Finite Volume Methods

Robert Eymard, Thierry Gallouët, Raphaèle Herbin

Finite Volume Methods

Robert Eymard[1], **Thierry Gallouët**[2] **and Raphaèle Herbin**[3]

January 2019. This manuscript is an update of the preprint
n0 97-19 du LATP, UMR 6632, Marseille, September 1997
which appeared in Handbook of Numerical Analysis,
P.G. Ciarlet, J.L. Lions eds, vol 7, pp 713-1020

[1]Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée, et Université de Paris XIII
[2]Ecole Normale Supérieure de Lyon
[3]Université de Provence, Marseille

Contents

1 Introduction **4**
1 Examples . . . 4
2 The finite volume principles for general conservation laws . . . 6
2.1 Time discretization . . . 7
2.2 Space discretization . . . 8
3 Comparison with other discretization techniques . . . 8
4 General guideline . . . 9

2 A one-dimensional elliptic problem **12**
5 A finite volume method for the Dirichlet problem . . . 12
5.1 Formulation of a finite volume scheme . . . 12
5.2 Comparison with a finite difference scheme . . . 14
5.3 Comparison with a mixed finite element method . . . 15
6 Convergence and error analysis for the Dirichlet problem . . . 16
6.1 Error estimate with C^2 regularity . . . 16
6.2 An error estimate using a finite difference technique . . . 19
7 General 1D elliptic equations . . . 21
7.1 Formulation of the finite volume scheme . . . 21
7.2 Error estimate . . . 23
7.3 The case of a point source term . . . 26
8 A semilinear elliptic problem . . . 27
8.1 Problem and Scheme . . . 27
8.2 Compactness results . . . 28
8.3 Convergence . . . 30

3 Elliptic problems in two or three dimensions **32**
9 Dirichlet boundary conditions . . . 32
9.1 Structured meshes . . . 33
9.2 General meshes and schemes . . . 37
9.3 Existence and estimates . . . 42
9.4 Convergence . . . 44
9.5 C^2 error estimate . . . 52
9.6 H^2 error estimate . . . 55
10 Neumann boundary conditions . . . 63
10.1 Meshes and schemes . . . 63
10.2 Discrete Poincaré inequality . . . 65
10.3 Error estimate . . . 69
10.4 Convergence . . . 72
11 General elliptic operators . . . 78
11.1 Discontinuous matrix diffusion coefficients . . . 78

11.2 Other boundary conditions . . . 82
12 Dual meshes and unknowns located at vertices . . . 84
12.1 The piecewise linear finite element method viewed as a finite volume method . . . 84
12.2 Classical finite volumes on a dual mesh . . . 85
12.3 "Finite Volume Finite Element" methods . . . 89
12.4 Generalization to the three dimensional case . . . 90
13 Mesh refinement and singularities . . . 91
13.1 Singular source terms and finite volumes . . . 91
13.2 Mesh refinement . . . 93
14 Compactness results . . . 93

4 Parabolic equations **97**
15 Introduction . . . 97
16 Meshes and schemes . . . 98
17 Error estimate for the linear case . . . 100
18 Convergence in the nonlinear case . . . 104
18.1 Solutions to the continuous problem . . . 104
18.2 Definition of the finite volume approximate solutions . . . 105
18.3 Estimates on the approximate solution . . . 106
18.4 Convergence . . . 113
18.5 Weak convergence and nonlinearities . . . 116
18.6 A uniqueness result for nonlinear diffusion equations . . . 118

5 Hyperbolic equations in the one dimensional case **122**
19 The continuous problem . . . 122
20 Numerical schemes in the linear case . . . 125
20.1 The centered finite difference scheme . . . 126
20.2 The upstream finite difference scheme . . . 126
20.3 The upwind finite volume scheme . . . 128
21 The nonlinear case . . . 133
21.1 Meshes and schemes . . . 133
21.2 L^∞-stability for monotone flux schemes . . . 136
21.3 Discrete entropy inequalities . . . 136
21.4 Convergence of the upstream scheme in the general case . . . 137
21.5 Convergence proof using BV . . . 141
22 Higher order schemes . . . 146
23 Boundary conditions . . . 147
23.1 A general convergence result . . . 147
23.2 A very simple example . . . 149
23.3 A simplifed model for two phase flows in pipelines . . . 150

6 Multidimensional nonlinear hyperbolic equations **153**
24 The continuous problem . . . 153
25 Meshes and schemes . . . 156
25.1 Explicit schemes . . . 157
25.2 Implicit schemes . . . 158
25.3 Passing to the limit . . . 158
26 Stability results for the explicit scheme . . . 160
26.1 L^∞ stability . . . 160
26.2 A "weak BV" estimate . . . 161
27 Existence of the solution and stability results for the implicit scheme . . . 164
27.1 Existence, uniqueness and L^∞ stability . . . 164

27.2 "Weak space BV" inequality . . . 167
27.3 "Time BV" estimate . . . 168
28 Entropy inequalities for the approximate solution . . . 172
28.1 Discrete entropy inequalities . . . 172
28.2 Continuous entropy estimates for the approximate solution . . . 174
29 Convergence of the scheme . . . 181
29.1 Convergence towards an entropy process solution . . . 182
29.2 Uniqueness of the entropy process solution . . . 183
29.3 Convergence towards the entropy weak solution . . . 187
30 Error estimate . . . 188
30.1 Statement of the results . . . 188
30.2 Preliminary lemmata . . . 190
30.3 Proof of the error estimates . . . 196
30.4 Remarks and open problems . . . 198
31 Boundary conditions . . . 198
32 Nonlinear weak-$\star$ convergence . . . 200
33 A stabilized finite element method . . . 203
34 Moving meshes . . . 204

7 Systems **207**
35 Hyperbolic systems of equations . . . 207
35.1 Classical schemes . . . 208
35.2 Rough schemes for complex hyperbolic systems . . . 210
35.3 Partial implicitation of explicit scheme . . . 213
35.4 Boundary conditions . . . 214
35.5 Staggered grids . . . 217
36 Incompressible Navier-Stokes Equations . . . 218
36.1 The continuous equation . . . 218
36.2 Structured staggered grids . . . 219
36.3 A finite volume scheme on unstructured staggered grids . . . 219
37 Flows in porous media . . . 222
37.1 Two phase flow . . . 222
37.2 Compositional multiphase flow . . . 223
37.3 A simplified case . . . 225
37.4 The scheme for the simplified case . . . 226
37.5 Estimates on the approximate solution . . . 229
37.6 Theorem of convergence . . . 232
38 Boundary conditions . . . 236
38.1 A two phase flow in a pipeline . . . 237
38.2 Two phase flow in a porous medium . . . 239
Bibliography

Chapter 1

Introduction

The finite volume method is a discretization method which is well suited for the numerical simulation of various types (elliptic, parabolic or hyperbolic, for instance) of conservation laws; it has been extensively used in several engineering fields, such as fluid mechanics, heat and mass transfer or petroleum engineering. Some of the important features of the finite volume method are similar to those of the finite element method, see ODEN [118]: it may be used on arbitrary geometries, using structured or unstructured meshes, and it leads to robust schemes. An additional feature is the local conservativity of the numerical fluxes, that is the numerical flux is conserved from one discretization cell to its neighbour. This last feature makes the finite volume method quite attractive when modelling problems for which the flux is of importance, such as in fluid mechanics, semi-conductor device simulation, heat and mass transfer... The finite volume method is locally conservative because it is based on a " balance" approach: a local balance is written on each discretization cell which is often called "control volume"; by the divergence formula, an integral formulation of the fluxes over the boundary of the control volume is then obtained. The fluxes on the boundary are discretized with respect to the discrete unknowns.

Let us introduce the method more precisely on simple examples, and then give a description of the discretization of general conservation laws.

1 Examples

Two basic examples can be used to introduce the finite volume method. They will be developed in details in the following chapters.

Example 1.1 (Transport equation) Consider first the linear transport equation

$$\left\{ \begin{array}{ll} u_t(x,t) + \mathrm{div}(\mathbf{v}u)(x,t) & = \quad 0, x \in \mathbb{R}^2, t \in \mathbb{R}_+, \\ u(x,0) = u_0(x), x \in \mathbb{R}^2 & \end{array} \right. \tag{1.1}$$

where u_t denotes the time derivative of u, $\mathbf{v} \in C^1(\mathbb{R}^2, \mathbb{R}^2)$, and $u_0 \in L^\infty(\mathbb{R}^2)$. Let $\mathcal{T}$ be a mesh of $\mathbb{R}^2$ consisting of polygonal bounded convex subsets of $\mathbb{R}^2$ and let $K \in \mathcal{T}$ be a "control volume", that is an element of the mesh $\mathcal{T}$. Integrating the first equation of (1.1) over K yields the following "balance equation" over K:

$$\int_K u_t(x,t)dx + \int_{\partial K} \mathbf{v}(x,t) \cdot \mathbf{n}_K(x) u(x,t) d\gamma(x) = 0, \forall t \in \mathbb{R}_+, \tag{1.2}$$

where $\mathbf{n}_K$ denotes the normal vector to ∂K, outward to K. Let $k \in \mathbb{R}_+^*$ be a constant time discretization step and let $t_n = nk$, for $n \in \mathbb{N}$. Writing equation (1.2) at time t_n, $n \in \mathbb{N}$ and discretizing the time

partial derivative by the Euler explicit scheme suggests to find an approximation $u^{(n)}(x)$ of the solution of (1.1) at time t_n which satisfies the following semi-discretized equation:

$$\frac{1}{k}\int_K (u^{(n+1)}(x) - u^{(n)}(x))dx + \int_{\partial K} \mathbf{v}(x,t_n)\cdot \mathbf{n}_K(x) u^{(n)}(x) d\gamma(x) = 0, \forall \mathbf{n} \in \mathbb{N}, \forall K \in \mathcal{T}, \tag{1.3}$$

where $d\gamma$ denotes the one-dimensional Lebesgue measure on ∂K and $u^{(0)}(x) = u(x,0) = u_0(x)$. We need to define the discrete unknowns for the (finite volume) space discretization. We shall be concerned here principally with the so-called "cell-centered" finite volume method in which each discrete unkwown is associated with a control volume. Let $(u_K^{(n)})_{K\in\mathcal{T},n\in\mathbb{N}}$ denote the discrete unknowns. For $K \in \mathcal{T}$, let $\mathcal{E}_K$ be the set of edges which are included in ∂K, and for $\sigma \subset \partial K$, let $\mathbf{n}_{K,\sigma}$ denote the unit normal to σ outward to K. The second integral in (1.3) may then be split as:

$$\int_{\partial K} \mathbf{v}(x,t_n)\cdot \mathbf{n}_K(x) u^{(n)}(x) d\gamma(x) = \sum_{\sigma\in\mathcal{E}_K} \int_\sigma \mathbf{v}(x,t_n)\cdot \mathbf{n}_{K,\sigma} u^{(n)}(x) d\gamma(x); \tag{1.4}$$

for $\sigma \subset \partial K$, let

$$v_{K,\sigma}^{(n)} = \int_\sigma \mathbf{v}(x,t_n)\mathbf{n}_{K,\sigma}(x) d\gamma(x).$$

Each term of the sum in the right-hand-side of (1.4) is then discretized as

$$F_{K,\sigma}^{(n)} = \begin{cases} v_{K,\sigma}^{(n)} u_K^{(n)} & \text{if } v_{K,\sigma}^{(n)} \geq 0, \\ v_{K,\sigma}^{(n)} u_L^{(n)} & \text{if } v_{K,\sigma}^{(n)} < 0, \end{cases} \tag{1.5}$$

where L denotes the neighbouring control volume to K with common edge σ. This "upstream" or "upwind" choice is classical for transport equations; it may be seen, from the mechanical point of view, as the choice of the "upstream information" with respect to the location of σ. This choice is crucial in the mathematical analysis; it ensures the stability properties of the finite volume scheme (see chapters 5 and 6). We have therefore derived the following finite volume scheme for the discretization of (1.1):

$$\begin{cases} \dfrac{\mathrm{m}(K)}{k}(u_K^{(n+1)} - u_K^{(n)}) + \displaystyle\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^{(n)} = 0, \forall K \in \mathcal{T}, \forall n \in \mathbb{N}, \\ u_K^{(0)} = \displaystyle\int_K u_0(x)dx, \end{cases} \tag{1.6}$$

where $\mathrm{m}(K)$ denotes the measure of the control volume K and $F_{K,\sigma}^{(n)}$ is defined in (1.5). This scheme is locally conservative in the sense that if σ is a common edge to the control volumes K and L, then $F_{K,\sigma} = -F_{L,\sigma}$. This property is important in several application fields; it will later be shown to be a key ingredient in the mathematical proof of convergence. Similar schemes for the discretization of linear or nonlinear hyperbolic equations will be studied in chapters 5 and 6.

Example 1.2 (Stationary diffusion equation) Consider the basic diffusion equation

$$\begin{cases} -\Delta u = f \text{ on } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{1.7}$$

Let $\mathcal{T}$ be a rectangular mesh. Let us integrate the first equation of (1.7) over a control volume K of the mesh; with the same notations as in the previous example, this yields:

$$\sum_{\sigma\in\mathcal{E}_K} \int_\sigma -\nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) = \int_K f(x)dx. \tag{1.8}$$

For each control volume $K \in \mathcal{T}$, let x_K be the center of K. Let σ be the common edge between the control volumes K and L. One way to approximate the flux $-\int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$ (although clearly not the only one), is to use a centered finite difference approximation:

$$F_{K,\sigma} = -\frac{\mathrm{m}(\sigma)}{d_\sigma}(u_L - u_K), \tag{1.9}$$

where $(u_K)_{K\in\mathcal{T}}$ are the discrete unknowns and d_σ is the distance between x_K and x_L. This finite difference approximation of the first order derivative $\nabla u \cdot \mathbf{n}$ on the edges of the mesh (where $\mathbf{n}$ denotes the unit normal vector) is consistent: the truncation error on the flux is of order h, where h is the maximum length of the edges of the mesh. We may note that the consistency of the flux holds because for any $\sigma = K|L$ common to the control volumes K and L, the line segment $[x_K x_L]$ is perpendicular to $\sigma = K|L$. Indeed, this is the case here since the control volumes are rectangular. This property is satisfied by other meshes which will bestudied hereafter. It is crucial for the discretization of diffusion operators.
In the case where the edge σ is part of the boundary, then d_σ denotes the distance between the center x_K of the control volume K to which σ belongs and the boundary. The flux $-\int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$, is then approximated by

$$F_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{d_\sigma} u_K, \tag{1.10}$$

Hence the finite volume scheme for the discretization of (1.7) is:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = \mathrm{m}(K) f_K, \forall K \in \mathcal{T}, \tag{1.11}$$

where $F_{K,\sigma}$ is defined by (1.9) and (1.10), and f_K denotes (an approximation of) the mean value of f on K. We shall see later (see chapters 2, 3 and 4) that the finite volume scheme is easy to generalize to a triangular mesh, whereas the finite difference method is not. As in the previous example, the finite volume scheme is locally conservative, since for any edge σ separating K from L, one has $F_{K,\sigma} = -F_{L,\sigma}$.

2 The finite volume principles for general conservation laws

The finite volume method is used for the discretization of conservation laws. We gave in the above section two examples of such conservation laws. Let us now present the discretization of general conservation laws by finite volume schemes. As suggested by its name, a conservation law expresses the conservation of a quantity $q(x,t)$. For instance, the conserved quantities may be the energy, the mass, or the number of moles of some chemical species. Let us first assume that the local form of the conservation equation may be written as

$$q_t(x,t) + \mathrm{div}\mathbf{F}(x,t) = f(x,t), \tag{2.1}$$

at each point x and each time t where the conservation of q is to be written. In equation (2.1), $(\cdot)_t$ denotes the time partial derivative of the entity within the parentheses, div represents the space divergence operator: $\mathrm{div}\mathbf{F} = \partial F_1/\partial x_1 + \cdots + \partial F_d/\partial x_d$, where $\mathbf{F} = (F_1, \ldots, F_d)^t$ denotes a vector function depending on the space variable x and on the time t, x_i is the i-th space coordinate, for $i = 1, \ldots, d$, and d is the space dimension, i.e. $d = 1, 2$ or 3; the quantity $\mathbf{F}$ is a flux which expresses a transport mechanism of q; the "source term" f expresses a possible volumetric exchange, due for instance to chemical reactions between the conserved quantities.

Thanks to the physicist's work, the problem can be closed by introducing constitutive laws which relate q, $\mathbf{F}$, f with some scalar or vector unknown $u(x,t)$, function of the space variable x and of the time t. For example, the components of u can be pressures, concentrations, molar fractions of the various chemical species by unit volume... The quantity q is often given by means of a known function $\bar{q}$ of $u(x,t)$, of the

space variable x and of the time t, that is $q(x,t) = \bar{q}(x,t,u(x,t))$. The quantity $\mathbf{F}$ may also be given by means of a function of the space variable x, the time variable t and of the unknown $u(x,t)$ and (or) by means of the gradient of u at point (x,t).... The transport equation of Example 1.1 is a particular case of (2.1) with $q(x,t) = u(x,t)$, $\mathbf{F}(x,t) = \mathbf{v}u(x,t)$ and $f(x,t) = f(x)$; so is the stationary diffusion equation of Example 1.2 with $q(x,t) = u(x)$, $\mathbf{F}(x,t) = -\nabla u(x)$, and $f(x,t) = f(x)$. The source term f may also be given by means of a function of x, t and $u(x,t)$.

Example 2.1 (The one-dimensional Euler equations) Let us consider as an example of a system of conservation laws the 1D Euler equations for equilibrium real gases; these equations may be written under the form (2.1), with

$$q = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} \text{ and } \mathbf{F} = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{pmatrix},$$

where ρ, u, E and p are functions of the space variable x and the time t, and refer respectively to the density, the velocity, the total energy and the pressure of the particular gas under consideration. The system of equations is closed by introducing the constitutive laws which relate p and E to the specific volume τ, with $\tau = \frac{1}{\rho}$ and the entropy s, through the constitutive laws:

$$p = \frac{\partial \varepsilon}{\partial \tau}(\tau, s) \text{ and } E = \rho(\varepsilon(\tau, s) + \frac{u^2}{2}),$$

where ε is the internal energy per unit mass, which is a given function of τ and s.

Equation (2.1) may be seen as the expression of the conservation of q in an infinitesimal domain; it is formally equivalent to the equation

$$\begin{aligned} \int_K q(x,t_2)dx - \int_K q(x,t_1)dx + \int_{t_1}^{t_2} \int_{\partial K} \mathbf{F}(x,t) \cdot \boldsymbol{n}_K(x) d\gamma(x) dt \\ = \int_{t_1}^{t_2} \int_K f(x,t) dx dt, \end{aligned} \tag{2.2}$$

for any subdomain K and for all times t_1 and t_2, where $\boldsymbol{n}_K(x)$ is the unit normal vector to the boundary ∂K, at point x, outward to K. Equation (2.2) expresses the conservation law in subdomain K between times t_1 and t_2. Here and in the sequel, unless otherwise mentionned, dx is the integration symbol for the d-dimensional Lebesgue measure in $\mathbb{R}^d$ and $d\gamma$ is the integration symbol for the $(d-1)$-dimensional Hausdorff measure on the considered boundary.

2.1 Time discretization

The time discretization of Equation (2.1) is performed by introducing an increasing sequence $(t_n)_{n\in\mathbb{N}}$ with $t_0 = 0$. For the sake of simplicity, only constant time steps will be considered here, keeping in mind that the generalization to variable time steps is straightforward. Let $k \in \mathbb{R}_+^\star$ denote the time step, and let $t_n = nk$, for $n \in \mathbb{N}$. It can be noted that Equation (2.1) could be written with the use of a space-time divergence. Hence, Equation (2.1) could be either discretized using a space-time finite volume discretization or a space finite volume discretization with a time finite difference scheme (the explicit Euler scheme, for instance). In the first case, the conservation law is integrated over a time interval and a space "control volume" as in the formulation (2.1). In the latter case, it is only integrated space wise, and the time derivative is approximated by a finite difference scheme; with the explicit Euler scheme, the term $(q)_t$ is therefore approximated by the differential quotient $(q^{(n+1)} - q^{(n)})/k$, and $q^{(n)}$ is computed with an approximate value of u at time t_n, denoted by $u^{(n)}$. Implicit and higher order schemes may also be used.

2.2 Space discretization

In order to perform a space finite volume discretization of equation (2.1), a mesh $\mathcal{T}$ of the domain Ω of $\mathbb{R}^d$, over which the conservation law is to be studied, is introduced. The mesh is such that $\overline{\Omega} = \cup_{K\in\mathcal{T}}\overline{K}$, where an element of $\mathcal{T}$, denoted by K, is an open subset of Ω and is called a control volume. Assumptions on the meshes will be needed for the definition of the schemes; they also depend on the type of equation to be discretized.

For the finite volume schemes considered here, the discrete unknowns at time t_n are denoted by $u_K^{(n)}$, $K \in \mathcal{T}$. The value $u_K^{(n)}$ is expected to be some approximation of u on the cell K at time t_n. The basic principle of the classical finite volume method is to integrate equation (2.1) over each cell K of the mesh $\mathcal{T}$. One obtains a conservation law under a nonlocal form (related to equation (2.2)) written for the volume K. Using the Euler time discretization, this yields

$$\int_K \frac{q^{(n+1)}(x) - q^{(n)}(x)}{k} dx + \int_{\partial K} \mathbf{F}(x, t_n) \cdot \boldsymbol{n}_K(x) d\gamma(x) = \int_K f(x, t_n) dx, \tag{2.3}$$

where $\boldsymbol{n}_K(x)$ is the unit normal vector to ∂K at point x, outward to K.

The remaining step in order to define the finite volume scheme is therefore the approximation of the "flux", $\mathbf{F}(x, t_n) \cdot \boldsymbol{n}_K(x)$, across the boundary ∂K of each control volume, in terms of $\{u_L^{(n)}, L \in \mathcal{T}\}$ (this flux approximation has to be done in terms of $\{u_L^{n+1}, L \in \mathcal{T}\}$ if one chooses the implicit Euler scheme instead of the explicit Euler scheme for the time discretization). More precisely, omitting the terms on the boundary of Ω, let $K|L = \overline{K} \cap \overline{L}$, with K, $L \in \mathcal{T}$, the exchange term (from K to L), $\int_{K|L} \mathbf{F}(x, t_n) \cdot \boldsymbol{n}_K(x) d\gamma(x)$, between the control volumes K and L during the time interval $[t_n, t_{n+1})$ is approximated by some quantity, $F_{K,L}^{(n)}$, which is a function of $\{u_M^{(n)}, M \in \mathcal{T}\}$ (or a function of $\{u_M^{n+1}, M \in \mathcal{T}\}$ for the implicit Euler scheme, or more generally a function of $\{u_M^{(n)}, M \in \mathcal{T}\}$ and $\{u_M^{n+1}, M \in \mathcal{T}\}$ if the time discretization is a one-step method). Note that $F_{K,L}^{(n)} = 0$ if the Hausdorff dimension of $\overline{K} \cap \overline{L}$ is less than $d - 1$ (e.g. $\overline{K} \cap \overline{L}$ is a point in the case $d = 2$ or a line segment in the case $d = 3$).

Let us point out that two important features of the classical finite volume method are

1. the conservativity, that is $F_{K,L}^{(n)} = -F_{L,K}^{(n)}$, for all K and $L \in \mathcal{T}$ and for all $n \in \mathbb{N}$.
2. the "consistency" of the approximation of $\mathbf{F}(x, t_n) \cdot \boldsymbol{n}_K(x)$, which has to be defined for each relation type between $\mathbf{F}$ and the unknowns.

These properties, together with adequate stability properties which are obtained by estimates on the approximate solution, will give some convergence properties of the finite volume scheme.

3 Comparison with other discretization techniques

The finite volume method is quite different from (but sometimes related to) the finite difference method or the finite element method. On these classical methods see e.g. Dahlquist and Björck [44], Thomée [144], Ciarlet [29], Ciarlet [30], Roberts and Thomas [126].

Roughly speaking, the principle of the finite difference method is, given a number of discretization points which may be defined by a mesh, to assign one discrete unknown per discretization point, and to write one equation per discretization point. At each discretization point, the derivatives of the unknown are replaced by finite differences through the use of Taylor expansions. The finite difference method becomes difficult to use when the coefficients involved in the equation are discontinuous (e.g. in the case of heterogeneous media). With the finite volume method, discontinuities of the coefficients will not be any problem if the mesh is chosen such that the discontinuities of the coefficients occur on the boundaries of the control volumes (see sections 7 and 11, for elliptic problems). Note that the finite volume scheme is often called "finite difference scheme" or "cell centered difference scheme". Indeed, in the finite volume

method, the finite difference approach can be used for the approximation of the fluxes on the boundary of the control volumes. Thus, the finite volume scheme differs from the finite difference scheme in that the finite difference approximation is used for the flux rather than for the operator itself.

