (2.57)$$

which is non negative ($A \geq 0$) since $\varphi_u(w) = \frac{|w-u|^p}{p}$ is convex and

$$B = \sum_{j,k^+} m_{jk} (w_j - w_k) \left(\frac{1}{s_k} \int_{\Omega_k} v_{kj}(u) dx - \frac{1}{l_{jk}} \int_{\Sigma_{jk}} v_{kj}(u) dx \right) \quad (2.58)$$

where $v_{kj}(u) = \frac{\varphi(u-w_k) - \varphi(u-w_j)}{w_k - w_j}$.

It comes from the series of transformation

$$\begin{aligned} E'(t) &= \int_{\mathcal{T}} \varphi'(u(t) - w_h(t)) (\partial_t u(t) - w'_h(t)) dx \\ &= - \sum_j \int_{\Omega_j} \varphi'(u(x,t) - w_j(t)) \mathbf{a} \cdot \nabla u dx \\ &\quad + \sum_j \left(\frac{1}{s_j} \int_{\Omega_j} \varphi'(u(x,t) - w_j(t)) dx \right) \left(\sum_{k^+} m_{jk} w_j(t) - \sum_{k^-} m_{jk} w_k(t) \right) \\ &= - \sum_j \int_{\Omega_j} \nabla \cdot (\mathbf{a} \varphi(u - w_j)) dx + \sum_{j,k^+} m_{jk} (w_j - w_k) \frac{1}{s_k} \int_{\Omega_k} \varphi'(u - w_k) dx \\ &= - \sum_{j,k^+} m_{jk} \left(\frac{1}{l_{jk}} \int_{\Sigma_{jk}} (\varphi(u - w_j) - \varphi(w_k)) d\sigma \right) \\ &\quad + \sum_{j,k^+} m_{jk} (w_j - w_k) \frac{1}{s_k} \int_{\Omega_k} \varphi'(u - w_k) dx \\ &= - \sum_{j,k^+} m_{jk} \left(\frac{1}{s_k} \int_{\Omega_k} (\varphi(u - w_j) - \varphi(u - w_k) - (w_j - w_k) \varphi'(u - w_k)) d\sigma \right) \end{aligned}$$

$$+ \sum_{j,k^+} m_{jk}(w_j - w_k) \left(\frac{1}{s_k} \int_{\Omega_k} v_{kj}(u) dx - \frac{1}{l_{jk}} \int_{\Sigma_{jk}} v_{kj}(u) dx \right).$$

It ends the proof.

The next step is to proof that A controls some of the terms inside B . We notice that the function which appears under the integral in A is

$$\begin{aligned} & \varphi_u(w_j) - \varphi_u(w_k) - (w_j - w_k) \varphi'_u(w_k) \\ &= -\frac{1}{2} (w_j - w_k)^2 \int_0^1 (1-t) \varphi''_u(w_k + t(w_j - w_k)) dt \\ &= -(p-1) \frac{1}{2} (w_j - w_k)^2 \int_0^1 (1-t) |u - w_k - t(w_j - w_k)|^{p-2} dt. \end{aligned} \quad (2.59)$$

The function under the integrals in B is

$$v_{kj}(u) = \int_0^1 \varphi'(u - w_k - t(w_j - w_k)) dt = \int_0^1 |u - w_k - t(w_j - w_k)|^{p-1} dt. \quad (2.60)$$

The gradient is

$$\nabla v_{kj}(u) = (p-1) \left(\int_0^1 |u - w_k - t(w_j - w_k)|^{p-3} (u - w_k - t(w_j - w_k)) dt \right) \nabla u. \quad (2.61)$$

In the next lemma we give sharp estimates for the integrals which appear in (2.59) and (2.61).

Lemma 32 *There exists a positive numbers $\alpha_p > 0$ such that for all real numbers a and b*

$$\alpha_p (|a|^{p-2} + |b|^{p-2}) \leq \int_0^1 (1-t) |ta + (1-t)b|^{p-2} dt. \quad (2.62)$$

If $ab \geq 0$, which means that a and b have the same signe, the inequality is trivial. Indeed assume for example that $a \geq 0$ and $b \geq 0$ then

$$\int_0^1 (1-t) |(1-t)a + tb|^{p-2} dt \geq |a|^{p-2} \int_0^1 (1-t)^{p-2} dt = \frac{1}{p-1} |a|^{p-2}$$

and

$$\int_0^1 (1-t) |(1-t)a + tb|^{p-2} dt \geq |b|^{p-2} \int_0^1 (1-t) t^{p-2} dt = \frac{1}{(p-1)(p-2)} |b|^{p-2}.$$

In this case

$$\int_0^1 (1-t) |(1-t)a + tb|^{p-2} dt \geq \frac{1}{2(p-1)(p-2)} (|a|^{p-2} + |b|^{p-2}).$$

Assume now that $ab < 0$, which is the case if $b < 0 < a$ for example. Set $\theta = \frac{a}{a-b} \in]0, 1[$ which is the solution of $(1-\theta)a + \theta b = 0$. Then

$$\int_0^1 (1-t) |(1-t)a + tb|^{p-2} dt = \int_0^\theta (1-t) |(1-t)a + tb|^{p-2} dt + \int_\theta^1 (1-t) |(1-t)a + tb|^{p-2} dt.$$

Assume that $\theta \geq \frac{1}{2}$ which means that $|a| \geq |b|$. Then

$$\begin{aligned} & \int_0^1 (1-t)|(1-t)a + tb|^{p-2} dt \geq \int_0^\theta (1-t)|(1-t)a + tb - (1-\theta)a - \theta b|^{p-2} dt \\ &= \int_0^\theta (1-t)|(\theta-t)(a-b)|^{p-2} dt = |a-b|^{p-2} \int_0^\theta ((1-\theta)|\theta-t|^{p-2} + |\theta-t|^{p-1}) dt \\ &= |a-b|^{p-2} \left((1-\theta) \frac{\theta^{p-1}}{p-1} + \frac{\theta^p}{p} \right) \geq |a-b|^{p-2} \theta^{p-2} \frac{\theta}{p} \geq \frac{|a|^{p-2}}{2p} \end{aligned}$$

since $\theta \geq 12$. Since it corresponds to $|a| \geq |b|$, it implies that

$$\int_0^1 (1-t)|(1-t)a + tb|^{p-2} dt \geq \frac{|a|^{p-2} + |b|^{p-2}}{4p}.$$

The last case is $\theta < \frac{1}{2}$ which means that $|b| > |a|$. In this case

$$\begin{aligned} & \int_0^1 (1-t)|(1-t)a + tb|^{p-2} dt \geq \int_\theta^1 (1-t)|(1-t)a + tb|^{p-2} dt \\ & \geq |a-b|^{p-2} \int_\theta^1 (1-t)(t-\theta)^{p-2} dt = |a-b|^{p-2} \int_\theta^1 ((1-\theta)(t-\theta)^{p-2} - (t-\theta)^{p-1}) dt \\ & \geq |a-b|^{p-2} \left(\frac{(1-\theta)^{p-1}}{p-1} - \frac{(1-\theta)^p}{p} \right) = |a-b|^{p-2} \frac{(1-\theta)^p}{p(p-1)} \\ & = |a-b|^{p-2} (1-\theta)^{p-2} \frac{(1-\theta)^2}{p(p-1)} \geq \frac{|b|^{p-2}}{4p(p-1)}. \end{aligned}$$

Since $|b| > |a|$ then we get the inequality

$$\int_0^1 (1-t)|(1-t)a + tb|^{p-2} dt \geq \int_\theta^1 (1-t)|(1-t)a + tb|^{p-2} dt \geq \frac{|a|^{p-2} + |b|^{p-2}}{8p(p-1)}.$$

The lemma is proved with $\alpha_p = \frac{1}{8p(p-1)}$.

Lemma 33 *There exists a positive numbers $\beta_p > 0$ such that for all real numbers a and b*

$$\int_0^1 |ta + (1-t)b|^{p-2} dt \leq \beta_p (|a|^{p-2} + |b|^{p-2}). \quad (2.63)$$

This second inequality is a consequence of the Hölder inequality

$$\begin{aligned} & \left(\int_0^1 |(1-t)a + tb|^{p-2} dt \right)^{\frac{1}{p-2}} \\ & \leq \left(\int_0^1 |(1-t)ta|^{p-2} dt \right)^{\frac{1}{p-2}} + \left(\int_0^1 |tb|^{p-2} dt \right)^{\frac{1}{p-2}} = \frac{1}{(p-1)^{\frac{1}{p-2}}} (|a| + |b|). \end{aligned}$$

Therefore

$$\int_0^1 |ta + (1-t)b|^{p-2} dt \leq \frac{1}{p-1} (|a| + |b|)^{p-2} \leq \frac{2^{p-2}}{p-1} (|a|^{p-2} + |b|^{p-2}).$$

Taking $\beta_p = \frac{2^{p-2}}{p-1}$, it ends the proof of the lemma.

Lemma 34 *One has the inequality*

$$A \geq \frac{(p-1)\alpha_p}{2} \sum_{j,k^+} m_{jk}(w_j - w_k)^2 \int_{\Omega_k} (|u - w_j|^{p-2} + |u - w_k|^{p-2}) dx$$

It is an consequence of the definition of A (2.57) and of the inequality (2.62).