The finite element method (see e.g. CIARLET [29]) is based on a variational formulation, which is written for both the continuous and the discrete problems, at least in the case of conformal finite element methods which are considered here. The variational formulation is obtained by multiplying the original equation by a "test function". The continuous unknown is then approximated by a linear combination of "shape" functions; these shape functions are the test functions for the discrete variational formulation (this is the so called "Galerkin expansion"); the resulting equation is integrated over the domain. The finite volume method is sometimes called a "discontinuous finite element method" since the original equation is multiplied by the characteristic function of each grid cell which is defined by $1_K(x) = 1$, if $x \in K$, $1_K(x) = 0$, if $x \notin K$, and the discrete unknown may be considered as a linear combination of shape functions. However, the techniques used to prove the convergence of finite element methods do not generally apply for this choice of test functions. In the following chapters, the finite volume method will be compared in more detail with the classical and the mixed finite element methods.

From the industrial point of view, the finite volume method is known as a robust and cheap method for the discretization of conservation laws (by robust, we mean a scheme which behaves well even for particularly difficult equations, such as nonlinear systems of hyperbolic equations and which can easily be extended to more realistic and physical contexts than the classical academic problems). The finite volume method is cheap thanks to short and reliable computational coding for complex problems. It may be more adequate than the finite difference method (which in particular requires a simple geometry). However, in some cases, it is difficult to design schemes which give enough precision. Indeed, the finite element method can be much more precise than the finite volume method when using higher order polynomials, but it requires an adequate functional framework which is not always available in industrial problems. Other more precise methods are, for instance, particle methods or spectral methods but these methods can be more expensive and less robust than the finite volume method.

4 General guideline

The mathematical theory of finite volume schemes has recently been undertaken. Even though we choose here to refer to the class of scheme which is the object of our study as the "finite volume" method, we must point out that there are several methods with different names (box method, control volume finite element methods, balance method to cite only a few) which may be viewed as finite volume methods. The name "finite difference" has also often been used referring to the finite volume method. We shall mainly quote here the works regarding the mathematical analysis of the finite volume method, keeping in mind that there exist numerous works on applications of the finite volume methods in the applied sciences, some references to which may be found in the books which are cited below.

Finite volume methods for convection-diffusion equations seem to have been first introduced in the early sixties by TICHONOV and SAMARSKII [142], SAMARSKII [130] and SAMARSKII [131].
The convergence theory of such schemes in several space dimensions has only recently been undertaken. In the case of vertex-centered finite volume schemes, studies were carried out by SAMARSKII, LAZAROV and MAKAROV [132] in the case of Cartesian meshes, HEINRICH [83], BANK and ROSE [7], CAI [20], CAI, MANDEL and MC CORMICK [21] and VANSELOW [149] in the case of unstructured meshes; see also MORTON and SÜLI [111], SÜLI [139], MACKENZIE, and MORTON [103], MORTON, STYNES and SÜLI [112] and SHASHKOV [136] in the case of quadrilateral meshes. Cell-centered finite volume schemes are addressed in MANTEUFFEL and WHITE [104], FORSYTH and SAMMON [69], WEISER and WHEELER [158] and LAZAROV, MISHEV and VASSILEVSKI [99] in the case of Cartesian meshes and in VASSILESKI, PETROVA and LAZAROV [150], HERBIN [84], HERBIN [85], LAZAROV and MISHEV [98], MISHEV [109] in the case of triangular or Voronoï meshes; let us also mention COUDIÈRE, VILA and VILLEDIEU [40] and COUDIÈRE, VILA and VILLEDIEU [41] where more general meshes are treated, with, however, a somewhat

technical geometrical condition. In the pure diffusion case,the cell centered finite volume method has also been analyzed with finite element tools: Agouzal, Baranger, Maitre and Oudin [4], Angermann [1], Baranger, Maitre and Oudin [8], Arbogast, Wheeler and Yotov [5], Angermann [1]. Semilinear convection-diffusion are studied in Feistauer, Felcman and Lukacova-Medvidova [62] with a combined finite element-finite volume method, Eymard, Gallouët and Herbin [55] with a pure finite volume scheme.

Concerning nonlinear hyperbolic conservation laws, the one-dimensional case is now classical; let us mention the following books on numerical methods for hyperbolic problems: Godlewski and Raviart [75], LeVeque [100], Godlewski and Raviart [76], Kröner [91], and references therein. In the multidimensional case, let us mention the convergence results which where obtained in Champier, Gallouët and Herbin [25], Kröner and Rokyta [92], Cockburn, Coquel and LeFloch [33] and the error estimates of Cockburn, Coquel and LeFloch [32] and Vila [155] in the case of an explicit scheme and Eymard, Gallouët, Ghilani and Herbin [52] in the case of explicit and implicit schemes. The proof of the error estimate of Eymard, Gallouët, Ghilani and Herbin [52], which is concerned with a flux of the form $\boldsymbol{v}(\boldsymbol{x},t)f(\boldsymbol{u})$ can easily be adapted for general fluxes of the form $F(\boldsymbol{x},t,\boldsymbol{u})$ Chainais-Hillairet [23].

The purpose of the following chapters is to lay out a mathematical framework for the convergence and error analysis of the finite volume method for the discretization of elliptic, parabolic or hyperbolic partial differential equations under conservative form, following the philosophy of the works of Champier, Gallouët and Herbin [25], Herbin [84], Eymard, Gallouët, Ghilani and Herbin [52] and Eymard, Gallouët and Herbin [55]. In order to do so, we shall describe the implementation of the finite volume method on some simple (linear or non-linear) academic problems, and develop the tools which are needed for the mathematical analysis. This approach helps determine the properties of finite volume schemes which lead to "good" schemes for complex applications.
Chapter 2 introduces the finite volume discretization of an elliptic operator in one space dimension. The resulting numerical scheme is compared to finite difference, finite element and mixed finite element methods in this particular case. An error estimate is given; this estimate is in fact contained in results shown later in the multidimensional case; however, with the one-dimensional case, one can already understand the basic principles of the convergence proof, and understand the difference with the proof of Manteuffel and White [104] or Forsyth and Sammon [69], which does not seem to generalize to the unstructured meshes. In particular, it is made clear that, although the finite volume scheme is not consistent in the finite difference sense since the truncation error does not tend to 0, the conservativity of the scheme, together with a consistent approximation of the fluxes and some "stability" allow the proof of convergence. The scheme and the error estimate are then generalized to the case of a more general elliptic operator allowing discontinuities in the diffusion coefficients. Finally, a semilinear problem is studied, for which a convergence result is proved. The principle of the proof of this result may be used for nonlinear problems in several space dimensions. It is used in Chapter 3 in order to prove convergence results for linear problems when no regularity on the exact solution is known.

In Chapter 3, the discretization of elliptic problems in several space dimensions by the finite volume method is presented. Structured meshes are shown to be an easy generalization of the one-dimensional case; unstructured meshes are then considered, for Dirichlet and Neumann conditions on the boundary of the domain. In both cases, admissible meshes are defined, and, following Eymard, Gallouët and Herbin [55], convergence results (with no regularity on the data) and error estimates assuming a C^2 or H^2 regular solution to the continuous problems are proved. As in the one-dimensional case, the conservativity of the scheme, together with a consistent approximation of the fluxes and some "stability" are used for the proof of convergence. In addition to the properties already used in the one-dimensional case, the multidimensional estimates require the use of a "discrete Poincaré" inequality which is proved in both Dirichlet and Neumann cases, along with some compactness properties which are also used and are given in the last section. It is then shown how to deal with matrix diffusion coefficients and more general boundary conditions. Singular sources and mesh refinement are also studied.

Chapter 4 deals with the discretization of parabolic problems. Using the same concepts as in Chapter 3, an error estimate is given in the linear case. A nonlinear degenerate parabolic problem is then studied, for which a convergence result is proved, thanks to a uniqueness result which is proved at the end of the chapter.

Chapter 5 introduces the finite volume discretization of a hyperbolic operator in one space dimension. Some basics on entropy weak solutions to nonlinear hyperbolic equations are recalled. Then the concept of stability of a scheme is explained on a simple linear advection problem, for which both finite difference and finite volume schemes are considered. Some well known schemes are presented with a finite volume formulation in the nonlinear case. A proof of convergence using a "weak BV inequality" which was found to be crucial in the multidimensional case (Chapter 6) is given in the one-dimensional case for the sake of clarity. For the sake of completeness, the proof of convergence based on "strong BV estimates" and the Lax-Wendroff theorem is also recalled, although it is not used for general meshes in the multidimensional case.

In Chapter 6, finite volume schemes for the discretization of multidimensional nonlinear hyperbolic conservation equations are studied. Under suitable assumptions, which are satisfied by several well known schemes, it is shown that the considered schemes are L^∞ stable (this is classical) but also satisfy some "weak BV inequality". This "weak BV" inequality is the key estimate to the proof of convergence of the schemes. Following Eymard, Gallouët, Ghilani and Herbin [52], both time implicit and explicit discretizations are considered. In the case of the implicit scheme, the existence of the solution must first be proved. The approximate solutions are shown to satisfy some discrete entropy inequalities. Using the weak BV estimate, the approximate solution is also shown to satisfy some continuous entropy inequalities. Introducing the concept of "entropy process solution" to the nonlinear hyperbolic equations (which is similar to the notion of measure valued solutions of DiPerna [46]), the approximate solutions are proved to converge towards an entropy process solution as the mesh size tends to 0. The entropy process solution is shown to be unique, and is therefore equal to the entropy weak solution, which concludes the convergence of the approximate solution towards the entropy weak solution. Finally error estimates are proved for both the explicit and implicit schemes.

The last chapter is concerned with systems of equations. In the case of hyperbolic systems which are considered in the first part, little is known concerning the continuous problem, so that the schemes which are introduced are only shown to be efficient by numerical experimentation. These "rough" schemes seem to be efficient for complex cases such as the Euler equations for real gases. The incompressible Navier-Stokes equations are then considered; after recalling the classical staggered grid finite volume formulation (see e.g. Patankar [123]), a finite volume scheme defined on a triangular mesh for the Stokes equation is studied. In the case of equilateral triangles, the tools of Chapter 3 allow to show that the approximate velocities converge to the exact velocities. Systems arising from modelling multiphase flow in porous media are then considered. The convergence of the approximate finite volume solution for a simplified case is then proved with the tools introduced in Chapter 6.

More precise references to recent works on the convergence of finite volume methods will be made in the following chapters. However, we shall not quote here the numerous works on applications of the finite volume methods in the applied sciences.

Chapter 2

A one-dimensional elliptic problem

The purpose of this chapter is to give some developments of the example 1.2 of the introduction in the one-dimensional case. The formalism needed to define admissible finite volume meshes is first given and applied to the Dirichlet problem. After some comparisons with other relevant schemes, convergence theorems and error estimates are provided. Then, the case of general linear elliptic equations is handled and finally, a first approach of a nonlinear problem is studied and introduces some compactness theorems in a quite simple framework; these compactenss theorems will be useful in further chapters.

5 A finite volume method for the Dirichlet problem

5.1 Formulation of a finite volume scheme

The principle of the finite volume method will be shown here on the academic Dirichlet problem, namely a second order differential operator without time dependent terms and with homogeneous Dirichlet boundary conditions. Let f be a given function from $(0,1)$ to $\mathbb{R}$, consider the following differential equation:

$$\begin{array}{lcll} -u_{xx}(x) & = & f(x), & x \in (0,1), \\ u(0) & = & 0, & \\ u(1) & = & 0. & \end{array} \tag{5.1}$$

If $f \in C([0,1],\mathbb{R})$, there exists a unique solution $u \in C^2([0,1],\mathbb{R})$ to Problem (5.1). In the sequel, this exact solution will be denoted by u. Note that the equation $-u_{xx} = f$ can be written in the conservative form $\mathrm{div}(\mathbf{F}) = f$ with $\mathbf{F} = -u_x$.
In order to compute a numerical approximation to the solution of this equation, let us define a mesh, denoted by $\mathcal{T}$, of the interval $(0,1)$ consisting of N cells (or control volumes), denoted by K_i, $i = 1,\ldots,N$, and N points of $(0,1)$, denoted by x_i, $i = 1,\ldots,N$, satisfying the following assumptions:

Definition 5.1 (Admissible one-dimensional mesh) An admissible mesh of $(0,1)$, denoted by $\mathcal{T}$, is given by a family $(K_i)_{i=1,\cdots,N}$, $N \in \mathbb{N}^\star$, such that $K_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, and a family $(x_i)_{i=0,\cdots,N+1}$ such that

$$x_0 = x_{\frac{1}{2}} = 0 < x_1 < x_{\frac{3}{2}} < \cdots < x_{i-\frac{1}{2}} < x_i < x_{i+\frac{1}{2}} < \cdots < x_N < x_{N+\frac{1}{2}} = x_{N+1} = 1.$$

One sets

$$\begin{array}{l} h_i = \mathrm{m}(K_i) = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}},\ i = 1,\ldots,N, \text{ and therefore } \displaystyle\sum_{i=1}^{N} h_i = 1, \\ h_i^- = x_i - x_{i-\frac{1}{2}}, h_i^+ = x_{i+\frac{1}{2}} - x_i,\ i = 1,\ldots,N, \\ h_{i+\frac{1}{2}} = x_{i+1} - x_i,\ i = 0,\ldots,N, \\ \mathrm{size}(\mathcal{T}) = h = \max\{h_i,\ i = 1,\ldots,N\}. \end{array}$$

The discrete unknowns are denoted by u_i, $i = 1, \ldots, N$, and are expected to be some approximation of u in the cell K_i (the discrete unknown u_i can be viewed as an approximation of the mean value of u over K_i, or of the value of $u(x_i)$, or of other values of u in the control volume $K_i \ldots$). The first equation of (5.1) is integrated over each cell K_i, as in (2.3) and yields

$$-u_x(x_{i+\frac{1}{2}}) + u_x(x_{i-\frac{1}{2}}) = \int_{K_i} f(x)dx, \qquad i = 1, \ldots, N.$$

A reasonable choice for the approximation of $-u_x(x_{i+\frac{1}{2}})$ (at least, for $i = 1, \ldots, N-1$) seems to be the differential quotient

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}.$$

This approximation is consistent in the sense that, if $u \in C^2([0,1], \mathbb{R})$, then there exists $C \in \mathbb{R}_+$ only depending on u such that

$$|R_{i+\frac{1}{2}}| = |F^\star_{i+\frac{1}{2}} + u_x(x_{i+\frac{1}{2}})| \leq Ch, \text{ where } F^\star_{i+\frac{1}{2}} = -\frac{u(x_{i+1}) - u(x_i)}{h_{i+\frac{1}{2}}}. \tag{5.2}$$

The quantity $R_{i+\frac{1}{2}}$ is called the consistency error .

Remark 5.1 (Using the mean value) Assume that x_i is the center of K_i. Let $\tilde{u}_i$ denote the mean value over K_i of the exact solution u to Problem (5.1). One may then remark that $|\tilde{u}_i - u(x_i)| \leq Ch_i^2$, with some C only depending on u; it follows easily that $(\tilde{u}_{i+1} - \tilde{u}_i)/h_{i+\frac{1}{2}} = u_x(x_{i+\frac{1}{2}}) + 0(h)$ also holds, for $i = 1, \ldots, N-1$ (recall that $h = \max\{h_i, i = 1, \ldots, N\}$). Hence the approximation of the flux is also consistent if the discrete unknowns u_i, $i = 1, \cdots, N$, are viewed as approximations of the mean value of u in the control volumes.

The Dirichlet boundary conditions are taken into account by using the values imposed at the boundaries to compute the fluxes on these boundaries. Taking these boundary conditions into consideration and setting $f_i = \frac{1}{h_i}\int_{K_i} f(x)dx$ for $i = 1, \ldots, N$ (in an actual computation, an approximation of f_i by numerical integration can be used), the finite volume scheme for problem (5.1) reads

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i,\ i = 1, \ldots, N \tag{5.3}$$

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}},\ i = 1, \ldots, N-1, \tag{5.4}$$

$$F_{\frac{1}{2}} = -\frac{u_1}{h_{\frac{1}{2}}}, \tag{5.5}$$

$$F_{N+\frac{1}{2}} = \frac{u_N}{h_{N+\frac{1}{2}}}. \tag{5.6}$$

Note that (5.4), (5.5), (5.6) may also be written

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}},\ i = 0, \ldots, N, \tag{5.7}$$

setting

$$u_0 = u_{N+1} = 0. \tag{5.8}$$

The numerical scheme (5.3)-(5.6) may be written under the following matrix form:

$$AU = b, \tag{5.9}$$

where $U = (u_1, \ldots, u_N)^t$, $b = (b_1, \ldots, b_N)^t$, with (5.8) and with A and b defined by

$$(AU)_i = \frac{1}{h_i}\Big(-\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}}\Big),\ i = 1, \ldots, N, \tag{5.10}$$

$$b_i = \frac{1}{h_i}\int_{K_i} f(x)dx,\ i = 1, \ldots, N, \tag{5.11}$$

Remark 5.2 There are other finite volume schemes for problem (5.1).

1. For instance, it is possible, in Definition 5.1, to take $x_1 \geq 0$, $x_N \leq 1$ and, for the definition of the scheme (that is (5.3)-(5.6)), to write (5.3) only for $i = 2, \ldots, N-1$ and to replace (5.5) and (5.6) by $u_1 = u_N = 0$ (note that (5.4) does not change). For this so-called "modified finite volume" scheme, it is also possible to obtain an error estimate as for the scheme (5.3)-(5.6) (see Remark 6.2). Note that, with this scheme, the union of all control volumes for which the "conservation law" is written is slightly different from $[0, 1]$ (namely $[x_{3/2}, x_{N-1/2}] \neq [0, 1]$) .

2. Another possibility is to take (primary) unknowns associated to the boundaries of the control volumes Keller [90], Courbet and Croisille [42]. We do not consider this case here.

5.2 Comparison with a finite difference scheme

With the same notations as in Section 5.1, consider that u_i is now an approximation of $u(x_i)$. It is interesting to notice that the expression

$$\eth_i^2 u = \frac{1}{h_i}(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}) = \frac{1}{h_i}\Big(-\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}}\Big)$$

is not a consistent approximation of $-u_{xx}(x_i)$ in the finite difference sense, that is the error made by replacing the derivative by a difference quotient (the truncation error Dahlquist and Björck [44]) does not tend to 0 as h tends to 0. Indeed, let $\overline{U} = \big(u(x_1), \ldots, u(x_N)\big)^t$; with the notations of (5.9)-(5.11), the truncation error may be defined as

$$r = A\overline{U} - b,$$

with $r = (r_1, \ldots, r_N)^t$. Note that for f regular enough, which is assumed in the sequel, $b_i = f(x_i) + 0(h)$. An estimate of r is obtained by using Taylor's expansion:

$$u(x_{i+1}) = u(x_i) + h_{i+\frac{1}{2}}u_x(x_i) + \frac{1}{2}h_{i+\frac{1}{2}}^2 u_{xx}(x_i) + \frac{1}{6}h_{i+\frac{1}{2}}^3 u_{xxx}(\xi_i),$$

for some $\xi_i \in (x_i, x_{i+1})$, which yields

$$r_i = -\frac{1}{h_i}\frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2}u_{xx}(x_i) + u_{xx}(x_i) + 0(h),\quad i = 1, \ldots, N,$$

which does not, in general tend to 0 as h tends to 0 (except in particular cases) as may be seen on the simple following example:

Example 5.1 Let $f \equiv 1$ and consider a mesh of $(0, 1)$, in the sense of Definition 5.1, satisfying $h_i = h$ for even i, $h_i = h/2$ for odd i and $x_i = (x_{i+1/2} + x_{i-1/2})/2$, for $i = 1, \ldots, N$. An easy computation shows that the truncation error r is such that

$$r_i = \begin{cases} -\frac{1}{4}, \text{ for even } i \\ +\frac{1}{2}, \text{ for odd } i. \end{cases}$$

Hence $\sup\{|r_i|, i = 1, \ldots, N\} \not\to 0$ as $h \to 0$.

Therefore, the scheme obtained from (5.3)-(5.6) is not consistent in the finite difference sense, even though it is consistent in the finite volume sense, that is, the numerical approximation of the fluxes is conservative and the truncation error on the fluxes tends to 0 as h tends to 0.
If, for instance, x_i is the center of K_i, for $i = 1, \ldots, N$, it is well known that for problem (5.1), the consistent finite difference scheme would be, omitting boundary conditions,

$$\frac{4}{2h_i + h_{i-1} + h_{i+1}}\Big[-\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}}\Big] = f(x_i),\ i = 2, \ldots, N-1, \tag{5.12}$$

Remark 5.3 Assume that x_i is, for $i = 1, \ldots, N$, the center of K_i and that the discrete unknown u_i of the finite volume scheme is considered as an approximation of the mean value $\tilde{u}_i$ of u over K_i (note that $\tilde{u}_i = u(x_i) + (h_i^2/24)u_{xx}(x_i) + 0(h^3)$, if $u \in C^3([0,1], \mathbb{R})$) instead of $u(x_i)$, then again, the finite volume scheme, considered once more as a finite difference scheme, is not consistent in the finite difference sense. Indeed, let $\tilde{R} = A\tilde{U} - b$, with $\tilde{U} = (\tilde{u}_1, \ldots, \tilde{u}_N)^t$, and $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_N)^t$, then, in general, $\tilde{R}_i$ does not go to 0 as h goes to 0. In fact, it will be shown later that the finite volume scheme, when seen as a finite difference scheme, is consistent in the finite difference sense if u_i is considered as an approximation of $u(x_i) - (h_i^2/8)u_{xx}(x_i)$. This is the idea upon which the first proof of convergence by Forsyth and Sammon in 1988 is based, see Forsyth and Sammon [69] and Section 6.2.

In the case of Problem (5.1), both the finite volume and finite difference schemes are convergent. The finite difference scheme (5.12) is convergent since it is stable, in the sense that $\|X\|_\infty \leq C\|AX\|_\infty$, for all $X \in \mathbb{R}^N$, where C is a constant and $\|X\|_\infty = \sup(|X_1|, \ldots, |X_N|)$, $X = (X_1, \ldots, X_N)^t$, and consistent in the usual finite difference sense. Since $A(\overline{U} - U) = R$, the stability property implies that $\|\overline{U} - U\|_\infty \leq C\|R\|_\infty$ which goes to 0, as h goes to 0, by definition of the consistency in the finite difference sense. The convergence of the finite volume scheme (5.3)-(5.6) needs some more work and is described in Section 6.1.

5.3 Comparison with a mixed finite element method

The finite volume method has often be thought of as a kind of mixed finite element method, since both methods involve the fluxes. However, we show here that, on the simple Dirichlet problem (5.1), the two methods yield two different schemes. For Problem (5.1), the discrete unknowns of the finite volume method are the valuesu_i, $i = 1, \ldots, N$. The finite volume method also introduces one discrete unknown at each of the control volume extremities, namely the numerical flux between the corresponding control volumes. And so indeed, the finite volume method for elliptic problems may appear closely related to the mixed finite element method. Recall that the mixed finite element method consists in introducing in Problem (5.1) the auxiliary variable $q = -u_x$, which yields the following system:

$$\begin{aligned} q + u_x &= 0, \\ q_x &= f; \end{aligned}$$

assuming $f \in L^2((0,1))$, a variational formulation of this system is:

$$q \in H^1((0,1)),\ u \in L^2((0,1)), \tag{5.13}$$

$$\int_0^1 q(x)p(x)dx = \int_0^1 u(x)p_x(x)dx,\ \forall\, p \in H^1((0,1)), \tag{5.14}$$

$$\int_0^1 q_x(x)v(x)dx = \int_0^1 f(x)v(x)dx,\ \forall\, v \in L^2((0,1)). \tag{5.15}$$

Considering an admissible mesh of $(0,1)$ (see Definition 5.1), the usual discretization of this variational formulation consists in taking the classical piecewise linear finite element functions for the approximation H of $H^1((0,1))$ and the piecewise constant finite element for the approximation L of $L^2((0,1))$. Then,

the discrete unknowns are $\{u_i, i = 1, \ldots, N\}$ and $\{q_{i+1/2}, i = 0, \ldots, N\}$ (u_i is an approximation of u in K_i and $q_{i+1/2}$ is an approximation of $-u_x(x_{i+1/2})$). The discrete equations are obtained by performing a Galerkin expansion of u and q with respect to the natural basis functions ψ_l, $l = 1, \ldots, N$ (spanning L), and $\varphi_{j+1/2}$, $j = 0, \ldots, N$ (spanning H) and by taking $p = \varphi_{i+1/2}$, $i = 0, \ldots, N$ in (5.14) and $v = \psi_k, k = 1, \ldots, N$ in (5.15). Let $h_0 = h_{N+1} = 0$, $u_0 = u_{N+1} = 0$ and $q_{-1/2} = q_{N+3/2} = 0$. Then the discrete system obtained by the mixed finite element method has $2N + 1$ unknowns and reads

$$q_{i+\frac{1}{2}}(\frac{h_i + h_{i+1}}{3}) + q_{i-\frac{1}{2}}(\frac{h_i}{6}) + q_{i+\frac{3}{2}}(\frac{h_{i+1}}{6}) = u_i - u_{i+1},\ i = 0, \ldots, N,$$

$$q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}} = \int_{K_i} f(x)dx,\ i = 1, \ldots, N.$$

Note that the unknowns $q_{i+1/2}$ cannot be eliminated from the system. The resolution of this system of equations does not give the same values $\{u_i, i = 1, \ldots, N\}$ than those obtained by using the finite volume scheme (5.3)-(5.6). In fact it is easily seen that, in this case, the finite volume scheme can be obtained from the mixed finite element scheme by using the following numerical integration for the left handside of (5.14):

$$\int_{K_i} g(x)dx = \frac{g(x_{i+1}) + g(x_i)}{2} h_i.$$

This is also true for some two-dimensional elliptic problems and therefore the finite volume error estimates for these problems may be obtained via the mixed finite element theory, see Agouzal, Baranger, Maitre and Oudin [4], Baranger, Maitre and Oudin [8].

6 Convergence and error analysis for the Dirichlet problem

6.1 Error estimate with C^2 regularity

We shall now prove the following error estimate, which will be generalized to more general elliptic problems and in higher space dimensions.

Theorem 6.1
Let $f \in C([0,1], \mathbb{R})$ and let $u \in C^2([0,1], \mathbb{R})$ be the (unique) solution of Problem (5.1). Let $\mathcal{T} = (K_i)_{i=1,\ldots,N}$ be an admissible mesh in the sense of Definition 5.1. Then, there exists a unique vector $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution to (5.3) -(5.6) and there exists $C \geq 0$, only depending on u, such that

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}} \leq C^2 h^2, \tag{6.1}$$

and

$$|e_i| \leq Ch,\ \forall i \in \{1, \ldots, N\}, \tag{6.2}$$

with $e_0 = e_{N+1} = 0$ and $e_i = u(x_i) - u_i$, for all $i \in \{1, \ldots, N\}$.

Proof
First remark that there exists a unique vector $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution to (5.3)-(5.6). Indeed, multiplying (5.3) by u_i and summing for $i = 1, \ldots, N$ gives

$$\frac{u_1^2}{h_{\frac{1}{2}}} + \sum_{i=1}^{N-1} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} + \frac{u_N^2}{h_{N+\frac{1}{2}}} = \sum_{i=1}^{N} u_i h_i f_i.$$

Therefore, if $f_i = 0$ for any $i \in \{1, \ldots, N\}$, then the unique solution to (5.3) is obtained by taking $u_i = 0$, for any $i \in \{1, \ldots, N\}$. This gives existence and uniqueness of $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution to (5.3) (with (5.4)-(5.6)).

One now proves (6.1). Let

$$\overline{F}_{i+\frac{1}{2}} = -u_x(x_{i+\frac{1}{2}}),\ i = 0, \ldots, N,$$

Integrating the equation $-u_{xx} = f$ over K_i yields

$$\overline{F}_{i+\frac{1}{2}} - \overline{F}_{i-\frac{1}{2}} = h_i f_i,\ i = 1, \ldots, N.$$

By (5.3), the numerical fluxes $F_{i+\frac{1}{2}}$ satisfy

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i,\ i = 1, \ldots, N.$$

Therefore, with $G_{i+\frac{1}{2}} = \overline{F}_{i+\frac{1}{2}} - F_{i+\frac{1}{2}}$,

$$G_{i+\frac{1}{2}} - G_{i-\frac{1}{2}} = 0,\ i = 1, \ldots, N.$$

Using the consistency of the fluxes (5.2), there exists $C > 0$, only depending on u, such that

$$F^\star_{i+\frac{1}{2}} = \overline{F}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}} \text{ and } |R_{i+\frac{1}{2}}|. \leq Ch, \tag{6.3}$$

Hence with $e_i = u(x_i) - u_i$, for $i = 1, \ldots, N$, and $e_0 = e_{N+1} = 0$, one has

$$G_{i+\frac{1}{2}} = -\frac{e_{i+1} - e_i}{h_{i+\frac{1}{2}}} - R_{i+\frac{1}{2}},\ i = 0, \ldots, N,$$

so that $(e_i)_{i=0,\ldots,N+1}$ satisfies

$$-\frac{e_{i+1} - e_i}{h_{i+\frac{1}{2}}} - R_{i+\frac{1}{2}} + \frac{e_i - e_{i-1}}{h_{i-\frac{1}{2}}} + R_{i-\frac{1}{2}} = 0,\ \forall i \in \{1, \ldots, N\}. \tag{6.4}$$

Multiplying (6.4) by e_i and summing over $i = 1, \ldots, N$ yields

$$-\sum_{i=1}^{N} \frac{(e_{i+1} - e_i)e_i}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{N} \frac{(e_i - e_{i-1})e_i}{h_{i-\frac{1}{2}}} = -\sum_{i=1}^{N} R_{i-\frac{1}{2}} e_i + \sum_{i=1}^{N} R_{i+\frac{1}{2}} e_i.$$

Noting that $e_0 = 0$, $e_{N+1} = 0$ and reordering by parts, this yields (with (6.3))

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}} \leq Ch \sum_{i=0}^{N} |e_{i+1} - e_i|. \tag{6.5}$$

The Cauchy-Schwarz inequality applied to the right hand side gives

$$\sum_{i=0}^{N} |e_{i+1} - e_i| \leq \Big(\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}}\Big)^{\frac{1}{2}} \Big(\sum_{i=0}^{N} h_{i+\frac{1}{2}}\Big)^{\frac{1}{2}}. \tag{6.6}$$

Since $\displaystyle\sum_{i=0}^{N} h_{i+\frac{1}{2}} = 1$ in (6.6) and from (6.5), one deduces (6.1).

Since, for all $i \in \{1, \ldots, N\}$, $e_i = \displaystyle\sum_{j=1}^{i} (e_j - e_{j-1})$, one can deduce, from (6.6) and (6.1) that (6.2) holds.

■

Remark 6.1 The error estimate given in this section does not use the discrete maximum principle (that is the fact that $f_i \geq 0$, for all $i = 1, \ldots, N$, implies $u_i \geq 0$, for all $i = 1, \ldots, N$), which is used in the proof of error estimates by the finite difference techniques, but the coerciveness of the elliptic operator, as in the proof of error estimates by the finite element techniques.

Remark 6.2

1. The above proof of convergence gives an error estimate of order h. It is sometimes possible to obtain an error estimate of order h^2. Indeed, this is the case, at least if $u \in C^4([0,1], \mathbb{R})$, if x_i is the center of K_i for all $i = 1, \ldots, N$. One obtains, in this case, $|e_i| \leq Ch^2$, for all $i \in \{1, \ldots, N\}$, where C only depends on u (see FORSYTH and SAMMON [69]).

2. It is also possible to obtain an error estimate for the modified finite volume scheme described in the first item of Remark 5.2 page 14. It is even possible to obtain an error estimate of order h^2 in the case $x_1 = 0$, $x_N = 1$ and assuming that $x_{i+1/2} = (1/2)(x_i + x_{i+1})$, for all $i = 1, \ldots, N-1$. In fact, in this case, one obtains $|R_{i+1/2}| \leq C_1 h^2$, for all $i = 1, \ldots, N-1$. Then, the proof of Theorem 6.1 gives (6.1) with h^4 instead of h^2 which yields $|e_i| \leq C_2 h^2$, for all $i \in \{1, \ldots, N\}$ (where C_1 and C_2 are only depending on u). Note that this modified finite volume scheme is also consistent in the finite difference sense. Then, the finite difference techniques yield also an error estimate on $|e_i|$, but only of order h.

3. It could be tempting to try and find error estimates with respect to the mean value of the exact solution on the control volumes rather than with respect to its value at some point of the control volumes. This is not such a good idea: indeed, if x_i is not the center of K_i (this will be the general case in several space dimensions), then one does not have (in general) $|\tilde{e}_i| \leq C_3 h^2$ (for some C_3 only depending on u) with $\tilde{e}_i = \tilde{u}_i - u_i$ where $\tilde{u}_i$ denotes the mean value of u over K_i.

Remark 6.3

1. If the assumption $f \in C([0,1], \mathbb{R})$ is replaced by the assumption $f \in L^2((0,1))$ in Theorem 6.1, then $u \in H^2((0,1))$ instead of $C^2([0,1], \mathbb{R})$, but the estimates of Theorem 6.1 still hold. In this case, the consistency of the fluxes must be obtained with a Taylor expansion with an integral remainder. This is feasible for C^2 functions, and since the remainder only depends on the H^2 norm, a density argument allows to conclude; see also Theorem 9.4 page 55 below and EYMARD, GALLOUËT and HERBIN [55].

2. If the assumption $f \in C([0,1], \mathbb{R})$ is replaced by the assumption $f \in L^1((0,1))$ in Theorem 6.1, then $u \in C^2([0,1], \mathbb{R})$ no longer holds and neither does $u \in H^2((0,1))$, but the convergence still holds; indeed there exists $C(u,h)$, only depending on u and h, such that $C(u,h) \to 0$, as $h \to 0$, and $|e_i| \leq C(u,h)$, for all $i = 1, \ldots, N$. The proof is similar to the one above, except that the estimate (6.3) is replaced by $|R_{i+1/2}| \leq C_1(u,h)$, for all $i = 0, \ldots, N$, with some $C_1(u,h)$, only depending on u and h, such that $C(u,h) \to 0$, as $h \to 0$.

Remark 6.4 Estimate (6.1) can be interpreted as a "discrete H_0^1" estimate on the error. A theoretical result which underlies the L^∞ estimate (6.2) is the fact that if Ω is an open bounded subset of $\mathbb{R}$, then $H_0^1(\Omega)$ is imbedded in $L^\infty(\Omega)$. This is no longer true in higher dimension. In two space dimensions, for instance, a discrete version of the imbedding of H_0^1 in L^p allows to obtain (see e.g. FIARD [65]) $\|e\|_p \leq Ch$, for all finite p, which in turn yields $\|e\|_\infty \leq Ch \ln h$ for convenient meshes (see Corollary 9.1 page 62).

The important features needed for the above proof seem to be the consistency of the approximation of the fluxes and the conservativity of the scheme; this conservativity is natural the fact that the scheme is obtained by integrating the equation over each cell, and the approximation of the flux on any interface is obtained by taking into account the flux balance (continuity of the flux in the case of no source term on the interface).
The above proof generalizes to other elliptic problems, such as a convection-diffusion equation of the form $-u_{xx} + au_x + bu = f$, and to equations of the form $-(\lambda u_x)_x = f$ where $\lambda \in L^\infty$ may be discontinuous, and is such that there exist α and β in $\mathbb{R}_+^\star$ such that $\alpha \leq \lambda \leq \beta$. These generalizations are studied in the next section. Other generalizations include similar problems in 2 (or 3) space dimensions, with

meshes consisting of rectangles (parallepipeds), triangles (tetrahedra), or general meshes of Voronoï type, and the corresponding evolutive (parabolic) problems. These generalizations will be addressed in further chapters.
Let us now give a proof of Estimate (6.2), under slightly different conditions, which uses finite difference techniques.

6.2 An error estimate using a finite difference technique

Convergence can be obtained via a method similar to that of the finite difference proof of convergence (following, for instance, FORSYTH and SAMMON [69], MANTEUFFEL and WHITE [104], FAILLE [58]). Most of these methods, are, however, limited to the finite volume method for Problem (5.1). Using the notations of Section 5.2 (recall that $\overline{U} = (u(x_1), \ldots, u(x_N))^t$, and $r = A\overline{U} - b = 0(1)$), the idea is to find $\overline{\overline{U}}$ "close" to $\overline{U}$, such that

$$A\overline{\overline{U}} = b + \overline{\overline{r}}, \text{ with } \overline{\overline{r}} = 0(h).$$

This value of $\overline{\overline{U}}$ was found in FORSYTH and SAMMON [69] and is such that $\overline{\overline{U}} = \overline{U} - V$, where

$$V = (v_1, \ldots, v_N)^t \text{ and } v_i = \frac{h_i^2 u_{xx}(x_i)}{8},\ i = 1, \ldots, N.$$

Then, one may decompose the truncation error as

$$r = A(\overline{U} - U) = AV + \overline{\overline{r}} \text{ with } \|V\|_\infty = 0(h^2) \text{ and } \overline{\overline{r}} = 0(h).$$

The existence of such a V is given in Lemma 6.1. In order to prove the convergence of the scheme, a stability property is established in Lemma 6.2.

Lemma 6.1 *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $(0,1)$, in the sense of Definition 5.1 page 12, such that x_i is the center of K_i for all $i = 1, \ldots, N$. Let $\alpha_{\mathcal{T}} > 0$ be such that $h_i > \alpha_{\mathcal{T}} h$ for all $i = 1, \ldots, N$ (recall that $h = \max\{h_1, \ldots, h_N\}$). Let $\overline{U} = (u(x_1), \ldots, u(x_N))^t \in \mathbb{R}^N$, where u is the solution to (5.1), and assume $u \in C^3([0,1], \mathbb{R})$. Let A be the matrix defining the numerical scheme, given in (5.10) page 14. Then there exists a unique $U = (u_1, \ldots, u_N)$ solution of (5.3)-(5.6) and there exists $\overline{\overline{r}}$ and $V \in \mathbb{R}^N$ such that*

$$r = A(\overline{U} - U) = AV + \overline{\overline{r}},\ \textit{with } \|V\|_\infty \leq Ch^2 \textit{ and } \|\overline{\overline{r}}\|_\infty \leq Ch,$$

where C only depends on u and $\alpha_{\mathcal{T}}$.

PROOF of Lemma 6.1

The existence and uniqueness of U is classical (it is also proved in Theorem 6.1).
For $i = 0, \ldots N$, define

$$R_{i+\frac{1}{2}} = -\frac{u(x_{i+1}) - u(x_i)}{h_{i+\frac{1}{2}}} + u_x(x_{i+\frac{1}{2}}).$$

Remark that

$$r_i = \frac{1}{h_i}(R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}}),\ \text{for } i = 0, \ldots, N, \tag{6.7}$$

where r_i is the $i-$th component of $r = A(\overline{U} - U)$.
The computation of $R_{i+\frac{1}{2}}$ yields

$$\begin{aligned} &R_{i+\frac{1}{2}} = -\tfrac{1}{4}(h_{i+1} - h_i)u_{xx}(x_{i+\frac{1}{2}}) + 0(h^2),\ i = 1, \ldots, N-1, \\ &R_{\frac{1}{2}} = -\tfrac{1}{4}h_1 u_{xx}(0) + 0(h^2),\ R_{N+\frac{1}{2}} = \tfrac{1}{4}h_N u_{xx}(1) + 0(h^2). \end{aligned}$$

Define $V = (v_1, \ldots, v_N)^t$ with $v_i = \frac{h_i^2 u_{xx}(x_i)}{8}$, $i = 1, \ldots, N$. Then,

$$-\frac{v_{i+1}-v_i}{h_{i+\frac{1}{2}}} = R_{i+\frac{1}{2}} + 0(h^2),\ i = 1,\dots,N-1,$$
$$-\frac{2v_1}{h_1} = R_{\frac{1}{2}} + 0(h^2),$$
$$\frac{2v_N}{h_N} = R_{N+\frac{1}{2}} + 0(h^2).$$

Since $h_i \geq \alpha_{\mathcal{T}} h$, for $i = 1,\dots,N$, replacing $R_{i+\frac{1}{2}}$ in (6.7) gives that $r_i = (AV)_i + 0(h)$, for $i = 1,\dots,N$, and $\|V\|_\infty = 0(h^2)$. Hence the lemma is proved. ∎

Lemma 6.2 (Stability) *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $[0,1]$ in the sense of Definition 5.1. Let A be the matrix defining the finite volume scheme given in (5.10). Then A is invertible and*

$$\|A\|_\infty^{-1} \leq \frac{1}{4}. \tag{6.8}$$

Proof of Lemma 6.2

First we prove a discrete maximum principle; indeed if $b_i \geq 0$, for all $i = 1,\dots,N$, and if U is solution of $AU = b$ then we prove that $u_i \geq 0$ for all $i = 1,\dots,N$.
Let $a = \min\{u_i, i = 0,\dots,N+1\}$ (recall that $u_0 = u_{N+1} = 0$) and $i_0 = \min\{i \in \{0,\dots,N+1\};\ u_i = a\}$. If $i_0 \neq 0$ and $i_0 \neq N+1$, then

$$\frac{1}{h_{i_0}}\Big(\frac{u_{i_0}-u_{i_0-1}}{h_{i_0-\frac{1}{2}}} - \frac{u_{i_0+1}-u_{i_0}}{h_{i_0+\frac{1}{2}}}\Big) = b_{i_0} \geq 0,$$

this is impossible since $u_{i_0+1} - u_{i_0} \geq 0$ and $u_{i_0} - u_{i_0-1} < 0$, by definition of i_0. Therefore, $i_0 = 0$ or $N+1$. Then, $a = 0$ and $u_i \geq 0$ for all $i = 1,\dots,N$.
Note that, by linearity, this implies that A is invertible.
Next, we shall prove that there exists $M > 0$ such that $\|A^{-1}\|_\infty \leq M$ (indeed, $M = 1/4$ is convenient). Let ϕ be defined on $[0,1]$ by $\phi(x) = \frac{1}{2}x(1-x)$. Then $-\phi_{xx}(x) = 1$ for all $x \in [0,1]$. Let $\Phi = (\phi_1,\dots,\phi_N)$ with $\phi_i = \phi(x_i)$; if A represented the usual finite difference approximation of the second order derivative, then we would have $A\Phi = 1$, since the difference quotient approximation of the second order derivative of a second order polynomial is exact ($\phi_{xxx} = 0$). Here, with the finite volume scheme (5.3)-(5.6), we have $A\Phi - \mathbf{1} = AW$ (where $\mathbf{1}$ denotes the vector of $\mathbb{R}^N$ the components of which are all equal to 1), with $W = (w_1,\dots,w_N) \in \mathbb{R}^N$ such that $W_i = -\frac{h_i^2}{8}$ (see proof of Lemma 6.1). Let $b \in \mathbb{R}^N$ and $AU = b$, since $A(\Phi - W) = \mathbf{1}$, we have

$$A(U - \|b\|_\infty(\Phi - W)) \leq 0,$$

this last inequality being meant componentwise. Therefore, by the above maximum principle, assuming, without loss of generality, that $h \leq 1$, one has

$$u_i \leq \|b\|_\infty(\phi_i - w_i),\ \text{so that } u_i \leq \frac{\|b\|_\infty}{4}.$$

(note that $\phi(x) \leq \frac{1}{8}$). But we also have

$$A(U + \|b\|_\infty(\Phi - W)) \geq 0,$$

and again by the maximum principle, we obtain

$$u_i \geq -\frac{\|b\|_\infty}{4}.$$

Hence $\|U\|_\infty \leq \frac{1}{4}\|b\|_\infty$. This shows that $\|A^{-1}\|_\infty \leq \frac{1}{4}$. ∎

This stability result, together with the existence of V given by Lemma 6.1, yields the convergence of the finite volume scheme, formulated in the next theorem.

Theorem 6.2 *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $[0,1]$ in the sense of Definition 5.1 page 12. Let $\alpha_{\mathcal{T}} \in \mathbb{R}_+^\star$ be such that $h_i \geq \alpha_{\mathcal{T}} h$, for all $i = 1, \ldots, N$ (recall that $h = \max\{h_1, \ldots, h_N\}$). Let $\overline{U} = (u(x_1), \ldots, u(x_N))^t \in \mathbb{R}^N$, and assume $u \in C^3([0,1], \mathbb{R})$ (recall that u is the solution to (5.1)). Let $U = (u_1, \ldots, u_N)$ be the solution given by the numerical scheme (5.3)-(5.6). Then there exists $C > 0$, only depending on $\alpha_{\mathcal{T}}$ and u, such that $\|U - \overline{U}\|_\infty \leq Ch$.*

Remark 6.5 In the proof of Lemma 6.2, it was shown that $A(\overline{U} - V) = b + 0(h)$; therefore, if, once again, the finite volume scheme is considered as a finite difference scheme, it is consistent, in the finite difference sense, when u_i is considered to be an approximation of $u(x_i) - (1/8)h_i^2 u_{xx}(x_i)$.

Remark 6.6 With the notations of Lemma 6.1, let r be the function defined by

$$r(x) = r_i, \quad \text{if } x \in K_i, \quad i = 1, \ldots, N,$$

the function r does not necessarily go to 0 (as h goes to 0) in the L^∞ norm (and even in the L^1 norm), but, thanks to the conservativity of the scheme, it goes to 0 in $L^\infty((0,1))$ for the weak-$\star$ topology, that is

$$\int_0^1 r(x)\varphi(x)dx \to 0, \quad \text{as} \quad h \to 0, \quad \forall \varphi \in L^1((0,1)).$$

This property will be called "weak consistency" in the sequel and may also be used to prove the convergence of the finite volume scheme (see Faille [58]).

The proof of convergence described above may be easily generalized to the two-dimensional Laplace equation $-\Delta u = f$ in two and three space dimensions if a rectangular or a parallepipedic mesh is used, provided that the solution u is of class C^3. However, it does not seem to be easily generalized to other types of meshes.

7 General 1D elliptic equations

7.1 Formulation of the finite volume scheme

This section is devoted to the formulation and to the proof of convergence of a finite volume scheme for a one-dimensional linear convection-diffusion equation, with a discontinuous diffusion coefficient. The scheme can be generalized in the two-dimensional and three-dimensional cases (for a space discretization which uses, for instance, simplices or parallelepipedes or a "Voronoï mesh", see Section 9.2 page 37) and to other boundary conditions.

Let $\lambda \in L^\infty((0,1))$ such that there exist $\underline{\lambda}$ and $\overline{\lambda} \in \mathbb{R}_+^\star$ with $\underline{\lambda} \leq \lambda \leq \overline{\lambda}$ a.e. and let $a, b, c, d \in \mathbb{R}$, with $b \geq 0$, and $f \in L^2((0,1))$. The aim, here, is to find an approximation to the solution, u, of the following problem:

$$-(\lambda u_x)_x(x) + a u_x(x) + b u(x) = f(x), \ x \in [0,1], \tag{7.1}$$

$$u(0) = c, \ u(1) = d. \tag{7.2}$$

The discontinuity of the coefficient λ may arise for instance for the permeability of a porous medium, the ratio between the permeability of sand and the permeability of clay being of an order of 10^3; heat conduction in a heterogeneous medium can also yield such discontinuities, since the conductivities of the different components of the medium may be quite different. Note that the assumption $b \geq 0$ ensures the existence of the solution to the problem.

Remark 7.1 Problem (7.1)-(7.2) has a unique solution u in the Sobolev space $H^1((0,1))$. This solution is continuous (on $[0,1]$) but is not, in general, of class C^2 (even if $\lambda(x) = 1$, for all $x \in [0,1]$). Note that one has $-\lambda u_x(x) = \int_0^x g(t)dt + C$, where C is some constant and $g = f - a u_x - bu \in L^1((0,1))$, so that λu_x is a continuous function and $u_x \in L^\infty((0,1))$.

Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh, in the sense of Definition 5.1 page 12, such that the discontinuities of λ coincide with the interfaces of the mesh.
The notations being the same as in section 5, integrating Equation (7.1) over K_i yields

$$-(\lambda u_x)(x_{i+\frac{1}{2}}) + (\lambda u_x)(x_{i-\frac{1}{2}}) + au(x_{i+\frac{1}{2}}) - au(x_{i-\frac{1}{2}}) + \int_{K_i} bu(x)dx = \int_{K_i} f(x)dx, \quad i = 1, \ldots, N.$$

Let $(u_i)_{i=1,\cdots,N}$ be the discrete unknowns. In the case $a \geq 0$, which will be considered in the sequel, the convective term $au(x_{i+1/2})$ is approximated by au_i ("upstream") because of stability considerations. Indeed, this choice always yields a stability result whereas the approximation of $au(x_{i+1/2})$ by $(a/2)(u_i + u_{i+1})$ (with the approximation of the other terms as it is done below) yields a stable scheme if $ah \leq 2\lambda$, for a uniform mesh of size h and a constant diffusion coefficient λ. The case $a \leq 0$ is easily handled in the same way by approximating $au(x_{i+1/2})$ by au_{i+1}. The term $\int_{K_i} bu(x)dx$ is approximated by bh_iu_i. Let us now turn to the approximation $H_{i+1/2}$ of $-\lambda u_x(x_{i+1/2})$. Let $\lambda_i = \frac{1}{h_i}\int_{K_i} \lambda(x)dx$; since $\lambda|_{K_i} \in C^1(\bar{K_i})$, there exists $c_\lambda \in \mathbb{R}_+$, only depending on λ, such that $|\lambda_i - \lambda(x)| \leq c_\lambda h$, $\forall\, x \in K_i$. In order that the scheme be conservative, the discretization of the flux at $x_{i+1/2}$ should have the same value on K_i and K_{i+1}. To this purpose, we introduce the auxiliary unknown $u_{i+1/2}$ (approximation of u at $x_{i+1/2}$). Since on K_i and K_{i+1}, λ is continuous, the approximation of $-\lambda u_x$ may be performed on each side of $x_{i+1/2}$ by using the finite difference principle:

$$H_{i+\frac{1}{2}} = -\lambda_i \frac{u_{i+\frac{1}{2}} - u_i}{h_i^+} \text{ on } K_i,\ i = 1, \ldots, N,$$

$$H_{i+\frac{1}{2}} = -\lambda_{i+1} \frac{u_{i+1} - u_{i+\frac{1}{2}}}{h_{i+1}^-} \text{ on } K_{i+1},\ i = 0, \ldots, N-1,$$

with $u_{1/2} = c$, and $u_{N+1/2} = d$, for the boundary conditions. (Recall that $h_i^+ = x_{i+1/2} - x_i$ and $h_i^- = x_i - x_{i-1/2}$). Requiring the two above approximations of $\lambda u_x(x_{i+1/2})$ to be equal (conservativity of the flux) yields the value of $u_{i+1/2}$ (for $i = 1, \ldots, N-1$):

$$u_{i+\frac{1}{2}} = \frac{u_{i+1}\dfrac{\lambda_{i+1}}{h_{i+1}^-} + u_i\dfrac{\lambda_i}{h_i^+}}{\dfrac{\lambda_{i+1}}{h_{i+1}^-} + \dfrac{\lambda_i}{h_i^+}} \tag{7.3}$$

which, in turn, allows to give the expression of the approximation $H_{i+\frac{1}{2}}$ of $\lambda u_x(x_{i+\frac{1}{2}})$:

$$\begin{aligned} &H_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u_{i+1} - u_i),\ i = 1, \ldots, N-1, \\ &H_{\frac{1}{2}} = -\frac{\lambda_1}{h_1^-}(u_1 - c), \\ &H_{N+\frac{1}{2}} = -\frac{\lambda_N}{h_N^+}(d - u_N) \end{aligned} \tag{7.4}$$

with

$$\tau_{i+\frac{1}{2}} = \frac{\lambda_i\lambda_{i+1}}{h_i^+\lambda_{i+1} + h_{i+1}^-\lambda_i},\ i = 1, \ldots, N-1. \tag{7.5}$$

Example 7.1 If $h_i = h$, for all $i \in \{1, \ldots, N\}$, and x_i is assumed to be the center of K_i, then $h_i^+ = h_i^- = \frac{h}{2}$, so that

$$H_{i+\frac{1}{2}} = -\frac{2\lambda_i\lambda_{i+1}}{\lambda_i + \lambda_{i+1}} \frac{u_{i+1} - u_i}{h},$$

and therefore the mean harmonic value of λ is involved.

The numerical scheme for the approximation of Problem (7.1)-(7.2) is therefore,

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + bh_i u_i = h_i f_i, \ \forall i \in \{1, \ldots, N\}, \tag{7.6}$$

with $f_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx$, for $i = 1, \ldots, N$, and where $(F_{i+\frac{1}{2}})_{i\in\{0,\ldots,N\}}$ is defined by the following expressions

$$F_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u_{i+1} - u_i) + au_i, \ \forall i \in \{1, \ldots, N-1\}, \tag{7.7}$$

$$F_{\frac{1}{2}} = -\frac{\lambda_1}{h_1^-}(u_1 - c) + ac, \ F_{N+\frac{1}{2}} = -\frac{\lambda_N}{h_N^+}(d - u_N) + au_N. \tag{7.8}$$

Remark 7.2 In the case $a \geq 0$, the choice of the approximation of $au(x_{i+1/2})$ by au_{i+1} would yield an unstable scheme, except for h small enough (when $a \leq 0$, the unstable scheme is au_i).

Taking (7.5), (7.7) and (7.8) into account, the numerical scheme (7.6) yields a system of N equations with N unknowns $u_1, \ldots, u_N$.

7.2 Error estimate

Theorem 7.1
Let $a, b \geq 0$, $c, d \in \mathbb{R}$, $\lambda \in L^\infty((0,1))$ such that $\underline{\lambda} \leq \lambda \leq \overline{\lambda}$ a.e. with some $\underline{\lambda}$, $\overline{\lambda} \in \mathbb{R}_+^\star$ and $f \in L^1((0,1))$. Let u be the (unique) solution of (7.1)-(7.2). Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh, in the sense of Definition 5.1, such that $\lambda \in C^1(\overline{K}_i)$ and $f \in C(\overline{K}_i)$, for all $i = 1, \cdots, N$. Let $\gamma = \max\{\|u_{xx}\|_{L^\infty(K_i)}, i = 1, \cdots, N\}$ and $\delta = \max\{\|\lambda\|_{L^\infty(K_i)}, i = 1, \cdots, N\}$. Then,

1. *there exists a unique vector $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution to (7.5)-(7.8),*
2. *there exists C, only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, such that*

$$\sum_{i=0}^{N} \tau_{i+\frac{1}{2}}(e_{i+1} - e_i)^2 \leq Ch^2, \tag{7.9}$$

where $\tau_{i+\frac{1}{2}}$ is defined in (7.5), and

$$|e_i| \leq Ch, \ \forall i \in \{1, \ldots, N\}, \tag{7.10}$$

with $e_0 = e_{N+1} = 0$ and $e_i = u(x_i) - u_i$, for all $i \in \{1, \ldots, N\}$.

Proof of Theorem 7.1

Step 1. Existence and uniqueness of the solution to (7.5)-(7.8).
Multiplying (7.6) by u_i and summing for $i = 1, \ldots, N$ yields that if $c = d = 0$ and $f_i = 0$ for any $i \in \{1, \ldots, N\}$, then the unique solution to (7.5)-(7.8) is obtained by taking $u_i = 0$, for any $i \in \{1, \ldots, N\}$. This yields existence and uniqueness of the solution to (7.5)-(7.8).

Step 2. Consistency of the fluxes.
Recall that $h = \max\{h_1, \ldots, h_N\}$. Let us first show the consistency of the fluxes.
Let $\overline{H}_{i+1/2} = -(\lambda u_x)(x_{i+1/2})$ and $H^\star_{i+1/2} = -\tau_{i+1/2}(u(x_{i+1}) - u(x_i))$, for $i = 0, \ldots, N$, with $\tau_{1/2} = \lambda_1/h_1^-$ and $\tau_{N+1/2} = \lambda_N/h_N^+$. Let us first show that there exists $C_1 \in \mathbb{R}_+^\star$, only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, such that

$$\begin{array}{l} H^\star_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + T_{i+\frac{1}{2}}, \\ |T_{i+\frac{1}{2}}| \leq C_1 h, \ i = 0, \ldots, N. \end{array} \tag{7.11}$$

In order to show this, let us introduce

$$H^{\star,-}_{i+\frac{1}{2}} = -\lambda_i \frac{u(x_{i+\frac{1}{2}}) - u(x_i)}{h_i^+} \text{ and } H^{\star,+}_{i+\frac{1}{2}} = -\lambda_{i+1} \frac{u(x_{i+1}) - u(x_{i+\frac{1}{2}})}{h_{i+1}^-}; \tag{7.12}$$

since $\lambda \in C^1(\bar{K}_i)$, one has $u \in C^2(\bar{K}_i)$; hence, there exists $C \in \mathbb{R}_+^\star$, only depending on γ and δ, such that

$$H^{\star,-}_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + R^-_{i+\frac{1}{2}}, \text{ where } |R^-_{i+\frac{1}{2}}| \leq Ch, \, i = 1, \ldots, N, \tag{7.13}$$

and

$$H^{\star,+}_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + R^+_{i+\frac{1}{2}}, \text{ where } |R^+_{i+\frac{1}{2}}| \leq Ch, \, i = 0, \ldots, N-1. \tag{7.14}$$

This yields (7.11) for $i = 0$ and $i = N$.
The following equality:

$$\overline{H}_{i+\frac{1}{2}} = H^{\star,-}_{i+\frac{1}{2}} - R^-_{i+\frac{1}{2}} = H^{\star,+}_{i+\frac{1}{2}} - R^+_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1, \tag{7.15}$$

yields that

$$u(x_{i+\frac{1}{2}}) = \frac{\frac{\lambda_{i+1}}{h_{i+1}^-} u(x_{i+1}) + \frac{\lambda_i}{h_i^+} u(x_i)}{\frac{\lambda_i}{h_i^+} + \frac{\lambda_{i+1}}{h_{i+1}^-}} + S_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1, \tag{7.16}$$

where

$$S_{i+\frac{1}{2}} = \frac{R^+_{i+\frac{1}{2}} - R^-_{i+\frac{1}{2}}}{\frac{\lambda_i}{h_i^+} + \frac{\lambda_{i+1}}{h_{i+1}^-}}$$

so that

$$|S_{i+\frac{1}{2}}| \leq \frac{1}{\underline{\lambda}} \frac{h_i^+ h_{i+1}^-}{h_i^+ + h_{i+1}^-} |R^+_{i+\frac{1}{2}} - R^-_{i+\frac{1}{2}}|.$$

Let us replace the expression (7.16) of $u(x_{i+1/2})$ in $H^{\star,-}_{i+1/2}$ defined by (7.12) (note that the computation is similar to that performed in (7.3)-(7.4)); this yields

$$H^{\star,-}_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u(x_{i+1}) - u(x_i)) - \frac{\lambda_i}{h_i^+} S_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1. \tag{7.17}$$

Using (7.15), this implies that $H^\star_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + T_{i+\frac{1}{2}}$ where

$$|T_{i+\frac{1}{2}}| \leq |R^-_{i+\frac{1}{2}}| + |R^+_{i+\frac{1}{2}} - R^-_{i+\frac{1}{2}}| \frac{\overline{\lambda}}{2\underline{\lambda}}.$$

Using (7.13) and (7.14), this last inequality yields that there exists C_1, only depending on $\overline{\lambda}, \underline{\lambda}, \gamma, \delta$, such that

$$|H^\star_{i+\frac{1}{2}} - \overline{H}_{i+\frac{1}{2}}| = |T_{i+\frac{1}{2}}| \leq C_1 h, \, i = 1, \ldots, N-1.$$

Therefore (7.11) is proved.
Define now the total exact fluxes;

$$\overline{F}_{i+\frac{1}{2}} = -(\lambda u_x)(x_{i+\frac{1}{2}}) + a u(x_{i+\frac{1}{2}}), \, \forall i \in \{0, \ldots, N\},$$

and define

$$F^\star_{i+\frac12} = -\tau_{i+\frac12}(u(x_{i+1}) - u(x_i)) + au(x_i),\ \forall i \in \{1,\dots,N-1\},$$

$$F^\star_{\frac12} = -\frac{\lambda_1}{h_1^-}(u(x_1) - c) + ac,\ F^\star_{N+\frac12} = -\frac{\lambda_N}{h_N^+}(d - u(x_N)) + au_N.$$

Then, from (7.11) and the regularity of u, there exists C_2, only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, such that

$$F^\star_{i+\frac12} = \overline{F}_{i+\frac12} + R_{i+\frac12},\ \text{with } |R_{i+\frac12}| \le C_2 h,\ i = 0,\dots,N. \tag{7.18}$$

Hence the numerical approximation of the flux is consistent.

Step 3. Error estimate.
Integrating Equation (7.1) over each control volume yields that

$$\overline{F}_{i+\frac12} - \overline{F}_{i-\frac12} + bh_i(u(x_i) + S_i) = h_i f_i,\ \forall i \in \{1,\dots,N\}, \tag{7.19}$$

where $S_i \in \mathbb{R}$ is such that there exists C_3 only depending on u such that $|S_i| \le C_3 h$, for $i = 1,\dots,N$. Using (7.18) yields that

$$F^\star_{i+\frac12} - F^\star_{i-\frac12} + bh_i(u(x_i) + S_i) = h_i f_i + R_{i+\frac12} - R_{i-\frac12},\ \forall i \in \{1,\dots,N\}. \tag{7.20}$$

Let $e_i = u(x_i) - u_i$, for $i = 1,\dots,N$, and $e_0 = e_{N+1} = 0$. Substracting (7.6) from (7.20) yields

$$-\tau_{i+\frac12}(e_{i+1} - e_i) + \tau_{i-\frac12}(e_i - e_{i-1}) + a(e_i - e_{i-1}) + bh_i e_i = -bh_i S_i + R_{i+\frac12} - R_{i-\frac12},\ \forall i \in \{1,\dots,N\}.$$

Let us multiply this equation by e_i, sum for $i = 1,\dots,N$, reorder the summations. Remark that

$$\sum_{i=1}^{N} e_i(e_i - e_{i-1}) = \frac12 \sum_{i=1}^{N+1} (e_i - e_{i-1})^2$$

and therefore

$$\sum_{i=0}^{N} \tau_{i+\frac12}(e_{i+1} - e_i)^2 + \frac{a}{2}\sum_{i=1}^{N+1}(e_i - e_{i-1})^2 + \sum_{i=1}^{N} bh_i e_i^2 = -\sum_{i=1}^{N} bh_i S_i e_i - \sum_{i=0}^{N} R_{i+\frac12}(e_{i+1} - e_i).$$

Since $|S_i| \le C_3 h$ and thanks to (7.18), one has

$$\sum_{i=0}^{N} \tau_{i+\frac12}(e_{i+1} - e_i)^2 \le \sum_{i=1}^{N} bC_3 h_i h|e_i| + \sum_{i=1}^{N} C_2 h|e_{i+1} - e_i|.$$

Remark that $|e_i| \le \sum_{j=1}^{N} |e_j - e_{j-1}|$. Denote by $A = \left(\sum_{i=0}^{N} \tau_{i+\frac12}(e_{i+1} - e_i)^2\right)^{\frac12}$ and $B = \left(\sum_{i=0}^{N} \frac{1}{\tau_{i+\frac12}}\right)^{\frac12}$. The Cauchy-Schwarz inequality yields

$$A^2 \le \sum_{i=1}^{N} bC_3 h_i h AB + C_2 h AB.$$

Now, since

$$\frac{1}{\tau_{i+\frac12}} \le \frac{\overline{\lambda}}{\underline{\lambda}^2}(h_{i+1}^- + h_i^+),\ \sum_{i=0}^{N}(h_{i+1}^- + h_i^+) = 1,\ \text{with } h_0^+ = h_{N+1}^- = 0,\ \text{and } \sum_{i=1}^{N} h_i = 1,$$

one obtains that $A \le C_4 h$, with C_4 only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, which yields Estimate (7.9). Applying once again the Cauchy-Schwarz inequality yields Estimate (7.10). ∎

7.3 The case of a point source term

In many physical problems, some discontinuous or point source terms appear. In the case where a source term exists at the interface $x_{i+1/2}$, the fluxes relative to K_i and K_{i+1} will differ because of this source term. The computation of the fluxes is carried out in a similar way, writing that the sum of the approximations of the fluxes must be equal to the source term at the interface. Consider again the one-dimensional conservation problem (7.1), (7.2) (with, for the sake of simplification, $a = b = c = d = 0$, we use below the notations of the previous section), but assume now that at $\underline{x} \in (0,1)$, a point source of intensity α exists. In this case, the problem may be written in the following way:

$$-(\lambda u_x(x))_x = f(x), \quad x \in (0,\underline{x}) \cup (\underline{x},1), \tag{7.21}$$

$$u(0) = 0, \tag{7.22}$$

$$u(1) = 0, \tag{7.23}$$

$$(\lambda u_x)^+(\underline{x}) - (\lambda u_x)^-(\underline{x}) = -\alpha, \tag{7.24}$$

where

$$(\lambda u_x)^+(\underline{x}) = \lim_{x\to\underline{x}, x>\underline{x}} (\lambda u_x)(x) \text{ and } (\lambda u_x)^-(\underline{x}) = \lim_{x\to\underline{x}, x<\underline{x}} (\lambda u_x)(x).$$

Equation (7.24) states that the flux is discontinuous at point $\underline{x}$. Another formulation of the problem is the following:

$$-(\lambda u_x)_x = g \text{ in } \mathcal{D}'((0,1)), \tag{7.25}$$

$$u(0) = 0, \tag{7.26}$$

$$u(1) = 0, \tag{7.27}$$

where $g = f + \alpha\delta_{\underline{x}}$, where $\delta_{\underline{x}}$ denotes the Dirac measure, which is defined by $< \delta_{\underline{x}}, \varphi >_{\mathcal{D}',\mathcal{D}} = \varphi(\underline{x})$, for any $\varphi \in \mathcal{D}((0,1)) = C_c^\infty((0,1), \mathbb{R})$, and $\mathcal{D}'((0,1))$ denotes the set of distributions on (0,1), i.e. the set of continuous linear forms on $\mathcal{D}((0,1))$.
Assuming the mesh to be such that $\underline{x} = x_{i+1/2}$ for some $i \in 1, \ldots, N-1$, the equation corresponding to the unknown u_i is $F^-_{i+1/2} - F_{i-1/2} = \int_{K_i} f(x)dx$, while the equation corresponding to the unknown u_{i+1} is $F_{i+3/2} - F^+_{i+1/2} = \int_{K_{i+1}} f(x)dx$. In order to compute the values of the numerical fluxes $F^\pm_{i+1/2}$, one must take the source term into account while writing the conservativity of the flux; hence at $x_{i+1/2}$, the two numerical fluxes at $x = \underline{x}$, namely $F^+_{i+\frac{1}{2}}$ and $F^-_{i+\frac{1}{2}}$, must satisfy, following Equation (7.24),

$$F^+_{i+\frac{1}{2}} - F^-_{i+\frac{1}{2}} = \alpha. \tag{7.28}$$

Next, the fluxes $F^+_{i+1/2}$ and $F^-_{i+1/2}$ must be expressed in terms of the discrete variables u_k, $k = 1, \ldots, N$; in order to do so, introduce the auxiliary variable $u_{i+1/2}$ (which will be eliminated later), and write

$$F^+_{i+\frac{1}{2}} = -\lambda_{i+1}\frac{u_{i+1} - u_{i+\frac{1}{2}}}{h^-_{i+1}}$$

$$F^-_{i+\frac{1}{2}} = -\lambda_i\frac{u_{i+\frac{1}{2}} - u_i}{h^+_i}.$$

Replacing these expressions in (7.28) yields

$$u_{i+\frac{1}{2}} = \frac{h^+_i h^-_{i+1}}{(h^-_{i+1}\lambda_i + h^+_i\lambda_{i+1})}[\frac{\lambda_{i+1}}{h^-_{i+1}}u_{i+1} + \frac{\lambda_i}{h^+_i}u_i + \alpha].$$

and therefore

$$F^+_{i+\frac{1}{2}} = \frac{h^+_i \lambda_{i+1}}{h^-_{i+1}\lambda_i + h^+_i \lambda_{i+1}}\alpha - \frac{\lambda_i \lambda_{i+1}}{h^-_{i+1}\lambda_i + h^+_i \lambda_{i+1}}(u_{i+1} - u_i)$$

$$F^-_{i+\frac{1}{2}} = \frac{-h^-_{i+1}\lambda_i}{h^-_{i+1}\lambda_i + h^+_i \lambda_{i+1}}\alpha - \frac{\lambda_i \lambda_{i+1}}{h^-_{i+1}\lambda_i + h^+_i \lambda_{i+1}}(u_{i+1} - u_i).$$

Note that the source term α is distributed on either side of the interface proportionally to the coefficient λ, and that, when $\alpha = 0$, the above expressions lead to

$$F^+_{i+\frac{1}{2}} = F^-_{i+\frac{1}{2}} = -\frac{\lambda_i \lambda_{i+1}}{h^-_{i+1}\lambda_i + h^+_i \lambda_{i+1}}(u_{i+1} - u_i).$$

Note that the error estimate given in Theorem 7.1 still holds in this case (under adequate assumptions).

8 A semilinear elliptic problem

8.1 Problem and Scheme

This section is concerned with the proof of convergence for some nonlinear problems. We are interested, as an example, by the following problem:

$$-u_{xx}(x) = f(x, u(x)),\ x \in (0,1), \tag{8.1}$$

$$u(0) = u(1) = 0, \tag{8.2}$$

with a function $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ such that

$$\begin{array}{c} f(x,s) \text{ is measurable with respect to } x \in (0,1) \text{ for all } s \in \mathbb{R} \\ \text{and continuous with respect to } s \in \mathbb{R} \text{ for a.e. } x \in (0,1), \end{array} \tag{8.3}$$

$$f \in L^\infty((0,1) \times \mathbb{R}). \tag{8.4}$$

It is possible to prove that there exists at least one weak solution to (8.1), (8.2), that is a function u such that

$$u \in H^1_0((0,1)),\ \int_0^1 u_x(x) v_x(x) dx = \int_0^1 f(x, u(x)) v(x) dx,\ \forall v \in H^1_0((0,1)). \tag{8.5}$$

Note that (8.5) is equivalent to "$u \in H^1_0((0,1))$ and $-u_{xx} = f(\cdot, u)$ in the distribution sense in $(0,1)$". The proof of the existence of such a solution is possible by using, for instance, the Schauder's fixed point theorem (see e.g. Deimling [45]) or by using the convergence theorem 8.1 which is proved in the sequel.

Let $\mathcal{T}$ be an admissible mesh of $[0,1]$ in the sense of Definition 5.1. In order to discretize (8.1), (8.2), let us consider the following (finite volume) scheme

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i(u_i),\ i = 1, \ldots, N, \tag{8.6}$$

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}},\ i = 0, \ldots, N, \tag{8.7}$$

$$u_0 = u_{N+1} = 0, \tag{8.8}$$

with $f_i(u_i) = \frac{1}{h_i}\int_{K_i} f(x, u_i) dx$, $i = 1, \ldots, N$.
The discrete unknowns are therefore $u_1, \ldots, u_N$. In order to give a convergence result for this scheme (Theorem 8.1), one first proves the existence of a solution to (8.6)-(8.8), a stability result, that is, an estimate on the solution of (8.6)-(8.8) (Lemma 8.1) and a compactness lemma (Lemma 8.2).

Lemma 8.1 (Existence and stability result) *Let $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ satisfying (8.3), (8.4) and $\mathcal{T}$ be an admissible mesh of $(0,1)$ in the sense of Definition 5.1. Then, there exists $(u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution of (8.6)-(8.8) and which satisfies:*

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq C, \tag{8.9}$$

for some $C \geq 0$ only depending on f.

Proof of Lemma 8.1

Define $M = \|f\|_{L^\infty((0,1)\times\mathbb{R})}$. The proof of estimate (8.9) is given in a first step, and the existence of a solution to (8.6)-(8.8) in a second step.

Step 1 (Estimate)

Let $V = (v_1, \ldots, v_N)^t \in \mathbb{R}^N$, there exists a unique $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution of (8.6)-(8.8) with $f_i(v_i)$ instead of $f_i(u_i)$ in the right hand-side (see Theorem 6.1 page 16). One sets $U = F(V)$, so that F is a continuous application from $\mathbb{R}^N$ to $\mathbb{R}^N$, and $(u_1, \ldots, u_N)$ is a solution to (8.6)-(8.8) if and only if $U = (u_1, \ldots, u_N)^t$ is a fixed point to F.

Multiplying (8.6) by u_i and summing over i yields

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq M \sum_{i=1}^{N} h_i |u_i|, \tag{8.10}$$

and from the Cauchy-Schwarz inequality, one has

$$|u_i| \leq \Big(\sum_{j=0}^{N} \frac{(u_{j+1} - u_j)^2}{h_{j+\frac{1}{2}}}\Big)^{\frac{1}{2}},\ i = 1, \ldots, N,$$

then (8.10) yields, with $C = M^2$,

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq C. \tag{8.11}$$

This gives, in particular, Estimate (8.9) if $(u_1, \ldots, u_N)^t \in \mathbb{R}^N$ is a solution of (8.6)-(8.8) (that is $u_i = v_i$ for all i).

Step 2 (Existence)

The application $F : \mathbb{R}^N \to \mathbb{R}^N$ defined above is continuous and, taking in $\mathbb{R}^N$ the norm

$$\|V\| = \Big(\sum_{i=0}^{N} \frac{(v_{i+1} - v_i)^2}{h_{i+\frac{1}{2}}}\Big)^{\frac{1}{2}},\ \text{for } V = (v_1, \ldots, v_N)^t,\ \text{with } v_0 = v_{N+1} = 0,$$

one has $F(B_M) \subset B_M$, where B_M is the closed ball of radius M and center 0 in $\mathbb{R}^N$. Then, F has a fixed point in B_M thanks to the Brouwer fixed point theorem (see e.g. Deimling [45]). This fixed point is a solution to (8.6)-(8.8). ∎

8.2 Compactness results

.

Lemma 8.2 (Compactness)
For an admissible mesh $\mathcal{T}$ of $(0,1)$ (see definition 5.1), let $(u_1,\dots,u_N)^t \in \mathbb{R}^N$ satisfy (8.9) for some $C \in \mathbb{R}$ (independent of $\mathcal{T}$) and let $u_\mathcal{T} : (0,1) \to \mathbb{R}$ be defined by $u_\mathcal{T}(x) = u_i$ if $x \in K_i$, $i = 1,\dots,N$. Then, the set $\{u_\mathcal{T}$, $\mathcal{T}$ admissible mesh of $(0,1)\}$ is relatively compact in $L^2((0,1))$. Furthermore, if $u_{\mathcal{T}_n} \to u$ in $L^2((0,1))$ and $\mathrm{size}(\mathcal{T}_n) \to 0$, as $n \to \infty$, then, $u \in H_0^1((0,1))$.