Lemma 35 *One has the inequality*

$$|B| \leq A + \gamma_p h \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^{p-2} \|\nabla u(t)\|_{L^p(\mathcal{T})}^2, \quad \gamma_p > 0.$$

In the integral (2.58) one sees a term which is the mean value in the cell of a given function $x \mapsto (w_k - w_j)v_{kj}(x)$ minus the mean value on the boundary. So one can apply the technical inequality of lemma 39 with $p = q = 2$. One obtains

$$|B| \leq \sum_{j,k^+} m_{jk}^{\frac{1}{2}} |w_j - w_k| \left(m_{jk}^{\frac{1}{2}} \left| \frac{1}{s_k} \int_{\Omega_k} v_{kj}(u) dx - \frac{1}{l_{jk}} \int_{\Sigma_{jk}} v_{kj}(u) dx \right| \right).$$

The inequality (2.65) with $p = q = 2$ shows that

$$m_{jk}^{\frac{1}{2}} \left| \frac{1}{s_k} \int_{\Omega_k} v_{kj}(u) dx - \frac{1}{l_{jk}} \int_{\Sigma_{jk}} v_{kj}(u) dx \right| \leq Ch^{\frac{1}{2}} \|v_{kj}\|_{L^2(\Omega_k)}.$$

Using now (2.61) and (2.63) one gets

$$\begin{aligned} & m_{jk}^{\frac{1}{2}} \left| \frac{1}{s_k} \int_{\Omega_k} v_{kj}(u) dx - \frac{1}{l_{jk}} \int_{\Sigma_{jk}} v_{kj}(u) dx \right| \\ & \leq C(p-1)\beta_p \left(\int_{\Omega_k} (|u - w_j|^{2p-4} + |u - w_k|^{2p-4}) |\nabla u|^2 dx \right)^{\frac{1}{2}} \end{aligned}$$

Therefore

$$|B| \leq C(p-1)\beta_p h \sum_{j,k^+} \left(\int_{\Omega_k} m_{jk} |w_j - w_k|^2 (|u - w_j|^{2p-4} + |u - w_k|^{2p-4}) |\nabla u|^2 dx \right)^{\frac{1}{2}}$$

The Cauchy-Schwarz inequality shows that

$$\begin{aligned} |B| & \leq \frac{\gamma}{2} \sum_{j,k^+} m_{jk}(w_j - w_k)^2 \int_{\Omega_k} (|u - w_j|^{p-2} + |u - w_k|^{p-2}) dx \\ & \quad + \frac{C(p-1)\beta_p h}{2\gamma} \sum_k \int_{\Omega_k} (|u - w_j|^{p-2} + |u - w_k|^{p-2}) |\nabla u|^2 dx. \end{aligned}$$

Let us take $\gamma = (p-1)\alpha_p$. Then $|B| \leq A + C \frac{\beta_p}{2\alpha_p} hD$ with

$$\begin{aligned} D & = \sum_k \int_{\Omega_k} (|u - w_j|^{p-2} + |u - w_k|^{p-2}) |\nabla u|^2 dx \\ & \leq 2 \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^{p-2} \|\nabla u(t)\|_{L^p(\mathcal{T})}^2. \end{aligned}$$

Theorem 3 *One has the inequality*

$$\|u(t) - u_h(t)\|_{L^p(\mathcal{T})} \leq \delta_p \left(th^{\frac{1}{2}} + h \right) \|\nabla u_0\|_{L^p(\mathcal{T})}. \quad (2.64)$$

One has

$$\frac{1}{p} \frac{d}{dt} \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^p \leq \gamma_p h \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^{p-2} \|\nabla u_0\|_{L^p(\mathcal{T})}^2,$$

that is

$$\frac{d}{dt} \|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^2 \leq 2\gamma_p h \|\nabla u_0\|_{L^p(\mathcal{T})}^2,$$

Integration in time yields

$$\|u(t) - u_h(t)\|_{L^p(\mathcal{T})}^2 \leq 2\gamma_p h \|\nabla u_0\|_{L^p(\mathcal{T})}^2 t + \|u(0) - u_h(0)\|_{L^p(\mathcal{T})}^2.$$

It ends the proof.

2.7 Approximation results

In this section we gather well known approximation results which are useful in the context of FV methods. The first inequality, lemma 36, is a very classical interpolation result. It measures the error at the initialization stage. The second inequality, lemma 39, is also very classical. This second inequality measures the systematic error due to the approximation of edges values by cell centered values. The method used to prove the result allows to consider small edges, which the common rule is a quadrangle cell degenerates to a triangle as in figure 2.7. It explains why we do not use the mapping to a reference cell because such a technique is not fully compatible for this kind of example.

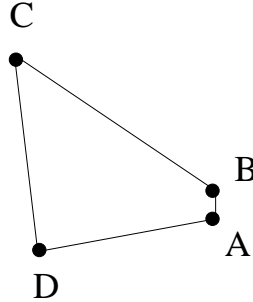


Figure 2.7: This quadrangle is close to a triangle, since the edge AB is small compared to the others.

In order to simplify the technical part of the proves, we consider a mesh with convex cells only. This is not an important restriction. The characteristic length of the mesh is h . It is by convention greater than all edges. The mesh is regular. We also assume that the number of neighbors per cell is bounded by a constant P which is independant of the mesh size h

$$\forall h > 0, \quad \text{card } I(j) \leq P \quad \forall j.$$

Lemma 36 *We consider a mesh in dimension $d = 2$ with convex cells. Let $u \in W^{1,p}(\Omega)$, $p \in [1, \infty]$. Let P_h be the projector in the mean. Then*

$$\|u - \Pi_h u\|_{L^p(\Omega)} \leq Ch \|\nabla u\|_{L^p(\Omega)},$$

where $C(Q) > 0$ is a constant that depends solely upon the quality factor of the mesh Q .

In the case $p = \infty$ the inequality is trivially satisfied with a constant equal to $2^0 = 1$. We give the proof for $p \in [1, \infty[$. One has

$$\|u - \Pi_h u\|_{L^p(\Omega)}^p = \sum_j \int_{x \in \Omega_j} \left| u(x) - \frac{1}{s_j} \int_{y \in \Omega_j} u(y) dy \right|^p dx.$$

Our first task is to estimate the term in parenthesis. One has

$$\begin{aligned} \int_{x \in \Omega_j} \left| u(x) - \frac{1}{s_j} \int_{y \in \Omega_j} u(y) dy \right|^p dx &= \frac{1}{s_j^p} \int_{x \in \Omega_j} \left| \int_{y \in \Omega_j} (u(x) - u(y)) dy \right|^p dx \\ &\leq s_j^{\frac{p}{q}-p} \int_{x \in \Omega_j} \int_{y \in \Omega_j} |u(x) - u(y)|^p dx dy \end{aligned}$$

where we have used the Hölder inequality for conjugate numbers

$$\left| \int_{y \in \Omega_j} (u(x) - u(y)) dy \right| \leq \left(\int_{y \in \Omega_j} |u(x) - u(y)|^p dy \right)^{\frac{1}{p}} s_j^{\frac{1}{q}}.$$

Let us define $z = x - y$ so that $z \in \Theta_j = \{x - y : \forall x, y \in \Omega_j\}$. Notice that $|z| \leq \text{diam}(\Omega_j) \leq h$ so the domain Θ_j has a measure bounded by $\text{diam}(\Omega_j)^2 = h^2$. For all $z \in \Theta_j$ we also define the set $\Omega_j^z = \{y \in \Omega_j : y + z \in \Omega_j\}$. Using these notations the double integral is

$$\begin{aligned} &\int_{z \in \Theta_j} \int_{y \in \Omega_j^z} |u(y+z) - u(y)|^p dz dy \\ &= \int_{z \in \Theta_j} \int_{y \in \Omega_j^z} \left| \int_0^1 \nabla u(y+tz) \cdot z dt \right|^p dz dy \\ &\leq \int_{z \in \Theta_j} \int_{y \in \Omega_j^z} \int_0^1 |\nabla u(y+tz)|^p |z|^p dt dz dy \\ &= \int_0^1 \int_{z \in \Theta_j} \left(\int_{y \in \Omega_j^z} |\nabla u(y+tz)|^p dy \right) |z|^p dz dt. \end{aligned}$$

Due to the convexity of cell Ω_j , then $y + tz = y + t(x - y) \in \Omega_j$. Therefore the term in parenthesis is bounded by $\|\nabla u\|_{L^p(\Omega_j)}^p$. We have already noticed the inequality $\text{meas}(\Theta_j) \leq h^2$. So

$$\int_{z \in \Theta_j} \int_{y \in \Omega_j^z} (u(y+z) - u(y))^p dz dy \leq h^{p+2} \|\nabla u\|_{L^p(\Omega_j)}^p.$$

Then

$$\int_{x \in \Omega_j} \left| u(x) - \frac{1}{s_j} \int_{y \in \Omega_j} u(y) dy \right|^p dx \leq s_j^{\frac{p}{q}-p} h^{p+2} \|\nabla u\|_{L^p(\Omega_j)}^p.$$

Since p and q are conjugate then $\frac{p}{q} - p = -1$. So

$$\|u - \Pi_h u\|_{L^p(\Omega)}^p \leq \sum_j \frac{h^2}{s_j} h^p \|\nabla u\|_{L^p(\Omega_j)}^p \leq \frac{h^2}{s_j} h^p \|\nabla u\|_{L^p(\Omega)}^p.$$

Noticing that $\frac{h^2}{s_j} \leq \frac{4}{\pi} Q^2$ the proof is ended.