Proof of Lemma 8.2

A possible proof is to use "classical" compactness results, replacing $u_\mathcal{T}$ by a continuous function, say $\overline{u}_\mathcal{T}$, piecewise affine, such that $\overline{u}_\mathcal{T}(x_i) = u_i$ for $i = 1,\dots,N$, and $\overline{u}_\mathcal{T}(0) = \overline{u}_\mathcal{T}(1) = 0$. The set $\{\overline{u}_\mathcal{T}$, $\mathcal{T}$ admissible mesh of $(0,1)\}$ is then bounded in $H_0^1((0,1))$, see Remark 9.9 page 49.
Another proof is given here, the interest of which is its simple generalization to multidimensional cases (such as the case of one unknown per triangle in 2 space dimensions, see Section 9.2 page 37 and Section 14 page 93) when the construction of such a function, $\overline{u}_\mathcal{T}$, "close" to $u_\mathcal{T}$ and bounded in $H_0^1((0,1))$ (independently of $\mathcal{T}$), is not so easy.

In order to have $u_\mathcal{T}$ defined on $\mathbb{R}$, one sets $u_\mathcal{T}(x) = 0$ for $x \notin [0,1]$. The proof may be decomposed into four steps.

Step 1. First remark that the set $\{u_\mathcal{T}$, $\mathcal{T}$ an admissible mesh of $(0,1)\}$ is bounded in $L^2(\mathbb{R})$. Indeed, this an easy consequence of (8.9), since one has, for all $x \in [0,1]$ (since $u_0 = 0$ and by the Cauchy-Schwarz inequality),

$$|u_\mathcal{T}(x)| \le \sum_{i=0}^{N} |u_{i+1} - u_i| \le \Big(\sum_{i=0}^{N} \frac{(u_{i+1}-u_i)^2}{h_{i+\frac{1}{2}}}\Big)^{\frac{1}{2}} \le C.$$

Step 2. Let $0 < \eta < 1$. One proves, in this step, that

$$\|u_\mathcal{T}(\cdot+\eta) - u_\mathcal{T}\|^2_{L^2(\mathbb{R})} \le C\eta(\eta+2h). \tag{8.12}$$

(Recall that $h = \mathrm{size}(\mathcal{T})$.)
Indeed, for $i \in \{0,\dots,N\}$ define $\chi_{i+1/2} : \mathbb{R} \to \mathbb{R}$, by $\chi_{i+1/2}(x) = 1$, if $x_{i+1/2} \in [x, x+\eta]$ and $\chi_{i+1/2}(x) = 0$, if $x_{i+1/2} \notin [x, x+\eta]$. Then, one has, for all $x \in \mathbb{R}$,

$$\begin{aligned}(u_\mathcal{T}(x+\eta) - u_\mathcal{T}(x))^2 &\le \Big(\sum_{i=0}^{N} |u_{i+1}-u_i|\chi_{i+\frac{1}{2}}(x)\Big)^2 \\ &\le \Big(\sum_{i=0}^{N} \frac{(u_{i+1}-u_i)^2}{h_{i+\frac{1}{2}}}\chi_{i+\frac{1}{2}}(x)\Big)\Big(\sum_{i=0}^{N}\chi_{i+\frac{1}{2}}(x)h_{i+\frac{1}{2}}\Big).\end{aligned} \tag{8.13}$$

Since $\sum_{i=0}^{N} \chi_{i+1/2}(x)h_{i+1/2} \le \eta + 2h$, for all $x \in \mathbb{R}$, and $\int_{\mathbb{R}} \chi_{i+1/2}(x)dx = \eta$, for all $i \in \{0,\dots,N\}$, integrating (8.13) over $\mathbb{R}$ yields (8.12).

Step 3. For $0 < \eta < 1$, Estimate (8.12) implies that

$$\|u_\mathcal{T}(\cdot+\eta) - u_\mathcal{T}\|^2_{L^2(\mathbb{R})} \le 3C\eta.$$

This gives (with Step 1), by the Kolmogorov compactness theorem (recalled in Section 14, see Theorem 14.1 page 94), the relative compactness of the set $\{u_\mathcal{T}$, $\mathcal{T}$ an admissible mesh of $(0,1)\}$ in $L^2((0,1))$ and also in $L^2(\mathbb{R})$ (since $u_\mathcal{T} = 0$ on $\mathbb{R} \setminus [0,1]$).

Step 4. In order to conclude the proof of Lemma 8.2, one may use Theorem 14.2 page 94, which we prove here in the one-dimensional case for the sake of clarity. Let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a sequence of admissible meshes of $(0,1)$ such that $\mathrm{size}(\mathcal{T}_n) \to 0$ and $u_{\mathcal{T}_n} \to u$, in $L^2((0,1))$, as $n \to \infty$. Note that $u_{\mathcal{T}_n} \to u$, in $L^2(\mathbb{R})$,

with $u = 0$ on $\mathbb{R} \setminus [0,1]$. For a given $\eta \in (0,1)$, let $n \to \infty$ in (8.12), with $u_{\mathcal{T}_n}$ instead of $u_{\mathcal{T}}$ (and size$(\mathcal{T}_n)$ instead of h). One obtains

$$\|\frac{u(\cdot+\eta)-u}{\eta}\|^2_{L^2(\mathbb{R})} \leq C. \tag{8.14}$$

Since $(u(\cdot+\eta)-u)/\eta$ tends to Du (the distribution derivative of u) in the distribution sense, as $\eta \to 0$, Estimate (8.14) yields that $Du \in L^2(\mathbb{R})$. Furthermore, since $u = 0$ on $\mathbb{R} \setminus [0,1]$, the restriction of u to $(0,1)$ belongs to $H^1_0((0,1))$. The proof of Lemma 8.2 is complete. ∎

.

8.3 Convergence

The following convergence result follows from lemmata 8.1 and 8.2.

Theorem 8.1 *Let $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ satisfying (8.3), (8.4). For an admissible mesh, $\mathcal{T}$, of $(0,1)$ (see Definition 5.1), let $(u_1, \ldots, u_N)^t \in \mathbb{R}^N$ be a solution to (8.6)-(8.8) (the existence of which is given by Lemma 8.1), and let $u_{\mathcal{T}} : (0,1) \to \mathbb{R}$ by $u_{\mathcal{T}}(x) = u_i$, if $x \in K_i$, $i = 1, \ldots, N$.*
Then, for any sequence $(\mathcal{T}_n)_{n\in\mathbb{N}}$ of admissible meshes such that size$(\mathcal{T}_n) \to 0$, *as $n \to \infty$, there exists a subsequence, still denoted by $(\mathcal{T}_n)_{n\in\mathbb{N}}$, such that $u_{\mathcal{T}_n} \to u$, in $L^2((0,1))$, as $n \to \infty$, where $u \in H^1_0((0,1))$ is a weak solution to (8.1), (8.2) (that is, a solution to (8.5)).*

Proof of Theorem 8.1

Let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a sequence of admissible meshes of $(0,1)$ such that size$(\mathcal{T}_n) \to 0$, as $n \to \infty$. By lemmata 8.1 and 8.2, there exists a subsequence, still denoted by $(\mathcal{T}_n)_{n\in\mathbb{N}}$, such that $u_{\mathcal{T}_n} \to u$, in $L^2((0,1))$, as $n \to \infty$, where $u \in H^1_0((0,1))$. In order to conclude, it only remains to prove that $-u_{xx} = f(\cdot, u)$ in the distribution sense in $(0,1)$.
To prove this, let $\varphi \in C^\infty_c((0,1))$. Let $\mathcal{T}$ be an admissible mesh of $(0,1)$, and $\varphi_i = \varphi(x_i)$, $i = 1, \ldots, N$, and $\varphi_0 = \varphi_{N+1} = 0$. If $(u_1, \ldots, u_N)$ is a solution to (8.6)-(8.8), multiplying (8.6) by φ_i and summing over $i = 1, \ldots, N$ yields

$$\int_0^1 u_{\mathcal{T}}(x)\psi_{\mathcal{T}}(x)dx = \int_0^1 f_{\mathcal{T}}(x)\varphi_{\mathcal{T}}(x)dx, \tag{8.15}$$

where

$$\psi_{\mathcal{T}}(x) = \frac{1}{h_i}\left(\frac{\varphi_i - \varphi_{i-1}}{h_{i-\frac{1}{2}}} - \frac{\varphi_{i+1} - \varphi_i}{h_{i+\frac{1}{2}}}\right),\ f_{\mathcal{T}}(x) = f(x, u_i) \text{ and } \varphi_{\mathcal{T}}(x) = \varphi_i, \text{ if } x \in K_i.$$

Note that, thanks to the regularity of the function φ,

$$\frac{\varphi_{i+1} - \varphi_i}{h_{i+\frac{1}{2}}} = \varphi_x(x_{i+\frac{1}{2}}) + R_{i+\frac{1}{2}},\ |R_{i+\frac{1}{2}}| \leq C_1 h,$$

with some C_1 only depending on φ, and therefore

$$\begin{aligned}\int_0^1 u_{\mathcal{T}}(x)\psi_{\mathcal{T}}(x)dx &= \sum_{i=1}^N \int_{K_i} \frac{u_i}{h_i}\Big(\varphi_x(x_{i-\frac{1}{2}}) - \varphi_x(x_{i+\frac{1}{2}})\Big)dx + \sum_{i=1}^N u_i(R_{i-\frac{1}{2}} - R_{i+\frac{1}{2}}) \\ &= \int_0^1 -u_{\mathcal{T}}(x)\theta_{\mathcal{T}}(x)dx + \sum_{i=0}^N R_{i+\frac{1}{2}}(u_{i+1} - u_i),\end{aligned}$$

with $u_0 = u_{N+1} = 0$, where the piecewise constant function

$$\theta_{\mathcal{T}} = \sum_{i=1,N} \frac{\varphi_x(x_{i+\frac{1}{2}}) - \varphi_x(x_{i-\frac{1}{2}})}{h_i} 1_{K_i}$$

tends to φ_{xx} as h tends to 0.
Let us consider (8.15) with $\mathcal{T}_n$ instead of $\mathcal{T}$; thanks to the Cauchy-Schwarz inequality, a passage to the limit as $n \to \infty$ gives, thanks to (8.9),

$$-\int_0^1 u(x)\varphi_{xx}(x)dx = \int_0^1 f(x,u(x))\varphi(x)dx,$$

and therefore $-u_{xx} = f(\cdot,u)$ in the distribution sense in $(0,1)$. This concludes the proof of Theorem 8.1. Note that the crucial idea of this proof is to use the property of consistency of the fluxes on the regular test function φ. ■

Remark 8.1 It is possible to give some extensions of the results of this section. For instance, Theorem 8.1 is true with an assumption of "sublinearity" on f instead of (8.4). Furthermore, in order to have both existence and uniqueness of the solution to (8.5) and a rate of convergence (of order h) in Theorem 8.1, it is sufficient to assume, instead of (8.3) and (8.4), that $f \in C^1([0,1]\times\mathbb{R},\mathbb{R})$ and that there exists $\gamma < 1$, such that $(f(x,s)-f(x,t))(s-t) \leq \gamma(s-t)^2$, for all $(x,s) \in [0,1]\times\mathbb{R}$.

Chapter 3

Elliptic problems in two or three dimensions

The topic of this chapter is the discretization of elliptic problems in several space dimensions by the finite volume method. The one-dimensional case which was studied in Chapter 2 is easily generalized to nonuniform rectangular or parallelipedic meshes. However, for general shapes of control volumes, the definition of the scheme (and the proof of convergence) requires some assumptions which define an "admissible mesh". Dirichlet and Neumann boundary conditions are both considered. In both cases, a discrete Poincaré inequality is used, and the stability of the scheme is proved by establishing estimates on the approximate solutions. The convergence of the scheme without any assumption on the regularity of the exact solution is proved; this result may be generalized, under adequate assumptions, to nonlinear equations. Then, again in both the Dirichlet and Neumann cases, an error estimate between the finite volume approximate solution and the C^2 or H^2 regular exact solution to the continuous problems are proved. The results are generalized to the case of matrix diffusion coefficients and more general boundary conditions. Section 12 is devoted to finite volume schemes written with unknowns located at the vertices. Some links between the finite element method, the "classical" finite volume method and the "control volume finite element" method introduced by FORSYTH [67] are given. Section 13 is devoted to the treatment of singular sources and to mesh refinement; under suitable assumption, it can be shown that error estimates still hold for "atypical" refined meshes. Finally, Section 14 is devoted to the proof of compactness results which are used in the proofs of convergence of the schemes.

9 Dirichlet boundary conditions

Let us consider here the following elliptic equation

$$-\Delta u(x) + \mathrm{div}(\mathbf{v}u)(x) + bu(x) = f(x), \quad x \in \Omega, \tag{9.1}$$

with Dirichlet boundary condition:

$$u(x) = g(x), \quad x \in \partial\Omega, \tag{9.2}$$

where

Assumption 9.1

1. *Ω is an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3,*
2. *$b \geq 0$,*
3. *$f \in L^2(\Omega)$,*

4. $\mathbf{v} \in C^1(\overline{\Omega}, \mathbb{R}^d); \mathrm{div}\mathbf{v} \geq 0$,

5. $g \in C(\partial\Omega, \mathbb{R})$ *is such that there exists* $\tilde{g} \in H^1(\Omega)$ *such that* $\overline{\gamma}(\tilde{g}) = g$ *a.e. on* $\partial\Omega$.

Here, and in the sequel, "polygonal" is used for both $d = 2$ and $d = 3$ (meaning polyhedral in the latter case) and $\overline{\gamma}$ denotes the trace operator from $H^1(\Omega)$ into $L^2(\partial\Omega)$. Note also that "a.e. on $\partial\Omega$" is a.e. for the $d-1$-dimensional Lebesgue measure on $\partial\Omega$.
Under Assumption 9.1, by the Lax-Milgram theorem, there exists a unique variational solution $u \in H^1(\Omega)$ of Problem (9.1)-(9.2). (For the study of elliptic problems and their discretization by finite element methods, see e.g. Ciarlet [29] and references therein). This solution satisfies $u = w + \tilde{g}$, where $\tilde{g} \in H^1(\Omega)$ is such that $\overline{\gamma}(\tilde{g}) = g$, a.e. on $\partial\Omega$, and w is the unique function of $H_0^1(\Omega)$ satisfying

$$\begin{array}{c} \displaystyle\int_\Omega \Big(\nabla w(x)\cdot\nabla\psi(x) + \mathrm{div}(\mathbf{v}w)(x)\psi(x) + bw(x)\psi(x)\Big)dx = \\ \displaystyle\int_\Omega \Big(-\nabla\tilde{g}(x)\cdot\nabla\psi(x) - \mathrm{div}(\mathbf{v}\tilde{g})(x)\psi(x) - b\tilde{g}(x)\psi(x) + f(x)\psi(x)\Big)dx,\ \forall\psi \in H_0^1(\Omega). \end{array} \tag{9.3}$$

9.1 Structured meshes

If Ω is a rectangle ($d = 2$) or a parallelepiped ($d = 3$), it may then be meshed with rectangular or parallelepipedic control volumes. In this case, the one-dimensional scheme may easily be generalized.

Rectangular meshes for the Laplace operator

Let us for instance consider the case $d = 2$, let $\Omega = (0,1)\times(0,1)$, and $f \in C^2(\Omega, \mathbb{R})$ (the three dimensional case is similar). Consider Problem (9.1)-(9.2) and assume here that $b = 0$, $\mathbf{v} = 0$ and $g = 0$ (the general case is considered later, on general unstructured meshes). The problem reduces to the pure diffusion equation:

$$\begin{array}{l} -\Delta u(x,y) = f(x,y),\ (x,y) \in \Omega, \\ u(x,y) = 0,\ (x,y) \in \partial\Omega. \end{array} \tag{9.4}$$

In this section, it is convenient to denote by (x, y) the current point of $\mathbb{R}^2$ (elsewhere, the notation x is used for a point or a vector of $\mathbb{R}^d$).
Let $\mathcal{T} = (K_{i,j})_{i=1,\cdots,N_1;j=1,\cdots,N_2}$ be an admissible mesh of $(0,1)\times(0,1)$, that is, satisfying the following assumptions (which generalize Definition 5.1)

Assumption 9.2 *Let* $N_1 \in \mathbb{N}^\star, N_2 \in \mathbb{N}^\star$, $h_1, \ldots, h_{N_1} > 0$, $k_1, \ldots, k_{N_2} > 0$ *such that*

$$\sum_{i=1}^{N_1} h_i = 1, \sum_{i=1}^{N_2} k_i = 1,$$

and let $h_0 = 0, h_{N_1+1} = 0, k_0 = 0, k_{N_2+1} = 0$. *For* $i = 1, \ldots, N_1$, *let* $x_{\frac{1}{2}} = 0$, $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h_i$, *(so that* $x_{N_1+\frac{1}{2}} = 1$*), and for* $j = 1, \ldots, N_2$, $y_{\frac{1}{2}} = 0$, $y_{j+\frac{1}{2}} = y_{j-\frac{1}{2}} + k_j$, *(so that* $y_{N_2+\frac{1}{2}} = 1$*) and*

$$K_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}].$$

Let $(x_i)_{i=0,N_1+1}$, *and* $(y_j)_{j=0,N_2+1}$, *such that*

$$x_{i-\frac{1}{2}} < x_i < x_{i+\frac{1}{2}},\ \ \textit{for } i = 1, \ldots, N_1,\ x_0 = 0,\ x_{N_1+1} = 1,$$

$$y_{j-\frac{1}{2}} < y_j < y_{j+\frac{1}{2}},\ \ \textit{for } j = 1, \ldots, N_2,\ y_0 = 0,\ y_{N_2+1} = 1,$$

and let $x_{i,j} = (x_i, y_j)$, *for* $i = 1, \ldots, N_1,$, $j = 1, \ldots, N_2$; *set*

$$h_i{}^- = x_i - x_{i-\frac{1}{2}},\ h_i{}^+ = x_{i+\frac{1}{2}} - x_i,\ \ \textit{for } i = 1, \ldots, N_1,\ h_{i+\frac{1}{2}} = x_{i+1} - x_i,\ \ \textit{for } i = 0, \ldots, N_1,$$

$$k_j{}^- = y_j - y_{j-\frac{1}{2}},\ k_j{}^+ = y_{j+\frac{1}{2}} - y_j,\ \ \textit{for } j = 1,\dots,N_2,\ k_{j+\frac{1}{2}} = y_{j+1} - y_j,\ \ \textit{for } j = 0,\dots,N_2.$$

Let $h = \max\{(h_i, i = 1,\cdots,N_1),(k_j, j = 1,\cdots,N_2)\}$.

As in the 1D case, the finite volume scheme is found by integrating the first equation of (9.4) over each control volume $K_{i,j}$, which yields

$$\left\{\begin{array}{l} -\displaystyle\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u_x(x_{i+\frac{1}{2}},y)dy + \int_{y_{i-\frac{1}{2}}}^{y_{i+\frac{1}{2}}} u_x(x_{i-\frac{1}{2}},y)dy \\ +\displaystyle\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_y(x,y_{j-\frac{1}{2}})dx - \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_y(x,y_{j+\frac{1}{2}})dx = \displaystyle\int_{K_{ij}} f(x,y)dx\,dy. \end{array}\right.$$

The fluxes are then approximated by differential quotients with respect to the discrete unknowns $(u_{i,j}, i = 1,\cdots,N_1, j = 1,\cdots,N_2)$ in a similar manner to the 1D case; hence the numerical scheme reads

$$F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j} + F_{i,j+\frac{1}{2}} - F_{i,j-\frac{1}{2}} = h_{i,j} f_{i,j},\ \forall\ (i,j) \in \{1,\dots,N_1\}\times\{1,\dots,N_2\}, \tag{9.5}$$

where $h_{i,j} = h_i \times k_j$, $f_{i,j}$ is the mean value of f over $K_{i,j}$, and

$$\begin{array}{l} F_{i+\frac{1}{2},j} = -\dfrac{k_j}{h_{i+\frac{1}{2}}}(u_{i+1,j} - u_{i,j}),\ \text{for } i = 0,\cdots,N_1, j = 1,\cdots,N_2, \\ F_{i,j+\frac{1}{2}} = -\dfrac{h_i}{k_{j+\frac{1}{2}}}(u_{i,j+1} - u_{i,j}),\ \text{for } i = 1,\cdots,N_1, j = 0,\cdots,N_2, \end{array} \tag{9.6}$$

$$u_{0,j} = u_{N_1+1,j} = u_{i,0} = u_{i,N_2+1} = 0,\ \text{ for } i = 1,\dots,N_1, j = 1,\dots,N_2. \tag{9.7}$$

The numerical scheme (9.5)-(9.7) is therefore clearly conservative and the numerical approximations of the fluxes can easily be shown to be consistent.

Proposition 9.1 (Error estimate) *Let $\Omega = (0,1)\times(0,1)$ and $f \in L^2(\Omega)$. Let u be the unique variational solution to (9.4). Under Assumptions 9.2, let $\zeta > 0$ be such that $h_i \geq \zeta h$ for $i = 1,\dots,N_1$ and $k_j \geq \zeta h$ for $j = 1,\dots,N_2$. Then, there exists a unique solution $(u_{i,j})_{i=1,\cdots,N_1,j=1,\cdots,N_2}$ to (9.5)-(9.7). Moreover, there exists $C > 0$ only depending on u, Ω and ζ such that*

$$\sum_{i,j} \frac{(e_{i+1,j} - e_{i,j})^2}{h_{i+\frac{1}{2}}} k_j + \sum_{i,j} \frac{(e_{i,j+1} - e_{i,j})^2}{k_{j+\frac{1}{2}}} h_i \leq Ch^2 \tag{9.8}$$

and

$$\sum_{i,j} (e_{i,j})^2 h_i k_j \leq Ch^2, \tag{9.9}$$

where $e_{i,j} = u(x_{i,j}) - u_{i,j}$, for $i = 1,\cdots,N_1, j = 1,\cdots,N_2$.

In the above proposition, since $f \in L^2(\Omega)$ and Ω is convex, it is well known that the variational solution u to (9.4) belongs to $H^2(\Omega)$. We do not give here the proof of this proposition since it is in fact included in Theorem 9.4 page 55 (see also Lazarov, Mishev and Vassilevski [99] where the case $u \in H^s$, $s \geq \frac{3}{2}$ is also studied).

In the case $u \in C^2(\overline{\Omega})$, the estimates (9.8) and (9.9) can be shown with the same technique as in the 1D case (see e.g. Fiard [65]). If $u \in C^2$ then the above estimates are a consequence of Theorem 9.3 page 52; in this case, the value C in (9.8) and (9.9) independent of ζ, and therefore the assumption $h_i \geq \zeta h$ for $i = 1,\dots,N_1$ and $k_j \geq \zeta h$ for $j = 1,\dots,N_2$ is no longer needed.

Relation (9.8) can be seen as an estimate of a "discrete H_0^1 norm" of the error, while relation (9.9) gives an estimate of the L^2 norm of the error.

Remark 9.1 Some slight modifications of the scheme (9.5)-(9.7) are possible, as in the first item of Remark 5.2 page 14. It is also possible to obtain, sometimes, an "h^2" estimate on the L^2 (or L^∞) norm of the error (that is "h^4" instead of "h^2" in (9.9)), exactly as in the 1D case, see Remark 6.2 page 18. In the case equivalent to the second case of Remark 6.2, the point $x_{i,j}$ is not necessarily the center of $K_{i,j}$.

When the mesh is no longer rectangular, the scheme (9.5)-(9.6) is not easy to generalize if keeping to a 5 points scheme. In particular, the consistency of the fluxes or the conservativity can be lost, see FAILLE [58], which yields a bad numerical behaviour of the scheme. One way to keep both properties is to introduce a 9-points scheme.

Quadrangular meshes: a nine-point scheme

Let Ω be an open bounded polygonal subset of $\mathbb{R}^2$, and f be a regular function from $\overline{\Omega}$ to $\mathbb{R}$. We still consider Problem 9.4, turning back to the usual notation x for the current point of $\mathbb{R}^2$,

$$\begin{array}{l} -\Delta u(x) = f(x),\ x \in \Omega, \\ u(x) = 0,\ x \in \partial\Omega. \end{array} \tag{9.10}$$

Let $\mathcal{T}$ be a mesh defined over Ω; then, integrating the first equation of (9.10) over any cell K of the mesh yields

$$-\int_{\partial K} \mathbf{grad} u \cdot \mathbf{n}_K = \int_K f,$$

where $\mathbf{n}_K$ is the normal to the boundary ∂K, outward to K. Let u_K denote the discrete unknown associated to the control volume $K \in \mathcal{T}$. In order to obtain a numerical scheme, if σ is a common edge to $K \in \mathcal{T}$ and $L \in \mathcal{T}$ (denoted by $K|L$) or if σ is an edge of $K \in \mathcal{T}$ belonging to $\partial\Omega$, the expression $\mathbf{grad} u \cdot \mathbf{n}_K$ must be approximated on σ by using the discrete unknowns. The study of the finite volume scheme in dimension 1 and the above straightforward generalization to the rectangular case showed that the fundamental properties of the method seem to be

1. conservativity: in the absence of any source term on $K|L$, the approximation of $\mathbf{grad} u \cdot \mathbf{n}_K$ on $K|L$ which is used in the equation associated with cell K is equal to the approximation of $-\mathbf{grad} u \cdot \mathbf{n}_L$ which is used in the equation associated with cell L. This property is naturally obtained when using a finite volume scheme.