We consider a convex cell in dimension $d = 2$. Let $u \in W^{1,p}(\Omega_j)$, $p \in [1, \infty]$. Let u_j be the mean value and let u_{jk} be the mean value on the edge

$$u_j = \frac{1}{s_j} \int_{\Omega_j} u(x) dx, \quad u_{jk} = \frac{1}{l_{jk}} \int_{\Sigma_{jk}} u(x) d\sigma, \quad \forall k.$$

We define A_j which is a measure of the difference in a local L^p norm

$$A_j = \left(\sum_{k \in I(j)} l_{jk} |u_{jk} - u_j|^p \right)^{\frac{1}{p}}.$$

We will show the inequality

$$A_j \leq C h^{\frac{1}{q}} \|\nabla u\|_{L^p(\Omega_j)}, \quad \frac{1}{p} + \frac{1}{q} = 1 \quad (2.65)$$

for some constant $C > 0$ is function of the quality factor of the cell. We will prove this estimate step by step.

First we consider one edge and we study the difference $\frac{l_{jk}|u_{jk} - u_{c,k}|}{s_j}$ where $u_{c,k}$ is the mean value of u in a small square inside the cell aligned with the edge jk , as depicted in figure 2.8.

The edge jk corresponds to the segment AB . By construction IJ is parallel to AB . This is always possible after convenient rotation of the square. Define

$$B_j = \left(\sum_{k \in I(j)} l_{jk} |u_{jk} - u_{c,k}|^p \right)^{\frac{1}{p}}.$$

Lemma 37 *One has the inequality*

$$B_j \leq C h^{\frac{1}{q}} \|\nabla u\|_{L^p(\Omega_j)} \quad (2.66)$$

where q is conjugate to p and for a convenient constant $C > 0$ which depends on p .

Let us first consider $\overline{u_{c,k}}$ which is the mean value of u along the segment $[HI]$. With local notations one has $u_{jk} - \overline{u_{c,k}} = \int_0^1 (u(A(\alpha)) - u(H(\alpha))) d\alpha$ where $A(\alpha) = A + \alpha(B - A)$ and $H(\alpha) = H + \alpha(I - H)$. So

$$u_{jk} - \overline{u_{c,k}} = \int_0^1 \int_0^\beta (\nabla u(M(\alpha, \beta)), A(\alpha) - H(\alpha)) d\beta d\alpha$$

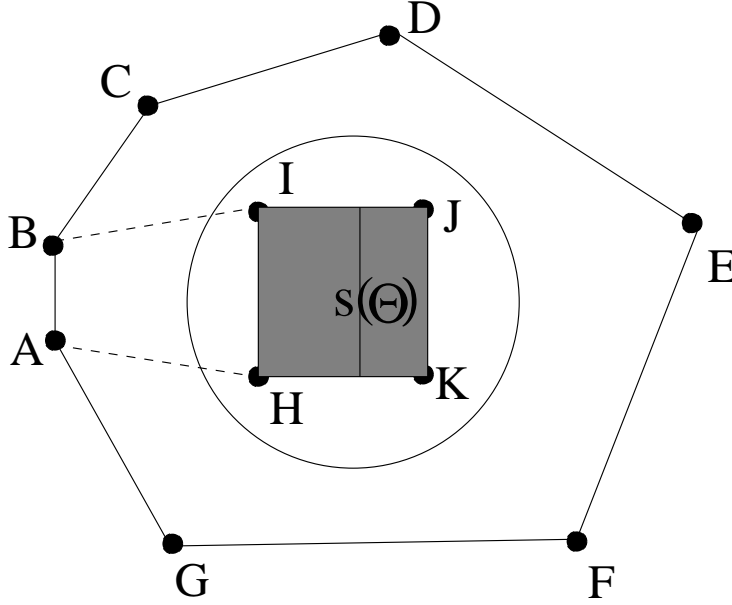


Figure 2.8: The convex cell $ABCDEFG$ contains a square with measure ch^2 . Due to convexity hypothesis, the domain $ABIH$ is inside the cell.

where $M(\alpha, \beta) = H(\alpha) + \beta(A(\alpha) - H(\alpha))$. Therefore

$$\begin{aligned} |u_{jk} - \overline{u_{c,k}}| &\leq Ch \int_0^\alpha \int_0^\beta |\nabla u(M(\alpha, \beta))| d\beta d\alpha \\ &\leq Ch \left(\int_0^\alpha \int_0^\beta |\nabla u(M(\alpha, \beta))|^p d\beta d\alpha \right)^{\frac{1}{p}} \leq Ch \left(\int_{[AHIB]} |\nabla u(\mathbf{x})|^p J dx \right)^{\frac{1}{p}} \end{aligned}$$

that is

$$|u_{jk} - \overline{u_{c,k}}| \leq Ch (\max J)^{\frac{1}{p}} \|\nabla u\|_{L^p(\Omega_j)}. \quad (2.67)$$

So we need to estimate $\max J$. Here the determinant of the Jacobian of the transformation $(\alpha, \beta) = \mathbf{x} = M(\alpha, \beta)$. To estimate J we choose a convenient reference frame such that the y axis is aligned with AB and HI . Then the transformation $(\alpha, \beta) = \mathbf{x} = (x_1, x_2)$ is characterized by

$$\begin{cases} x_1 = h_1 + \alpha(i_1 - h_1) + \beta(a_1 + \alpha(b_1 - a_1) - h_1 - \alpha(i_1 - h_1)), \\ x_2 = h_2 + \alpha(i_2 - h_2) + \beta(a_2 + \alpha(b_2 - a_2) - h_2 - \alpha(i_2 - h_2)) \end{cases}$$

that is after simplifications

$$\begin{cases} x_1 = h_1 + \beta(a_1 - h_1), \\ x_2 = h_2 + \alpha(i_2 - h_2) + \beta(a_2 - h_2) + \alpha\beta(b_2 - a_2) - i_2 + h_2. \end{cases}$$

Therefore

$$\nabla(x_1, x_2)_{(\alpha, \beta)} = \begin{pmatrix} 0 & a_1 - h_1 \\ (1 - \beta)(i_2 - h_2) + \beta(b_2 - a_2) & \dots \end{pmatrix}$$

and

$$|\det(\nabla(x_1, x_2)_{(\alpha, \beta)})| \geq (h_1 - a_1) \times \min(b_2 - a_2, i_2 - h_2).$$

Going back to the figure 2.8 and using the fact that the mesh is regular, we obtain

$$|\det(\nabla(x_1, x_2)_{(\alpha, \beta)})| \geq C'h \min(h, l_{jk}) = C'hl_{jk}$$

for some constant $C' > 0$ independent of h (recall that $l_{jk} \leq h$ by convention).

Thus $J = \det(\nabla(x_1, x_2)_{(\alpha, \beta)})^{-1} \leq \frac{1}{C'hl_{jk}}$. Therefore

$$l_{jk} |u_{jk} - \overline{u_{c,k}}|^p \leq l_{jk} \frac{(Ch)^p}{C'hl_{jk}} \|\nabla u\|_{L^p(\Omega_j)}^p \leq \tilde{C}h^{p-1} \|\nabla u\|_{L^p(\Omega_j)}^p. \quad (2.68)$$

The same inequality holds if we replace the segment $[HI]$ by any parallel vertical segment $s(\theta)$ between $[HI]$ and $[KJ]$, as described in figure 2.8. The mean value of u on this segment is denoted $u_{c,k}(\theta)$ where $0 \leq \theta \leq 1$ is the parameter that controls the position of the segment $s(\theta)$. With these notations $s(0) = [HI]$ and $s(1) = [KJ]$. One also have $u_{c,k}(0) = \overline{u_{c,k}}$ and $u_{c,k} = \int_0^1 u_{c,k}(\theta) d\theta$. Using the same kind of analysis as in the proof in inequality (2.68) one gets the estimate

$$l_{jk} |u_{jk} - u_{c,k}(\theta)|^p \leq \widehat{C}h^{p-1} \|\nabla u\|_{L^p(\Omega_j)}^p, \quad 0 \leq \theta \leq 1.$$

The constant \widehat{C} is uniform for all θ . Using the Hölder inequality one gets

$$l_{jk} |u_{jk} - u_{c,k}|^p \leq l_{jk} \int_0^1 |u_{jk} - u_{c,k}(\theta)|^p d\theta \leq \widehat{C}h^{p-1} \|\nabla u\|_{L^p(\Omega_j)}^p.$$

After summation over all neighbors (which are in finite number by hypothesis) we obtain the claim. It ends the proof.

Next we define

$$C_j = \left(\sum_{k \in I(j)} l_{jk} |u_{c,k} - u_j|^p \right)^{\frac{1}{p}}.$$

Lemma 38 *One has the inequality*

$$C_j \leq Ch^{\frac{1}{q}} \|\nabla u\|_{L^p(\Omega_j)}$$

where q is conjugate to p and for a convenient constant $C > 0$.

We set $s_{c,k}$ the area of the square $[HIJK]$. One has

$$u_{c,k} - u_j = \frac{1}{s_{c,k} s_j} \int_{x \in \Omega_j} \int_{y \in [HIJK]} (u(x) - u(y)) dx dy$$

thus using the Hölder inequality

$$|u_{c,k} - u_j| \leq \frac{1}{(s_{c,k} s_j)^{\frac{1}{p}}} \left(\int_{x \in \Omega_j} \int_{y \in [HIJK]} |u(x) - u(y)|^p dx dy \right)^{\frac{1}{p}}.$$

Using now the same inequalities as in the prof of lemma 36 the result is proved.

Lemma 39 Consider A_j defined in (2.65). One has the inequality

$$A_j \leq Ch^{\frac{1}{q}} \|\nabla u\|_{L^p(\Omega_j)}$$

where $C > 0$ is a constant that depends solely upon the quality factor of the cell.

It comes from $A_j \leq B_j + C_j$.

Next we define $A = \left(\sum_j A_j^p\right)^{\frac{1}{p}}$.

Lemma 40 One has the inequality

$$A \leq Ch^{\frac{1}{q}} \|\nabla u\|_{L^p(\Omega)}$$

where $C > 0$ is a universal constant.

This is evident from lemma (39).

Chapter 3

Non linear schemes for advection

Linear schemes for advection are formally first order in space and time. It is highly desirable for applications to design numerical methods with better accuracy. It has been observed that non linear schemes based on the seminal ideas of Boris-and-Book and Van Leer are good competitors. At least they offer in practice a major enhancement of the quality of the result for a very acceptable cost.

In what follows we first construct the famous family of TVD schemes in dimension one and we develop an original method for the study of convergence. We will see that TVD schemes have a control of the oscillations which is by construction a consequence of the maximum principle. We will present practically used schemes including MUSCL schemes, TVD schemes and the Sweby diagram, the Lagoutiere's procedure and finally the repair method of Shashkov and Wendroff. By inspection of all these methods, the beginner may think that it is a jungle, nevertheless we will try to give a deductive and comprehensive presentation of all these methods.

3.1 Theory in 1D

First we make the analysis on a equidistribute cartesian grid. We consider the general form of an explicit Finite Volume scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n}{\Delta x} = 0, \quad a > 0. \quad (3.1)$$

We will use natural notations on a cartesian mesh. For example $x_{j+\frac{1}{2}} = x_j + \frac{1}{2}\Delta x$ is the position of the interface between cell j and cell $j + 1$. For the sake of simplicity we note $u = u^n$, $\bar{u} = u^{n+1}$ and $u_{j+\frac{1}{2}}^* = u_{j+\frac{1}{2}}^n$ for all j .

With these notations (3.1) rewrites

$$\frac{\bar{u}_j - u_j}{\Delta t} + a \frac{u_{j+\frac{1}{2}}^* - u_{j-\frac{1}{2}}^*}{\Delta x} = 0, \quad a > 0$$

or equivalently

$$\bar{u}_j = u_j - \nu \left(u_{j+\frac{1}{2}}^* - u_{j-\frac{1}{2}}^* \right), \quad \nu = a \frac{\Delta t}{\Delta x}. \quad (3.2)$$

At this stage of the construction the fluxes are $u_{j+\frac{1}{2}}^*$ the unknowns of the problem. The design principle of these schemes is to decide that the maximum principle must hold in a sense or another. The form which is commonly retained is the following

$$\min(u_j, u_{j-1}) \equiv m_j \leq \bar{u}_j \leq M_j \equiv \max(u_j, u_{j-1}). \quad (3.3)$$

This formulation is adapted to a positive velocity $a > 0$. If the velocity is negative one must of course take the other side that is $\min(u_j, u_{j+1}) \equiv m_j \leq \bar{u}_j \leq M_j \equiv \max(u_j, u_{j+1})$. It is easy to justify (3.3). First the exact solution moves from the left to the right. Second the basic scheme (1.4) satisfies this inequality since $\bar{u}_j = (1 - \nu)u_j + \nu u_{j-1}$. From now on we use only (3.3) and we detail the consequences.

3.1.1 The Muscl method

This method is very popular. It is based on seminal ideas from Van Leer and the theory of Roe and Sweby.

The fundamental idea is based on the following consideration. The numerical solution u_j can be seen as the mean value of the exact solution in the cell at the beginning of the time step. Thus the upwind scheme

$$\frac{\bar{u}_j - u_j}{\Delta t} + a \frac{u_j - u_{j-1}}{\Delta x} = 0, \quad \text{that is } u_{j+\frac{1}{2}}^* = u_j$$

can be decomposed in two steps. In the first step one reconstructs an approximate solution from the knowledge of the mean value (u_j). In a second step one advects the reconstructed solution and projects it onto the mesh. This is explained in figure 3.1. Of course a constant per cell P^0 reconstruction is a poor approximation from the point of view of accuracy. So the idea is to reconstruct a better approximation.

In this direction a first competitor is the linear reconstruction

$$u_j(x) = u_j + d_j(x - x_j)$$

where the approximation of the derivative is centered

$$d_j = \frac{\frac{1}{2}(u_{j+1} + u_j) - \frac{1}{2}(u_j + u_{j-1})}{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}}$$

This is not the definition of the local derivative that is retained in practice. One prefers to localize the derivative. It means that the linear reconstruction that is retained is

$$u_j(x) = u_j + d_{j+\frac{1}{2}}(x - x_j) \quad \text{with } d_{j+\frac{1}{2}} = \frac{u_{j+1} - u_j}{\Delta x} \quad \text{for } x_j \leq x \leq x_{j+1}. \quad (3.4)$$

Then one advects the reconstructed profile and projects it onto the mesh. The total mass which passes through $x_{j+\frac{1}{2}}$ is

$$\int_{x_{j+\frac{1}{2}} - a\Delta t}^{x_{j+\frac{1}{2}}} u_j(x) dx = a\Delta t u_j(y_j), \quad y_j = x_{j+\frac{1}{2}} - \frac{1}{2}a\Delta t.$$

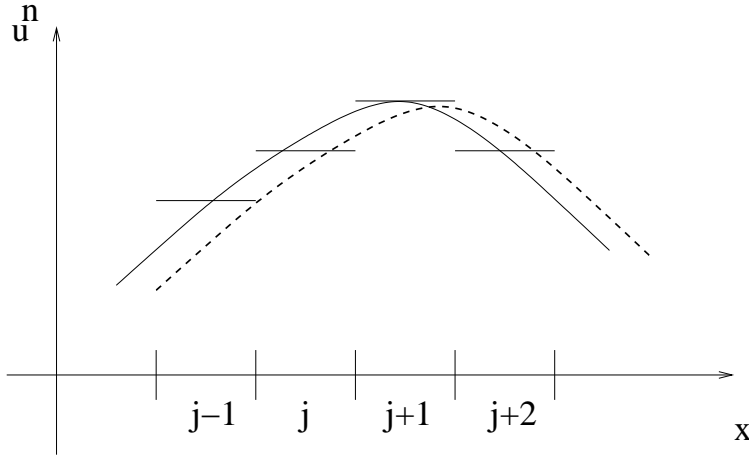


Figure 3.1: Reconstruction of an approximate solution (the curve) from the discrete solution (the steps). The approximate solution after advection is the dashed curve.

This is an exact formula since the function u_j is piecewise linear. So

$$\begin{aligned} \int_{x_{j+\frac{1}{2}}-a\Delta t}^{x_{j+\frac{1}{2}}} u_j(x) dx &= a\Delta t \left(u_j + d_{j+\frac{1}{2}} \left(x_{j+\frac{1}{2}} - \frac{1}{2}a\Delta t - x_j \right) \right) \\ &= a\Delta t \left(u_j + \frac{1}{2}(1-\nu)(u_{j+1} - u_j) \right). \end{aligned}$$

Finally one gets

$$\Delta x \bar{u}_j = \Delta x u_j - \int_{x_{j+\frac{1}{2}}-a\Delta t}^{x_{j+\frac{1}{2}}} u_j(x) dx + \int_{x_{j-\frac{1}{2}}-a\Delta t}^{x_{j+\frac{1}{2}}} u_{j-1}(x) dx$$

that is after all simplifications

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}(1-\nu)(u_{j+1} - u_j), \quad \forall j.$$

This is the Lax-Wendroff flux.

Lemma 41 *The Lax-Wendroff scheme is second order in time and space. It is stable in L^2 under CFL: $\nu \leq 1$. It does not satisfy the maximum principle.*

The stability proof is easy in Fourier. The fact that it does not satisfy the maximum principle is a consequence of a famous theorem proved by Godunov.