2. consistency of the fluxes: taking for u_K the value of u in a fixed point of K (for instance, the center of gravity of K), where u is a regular function, the difference between $\mathbf{grad} u \cdot \mathbf{n}_K$ and the chosen approximation of $\mathbf{grad} u \cdot \mathbf{n}_K$ is of an order less or equal to that of the mesh size. This need of consistency will be discussed in more detail: see remarks 9.2 page 37 and 9.8 page 48

Several computer codes use the following "natural" extension of (9.6) for the approximation of $\mathbf{grad} u \cdot \mathbf{n}_K$ on $\overline{K} \cap \overline{L}$:

$$\mathbf{grad} u \cdot \mathbf{n}_K = \frac{u_L - u_K}{d_{K|L}},$$

where $d_{K|L}$ is the distance between the center of the cells K and L. This choice, however simple, is far from optimal, at least in the case of a general (non rectangular) mesh, because the fluxes thus obtained are not consistent; this yields important errors, especially in the case where the mesh cells are all oriented in the same direction, see FAILLE [58], FAILLE [59]. This problem may be avoided by modifying the approximation of $\mathbf{grad} u \cdot \mathbf{n}_K$ so as to make it consistent. However, one must be careful, in doing so, to maintain the conservativity of the scheme. To this purpose, a 9-points scheme was developped, which is denoted by FV9.
Let us describe now how the flux $\mathbf{grad} u \cdot \mathbf{n}_K$ is approximated by the FV9 scheme. Assume here, for the sake of clarity, that the mesh $\mathcal{T}$ is structured; indeed, it consists in a set of quadrangular cells $\{K_{i,j}, i = 1, \ldots, N; j = 1, \ldots, M\}$. As shown in Figure 3.1, let $C_{i,j}$ denote the center of gravity of the cell

$K_{i,j}$, $\sigma_{i,j-1/2}$, $\sigma_{i+1/2,j}$, $\sigma_{i,j+1/2}$, $\sigma_{i-1/2,j}$ the four edges to $K_{i,j}$ and $\eta_{i,j-1/2}$, $\eta_{i+1/2,j}$, $\eta_{i,j+1/2}$, $\eta_{i-1/2,j}$ their respective orthogonal bisectors. Let $\zeta_{i,j-1/2}$, (resp. $\zeta_{i+1/2,j}$, $\zeta_{i,j+1/2}$, $\zeta_{i-1/2,j}$) be the lines joining points $C_{i,j}$ and $C_{i,j-1}$ (resp. $C_{i,j}$ and $C_{i+1,j}$, $C_{i,j}$ and $C_{i,j+1}$, $C_{i-1,j}$ and $C_{i,j}$).

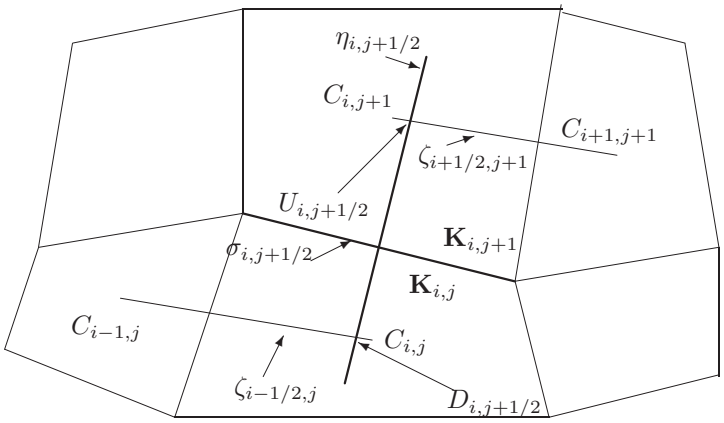


Figure 3.1: FV9 scheme

Consider for instance the edge $\sigma_{i,j+1/2}$ which lies between the cells $K_{i,j}$ and $K_{i,j+1}$ (see Figure 3.1). In order to find an approximation of $\mathbf{grad}u \cdot \mathbf{n}_K$, for $K = K_{i,j}$, at the center of this edge, we shall first derive an approximation of u at the two points $U_{i,j+1/2}$ and $D_{i,j+1/2}$ which are located on the orthogonal bisector $\eta_{i,j+1/2}$ of the edge $\sigma_{i,j+1/2}$, on each side of the edge. Let $\phi_{i,j+1/2}$ be the approximation of $-\mathbf{grad}u \cdot \mathbf{n}_K$ at the center of the edge $\sigma_{i,j+1/2}$. A natural choice for $\phi_{i,j+1/2}$ consists in taking

$$\phi_{i,j+1/2} = -\frac{u^U_{i,j+1/2} - u^D_{i,j+1/2}}{d(U_{i,j+1/2}, D_{i,j+1/2})}, \tag{9.11}$$

where $u^U_{i,j+1/2}$ and $u^D_{i,j+1/2}$ are approximations of u at $U_{i,j+1/2}$ and $D_{i,j+1/2}$, and $d(U_{i,j+1/2}, D_{i,j+1/2})$ is the distance between points $U_{i,j+1/2}$ and $D_{i,j+1/2}$.
The points $U_{i,j+1/2}$ and $D_{i,j+1/2}$ are chosen so that they are located on the lines ζ which join the centers of the neighbouring cells. The points $U_{i,j+1/2}$ and $D_{i,j+1/2}$ are therefore located at the intersection of the orthogonal bisector $\eta_{i,j+1/2}$ with the adequate ζ lines, which are chosen according to the geometry of the mesh. More precisely,

$$\begin{array}{lll}
U_{i,j+1/2} & = \eta_{i,j+1/2} \cap \zeta_{i-1/2,j+1} & \text{if } \eta_{i,j+1/2} \text{ is to the left of } C_{i,j+1} \\
 & = \eta_{i,j+1/2} \cap \zeta_{i+1/2,j+1} & \text{otherwise} \\
D_{i,j+1/2} & = \eta_{i,j+1/2} \cap \zeta_{i-1/2,j} & \text{if } \eta_{i,j+1/2} \text{ is to the left of } C_{i,j} \\
 & = \eta_{i,j+1/2} \cap \zeta_{i+1/2,j} & \text{otherwise}
\end{array}$$

In order to satisfy the property of consistency of the fluxes, a second order approximation of u at points $U_{i,j+1/2}$ and $D_{i,j+1/2}$ is required. In the case of the geometry which is described in Figure 3.1, the following linear approximations of $u^U_{i,j+1/2}$ and $u^D_{i,j+1/2}$ can be used in (9.11);

$$u^U_{i,j+1/2} = \alpha u_{i+1,j+1} + (1-\alpha)u_{i,j+1} \qquad \text{where } \alpha = \frac{d(C_{i,j+1}, U_{i,j+1/2})}{d(C_{i,j+1}, C_{i+1,j+1})}$$

$$u^D_{i,j+1/2} = \beta u_{i-1,j} + (1-\beta)u_{i,j} \qquad \text{where } \beta = \frac{d(C_{i,j}, D_{i,j+1/2})}{d(C_{i-1,j}, C_{i,j})}$$

The approximation of $\mathbf{grad} u \cdot \mathbf{n}_K$ at the center of a "vertical" edge $\sigma_{i+1/2,j}$ is performed in a similar way, by introducing the points $R_{i+1/2,j}$ intersection of the orthogonal bisector $\eta_{i+1/2,j}$ and, according to the geometry, of the line $\zeta_{i,j-1/2}$ or $\zeta_{i,j+1/2}$, and $L_{i+1/2,j}$ intersection of $\eta_{i+1/2,j}$ and $\zeta_{i+1,j-1/2}$ or $\zeta_{i+1,j+1/2}$. Note that the outmost grid cells require a particular treatment (see FAILLE [58]).
The scheme which is described above is stable under a geometrical condition on the family of meshes which is considered. Since the fluxes are consistent and the scheme is conservative, it also satisfies a property of "weak consistency", that is, as in the one dimensional case (see remark 6.6 page 21 of Section 7), the exact solution of (9.10) satisfies the numerical scheme with an error which tends to 0 in $L^\infty(\Omega)$ for the weak-$\star$ topology. Under adequate restrictive assumptions, the convergence of the scheme can be deduced, see FAILLE [58].
Numerical tests were performed for the Laplace operator and for operators of the type $-\mathrm{div}(\ \Lambda\ \mathbf{grad}.)$, where Λ is a variable and discontinuous matrix (see FAILLE [58]); the discontinuities of Λ are treated in a similar way as in the 1D case (see Section 7). Comparisons with solutions which were obtained by the bilinear finite element method, and with known analytical solutions, were performed. The results given by the VF9 scheme and by the finite element scheme were very similar.
The two drawbacks of this method are the fact that it is a 9-points scheme, and therefore computationally expensive, and that it yields a nonsymmetric matrix even if the original continuous operator is symmetric. Also, its generalization to three dimensions is somewhat complex.

Remark 9.2 The proof of convergence of this scheme is hindered by the lack of consistency for the discrete adjoint operator (see Section 9.4). An error estimate is also difficult to obtain because the numerical flux at an interface $K|L$ cannot be written under the form $\tau_{K|L}(u_K - u_L)$ with $\tau_{K|L} > 0$. Note, however, that under some geometrical assumptions on the mesh, see FAILLE [58] and COUDIÈRE, VILA and VILLEDIEU [41], error estimates may be obtained.

9.2 General meshes and schemes

Let us now turn to the discretization of convection-diffusion problems on general structured or non structured grids, consisting of any polygonal (recall that we shall call "polygonal" any polygonal domain of $\mathbb{R}^2$ or polyhedral domain or $\mathbb{R}^3$) control volumes (satisfying adequate geometrical conditions which are stated in the sequel) and not necessarily ordered in a Cartesian grid. The advantage of finite volume schemes using non structured meshes is clear for convection-diffusion equations. On one hand, the stability and convergence properties of the finite volume scheme (with an upstream choice for the convective flux) ensure a robust scheme for any admissible mesh as defined in Definitions 9.1 page 37 and 10.1 page 63 below, without any need for refinement in the areas of a large convection flux. On the other hand, the use of a non structured mesh allows the computation of a solution for any shape of the physical domain.

We saw in the previous section that a consistent discretization of the normal flux $-\nabla u \cdot \boldsymbol{n}$ over the interface of two control volumes K and L may be performed with a differential quotient involving values of the unknown located on the orthogonal line to the interface between K and L, on either side of this interface. This remark suggests the following definition of admissible finite volume meshes for the discretization of diffusion problems. We shall only consider here, for the sake of simplicity, the case of polygonal domains. The case of domains with a regular boundary does not introduce any supplementary difficulty other than complex notations. The definition of admissible meshes and notations introduced in this definition are illustrated in Figure 3.2

Definition 9.1 (Admissible meshes) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$, or 3. An admissible finite volume mesh of Ω, denoted by $\mathcal{T}$, is given by a family of "control volumes", which

are open polygonal convex subsets of Ω , a family of subsets of $\overline{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$, denoted by $\mathcal{E}$ (these are the edges (two-dimensional) or sides (three-dimensional) of the control volumes), with strictly positive $(d-1)$-dimensional measure, and a family of points of Ω denoted by $\mathcal{P}$ satisfying the following properties (in fact, we shall denote, somewhat incorrectly, by $\mathcal{T}$ the family of control volumes):

(i) The closure of the union of all the control volumes is $\overline{\Omega}$;

(ii) For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \overline{K} \setminus K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$. Furthermore, $\mathcal{E} = \cup_{K \in \mathcal{T}} \mathcal{E}_K$.

(iii) For any $(K, L) \in \mathcal{T}^2$ with $K \neq L$, either the $(d-1)$-dimensional Lebesgue measure of $\overline{K} \cap \overline{L}$ is 0 or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which will then be denoted by $K|L$.

(iv) The family $\mathcal{P} = (x_K)_{K \in \mathcal{T}}$ is such that $x_K \in \overline{K}$ (for all $K \in \mathcal{T}$) and, if $\sigma = K|L$, it is assumed that $x_K \neq x_L$, and that the straight line $\mathcal{D}_{K,L}$ going through x_K and x_L is orthogonal to $K|L$.

(v) For any $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega$, let K be the control volume such that $\sigma \in \mathcal{E}_K$. If $x_K \notin \sigma$, let $\mathcal{D}_{K,\sigma}$ be the straight line going through x_K and orthogonal to σ, then the condition $\mathcal{D}_{K,\sigma} \cap \sigma \neq \emptyset$ is assumed; let $y_\sigma = \mathcal{D}_{K,\sigma} \cap \sigma$.

In the sequel, the following notations are used.
The mesh size is defined by: $\text{size}(\mathcal{T}) = \sup\{\text{diam}(K),\ K \in \mathcal{T}\}$.
For any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$, $\text{m}(K)$ is the d-dimensional Lebesgue measure of K (it is the area of K in the two-dimensional case and the volume in the three-dimensional case) and $\text{m}(\sigma)$ the $(d-1)$-dimensional measure of σ.
The set of interior (resp. boundary) edges is denoted by $\mathcal{E}_{\text{int}}$ (resp. $\mathcal{E}_{\text{ext}}$), that is $\mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ (resp. $\mathcal{E}_{\text{ext}} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$).
The set of neighbours of K is denoted by $\mathcal{N}(K)$, that is $\mathcal{N}(K) = \{L \in \mathcal{T};\ \exists \sigma \in \mathcal{E}_K,\ \overline{\sigma} = \overline{K} \cap \overline{L}\}$.
If $\sigma = K|L$, we denote by d_σ or $d_{K|L}$ the Euclidean distance between x_K and x_L (which is positive) and by $d_{K,\sigma}$ the distance from x_K to σ.
If $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}$, let d_σ denote the Euclidean distance between x_K and y_σ (then, $d_\sigma = d_{K,\sigma}$).
For any $\sigma \in \mathcal{E}$; the "transmissibility" through σ is defined by $\tau_\sigma = \text{m}(\sigma)/d_\sigma$ if $d_\sigma \neq 0$.
In some results and proofs given below, there are summations over $\sigma \in \mathcal{E}_0$, with $\mathcal{E}_0 = \{\sigma \in \mathcal{E};\ d_\sigma \neq 0\}$. For simplicity, (in these results and proofs) $\mathcal{E} = \mathcal{E}_0$ is assumed.

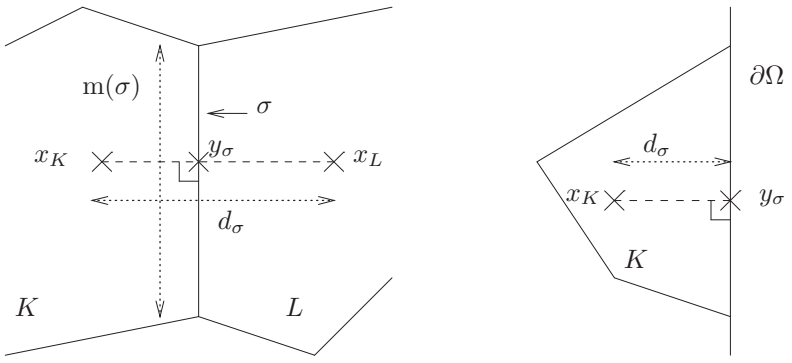


Figure 3.2: Admissible meshes

Remark 9.3 (i) The definition of y_σ for $\sigma \in \mathcal{E}_{\text{ext}}$ requires that $y_\sigma \in \sigma$. However, In many cases, this condition may be relaxed. The condition $x_K \in \overline{K}$ may also be relaxed as described, for instance, in Example 9.1 below.

(ii) The condition $x_K \neq x_L$ if $\sigma = K|L$, is in fact quite easy to satisfy: two neighbouring control volumes K, L which do not satisfy it just have to be collapsed into a new control volume M with $x_M = x_K = x_L$, and the edge $K|L$ removed from the set of edges. The new mesh thus obtained is admissible.

Example 9.1 (Triangular meshes) Let Ω be an open bounded polygonal subset of $\mathbb{R}^2$. Let $\mathcal{T}$ be a family of open triangular disjoint subsets of Ω such that two triangles having a common edge have also two common vertices. Assume that all angles of the triangles are less than $\pi/2$. This last condition is sufficient for the orthogonal bisectors to intersect inside each triangle, thus naturally defining the points $x_K \in K$. One obtains an admissible mesh. In the case of an elliptic operator, the finite volume scheme defined on such a grid using differential quotients for the approximation of the normal flux yields a 4-point scheme HERBIN [84]. This scheme does not lead to a finite difference scheme consistent with the continuous diffusion operator (using a Taylor expansion). The consistency is only verified for the approximation of the fluxes, but this, together with the conservativity of the scheme yields the convergence of the scheme, as it is proved below.
Note that the condition that all angles of the triangles are less than $\pi/2$ (which yields $x_K \in K$) may be relaxed (at least for the triangles the closure of which are in Ω) to the so called "strict Delaunay condition" which is that the closure of the circumscribed circle to each triangle of the mesh does not contain any other triangle of the mesh. For such a mesh, the point x_K (which is the intersection of the orthogonal bisectors of the edges of K) is not always in K, but the scheme (9.17)-(9.19) is convenient since (9.18) yields a consistent approximation of the diffusion fluxes and since the transmissibilities (denoted by $\tau_{K|L}$) are positive.

Example 9.2 (Voronoï meshes) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$. An admissible finite volume mesh can be built by using the so called "Voronoï" technique. Let $\mathcal{P}$ be a family of points of $\overline{\Omega}$. For example, this family may be chosen as $\mathcal{P} = \{(k_1 h, \ldots, k_d h),\ k_1, \ldots k_d \in \mathbb{Z}\} \cap \Omega$, for a given $h > 0$. The control volumes of the Voronoï mesh are defined with respect to each point x of $\mathcal{P}$ by

$$K_x = \{y \in \Omega, |x - y| < |z - y|,\ \forall z \in \mathcal{P},\ z \neq x\},$$

where $|x - y|$ denotes the Euclidean distance between x and y. Voronoï meshes are admissible in the sense of Definition 9.1 if the assumption "on the boundary", namely part (v) of Definition 9.1, is satisfied. Indeed, this is true, in particular, if the number of points $x \in \mathcal{P}$ which are located on $\partial\Omega$ is "large enough". Otherwise, the assumption (v) of Definition 9.1 may be replaced by the weaker assumption "$d(y_\sigma, \sigma) \leq \text{size}(\mathcal{T})$ for any $\sigma \in \mathcal{E}_{\text{ext}}$" which is much easier to satisfy. Note also that a slight modification of the treatment of the boundary conditions in the finite volume scheme (9.20)-(9.23) page 42 allows us to obtain convergence and error estimates results (as in theorems 9.1 page 45 and 9.3 page 52) for all Voronoï meshes. This modification is the obvious generalization of the scheme described in the first item of Remark 5.2 page 14 for the 1D case. It consists in replacing, for $K \in \mathcal{T}$ such that $\mathcal{E}_K \cap \mathcal{E}_{\text{ext}} \neq \emptyset$, the equation (9.20), associated to this control volume, by the equation $u_K = g(z_K)$, where z_K is some point on $\partial\Omega \cap \partial K$. In fact, Voronoï meshes often satisfy the following property:

$$\mathcal{E}_K \cap \mathcal{E}_{\text{ext}} \neq \emptyset \Rightarrow x_K \in \partial\Omega$$

and the mesh is therefore admissible in the sense of Definition 9.1 (then, the scheme (9.20)-(9.23) page 42 yields $u_K = g(x_K)$ if $K \in \mathcal{T}$ is such that $\mathcal{E}_K \cap \mathcal{E}_{\text{ext}} \neq \emptyset$).
An advantage of the Voronoï method is that it easily leads to meshes on non polygonal domains Ω.

Let us now introduce the space of piecewise constant functions associated to an admissible mesh and some "discrete H_0^1" norm for this space. This discrete norm will be used to obtain stability properties which are given by some estimates on the approximate solution of a finite volume scheme.

Definition 9.2 (Discrete space and norm) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3, and $\mathcal{T}$ be an admissible finite volume mesh in the sense of Definition 9.1 page 37. . Let $X(\mathcal{T})$ as the set of functions from Ω to $\mathbb{R}$ which are constant over each control volume of the mesh.

For $u \in X(\mathcal{T})$, define the discrete H_0^1 norm by

$$\|u\|_{1,\mathcal{T}} = \Big(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma u)^2\Big)^{\frac{1}{2}}, \tag{9.12}$$

where $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$ and $D_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$, $D_\sigma u = |u_K|$ if $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$, and where u_K denotes the value taken by u on the control volume K and the sets $\mathcal{E}$, $\mathcal{E}_{\rm int}$, $\mathcal{E}_{\rm ext}$ and $\mathcal{E}_K$ are defined in Definition 9.1 page 37.

The discrete H_0^1 norm is used in the following sections to prove the congergence of finite volume schemes and, under some regularity conditions, to give error estimates. It is related to the H_0^1 norm, see the convergence of the norms in Theorem 9.1. One of the tools used below is the following "discrete Poincaré inequality" which may also be found in Temam [141]:

Lemma 9.1 (Discrete Poincaré inequality) *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3, $\mathcal{T}$ an admissible finite volume mesh in the sense of Definition 9.1 and $u \in X(\mathcal{T})$ (see Definition 9.2), then*

$$\|u\|_{L^2(\Omega)} \le \mathrm{diam}(\Omega)\|u\|_{1,\mathcal{T}}, \tag{9.13}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 9.2 page 39.

Remark 9.4 (Dirichlet condition on part of the boundary) *This lemma gives a discrete Poincaré inequality for Dirichlet boundary conditions on the boundary $\partial\Omega$. In the case of a Dirichlet condition on part of the boundary only, it is still possible to prove a Discrete boundary condition provided that the polygonal bounded open set Ω is also connex, thanks to Lemma 9.1 page 40 proven in the sequel.*

Proof of Lemma 9.1
For $\sigma \in \mathcal{E}$, define χ_σ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\{0,1\}$ by $\chi_\sigma(x,y) = 1$ if $\sigma \cap [x,y] \neq \emptyset$ and $\chi_\sigma(x,y) = 0$ otherwise.

Let $u \in X(\mathcal{T})$. Let $\boldsymbol{d}$ be a given unit vector. For all $x \in \Omega$, let $\mathcal{D}_x$ be the semi-line defined by its origin, x, and the vector $\boldsymbol{d}$. Let $y(x)$ such that $y(x) \in \mathcal{D}_x \cap \partial\Omega$ and $[x, y(x)] \subset \overline{\Omega}$, where $[x, y(x)] = \{tx + (1-t)y(x), t \in [0,1]\}$ (i.e. $y(x)$ is the first point where $\mathcal{D}_x$ meets $\partial\Omega$).

Let $K \in \mathcal{T}$. For a.e. $x \in K$, one has

$$|u_K| \le \sum_{\sigma\in\mathcal{E}} D_\sigma u\, \chi_\sigma(x, y(x)),$$

where the notations $D_\sigma u$ and u_K are defined in Definition 9.2 page 39. We write the above inequality for a.e $x \in \Omega$ and not for all $x \in \Omega$ in order to account for the cases where an edge or a vertex of the mesh is included in the semi-line $[x, y(x)]$; in both cases one may not write the above inequality, but there are only a finite number of edges and vertices, and since $\boldsymbol{d}$ is fixed, the above inequality may be written almost everywhere.
Let $c_\sigma = |\boldsymbol{d} \cdot \boldsymbol{n}_\sigma|$ (recall that $\xi \cdot \eta$ denotes the usual scalar product of ξ and η in $\mathbb{R}^d$). By the Cauchy-Schwarz inequality, the above inequality yields:

$$|u_K|^2 \le \sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma u)^2}{d_\sigma c_\sigma} \chi_\sigma(x, y(x)) \sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x, y(x)), \text{ for a.e. } x \in K. \tag{9.14}$$

Let us show that, for a.e. $x \in \Omega$,

$$\sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x, y(x)) \le \mathrm{diam}(\Omega). \tag{9.15}$$

Let $x \in K$, $K \in \mathcal{T}$, such that $\sigma \cap [x, y(x)]$ contains at most one point, for all $\sigma \in \mathcal{E}$, and $[x, y(x)]$ does not contain any vertex of $\mathcal{T}$ (proving (9.15) for such points x leads to (9.15) a.e. on Ω, since $\boldsymbol{d}$ is fixed).