Theorem 4 *A linear scheme for advection which satisfies the maximum principle is only first order.*

That is why a natural idea is to bound the derivative $d_{j+\frac{1}{2}}$ in order to recover the maximum principle. Let us decide that $d_{j+\frac{1}{2}}$ will be multiplied by a factor $\varphi_{j+\frac{1}{2}}$. Then the flux will be defined by the formula

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}(1-\nu)(u_{j+1} - u_j)\varphi_{j+\frac{1}{2}}, \quad \forall j. \quad (3.5)$$

Definition 16 The correction term $\varphi_{j+\frac{1}{2}}$ is called a limiter.

A priori one may think that $0 \leq \varphi_{j+\frac{1}{2}} \leq 1$ but we will see at the end of the analysis that it is not mandatory. A first idea is to seek the correction term $\varphi_{j+\frac{1}{2}}$ as a function of the local ratio of slope

$$\varphi_{j+\frac{1}{2}} = \varphi(r_{j+\frac{1}{2}}), \quad r_{j+\frac{1}{2}} = \frac{u_j - u_{j-1}}{u_{j+1} - u_j}.$$

A first requirement is

$$\varphi(1) = 1 \tag{3.6}$$

so that the flux (3.5) is very close to the Lax-Wendroff flux by a continuity argument. A **second requirement** is to select formulas such that

$$\varphi(r) = r\varphi\left(\frac{1}{r}\right) \tag{3.7}$$

because it is a way to be symmetric with respect to the change $j-1 \leftrightarrow j+1$, that is $r_{j+\frac{1}{2}} \leftrightarrow \left(r_{j+\frac{1}{2}}\right)^{-1}$. A **third requirement** is to ask that

$$\varphi(r) = 0 \quad \forall r \leq 0 \tag{3.8}$$

The reason is that negative r corresponds to a maximum or a minimum of the numerical profile. The Lax-Wendroff scheme enhances extrema. More generally an extremum requires a low order scheme to be sure that the maximum principle holds locally. The use of the upwind flux is guaranteed by (3.8).

Definition 17 The minmod function $(a, b) \mapsto \text{minmod}(a, b)$ is defined by

- If $ab \leq 0$ then $\text{minmod}(a, b) = 0$.
- If $a > 0$ and $b > 0$, then $\text{minmod}(a, b) = \min(a, b)$.
- If $a < 0$ and $b < 0$, then $\text{minmod}(a, b) = \max(a, b)$.

By recurrence it defines the minmod function $\mathbb{R}^p \rightarrow \mathbb{R}$ for all $p \geq 2$ by

$$\text{minmod}(\mathbf{a}) = \text{minmod}(\text{minmod}(\mathbf{b}), c)$$

where $\mathbf{a} \in \mathbb{R}^p$ is arbitrary, and $\mathbf{a} = (\mathbf{b}, c)$ with $\mathbf{b} \in \mathbb{R}^{p-1}$ and $c \in \mathbb{R}$.

Lemma 42 Assume the function $r \mapsto \varphi(r)$ is such that

$$0 \leq \varphi(r) \leq 2\text{minmod}(1, r) \tag{3.9}$$

then the maximum principle holds.

The scheme rewrites under the form

$$\begin{aligned} \bar{u}_j &= u_j - \nu \left(u_j + \frac{1}{2}(1-\nu)(u_{j+1} - u_j)\varphi_{j+\frac{1}{2}} - u_{j-1} - \frac{1}{2}(1-\nu)(u_j - u_{j-1})\varphi_{j-\frac{1}{2}} \right) \\ &= u_j - \nu \left(1 + \frac{1}{2}(1-\nu) \left(\frac{\varphi_{j+\frac{1}{2}}}{r_{j+\frac{1}{2}}} - \varphi_{j-\frac{1}{2}} \right) \right) (u_j - u_{j-1}) \end{aligned}$$

or equivalently

$$\bar{u}_j = (1 - C_j)u_j + C_j u_{j-1}, \quad C_j = \nu + \frac{\nu(1 - \nu)}{2} \left(\frac{\varphi_{j+\frac{1}{2}}}{r_{j+\frac{1}{2}}} - \varphi_{j-\frac{1}{2}} \right).$$

So the maximum principle holds under the condition that $0 \leq C_j \leq 1$ that is

$$0 \leq \nu + \frac{\nu(1 - \nu)}{2} \left(\frac{\varphi_{j+\frac{1}{2}}}{r_{j+\frac{1}{2}}} - \varphi_{j-\frac{1}{2}} \right) \leq 1.$$

Assume (3.9), then $0 \leq \varphi_{j-\frac{1}{2}} \leq 2$ and $1 - \frac{1-\nu}{2}\varphi_{j-\frac{1}{2}} \geq 1 - (1-\nu) \geq 0$ so $0 \leq C_j$.

Similarly (3.9) implies $0 \leq \varphi_{j+\frac{1}{2}} \leq 2r_{j+\frac{1}{2}}$ so $1 + \frac{1-\nu}{2}\frac{\varphi_{j+\frac{1}{2}}}{r_{j+\frac{1}{2}}} \leq 1 + (1-\nu) = 2 - \nu$.

Therefore

$$\nu + \frac{\nu(1 - \nu)}{2} \frac{\varphi_{j+\frac{1}{2}}}{r_{j+\frac{1}{2}}} \leq 2\nu - \nu^2 \leq 1, \quad \forall \nu \in [0, 1].$$

It ends the proof.

An incredibly large number of different formulas have been proposed by various authors to satisfy all these requirements. We limit the presentation to the minmod flux and the superbee flux.

Definition 18 *The minmod flux corresponds to*

$$\varphi(r) = \min\text{mod}(1, r). \quad (3.10)$$

Definition 19 *The Superbee flux corresponds to*

$$\varphi(r) = \min\text{mod}(1, \max(1, 2r), \max(2, r)). \quad (3.11)$$

3.1.2 The construction of Lagoutière

The design principle is different even if the fluxes are the same. The idea is that a large $\varphi_{j+\frac{1}{2}}$ helps to counterbalance the intrinsic diffusivity of the upwind flux. So we abandon the second order requirement (3.7). Another interest of this method is the algebra which is completely different.

Plugging (3.2) in (3.3) we get the equivalent formulation of the maximum principle

$$m_j \leq u_j - \nu \left(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}} \right) \leq M_j$$

that is

$$\frac{1}{\nu}(u_j - M_j) + u_{j-\frac{1}{2}} \leq u_{j+\frac{1}{2}} \leq \frac{1}{\nu}(u_j - m_j) + u_{j-\frac{1}{2}}. \quad (3.12)$$

It is not possible to determine $u_{j+\frac{1}{2}}$ independatly of $u_{j-\frac{1}{2}}$, because the inequalities couple these quantities. The idea that was proposed first by F. Lagoutière is to decide a priori that the fluxes must satisfy another double inequality

$$u_{j+\frac{1}{2}} \in [m_{j+1}, M_{j+1}], \quad \forall j. \quad (3.13)$$

This is absolutely not evident that (3.13) is a good choice. However it will a consequence of the following analysis. We begin to notice that a sufficient condition for (3.12) to hold is

$$u_{j+\frac{1}{2}} \in [a_j, b_j] \quad (3.14)$$

where

$$a_j = \frac{1}{\nu}(u_j - M_j) + M_j \text{ and } b_j = \frac{1}{\nu}(u_j - m_j) + m_j.$$

Lemma 43 *Assume the CFL like condition $\nu \in]0, 1]$. Then the interval defined by inequalities (3.13-3.14) is not empty, more precisely*

$$u_j \in [m_{j+1}, M_{j+1}] \cap [a_j, b_j] \neq \emptyset. \quad (3.15)$$

Notice that $u_{j+\frac{1}{2}} = u_j$ gives back the upwind scheme. We only have to prove that $a_j \leq u_j \leq b_j$. The inequality $a_j \leq u_j$ reduces to

$$\frac{1}{\nu}(u_j - M_j) + M_j \leq u_j \text{ that is } \left(\frac{1}{\nu} - 1\right)(u_j - M_j) \leq 0.$$

This inequality is always true since $\left(\frac{1}{\nu} - 1\right) \geq 0$ due the CFL hypothesis and $u_j - M_j \leq 0$. Similarly the inequality $u_j \leq b_j$ is equivalent to

$$u_j \leq \frac{1}{\nu}(u_j - m_j) + m_j \text{ that is } \left(\frac{1}{\nu} - 1\right)(u_j - m_j) \geq 0$$

which is always true unde CFL. The proof is ended.

Lemma 44 *Any flux (3.15) can be rewritten as (3.5) with a limiter*

$$0 \leq \varphi(r) \leq \min\text{mod}\left(\frac{2}{\nu}, \frac{2r}{1-\nu}\right).$$

This is evident.

Definition 20 *The downwind flux is defined by: the flux $u_{j+\frac{1}{2}}$ is the closest value to the downwind value of the unknown (that is u_{j+1}) with the constraint that \bar{u}_j satisfies the maximum principle (3.3).*

3.1.3 Convergence

The inequality (3.16), which is due to Harten, enlightes the fact that the maximum principle entails the control of oscillations.

Lemma 45 *Consider a scheme (3.1) for which moreover the maximum principle (3.3) holds. Then one has the inequality*

$$\sum_{j \in \mathbb{Z}} |\bar{u}_j - \bar{u}_{j-1}| \leq \sum_{j \in \mathbb{Z}} |u_j - u_{j-1}|. \quad (3.16)$$

The maximum principle (3.3) rewrites $\bar{u}_j = u_j + C_j(u_{j-1} - u_j)$ with $C_j \in [0, 1]$. So $\bar{u}_j - \bar{u}_{j-1} = (1 - C_j)(u_j - u_{j-1}) + C_{j-1}(u_{j-1} - u_{j-2})$ and

$$|\bar{u}_j - \bar{u}_{j-1}| \leq (1 - C_j)|u_j - u_{j-1}| + C_{j-1}|u_{j-1} - u_{j-2}|$$

Then

$$\sum_j |\bar{u}_j - \bar{u}_{j-1}|$$

$$\leq \sum_j (1 - C_j) |u_j - u_{j-1}| + \sum_j C_{j-1} |u_{j-1} - u_{j-2}| = \sum_j |u_j - u_{j-1}|.$$

It ends the proof.

In order to exploit the previous inequality we define the space of functions with bounded total variation in dimension one. This approach is a simplification of the general approach developed in section 1.4.3. The little difference is that we deal with bounded functions.

Definition 21 Consider $u \in L^\infty(\mathbb{R})$. We define the total variation of u

$$TV(u) = \limsup_{\varepsilon > 0} \int_{-\infty}^{\infty} \frac{|u(x) - u(x - \varepsilon)|}{\varepsilon}. \quad (3.17)$$

The space of all bounded functions with a finite total variation $TV(u) < \infty$ is

$$\mathcal{VB} = \{u \in L^\infty(\mathbb{R}), TV(u) < \infty\}.$$

By construction the continuous variation (3.17) is very close to the discrete variation (3.16). That is why this criterion is well adapted to the discussion of (3.16).

Functions in \mathcal{VB} have a certain number of interesting properties. A first property is the following. For $u \in \mathcal{VB}$ one has the inequality

$$\int_{-\infty}^{\infty} \frac{|u(x) - u(x - \varepsilon)|}{\varepsilon} \leq TV(u), \quad \forall \varepsilon > 0. \quad (3.18)$$

Indeed $|u(x) - u(x - \varepsilon)| \leq |u(x) - u(x - \frac{1}{2}\varepsilon)| + |u(x - \frac{1}{2}\varepsilon) - u(x - \varepsilon)|$. So

$$\int_{-\infty}^{\infty} \frac{|u(x) - u(x - \varepsilon)|}{\varepsilon} \leq \int_{-\infty}^{\infty} \frac{|u(x) - u(x - \frac{1}{2}\varepsilon)|}{\frac{1}{2}\varepsilon}.$$

By iteration one gets (3.18). A second property is

$$TV(u) = \int_{-\infty}^{\infty} |u'(x)| = \|u'\|_1 \quad \forall u \in W^{1,1}(\mathbb{R}).$$

Let us now consider $v \in \mathcal{VT}$ and its regularization (2.16)

$$v_\alpha(x) = \frac{1}{\alpha} \int_{-\infty}^{\infty} \varphi\left(\frac{x-y}{\alpha}\right) v(y) dy, \quad 0 < \alpha. \quad (3.19)$$

Lemma 46 One has the inequalities

- a) $\|v - v_\alpha\|_1 \leq \alpha TV(v)$,
- b) $TV(v_\alpha) = \|v'_\alpha\|_1 \leq TV(v)$
- c) $\|v''_\alpha\|_1 \leq \frac{TV(v)}{\alpha} \times \int_x |\varphi'(x)| dx$.

This is essentially the same proof as for lemma 22 but in L^1 and not in L^∞ .

By construction

$$v_\alpha(x) - v(x) = \frac{1}{\alpha} \int_{x-\alpha}^{x+\alpha} \varphi\left(\frac{x-y}{\alpha}\right) (v(y) - v(x)) dy$$

$$= \frac{1}{\alpha} \int_{t=-\alpha}^{\alpha} \varphi\left(\frac{x+t}{\alpha}\right) (v(x+t) - v(x)) dt.$$

So

$$\begin{aligned} \|v_\alpha - v\|_1 &= \int_{x=-\infty}^{\infty} |v_\alpha(x) - v(x)| \\ &\leq \int_{t=-\alpha}^{\alpha} \int_{x=-\infty}^{\infty} \varphi\left(\frac{x+t}{\alpha}\right) \frac{|v(x+t) - v(x)|}{\alpha} dx dt \\ &\leq \int_{t=-\alpha}^{\alpha} \int_{x=-\infty}^{\infty} \frac{|v(x+t) - v(x)|}{\alpha} dx dt = \int_{t=-\alpha}^{\alpha} \frac{TV(v)}{\alpha} dt = \alpha TV(v) \end{aligned}$$

which shows the first inequality.

One also has

$$v_\alpha(x) = \frac{1}{\alpha} \int_{-\infty}^{\infty} \varphi\left(\frac{z}{\alpha}\right) v(x-z) dz.$$

Therefore

$$\begin{aligned} &\int_{x=-\infty}^{\infty} \frac{|v_\alpha(x) - v_\alpha(x-\varepsilon)|}{\varepsilon} dx \\ &\leq \frac{1}{\alpha} \int_{x=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \varphi\left(\frac{z}{\alpha}\right) \frac{|v(x-z) - v(x-\varepsilon-z)|}{\varepsilon} dz dx \\ &\leq \frac{1}{\alpha} \int_{z=-\infty}^{\infty} \varphi\left(\frac{z}{\alpha}\right) \int_{x=-\infty}^{\infty} \frac{|v(x-z) - v(x-\varepsilon-z)|}{\varepsilon} dx dz \\ &\leq \frac{1}{\alpha} \int_{z=-\infty}^{\infty} \varphi\left(\frac{z}{\alpha}\right) TV(v) dz = TV(v). \end{aligned}$$

Passing to the limit $\varepsilon \rightarrow^+$, it shows the second inequality.

Concerning the third inequality one has $(v_\alpha)'(x) = \frac{1}{\alpha^2} \int_{-\infty}^{\infty} \varphi'\left(\frac{z}{\alpha}\right) v(x-z) dz$.

By comparison with the previous expression one gets

$$\begin{aligned} &\int_{x=-\infty}^{\infty} \frac{|(v_\alpha(x))' - (v_\alpha)'(x-\varepsilon)|}{\varepsilon} dx \\ &\leq \frac{1}{\alpha^2} \int_{z=-\infty}^{\infty} \left| \varphi'\left(\frac{z}{\alpha}\right) \right| TV(v) dz = \frac{TV(v)}{\alpha} \times \int_x |\varphi'(x)| dx. \end{aligned}$$

It ends the proof.

Lemma 47 Consider an initial data $u_0 \in \mathcal{VT}$ together with the mean value initialization $u_j^0 = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x) dx$. Consider any scheme (3.1) satisfying the maximum principle (3.3). Then one has the inequality

$$\sum_j |u_j^{n+1} - u_j^n| \leq \Delta x VT(u_0).$$

It is a consequence of lemma 45 and lemma 46.

Nous montrons la convergence des schémas TVD de la section précédente.

Soit $v_j^n = \frac{1}{\Delta x} \int_{j\Delta x}^{(j+1)\Delta x} u(n\Delta t, x) dx$ la moyenne de la solution exacte dans la maille j et au temps $n\Delta t$. On pose $v^n = (v_j^n)$.

Theorem 5 Consider an initial data $u_0 \in \mathcal{VT}$ together with the mean value initialization $u_j^0 = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x) dx$. Consider any scheme (3.1) satisfying the maximum principle (3.3). Assume the flux satisfies (3.5) or that $u_{j+\frac{1}{2}}$ is in the interval defined in (3.15). Define v^n which is the mean value of the exact solution

$$v_j^n = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(n\Delta t, x) dx.$$

One has the inequality

$$\|u^n - v^n\|_1 \leq C\sqrt{T\Delta x}, \quad n\Delta t \leq T. \quad (3.20)$$

Let w^n be the discrete solution of the upwind scheme with the same initialization $w^0 = u^0$. One has

$$\|u^n - v^n\|_1 \leq \|u^n - w^n\|_1 + \|w^n - v^n\|_1. \quad (3.21)$$

Let

$$r_j^n = \frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^n - u_{j-1}^n}{\Delta x} = \left(a \frac{u_j^n - u_{j+\frac{1}{2}}^*}{\Delta x} \right) - \left(a \frac{u_{j-1}^n - u_{j-\frac{1}{2}}^*}{\Delta x} \right)$$

be the truncation error between u^n and w^n . Let T be the right shift operator, that is $Tz_j = z_{j-1}$. One has $r_n = (I - T)s^n$ and $s_j^n = a \frac{u_j^n - u_{j+\frac{1}{2}}^*}{\Delta x}$. By construction $\|s^n\|_1 \leq a\Delta x \sum_j \frac{|u_j^0 - u_{j-1}^0|}{\Delta x}$. However

$$\begin{aligned} |u_j^0 - u_{j-1}^0| &= \left| \frac{1}{\Delta x} \int_{j\Delta x}^{(j+1)\Delta x} (u_0(x) - u_0(x - \Delta x)) dx \right| \\ &\leq \frac{1}{\Delta x} \int_{j\Delta x}^{(j+1)\Delta x} |u_0(x) - u_0(x - \Delta x)| dx. \end{aligned}$$

Therefore $\|s^n\|_1 \leq aVT(u_0)$. Set $e^n = u^n - w^n$ with the recurrence

$$e^0 = 0, \quad e^{n+1} = ((1 - \nu)I + \nu T) e^n + \Delta t(I - T)s^n.$$

It satisfies the formalism developed in section 2.4. Therefore the result holds.