There exists $\sigma_x \in \mathcal{E}_{\rm ext}$ such that $y(x) \in \sigma_x$. Then, using the fact that the control volumes are convex, one has:

$$\sum_{\sigma\in\mathcal{E}} \chi_\sigma(x,y(x)) d_\sigma c_\sigma = |(x_K - x_{\sigma_x})\cdot \boldsymbol{d}|.$$

Since x_K and $x_{\sigma_x} \in \overline{\Omega}$ (see Definition 9.1), this gives (9.15).
Let us integrate (9.14) over Ω; (9.15) gives

$$\sum_{K\in\mathcal{T}} \int_K |u_K|^2 dx \le \mathrm{diam}(\Omega) \sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma u)^2}{d_\sigma c_\sigma} \int_\Omega \chi_\sigma(x,y(x)) dx.$$

Since $\int_\Omega \chi_\sigma(x,y(x))dx \le \mathrm{diam}(\Omega)\mathrm{m}(\sigma)c_\sigma$, this last inequality yields

$$\sum_{K\in\mathcal{T}} \int_K |u_K|^2 dx \le (\mathrm{diam}(\Omega))^2 \sum_{\sigma\in\mathcal{E}} |D_\sigma u|^2 \frac{\mathrm{m}(\sigma)}{d_\sigma} dx.$$

Hence the result. ∎

Let $\mathcal{T}$ be an admissible mesh. Let us now define a finite volume scheme to discretize (9.1), (9.2) page 32. Let

$$f_K = \frac{1}{\mathrm{m}(K)} \int_K f(x)dx, \forall K \in \mathcal{T}. \tag{9.16}$$

Let $(u_K)_{K\in\mathcal{T}}$ denote the discrete unknowns. In order to describe the scheme in the most general way, one introduces some auxiliary unknowns (as in the 1D case, see Section 7), namely the fluxes $F_{K,\sigma}$, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, and some (expected) approximation of u in σ, denoted by u_σ, for all $\sigma \in \mathcal{E}$. For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, let $\mathbf{n}_{K,\sigma}$ denote the normal unit vector to σ outward to K and $v_{K,\sigma} = \int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma} d\gamma(x)$. Note that $d\gamma$ is the integration symbol for the $(d-1)$-dimensional Lebesgue measure on the considered hyperplane. In order to discretize the convection term $\mathrm{div}(\mathbf{v}(x)u(x))$ in a stable way (see Section 7 page 21), let us define the upstream choice $u_{\sigma,+}$ of u on an edge σ with respect to $\mathbf{v}$ in the following way. If $\sigma = K|L$, then $u_{\sigma,+} = u_K$ if $v_{K,\sigma} \ge 0$, and $u_{\sigma,+} = u_L$ otherwise; if $\sigma \subset K \cap \partial\Omega$, then $u_{\sigma,+} = u_K$ if $v_{K,\sigma} \ge 0$ and $u_{\sigma,+} = g(y_\sigma)$ otherwise.
Let us first assume that the points x_K are located in the interior of each control volume, and are therefore not located on the edges, hence $d_{K,\sigma} > 0$ for any $\sigma \in \mathcal{E}_K$, where $d_{K,\sigma}$ is the distance from x_K to σ. A finite volume scheme can be defined by the following set of equations:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + b\mathrm{m}(K)u_K = \mathrm{m}(K)f_K,\ \forall K \in \mathcal{T}, \tag{9.17}$$

$$F_{K,\sigma} = -\tau_{K|L}(u_L - u_K),\ \forall \sigma \in \mathcal{E}_{\rm int},\ \text{if } \sigma = K|L, \tag{9.18}$$

$$F_{K,\sigma} = -\tau_\sigma(g(y_\sigma) - u_K),\ \forall \sigma \in \mathcal{E}_{\rm ext} \text{ such that } \sigma \in \mathcal{E}_K. \tag{9.19}$$

In the general case, the center of the cell may be located on an edge. This is the case for instance when constructing Voronoï meshes with some of the original points located on the boundary $\partial\Omega$. In this case, the following formulation of the finite volume scheme is valid, and is equivalent to the above scheme if no cell center is located on an edge:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + b\mathrm{m}(K)u_K = \mathrm{m}(K)f_K,\ \forall K \in \mathcal{T}, \tag{9.20}$$

$$F_{K,\sigma} = -F_{L,\sigma},\ \forall \sigma \in \mathcal{E}_{\rm int},\ \text{if } \sigma = K|L, \tag{9.21}$$

$$F_{K,\sigma} d_{K,\sigma} = -\mathrm{m}(\sigma)(u_\sigma - u_K),\ \forall \sigma \in \mathcal{E}_K,\ \forall K \in \mathcal{T}, \tag{9.22}$$

$$u_\sigma = g(y_\sigma),\ \forall \sigma \in \mathcal{E}_{\mathrm{ext}}. \tag{9.23}$$

Note that (9.20)-(9.23) always lead, after an easy elimination of the auxiliary unknowns, to a linear system of N equations with N unknowns, namely the $(u_K)_{K\in\mathcal{T}}$, with $N = \mathrm{card}(\mathcal{T})$.

Remark 9.5

1. Note that one may have, for some $\sigma \in \mathcal{E}_K$, $x_K \in \sigma$, and therefore, thanks to (9.22), $u_\sigma = u_K$.
2. The choice $u_\sigma = g(y_\sigma)$ in (9.23) needs some discussion. Indeed, this choice is possible since g is assumed to belong to $C(\partial\Omega, \mathbb{R})$ and then is everywhere defined on $\partial\Omega$. In the case where the solution to (9.1), (9.2) page 32 belongs to $H^2(\Omega)$ (which yields $g \in C(\partial\Omega, \mathbb{R})$), it is clearly a good choice since it yields the consistency of fluxes (even though an error estimate also holds with other choices for u_σ, the choice given below is, for instance, possible). If $g \in H^{1/2}$ (and not continuous), the value $g(y_\sigma)$ is not necessarily defined. Then, another choice for u_σ is possible, for instance,

$$u_\sigma = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma g(x) d\gamma(x).$$

   With this latter choice for u_σ, a convergence result also holds, see Theorem 9.2.

For the sake of simplicity, it is assumed in Definition 9.1 that $x_K \neq x_L$, for all K, $L \in \mathcal{T}$. This condition may be relaxed; it simply allows an easy expression of the numerical flux $F_{K,\sigma} = -\tau_{K|L}(u_L - u_K)$ if $\sigma = K|L$.

9.3 Existence and estimates

Let us first prove the existence of the approximate solution and an estimate on this solution. This estimate ensures the stability of the scheme and will be obtained by using the discrete Poincaré inequality (9.13) and will yield convergence thanks to a compactness theorem given in Section 14 page 93.

Lemma 9.2 (Existence and estimate) *Under Assumptions 9.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 9.1 page 37; there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to equations (9.20)-(9.23). Furthermore, assuming $g = 0$ and defining $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2) by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, and for any $K \in \mathcal{T}$, the following estimate holds:*

$$\|u_\mathcal{T}\|_{1,\mathcal{T}} \leq \mathrm{diam}(\Omega) \|f\|_{L^2(\Omega)}, \tag{9.24}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 9.2.

Proof of Lemma 9.2

Equations (9.20)-(9.23) lead, after an easy elimination of the auxiliary unknowns, to a linear system of N equations with N unknowns, namely the $(u_K)_{K\in\mathcal{T}}$, with $N = \mathrm{card}(\mathcal{T})$.

Step 1 (existence and uniqueness)
Assume that $(u_K)_{K\in\mathcal{T}}$ satisfies this linear system with $g(y_\sigma) = 0$ for any $\sigma \in \mathcal{E}_{\mathrm{ext}}$, and $f_K = 0$ for all $K \in \mathcal{T}$. Let us multiply (9.20) by u_K and sum over K; from (9.21) and (9.22) one deduces

$$b \sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^2 + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} u_K + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} u_K = 0, \tag{9.25}$$

which gives, reordering the summation over the set of edges

$$b\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^2+\sum_{\sigma\in\mathcal{E}}\tau_\sigma(D_\sigma u)^2+\sum_{\sigma\in\mathcal{E}}v_\sigma\Big(u_{\sigma,+}-u_{\sigma,-}\Big)u_{\sigma,+}=0, \tag{9.26}$$

where
$|D_\sigma u|=|u_K-u_L|$, if $\sigma=K|L$ and $|D_\sigma u|=|u_K|$, if $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}$;
$v_\sigma=|\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}d\gamma(x)|$, $\mathbf{n}$ being a unit normal vector to σ;
$u_{\sigma,-}$ is the downstream value to σ with respect to $\mathbf{v}$, i.e. if $\sigma=K|L$, then $u_{\sigma,-}=u_K$ if $v_{K,\sigma}\le 0$, and $u_{\sigma,-}=u_L$ otherwise; if $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}$, then $u_{\sigma,-}=u_K$ if $v_{K,\sigma}\le 0$ and $u_{\sigma,-}=u_\sigma$ if $v_{K,\sigma}>0$.
Note that $u_\sigma=0$ if $\sigma\in\mathcal{E}_{\rm ext}$.

Now, remark that

$$\sum_{\sigma\in\mathcal{E}}v_\sigma u_{\sigma,+}(u_{\sigma,+}-u_{\sigma,-})=\frac{1}{2}\sum_{\sigma\in\mathcal{E}}v_\sigma\Big((u_{\sigma,+}-u_{\sigma,-})^2+(u_{\sigma,+}^2-u_{\sigma,-}^2)\Big) \tag{9.27}$$

and, thanks to the assumption div$\mathbf{v}\ge 0$,

$$\sum_{\sigma\in\mathcal{E}}v_\sigma(u_{\sigma,+}^2-u_{\sigma,-}^2)=\sum_{K\in\mathcal{T}}\Big(\int_{\partial K}\mathbf{v}(x)\cdot\mathbf{n}_K d\gamma(x)\Big)u_K^2=\int_\Omega(\mathrm{div}\mathbf{v}(x))u_{\mathcal{T}}^2(x)dx\ge 0. \tag{9.28}$$

Hence,

$$b\|u_{\mathcal{T}}\|_{L^2(\Omega)}^2+\|u_{\mathcal{T}}\|_{1,\mathcal{T}}^2=b\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^2+\sum_{\sigma\in\mathcal{E}}\tau_\sigma(D_\sigma u)^2\le 0, \tag{9.29}$$

One deduces, from (9.29), that $u_K=0$ for all $K\in\mathcal{T}$.
This proves the existence and the uniqueness of the solution $(u_K)_{K\in\mathcal{T}}$, of the linear system given by (9.20)-(9.23), for any $\{g(y_\sigma),\ \sigma\in\mathcal{E}_{\rm ext}\}$ and $\{f_K,\ K\in\mathcal{T}\}$.

Step 2 (estimate)
Assume $g=0$. Multiply (9.20) by u_K, sum over K; then, thanks to (9.21), (9.22), (9.27) and (9.28) one has

$$b\|u_{\mathcal{T}}\|_{L^2(\Omega)}^2+\|u_{\mathcal{T}}\|_{1,\mathcal{T}}^2\le\sum_{K\in\mathcal{T}}\mathrm{m}(K)f_K u_K.$$

By the Cauchy-Schwarz inequality, this inequality yields

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}}^2\le(\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^2)^{\frac{1}{2}}(\sum_{K\in\mathcal{T}}\mathrm{m}(K)f_K^2)^{\frac{1}{2}}\le\|f\|_{L^2(\Omega)}\|u_{\mathcal{T}}\|_{L^2(\Omega)}.$$

Thanks to the discrete Poincaré inequality (9.13), this yields $\|u_{\mathcal{T}}\|_{1,\mathcal{T}}\le\|f\|_{L^2(\Omega)}\mathrm{diam}(\Omega)$, which concludes the proof of the lemma. ∎

Let us now state a discrete maximum principle which is satisfied by the scheme (9.20)-(9.23); this is an interesting stability property, even though it will not be used in the proofs of the convergence and error estimate.

Proposition 9.2 *Under Assumption 9.1 page 32, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 9.1 page 37, let $(f_K)_{K\in\mathcal{T}}$ be defined by (9.16). If $f_K\ge 0$ for all $K\in\mathcal{T}$, and $g(y_\sigma)\ge 0$, for all $\sigma\in\mathcal{E}_{\rm ext}$, then the solution $(u_K)_{K\in\mathcal{T}}$ of (9.20)-(9.23) satisfies $u_K\ge 0$ for all $K\in\mathcal{T}$.*

PROOF of Proposition 9.2
Assume that $f_K \geq 0$ for all $K \in \mathcal{T}$ and $g(y_\sigma) \geq 0$ for all $\sigma \in \mathcal{E}_{\rm ext}$. Let $a = \min\{u_K, K \in \mathcal{T}\}$. Let K_0 be a control volume such that $u_{K_0} = a$. Assume first that K_0 is an "interior" control volume, in the sense that $\mathcal{E}_K \subset \mathcal{E}_{\rm int}$, and that $u_{K_0} \leq 0$. Then, from (9.20),

$$\sum_{\sigma \in \mathcal{E}_{K_0}} F_{K_0,\sigma} + \sum_{\sigma \in \mathcal{E}_{K_0}} v_{K_0,\sigma} u_{\sigma,+} \geq 0; \tag{9.30}$$

since for any neighbour L of K_0 one has $u_L \geq u_{K_0}$, then, noting that div$\mathbf{v} \geq 0$, one must have $u_L = u_{K_0}$ for any neighbour L of K. Hence, setting $B = \{K \in \mathcal{T},\, u_K = a\}$, there exists $K \in B$ such that $\mathcal{E}_K \not\subset \mathcal{E}_{\rm int}$, that is K is a control volume "neighbouring the boundary".
Assume then that K_0 is a control volume neighbouring the boundary and that $u_{K_0} = a < 0$. Then, for an edge $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$, relations (9.22) and (9.23) yield $g(y_\sigma) < 0$, which is in contradiction with the assumption. Hence Proposition 9.2 is proved. ∎

Remark 9.6 The maximum principle immediately yields the existence and uniqueness of the solution of the numerical scheme (9.20)-(9.23), which was proved directly in Lemma 9.2.

9.4 Convergence

Let us now show the convergence of approximate solutions obtained by the above finite volume scheme when the size of the mesh tends to 0. One uses Lemma 9.2 together with the compactness theorem 14.2 given at the end of this chapter to prove the convergence result. In order to use Theorem 14.2, one needs the following lemma.

Lemma 9.3 *Let Ω be an open bounded set of $\mathbb{R}^d$, $d = 2$ or 3. Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 9.1 page 37 and $u \in X(\mathcal{T})$ (see Definition 9.2). One defines $\tilde{u}$ by $\tilde{u} = u$ a.e. on Ω, and $\tilde{u} = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$. Then there exists $C > 0$, only depending on Ω, such that*

$$\|\tilde{u}(\cdot + \eta) - \tilde{u}\|^2_{L^2(\mathbb{R}^d)} \leq \|u\|^2_{1,\mathcal{T}} |\eta|(|\eta| + C\,\mathrm{size}(\mathcal{T})), \forall \eta \in \mathbb{R}^d. \tag{9.31}$$

PROOF of Lemma 9.3
For $\sigma \in \mathcal{E}$, define χ_σ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\{0,1\}$ by $\chi_\sigma(x,y) = 1$ if $[x,y] \cap \sigma \neq \emptyset$ and $\chi_\sigma(x,y) = 0$ if $[x,y] \cap \sigma = \emptyset$.
Let $\eta \in \mathbb{R}^d$, $\eta \neq 0$. One has

$$|\tilde{u}(x+\eta) - \tilde{u}(x)| \leq \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) |D_\sigma u|, \quad \text{for a.e. } x \in \Omega$$

(see Definition 9.2 page 39 for the definition of $D_\sigma u$).
This gives, using the Cauchy-Schwarz inequality,

$$|\tilde{u}(x+\eta) - \tilde{u}(x)|^2 \leq \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) \frac{|D_\sigma u|^2}{d_\sigma c_\sigma} \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) d_\sigma c_\sigma, \quad \text{for a.e. } x \in \mathbb{R}^d, \tag{9.32}$$

where $c_\sigma = |\boldsymbol{n}_\sigma \cdot \frac{\eta}{|\eta|}|$, and $\boldsymbol{n}_\sigma$ denotes a unit normal vector to σ.

Let us now prove that there exists $C > 0$, only depending on Ω, such that

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) d_\sigma c_\sigma \leq |\eta| + C\,\mathrm{size}(\mathcal{T}), \tag{9.33}$$

for a.e. $x \in \mathbb{R}^d$.

Let $x \in \mathbb{R}^d$ such that $\sigma \cap [x, x+\eta]$ contains at most one point, for all $\sigma \in \mathcal{E}$, and $[x, x+\eta]$ does not contain any vertex of $\mathcal{T}$ (proving (9.33) for such points x gives (9.33) for a.e. $x \in \mathbb{R}^d$, since η is fixed). Since Ω is not assumed to be convex, it may happen that the line segment $[x, x+\eta]$ is not included in $\overline{\Omega}$. In order to deal with this, let $y, z \in [x, x+\eta]$ such that $y \neq z$ and $[y, z] \subset \overline{\Omega}$; there exist K, $L \in \mathcal{T}$ such that $y \in \overline{K}$ and $z \in \overline{L}$. Hence,

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(y,z) d_\sigma c_\sigma = |(y_1 - z_1) \cdot \frac{\eta}{|\eta|}|,$$

where $y_1 = x_K$ or y_σ with $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$ and $z_1 = x_L$ or $y_{\tilde{\sigma}}$ with $\tilde{\sigma} \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_L$, depending on the position of y and z in $\overline{K}$ or $\overline{L}$ respectively.
Since $y_1 = y + y_2$, with $|y_2| \leq \mathrm{size}(\mathcal{T})$, and $z_1 = z + z_2$, with $|z_2| \leq \mathrm{size}(\mathcal{T})$, one has

$$|(y_1 - z_1) \cdot \frac{\eta}{|\eta|}| \leq |y - z| + |y_2| + |z_2| \leq |y - z| + 2\,\mathrm{size}(\mathcal{T})$$

and

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(y,z) d_\sigma c_\sigma \leq |y - z| + 2\,\mathrm{size}(\mathcal{T}). \tag{9.34}$$

Note that this yields (9.33) with $C = 2$ if $[x, x+\eta] \subset \overline{\Omega}$.
Since Ω has a finite number of sides, the line segment $[x, x+\eta]$ intersects $\partial\Omega$ a finite number of times; hence there exist $t_1, \ldots, t_n$ such that $0 \leq t_1 < t_2 < \ldots < t_n \leq 1$, $n \leq N$, where N only depends on Ω (indeed, it is possible to take $N = 2$ if Ω is convex and N equal to the number of sides of Ω for a general Ω) and such that

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) d_\sigma c_\sigma = \sum_{\substack{i=1,n-1 \\ \mathrm{odd} i}} \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x_i, x_{i+1}) d_\sigma c_\sigma,$$

with $x_i = x + t_i \eta$, for $i = 1, \ldots, n$, $x_i \in \partial\Omega$ if $t_i \notin \{0, 1\}$ and $[x_i, x_{i+1}] \subset \overline{\Omega}$ if i is odd.
Then, thanks to (9.34) with $y = x_i$ and $z = x_{i+1}$, for $i = 1, \ldots, n-1$, one has (9.33) with $C = 2(N-1)$ (in particular, if Ω is convex, $C = 2$ is convenient for (9.33) and therefore for (9.31) as we shall see below).

In order to conclude the proof of Lemma 9.3, remark that, for all $\sigma \in \mathcal{E}$,

$$\int_{\mathbb{R}^d} \chi_\sigma(x, x+\eta) dx \leq \mathrm{m}(\sigma) c_\sigma |\eta|.$$

Therefore, integrating (9.32) over $\mathbb{R}^d$ yields, with (9.33),

$$\|\tilde{u}(\cdot + \eta) - \tilde{u}\|^2_{L^2(\mathbb{R}^d)} \leq \big(\sum_{\sigma \in \mathcal{E}} \frac{\mathrm{m}(\sigma)}{d_\sigma} |D_\sigma u|^2\big) |\eta| (|\eta| + C\,\mathrm{size}(\mathcal{T})).$$

■

We are now able to state the convergence theorem. We shall first prove the convergence result in the case of homogeneous Dirichlet boundary conditions, i.e. $g = 0$; thenonhomogenous case is then considered (see Theorem 9.2 page 51), following Eymard, Gallouët and Herbin [55].

Theorem 9.1 (Convergence, homogeneous Dirichlet boundary conditions) *Under Assumption 9.1 page 32 with $g = 0$, let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 9.1 page 37). Let$(u_K)_{K \in \mathcal{T}}$ be the solution of the system given by equations (9.20)-(9.23) (existence and uniqueness of $(u_K)_{K \in \mathcal{T}}$ are given in Lemma 9.2). Define $u_\mathcal{T} \in X(\mathcal{T})$ by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, and for any $K \in \mathcal{T}$. Then $u_\mathcal{T}$ converges in $L^2(\Omega)$ to the unique variational solution $u \in H^1_0(\Omega)$ of Problem (9.1), (9.2) as* $\mathrm{size}(\mathcal{T}) \to 0$. *Furthermore $\|u_\mathcal{T}\|_{1,\mathcal{T}}$ converges to $\|u\|_{H^1_0(\Omega)}$ as* $\mathrm{size}(\mathcal{T}) \to 0$.

Remark 9.7

1. In Theorem 9.1, the hypothesis $f \in L^2(\Omega)$ is not necessary. It is used essentially to obtain a bound on $\|u_\mathcal{T}\|_{1,\mathcal{T}}$. In order to pass to the limit, the hypothesis "$f \in L^1(\Omega)$" is sufficient. Then, in Theorem 9.1, the hypothesis $f \in L^2(\Omega)$ can be replaced by $f \in L^p(\Omega)$ for some $p > 1$, if $d = 2$, and for $p \geq \frac{6}{5}$, if $d = 3$, provided that the meshes satisfy, for some fixed $\zeta > 0$, $d_{K,\sigma} \geq \zeta d_\sigma$, for all $\sigma \in \mathcal{E}_K$ and for all control volumes K. Indeed, one obtains, in this case, a bound on $\|u_\mathcal{T}\|_{1,\mathcal{T}}$ by using a "discrete Sobolev inequality" (proved in Lemma 9.5 page 60).

   It is also possible to obtain convergence results, towards a "very weak solution" of Problem (9.1), (9.2), with only $f \in L^1(\Omega)$, by working with some discrete equivalent of the $W_0^{1,q}$-norm, with $q < \frac{d}{d-1}$. This is not detailed here.

2. In Theorem 9.1, it is also possible to prove convergence results when $f(x)$ (resp. $\mathbf{v}(x)$) is replaced by some nonlinear function $f(x, u(x))$, (resp. $\mathbf{v}(x, u(x))$ under adequate assumptions, see [55].

Proof of Theorem 9.1

Let Y be the set of approximate solutions, that is the set of $u_\mathcal{T}$ where $\mathcal{T}$ is an admissible mesh in the sense of Definition 9.1 page 37. First, we want to prove that $u_\mathcal{T}$ tends to the unique solution (in $H_0^1(\Omega)$) to (9.3) as size$(\mathcal{T}) \to 0$.

Thanks to Lemma 9.2 and to the discrete Poincaré inequality (9.13), there exists $C_1 \in \mathbb{R}$, only depending on Ω and f, such that $\|u_\mathcal{T}\|_{1,\mathcal{T}} \leq C_1$ and $\|u_\mathcal{T}\|_{L^2(\Omega)} \leq C_1$ for all $u_\mathcal{T} \in Y$. Then, thanks to Lemma 9.3 and to the compactness result given in Theorem 14.2 page 94, the set Y is relatively compact in $L^2(\Omega)$ and any possible limit (in $L^2(\Omega)$) of a sequence $(u_{\mathcal{T}_n})_{n\in\mathbb{N}} \subset Y$ (such that size$(\mathcal{T}_n) \to 0$) belongs to $H_0^1(\Omega)$. Therefore, thanks to the uniqueness of the solution (in $H_0^1(\Omega)$) of (9.3), it is sufficient to prove that if $(u_{\mathcal{T}_n})_{n\in\mathbb{N}} \subset Y$ converges towards some $u \in H_0^1(\Omega)$, in $L^2(\Omega)$, and size$(\mathcal{T}_n) \to 0$ (as $n \to \infty$), then u is the solution to (9.3). We prove this result below, omiting the index n, that is assuming $u_\mathcal{T} \to u$ in $L^2(\Omega)$ as size$(\mathcal{T}) \to 0$.