3.1.4 The repair method of Shashkov and Wendroff

The previous methods are **a priori** methods, that is one wants to predict an optimal value of the flux $au_{j+\frac{1}{2}}$ at the interface $u_{j+\frac{1}{2}}$ by considering the worst prediction for $au_{j-\frac{1}{2}}$. This is particularly evident in the Lagoutiere's formalism where the first to ask is to get some a priori estimates on $u_{j-\frac{1}{2}}$. The idea of Shashkov and Wendroff is different, it is fundamentally an **a posteriori** method.

The repair strategy amounts to: 1) compute a prediction of the numerical solution at time step $n+1$ using a "reasonable" and "local" scheme, as instance this scheme can be a high order non monotone scheme or a highly anti-dissipative scheme, 2) check if the new value satisfies a local maximum principle, 3) if the

new value does not satisfy the local maximum principle, then repair it. Repairing means changing the value of the unknown for 3) to be fulfilled. A difficulty is that one wants the total mass to be preserved. So we need to describe in details how to redistribute the mass of the repaired quantity, such as the total mass is preserved.

This family of algorithms can be local if one redistributes the mass in a local box around the cell that needs to be repaired, or global if one redistributes the mass in the entire domain.

The prediction scheme

First, one computes the new value of the unknown using the finite volume and conservative scheme

$$\frac{\bar{u}_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^n + c_{j+\frac{1}{2}}^n - u_{j-1}^n - c_{j-\frac{1}{2}}^n}{\Delta x} = 0, \quad (3.22)$$

where u_j^n is the mean value of $u(x)$ in cell j : $[x_{x-\frac{1}{2}}; x_{x+\frac{1}{2}}]$. The Courant or CFL number is a priori less than one ($a > 0$) $\nu = a \frac{\Delta t}{\Delta x} \leq 1$. The scheme is *a priori* different from the upwind scheme. Thus all the difference between the upwind scheme and the scheme used is embedded in the definition of the correction flux $c_{j+\frac{1}{2}}^n$ for all j . This correction flux can be either a linear or a non linear function of (u_j^n) . The only assumption is

$$\exists C_1 > 0, \exists k \in \mathbf{N}, |c_{j+\frac{1}{2}}^n| \leq C_1 \sum_{j-k \leq q \leq j+k} |u_q - u_{q-1}|. \quad (3.23)$$

This hypothesis essentially means that the flux is defined as the upwind flux plus a correction. Of course the correction is zero if the numerical profile is flat (that is if $u_q - u_{q-1} \equiv 0$ in a neighborhood of cell j): (3.23) is compatible with such a principle. The hypothesis is true for the Lax-Wendroff scheme as instance, moreover all non linear TVD algorithms satisfy it.

Correction step

The spirit of this repair algorithm is to compare \bar{u}_j^{n+1} with

$$M_j^n = \max(u_j^n, u_{j-1}^n) \text{ and } m_j^n = \min(u_j^n, u_{j-1}^n), \quad (3.24)$$

that is one checks if $m_j^n \leq \bar{u}_j^{n+1} \leq M_j^n$ is true or not. Suppose $\bar{u}_j^{n+1} > M_j^n$, then one has to modify the value of \bar{u}_j^{n+1} and redistribute the mass “around”. In the convergence analysis of the method, we discovered that it is better at the theoretical level not to redistribute the mass globally but locally at least in a box of size $p \in \mathbf{N}$ around the current cell. This is why we have introduced a new step in the repair algorithm to be able to ensure that the redistribution of mass can be made in the box of size p . Since p is a parameter of the method, one recovers the global repair by setting $p \approx +\infty$.

So let us define boxes of size p . Each box is the collection of cells j such that $rp \leq j \leq (r+1)p - 1$ where $r \in \mathbf{Z}$. The mathematical definition of these boxes B_r is

$$B_r = \{j; rp \leq j \leq (r+1)p - 1\}, \quad r \in \mathbf{Z}. \quad (3.25)$$

However it is also possible to use boxes of different sizes, provided the size is smaller than the predefined maximal box's size p . It is also possible to use moving boxes, that is the starting point of each box is different from one time step to the other. For the simplicity of the mathematical exposure we use only (3.25).

To make the correction we first need to compute

$$b_r^M = \left(\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - M_j^n) \right) - \nu(c_{(r+1)p-\frac{1}{2}}^n - c_{rp-\frac{1}{2}}^n) \quad (3.26)$$

$$b_r^m = \left(\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - m_j^n) \right) - \nu(c_{(r+1)p-\frac{1}{2}}^n - c_{rp-\frac{1}{2}}^n) \quad (3.27)$$

Most presumably $b_r^M \leq 0$ (resp. $b_r^m \geq 0$), since this is the result of a comparison between $\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - M_j^n) \leq 0$ (resp. $\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - m_j^n) \geq 0$) and $\nu(c_{(r+1)p-\frac{1}{2}}^n - c_{rp-\frac{1}{2}}^n)$. Moreover if p is large enough and ν small enough (i.e. the time step is small) then $b_r^M \leq 0$ (resp. $b_r^m \geq 0$) is probably stating true. The correction is here to ensure that $b_r^M \leq 0$ and $b_r^m \geq 0$ are always satisfied. The idea being that if these inequalities are not satisfied then we multiply the value of the fluxes by a small number such that $b_r^M \leq 0$ and $b_r^m \geq 0$ are fulfilled.

So we define

$$d_{j-\frac{1}{2}}^n = \mu_{j-\frac{1}{2}}^n c_{j-\frac{1}{2}}^n \quad \text{where } \mu_{j-\frac{1}{2}}^n \in [0, 1]. \quad (3.28)$$

The coefficient $\mu_{j-\frac{1}{2}}^n$ has to be computed to give a corrected value of the flux. The constraint $\mu_{j-\frac{1}{2}}^n \in [0, 1]$ appears natural from the consistency point of view. We expect that the definition of these $\mu_{j-\frac{1}{2}}^n$ will be the closest as possible to 1, so that the corrected flux $d_{j-\frac{1}{2}}^n$ is almost equal to the flux of the prediction scheme. We need to check

$$\tilde{b}_r^M - \nu(\mu_{(r+1)p-\frac{1}{2}}^n c_{(r+1)p-\frac{1}{2}}^n - \mu_{rp-\frac{1}{2}}^n c_{rp-\frac{1}{2}}^n) \leq 0, \quad (3.29)$$

$$\tilde{b}_r^m - \nu(\mu_{(r+1)p-\frac{1}{2}}^n c_{(r+1)p-\frac{1}{2}}^n - \mu_{rp-\frac{1}{2}}^n c_{rp-\frac{1}{2}}^n) \geq 0, \quad (3.30)$$

where by definition $\tilde{b}_r^M = \sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - M_j^n)$ is not positive and $\tilde{b}_r^m = \sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - m_j^n)$ is not negative. One feasible strategy can be derived as:

Analysis of (3.29)

if $c_{(r+1)p-\frac{1}{2}}^n \geq 0$ and $c_{rp-\frac{1}{2}}^n \geq 0$: Then (3.29) is true once $\nu(\mu_{rp-\frac{1}{2}}^n c_{rp-\frac{1}{2}}^n) \leq -\tilde{b}_r^M$. Thus we define $\varphi_r^{1,-}, \varphi_r^{1,+}$ such that

$$\mu_{rp-\frac{1}{2}}^n \leq \varphi_r^{1,-} = \frac{-\tilde{b}_r^M}{c_{rp-\frac{1}{2}}^n} \quad \text{and} \quad \varphi_r^{1,+} = +\infty. \quad (3.31)$$

if $c_{(r+1)p-\frac{1}{2}}^n < 0$ and $c_{rp-\frac{1}{2}}^n < 0$: Then (3.29) is true once $\nu(-\mu_{(r+1)p-\frac{1}{2}}^n c_{(r+1)p-\frac{1}{2}}^n) \leq -\tilde{b}_r^M$. Thus we define $\varphi_r^{2,-}, \varphi_r^{2,+}$ such that

$$\varphi_r^{2,-} = +\infty \quad \text{and} \quad \mu_{(r+1)p-\frac{1}{2}}^n \leq \varphi_r^{2,+} = \frac{-\tilde{b}_r^M}{-c_{(r+1)p-\frac{1}{2}}^n}. \quad (3.32)$$

if $c_{(r+1)p-\frac{1}{2}}^n \geq 0$ and $c_{rp-\frac{1}{2}}^n < 0$: Then (3.29) is true without condition. Thus $\varphi_r^{3,-} = \varphi_r^{3,+} = +\infty$

if $c_{(r+1)p-\frac{1}{2}}^n < 0$ and $c_{rp-\frac{1}{2}}^n \geq 0$: Then it is not possible to simplify the inequality (3.29). Thus we impose $\varphi_r^{4,-} = \varphi_r^{4,+}$ and

$$\mu_{(r+1)p-\frac{1}{2}}^n, \mu_{rp-\frac{1}{2}}^n \leq \varphi_r^{4,-} \frac{-\tilde{b}_r^M}{c_{rp-\frac{1}{2}}^n - c_{(r+1)p-\frac{1}{2}}^n} \quad (3.33)$$

Analysis of (3.30)

if $c_{(r+1)p-\frac{1}{2}}^n \geq 0$ and $c_{rp-\frac{1}{2}}^n \geq 0$: Then (3.30) is true once $\nu(\mu_{(r+1)p-\frac{1}{2}}^n c_{(r+1)p-\frac{1}{2}}^n) \leq \tilde{b}_r^m$. Thus we impose that

$$\psi_r^{1,-} = +\infty \text{ and } \mu_{(r+1)p-\frac{1}{2}}^n \leq \psi_r^{1,+} \frac{\tilde{b}_r^m}{c_{(r+1)p-\frac{1}{2}}^n}. \quad (3.34)$$

if $c_{(r+1)p-\frac{1}{2}}^n < 0$ and $c_{rp-\frac{1}{2}}^n < 0$: Then (3.30) is true once $\nu(-\mu_{rp-\frac{1}{2}}^n c_{rp-\frac{1}{2}}^n) \leq \tilde{b}_r^m$. Thus we impose that

$$\mu_{rp-\frac{1}{2}}^n \leq \psi_r^{2,-} = \frac{\tilde{b}_r^m}{-c_{rp-\frac{1}{2}}^n} \text{ and } \psi_r^{2,+} = +\infty. \quad (3.35)$$

if $c_{(r+1)p-\frac{1}{2}}^n \geq 0$ and $c_{rp-\frac{1}{2}}^n < 0$: Then it is not possible to simplify the inequality (3.30). Thus we impose $\psi_r^{3,-} = \psi_r^{3,+}$ that

$$\mu_{(r+1)p-\frac{1}{2}}^n, \mu_{rp-\frac{1}{2}}^n \leq \psi_r^{3,-} = \frac{-\tilde{b}_r^M}{-c_{(r+1)p-\frac{1}{2}}^n + c_{rp-\frac{1}{2}}^n} \quad (3.36)$$

if $c_{(r+1)p-\frac{1}{2}}^n < 0$ and $c_{rp-\frac{1}{2}}^n \geq 0$: Then (3.30) is true without condition. Thus $\psi_r^{4,-} = \psi_r^{4,+} = +\infty$

Let us consider each of the cases considered in inequalities (3.31) to (3.36). We gather the restrictions it imposes for all $\mu_{rp-\frac{1}{2}}^n$. The mathematical definition of the correction algorithm is the following.

Definition 22 *Let us define the corrected fluxes at the boundaries of the boxes*

$$d_{(r+1)p-\frac{1}{2}}^n = \min(1, \min_{1 \leq l \leq 4} \varphi_r^{l,+}, \min_{1 \leq l \leq 4} \psi_r^{l,+}, \min_{1 \leq l \leq 4} \varphi_{r+1}^{l,-}, \min_{1 \leq l \leq 4} \psi_{r+1}^{l,-}) \times c_{(r+1)p-\frac{1}{2}}^n. \quad (3.37)$$

Inside the boxes we do not correct, that is

$$d_{j-\frac{1}{2}}^n = c_{j-\frac{1}{2}}^n, \quad \forall j \neq rp. \quad (3.38)$$

The next step consists in the computation of the new prediction \hat{u}_j^{n+1} with the corrected flux:

$$\frac{\hat{u}_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_j^n + d_{j+\frac{1}{2}}^n - u_{j-1}^n - d_{j-\frac{1}{2}}^n}{\Delta x} = 0. \quad (3.39)$$

Lemma 48 *One has the inequalities after the correction step*

$$\left(\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - M_j^n) \right) - \nu \left(d_{(r+1)p-\frac{1}{2}}^n - d_{rp-\frac{1}{2}}^n \right) \leq 0 \quad (3.40)$$

$$\left(\sum_{j \in B_r} (u_j^n - \nu(u_j^n - u_{j-1}^n) - m_j^n) \right) - \nu \left(d_{(r+1)p-\frac{1}{2}}^n - d_{rp-\frac{1}{2}}^n \right) \geq 0. \quad (3.41)$$

The proof is performed by considering all cases (3.31-3.36) separately.

Inequalities (3.40-3.41) will be crucial in the analysis of the repairing procedure.

Repairing

As already mentioned, to repair a value means successively to compare with the local maximum and minimum, to truncate if needed, then to redistribute the excess of mass on all surrounding cells. Using mathematical notations, one gets

$$\begin{cases} \text{if } \hat{u}_j^{n+1} > M_j^n & \text{then } \mathbf{u}_j^{n+1} = M_j^n & \text{and } \Delta m_j^n = \hat{u}_j^{n+1} - M_j^n > 0, \\ \text{if } \hat{u}_j^{n+1} < m_j^n & \text{then } \mathbf{u}_j^{n+1} = m_j^n & \text{and } \Delta m_j^n = \hat{u}_j^{n+1} - m_j^n < 0, \\ \text{else} & \text{then } \mathbf{u}_j^{n+1} = \hat{u}_j^{n+1} & \text{and } \Delta m_j^n = 0. \end{cases} \quad (3.42)$$

The total mass in box B_r of the new unknown \mathbf{u}_j^{n+1} may as well be different from the correct mass, so one defines the default of mass as:

$$\Delta M_r^n = \sum_{rp \leq j \leq (r+1)p-1} \Delta m_j^n. \quad (3.43)$$

This default of mass may be positive or negative. So we need to redistribute it on the box to get at least a conservative algorithm. We consider

$$\begin{cases} \text{if } \Delta M_r^n > 0 & \text{then } u_j^{n+1} = \mathbf{u}_j^{n+1} + \lambda_r^n (M_j^n - \mathbf{u}_j^{n+1}), \\ \text{if } \Delta M_r^n < 0 & \text{then } u_j^{n+1} = \mathbf{u}_j^{n+1} + \lambda_r^n (m_j^n - \mathbf{u}_j^{n+1}), \\ \text{if } \Delta M_r^n = 0 & \text{then } u_j^{n+1} = \mathbf{u}_j^{n+1}, \end{cases} \quad (3.44)$$

where the coefficient λ_r^n is set to

$$\begin{cases} \text{if } \Delta M_r^n > 0 & \lambda_r^n = \frac{\Delta M_r^n}{\sum_{rp \leq j \leq (r+1)p-1} (M_j^n - \mathbf{u}_j^{n+1})}, \\ \text{if } \Delta M_r^n < 0 & \lambda_r^n = \frac{\Delta M_r^n}{\sum_{rp \leq j \leq (r+1)p-1} (m_j^n - \mathbf{u}_j^{n+1})}, \\ \text{if } \Delta M_r^n = 0 & \lambda_r^n = 0. \end{cases} \quad (3.45)$$

The repair algorithm that we analyze in this paper consists of equations (3.22) to (3.45).

Properties

Before proving our main convergence theorem, we state the stability lemma

Lemma 49 *Whatever the value of $p \in \mathbf{N}^*$ is, for all time step n , the repair algorithm is such that the total mass is preserved (conservation)*

$$\sum_j \bar{u}_j = \sum_j u_j, \quad (3.46)$$

the maxima and minima are respected (maximum principle)

$$m_j^n \leq u_j^{n+1} \leq M_j^n, \quad \forall j. \quad (3.47)$$

Bibliography

- [1] Després B. Lax theorem and finite volume schemes. *Math. of Comp.*, 73:1203–1234, 2001.
- [2] H. Brezis. *Analyse fonctionnelle*. Masson, 1983.
- [3] P. G. Ciarlet. *The finite element methods for elliptic problems*. North-Holland, 1980.
- [4] Jean-Michel Ghidaglia Daniel Bouche and Frédéric Pascal. Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation. *SIAM J. Numer. Anal.*, 2005.
- [5] A. Ern and J.-L. Guermond. *Eléments Finis : théorie, applications, mise en oeuvre*. SMAI, Math. et App., Springer Verlag, 2002.
- [6] V. Girault and P.A. Raviart. *Finite elements methods for Navier Stokes equations*, volume 5. Springer Series SCM, 1986.
- [7] S. Godunov and V.S. Ryabenck’ii. *Introduction to the theory of differential schemes*. Fizmatgiz, Moscow, 1962.
- [8] R. Herbin. *Analyse numérique des équations aux dérivées partielles (electronic)*. University Aix Marseille I.
- [9] K. Friedrichs R. Courant and H. Lewy. Uber die partiellen differenzgleichungen der methematischen physik. *Mathematische Annalen*, 100:32–74, 1928.
- [10] T. Gallouet R. Eymard and R. Herbin. *Finite Volume methods in Handbook of Numerical Analysis*. 2000.
- [11] P. A. Raviart and J. M. Thomas. *Introduction à l’analyse numérique des EDP*. Masson, 1983.
- [12] R. D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*. Interscience Publishers, 1957.