Let $\psi \in C_c^\infty(\Omega)$ and let size$(\mathcal{T})$ be small enough so that $\psi(x) = 0$ if $x \in K$ and $K \in \mathcal{T}$ is such that $\partial K \cap \partial\Omega \neq \emptyset$. Multiplying (9.20) by $\psi(x_K)$, and summing the result over $K \in \mathcal{T}$ yields

$$T_1 + T_2 + T_3 = T_4, \tag{9.35}$$

with

$$T_1 = b \sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K \psi(x_K),$$

$$T_2 = -\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(u_L - u_K)\psi(x_K),$$

$$T_3 = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} \psi(x_K),$$

$$T_4 = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\psi(x_K) f_K.$$

First remark that, since $u_\mathcal{T}$ tends to u in $L^2(\Omega)$,

$$T_1 \to b \int_\Omega u(x)\psi(x)dx \text{ as size}(\mathcal{T}) \to 0.$$

Similarly,

$$T_4 \to \int_\Omega f(x)\psi(x)dx \text{ as size}(\mathcal{T}) \to 0.$$

Let us now turn to the study of T_2;

$$T_2 = -\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(u_L-u_K)(\psi(x_K)-\psi(x_L)).$$

Consider the following auxiliary expression:

$$\begin{aligned} T_2' &= \int_\Omega u_\mathcal{T}(x)\Delta\psi(x)dx \\ &= \sum_{K\in\mathcal{T}} u_K \int_K \Delta\psi(x)dx \\ &= \sum_{K|L\in\mathcal{E}_{\rm int}} (u_K-u_L)\int_{K|L} \nabla\psi(x)\cdot \boldsymbol{n}_{K,L}d\gamma(x). \end{aligned}$$

Since $u_\mathcal{T}$ converges to u in $L^2(\Omega)$, it is clear that T_2' tends to $\displaystyle\int_\Omega u(x)\Delta\psi(x)\,dx$ as size$(\mathcal{T})$ tends to 0.
Define

$$R_{K,L} = \frac{1}{{\rm m}(K|L)}\int_{K|L}\nabla\psi(x)\cdot\boldsymbol{n}_{K,L}d\gamma(x) - \frac{\psi(x_L)-\psi(x_K)}{d_{K|L}},$$

where $\boldsymbol{n}_{K,L}$ denotes the unit normal vector to $K|L$, outward to K, then

$$\begin{aligned} |T_2+T_2'| &= |\sum_{K|L\in\mathcal{E}_{\rm int}} {\rm m}(K|L)(u_K-u_L)R_{K,L}| \\ &\le \Big[\sum_{K|L\in\mathcal{E}_{\rm int}} {\rm m}(K|L)\frac{(u_K-u_L)^2}{d_{K|L}} \sum_{K|L\in\mathcal{E}_{\rm int}} {\rm m}(K|L)d_{K|L}(R_{K,L})^2\Big]^{1/2}, \end{aligned}$$

Regularity properties of the function ψ give the existence of $C_2\in\mathbb{R}$, only depending on ψ, such that $|R_{K,L}|\le C_2{\rm size}(\mathcal{T})$. Therefore, since

$$\sum_{K|L\in\mathcal{E}_{\rm int}} {\rm m}(K|L)d_{K|L} \le d{\rm m}(\Omega),$$

from Estimate (9.24), we conclude that $T_2+T_2'\to 0$ as size$(\mathcal{T})\to 0$.
Let us now show that T_3 tends to $-\int_\Omega \mathbf{v}(x)u(x)\nabla\psi(x)dx$ as size$(\mathcal{T})\to 0$. Let us decompose $T_3 = T_3'+T_3''$ where

$$T_3' = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}(u_{\sigma,+}-u_K)\psi(x_K)$$

and

$$T_3'' = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}u_K\psi(x_K) = \int_\Omega {\rm div}\mathbf{v}(x)u_\mathcal{T}(x)\psi_\mathcal{T}(x)dx,$$

where $\psi_\mathcal{T}$ is defined by $\psi_\mathcal{T}(x)=\psi(x_K)$ if $x\in K$, $K\in\mathcal{T}$. Since $u_\mathcal{T}\to u$ and $\psi_\mathcal{T}\to\psi$ in $L^2(\Omega)$ as size$(\mathcal{T})\to 0$ (indeed, $\psi_\mathcal{T}\to\psi$ uniformly on Ω as size$(\mathcal{T})\to 0$) and since div$\mathbf{v}\in L^\infty(\Omega)$, one has

$$T_3''\to\int_\Omega {\rm div}\mathbf{v}(x)u(x)\psi(x)dx \text{ as size}(\mathcal{T})\to 0.$$

Let us now rewrite T_3' as $T_3' = T_3'''+r_3$ with

$$T_3''' = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(u_{\sigma,+}-u_K)\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}\psi(x)d\gamma(x)$$

and

$$r_3 = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(u_{\sigma,+}-u_K)\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}(\psi(x_K)-\psi(x))d\gamma(x).$$

Thanks to the regularity of $\mathbf{v}$ and ψ, there exists C_3 only depending on $\mathbf{v}$ and ψ such that

$$|r_3| \le C_3\text{size}(\mathcal{T})\sum_{K|L\in\mathcal{E}_{\rm int}}|u_K-u_L|\text{m}(K|L),$$

which yields, with the Cauchy-Schwarz inequality,

$$|r_3| \le C_3\text{size}(\mathcal{T})(\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}|u_K-u_L|^2)^{\frac{1}{2}}(\sum_{K|L\in\mathcal{E}_{\rm int}}\text{m}(K|L)d_{K|L})^{\frac{1}{2}},$$

from which one deduces, with Estimate (9.24), that $r_3\to 0$ as $\text{size}(\mathcal{T})\to 0$.
Next, remark that

$$T_3''' = -\sum_{K\in\mathcal{T}}u_K\sum_{\sigma\in\mathcal{E}_K}\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}\psi(x)d\gamma(x) = -\sum_{K\in\mathcal{T}}u_K\int_K \text{div}(\mathbf{v}(x)\psi(x))dx.$$

This implies (since $u_\mathcal{T}\to u$ in $L^2(\Omega)$) that $T_3'''\to -\int_\Omega \text{div}(\mathbf{v}(x)\psi(x))u(x)dx$, so that T_3' has the same limit and $T_3\to -\int_\Omega \mathbf{v}(x)\cdot\nabla\psi(x)u(x)dx$.

Hence, letting $\text{size}(\mathcal{T})\to 0$ in (9.35) yields that the function $u\in H_0^1(\Omega)$ satisfies

$$\int_\Omega\Big(bu(x)\psi(x)-u(x)\Delta\psi(x)-\mathbf{v}(x)u(x)\nabla\psi(x)-f(x)\psi(x)\Big)dx=0,\ \forall\psi\in C_c^\infty(\Omega),$$

which, in turn, yields (9.3) thanks to the fact that $u\in H_0^1(\Omega)$, and to the density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$. This concludes the proof of $u_\mathcal{T}\to u$ in $L^2(\Omega)$ as $\text{size}(\mathcal{T})\to 0$, where u is the unique solution (in $H_0^1(\Omega)$) to (9.3).

S Let us now prove that $\|u_\mathcal{T}\|_{1,\mathcal{T}}$ tends to $\|u\|_{H_0^1(\Omega)}$ in the pure diffusion case, i.e. assuming $b=0$ and $\mathbf{v}=0$. Since

$$\|u_\mathcal{T}\|_{1,\mathcal{T}}^2 = \int_\Omega f_\mathcal{T}(x)u_\mathcal{T}(x)dx\to\int_\Omega f(x)u(x)dx\text{ as size}(\mathcal{T})\to 0,$$

where $f_\mathcal{T}$ is defined from Ω to $\mathbb{R}$ by $f_\mathcal{T}(x)=f_K$ a.e. on K for all $K\in\mathcal{T}$, it is easily seen that

$$\|u_\mathcal{T}\|_{1,\mathcal{T}}^2\to\int_\Omega f(x)u(x)dx=\|u\|_{H_0^1(\Omega)}^2\text{ as size}(\mathcal{T})\to 0.$$

This concludes the proof of Theorem 9.1. ∎

Remark 9.8 (Consistency for the adjoint operator) The proof of Theorem 9.1 uses the property of consistency of the (diffusion) fluxes on the test functions. This property consists in writing the consistency of the fluxes for the adjoint operator to the discretized Dirichlet operator. This consistency is achieved thanks to that of fluxes for the discretized Dirichlet operator and to the fact that this operator is self adjoint. In fact, any discretization of the Dirichlet operator giving "L^2-stability" and consistency of fluxes on its adjoint, yields a convergence result (see also Remark 9.2 page 37). On the contrary, the error estimates proved in sections 9.5 and 9.6 directly use the consistency for the discretized Dirichlet operator itself.

Remark 9.9 (Finite volume schemes and H^1 approximate solutions)
In the above proof, we showed that a sequence of approximate solutions (which are piecewise constant functions) converges in $L^2(\Omega)$ to a limit which is in $H^1_0(\Omega)$. An alternative to the use of Theorem 14.2 is the construction of a bounded sequence in $H^1(\mathbb{R}^d)$ from the sequence of approximate solutions. This can be performed by convoluting the approximate solution with a mollifier "of size size($\mathcal{T}$)". Using Rellich's compactness theorem and the weak sequential compactness of the bounded sets of H^1, one obtains that the limit of the sequence of approximate solutions is in H^1_0.

Let us now deal with the case of non homogeneous Dirichlet boundary conditions, in which case $g \in H^{1/2}(\partial\Omega)$ is no longer assumed to be 0. The proof uses the following preliminary result:

Lemma 9.4 *Let Ω be an open bounded polygonal subset of $\mathbb{R}^2$, $\tilde{g} \in H^1(\Omega)$ and $g = \overline{\gamma}(\tilde{g})$ (recall that $\overline{\gamma}$ is the "trace" operator from $H^1(\Omega)$ to $H^{1/2}(\partial\Omega)$). Let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 9.1 page 37) such that, for some $\zeta > 0$, the inequality $d_{K,\sigma} \geq \zeta \mathrm{diam}(K)$ holds for all control volumes $K \in \mathcal{T}$ and for all $\sigma \in \mathcal{E}_K$, and let $M \in \mathbb{N}$ be such that $\mathrm{card}(\mathcal{E}_K) \leq M$ for all $K \in \mathcal{T}$. Let us define $\tilde{g}_K$ for all $K \in \mathcal{T}$ by*

$$\tilde{g}_K = \frac{1}{\mathrm{m}(K)} \int_K \tilde{g}(x)dx$$

and $\tilde{g}_\sigma$ for all $\sigma \in \mathcal{E}_{\mathrm{ext}}$ by

$$\tilde{g}_\sigma = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma g(x)d\gamma(x).$$

Let us define

$$\mathcal{N}(\tilde{g}, \mathcal{T}) = \Big(\sum_{\sigma = K|L \in \mathcal{E}_{\mathrm{int}}} \tau_{K|L}(\tilde{g}_K - \tilde{g}_L)^2 + \sum_{\sigma \in \mathcal{E}_{\mathrm{ext}}} \tau_\sigma (\tilde{g}_{K(\sigma)} - \tilde{g}_\sigma)^2 \Big)^{\frac{1}{2}}, \tag{9.36}$$

where $K(\sigma) = K$ if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$. Then there exists $C \in \mathbb{R}_+$, only depending on ζ and M, such that

$$\mathcal{N}(\tilde{g}, \mathcal{T}) \leq C \|\tilde{g}\|_{H^1(\Omega)}. \tag{9.37}$$

Proof of Lemma 9.4

Lemma 9.4 is given in the two dimensional case, an analogous result is possible in the three dimensional case. Let Ω, $\tilde{g}$, $\mathcal{T}$, ζ, M satisfying the hypotheses of Lemma 9.4. By a classical argument of density, one may assume that $\tilde{g} \in C^1(\overline{\Omega}, \mathbb{R})$.
A first step consists in proving that there exists $C_1 \in \mathbb{R}_+$, only depending on ζ, such that

$$(\tilde{g}_K - \tilde{g}_\sigma)^2 \leq C_1 \frac{\mathrm{diam}(K)}{\mathrm{m}(\sigma)} \int_K |\nabla \tilde{g}(x)|^2 dx, \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K, \tag{9.38}$$

where $\tilde{g}_K$ (resp. $\tilde{g}_\sigma$) is the mean value of $\tilde{g}$ on K (resp. σ), for $K \in \mathcal{T}$ (resp. $\sigma \in \mathcal{E}$). Indeed, without loss of generality, one assumes that $\sigma = \{0\} \times J_0$, with J_0 is a closed interval of $\mathbb{R}$ and $K \subset \mathbb{R}_+ \times \mathbb{R}$.
Let $\alpha = \max\{x_1, x = (x_1, x_2)^t \in \overline{K}\}$ and $a = (\alpha, \beta)^t \in \overline{K}$. In the following, a is fixed. For all $x_1 \in (0, \alpha)$, let $J(x_1) = \{x_2 \in \mathbb{R}, \text{ such that } (x_1, x_2)^t \in \overline{K}\}$, so that $J_0 = J(0)$.
For a.e. $x = (x_1, x_2)^t \in K$ and a.e., for the 1-Lebesgue measure, $y = (0, \overline{y})^t \in \sigma$ (with $\overline{y} \in J_0$), one sets $z(x, y) = ta + (1-t)y$ with $t = \frac{x_1}{\alpha}$. Note that, since $\overline{K}$ is convex, $z(x, y) \in \overline{K}$ and $z(x, y) = (x_1, z_2(x_1, \overline{y}))^t$, with $z_2(x_1, \overline{y}) = \frac{x_1}{\alpha}\beta + (1 - \frac{x_1}{\alpha})\overline{y}$.
One has, using the Cauchy-Schwarz inequality,

$$(\tilde{g}_K - \tilde{g}_\sigma)^2 \leq \frac{2}{\mathrm{m}(K)\mathrm{m}(\sigma)}(A + B), \tag{9.39}$$

where

$$A = \int_K \int_\sigma \big(\tilde{g}(x) - \tilde{g}(z(x,y))\big)^2 d\gamma(y)dx,$$

and

$$B = \int_K \int_\sigma \big(\tilde{g}(z(x,y)) - \tilde{g}(y)\big)^2 d\gamma(y)dx.$$

Let us now obtain a bound of A. Let $D_i\tilde{g}$, $i = 1$ or 2, denote the partial derivative of $\tilde{g}$ w.r.t. the components of $x = (x_1, x_2)^t \in \mathbb{R}^2$. Then,

$$A = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \big(\int_{z_2(x_1,\overline{y})}^{x_2} D_2\tilde{g}(x_1, s)ds\big)^2 d\overline{y}dx_2dx_1.$$

The Cauchy-Schwarz inequality yields

$$A \leq \mathrm{diam}(K) \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \int_{J(x_1)} \big(D_2\tilde{g}(x_1, s)\big)^2 ds d\overline{y}dx_2dx_1$$

and therefore

$$A \leq \mathrm{diam}(K)^3 \int_K \big(D_2\tilde{g}(x)\big)^2 dx. \tag{9.40}$$

One now turns to the study of B, which can be rewritten as

$$B = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \big(\int_0^{x_1} [D_1\tilde{g}(s, z_2(s,\overline{y})) + \frac{\beta - \overline{y}}{\alpha} D_2\tilde{g}(s, z_2(s,\overline{y}))]ds\big)^2 d\overline{y}dx_2dx_1.$$

The Cauchy-Schwarz inequality and the fact that $\alpha \geq \zeta \mathrm{diam}(K)$ give that

$$B \leq 2\mathrm{diam}(K)(B_1 + \frac{1}{\zeta^2} B_2), \tag{9.41}$$

with

$$B_i = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \int_0^{x_1} \big(D_i\tilde{g}(s, z_2(s,\overline{y}))\big)^2 ds d\overline{y}dx_2dx_1,\ i = 1, 2.$$

First, using Fubini's theorem, one has

$$B_i = \int_{J(0)} \int_0^\alpha \big(D_i\tilde{g}(s, z_2(s,\overline{y}))\big)^2 \int_s^\alpha \int_{J(x_1)} dx_2dx_1 ds d\overline{y}.$$

Therefore

$$B_i \leq \mathrm{diam}(K) \int_0^\alpha \int_{J(0)} \big(D_i\tilde{g}(s, z_2(s,\overline{y}))\big)^2 (\alpha - s) d\overline{y}ds.$$

Then, with the change of variables $z_2 = z_2(s,\overline{y})$, one gets

$$B_i \leq \mathrm{diam}(K) \int_0^\alpha \int_{J(s)} \big(D_i\tilde{g}(s, z_2)\big)^2 \frac{\alpha - s}{1 - \frac{s}{\alpha}} dz_2ds.$$

Hence

$$B_i \leq \mathrm{diam}(K)^2 \int_K \big(D_i\tilde{g}(x)\big)^2 dx. \tag{9.42}$$

Using the fact that $\mathrm{m}(K) \geq \pi\zeta^2\big(\mathrm{diam}(K)\big)^2$, (9.39), (9.40), (9.41) and (9.42), one concludes (9.38).

In order to conclude the proof of (9.37), one remarks that

$$\Big(\mathcal{N}(\tilde{g},\mathcal{T})\Big)^2 \leq 2\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma(\tilde{g}_K-\tilde{g}_\sigma)^2.$$

Because, for all $K\in\mathcal{T}$ and $\sigma\in\mathcal{E}_K$, $d_\sigma\geq\zeta\mathrm{diam}(K)$, one gets thanks to (9.38), that

$$\Big(\mathcal{N}(\tilde{g},\mathcal{T})\Big)^2 \leq 2\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{C_1}{\zeta}\int_K|\nabla\tilde{g}(x)|^2dx.$$

The above inequality shows that

$$\Big(\mathcal{N}(\tilde{g},\mathcal{T})\Big)^2 \leq 2M\frac{C_1}{\zeta}\int_\Omega|\nabla\tilde{g}(x)|^2dx,$$

which implies (9.37). ∎

Theorem 9.2 (Convergence, non homogeneous Dirichlet boundary condition)
Assume items 1, 2, 3 and 4 of Assumption 9.1 page 32 and $g\in H^{1/2}(\partial\Omega)$. Let $\zeta\in\mathbb{R}_+$ and $M\in\mathbb{N}$ be given values. Let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 9.1 page 37) such that $d_{K,\sigma}\geq\zeta\mathrm{diam}(K)$ for all control volumes $K\in\mathcal{T}$ and for all $\sigma\in\mathcal{E}_K$, and $\mathrm{card}(\mathcal{E}_K)\leq M$ for all $K\in\mathcal{T}$. Let $(u_K)_{K\in\mathcal{T}}$ be the solution of the system given by equations (9.20)-(9.22) and

$$u_\sigma=\frac{1}{\mathrm{m}(\sigma)}\int_\sigma g(x)d\gamma(x),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}. \tag{9.43}$$

(note that the proofs of existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ which were given in Lemma 9.2 page 42 remain valid). Define $u_\mathcal{T}\in X(\mathcal{T})$ by $u_\mathcal{T}(x)=u_K$ for a.e. $x\in K$ and for any $K\in\mathcal{T}$. Then, $u_\mathcal{T}$ converges, in $L^2(\Omega)$, to the unique variational solution $u\in H^1(\Omega)$ of Problem (9.1), (9.2) as $\mathrm{size}(\mathcal{T})\to 0$.

Proof of Theorem 9.2
The proof is only detailed for the case $b=0$ and $\mathbf{v}=0$ (the extension of the proof to the general case is straightforward using the proof of Theorem 9.1 page 45). Let $\tilde{g}\in H^1(\Omega)$ be such that the trace of $\tilde{g}$ on $\partial\Omega$ is equal to g. One defines $\tilde{u}_\mathcal{T}\in X(\mathcal{T})$ by $\tilde{u}_\mathcal{T}=u_\mathcal{T}-\tilde{g}_\mathcal{T}$ where $\tilde{g}_\mathcal{T}\in X(\mathcal{T})$ is defined by $\tilde{g}(x)=\frac{1}{\mathrm{m}(K)}\int_K\tilde{g}(y)dy$ for all $x\in K$ and all $K\in\mathcal{T}$. Then $(\tilde{u}_K)_{K\in\mathcal{T}}$ satisfies

$$\sum_{\sigma\in\mathcal{E}_K}\tilde{F}_{K,\sigma}=\mathrm{m}(K)f_K-\sum_{\sigma\in\mathcal{E}_K}G_{K,\sigma},\ \forall K\in\mathcal{T}, \tag{9.44}$$

$$\tilde{F}_{K,\sigma}=-\tau_{K|L}(\tilde{u}_L-\tilde{u}_K),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if } \sigma=K|L, \tag{9.45}$$

$$\tilde{F}_{K,\sigma}=\tau_\sigma(\tilde{u}_K),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}\ \text{such that}\ \sigma\in\mathcal{E}_K. \tag{9.46}$$

$$G_{K,\sigma}=-\tau_{K|L}(\tilde{g}_L-\tilde{g}_L),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if } \sigma=K|L, \tag{9.47}$$

$$G_{K,\sigma}=-\tau_\sigma(\tilde{g}_\sigma-\tilde{g}_L),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}\ \text{such that}\ \sigma\in\mathcal{E}_K, \tag{9.48}$$

where $\tilde{g}_\sigma=\frac{1}{\mathrm{m}(\sigma)}\int_\sigma g(x)d\gamma(x)$ Multiplying (9.44) by $\tilde{u}_K$, summing over $K\in\mathcal{T}$, gathering by edges in the right hand side and using the Cauchy-Schwarz inequality yields

$$\|\tilde{u}_\mathcal{T}\|_{1,\mathcal{T}}^2\leq\sum_{K\in\mathcal{T}}\mathrm{m}(K)f_K\tilde{u}_K+\mathcal{N}(\tilde{g},\mathcal{T})\|\tilde{u}_\mathcal{T}\|_{1,\mathcal{T}},$$

from the definition (9.36) page 49 of $\mathcal{N}(\tilde{g},\mathcal{T})$ and Definition 9.2 page 39 of $\|\cdot\|_{1,\mathcal{T}}$. Therefore, thanks to Lemma 9.4 page 49 and the discrete Poincaré inequality (9.13), there exists $C_1\in\mathbb{R}$, only depending

on Ω, $\|\tilde{g}\|_{H^1(\Omega)}$, ζ, M and f, such that $\|\tilde{u}_\mathcal{T}\|_{1,\mathcal{T}} \leq C_1$ and $\|\tilde{u}_\mathcal{T}\|_{L^2(\Omega)} \leq C_1$. Let us now prove that $\tilde{u}_\mathcal{T}$ converges in $L^2(\Omega)$, as size$(\mathcal{T}) \to 0$, towards the unique solution in $H^1_0(\Omega)$ to (9.3). We proceed as in Theorem 9.1 page 45. Using Lemma 9.3, the compactness result given in Theorem 14.2 page 94 and the uniqueness of the solution (in $H^1_0(\Omega)$) of (9.3), it is sufficient to prove that if $\tilde{u}_\mathcal{T}$ converges towards some $\tilde{u} \in H^1_0(\Omega)$, in $L^2(\Omega)$ as size$(\mathcal{T}) \to 0$, then $\tilde{u}$ is the solution to (9.3). In order to prove this result, let us introduce the function $\tilde{g}_\mathcal{T}$ defined by

$$\tilde{g}_\mathcal{T}(x) = \frac{1}{\mathrm{m}(K)} \int_K \tilde{g}(y)dy, \, \forall x \in K, \, \forall K \in \mathcal{T},$$

which converges to $\tilde{g}$ in $L^2(\Omega)$, as size$(\mathcal{T}) \to 0$. Then the function $u_\mathcal{T}$ converges in $L^2(\Omega)$, as size$(\mathcal{T}) \to 0$ to $u = \tilde{u} + \tilde{g} \in H^1(\Omega)$ and the proof that $\tilde{u}$ is the unique solution of (9.3) is identical to the corresponding part in the proof of Theorem 9.1 page 45. This completes the proof of Theorem 9.2. ∎

Remark 9.10 (Lipschitz continuous boundary data) A simpler proof of convergence for the finite volume scheme with non homogeneous Dirichlet boundary condition is possible if g is the trace of a Lipschitz-continuous function $\tilde{g}$. In thiscase, ζ and M do not have to be introduced and Lemma 9.4 is not used. The scheme is defined with $u_\sigma = g(y_\sigma)$ instead of the average value of g on σ, and the proof uses $\tilde{g}(x_K)$ instead of the average value of $\tilde{g}$ on K.

9.5 C^2 error estimate

Under adequate regularity assumptions on the solution of Problem (9.1)-(9.2), one may prove that the error between the exact solution and the approximate solution given by the finite volume scheme (9.20)-(9.23) is of order size$(\mathcal{T}) = \sup_{K\in\mathcal{T}} \mathrm{diam}(K)$, in a certain sense which we give in the following theorem:

Theorem 9.3 *Under Assumption 9.1 page 32, let $\mathcal{T}$ be an admissible mesh as defined in Definition 9.1 page 37 and $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2 page 39) be defined a.e.in Ω by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$, where $(u_K)_{K\in\mathcal{T}}$ is the solution to (9.20)-(9.23). Assume that the unique variational solution u of Problem (9.1)-(9.2) satisfies $u \in C^2(\overline{\Omega})$. Let, for each $K \in \mathcal{T}$, $e_K = u(x_K) - u_K$, and $e_\mathcal{T} \in X(\mathcal{T})$ defined by $e_\mathcal{T}(x) = e_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$.*
Then, there exists $C > 0$ only depending on u, $\mathbf{v}$ and Ω such that

$$\|e_\mathcal{T}\|_{1,\mathcal{T}} \leq C\mathrm{size}(\mathcal{T}), \tag{9.49}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H^1_0 norm defined in Definition 9.2,

$$\|e_\mathcal{T}\|_{L^2(\Omega)} \leq C\mathrm{size}(\mathcal{T}) \tag{9.50}$$

and

$$\begin{aligned} &\sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}} \\ \sigma=K|L}} \mathrm{m}(\sigma)d_\sigma \Big(\frac{u_L - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)\Big)^2 + \\ &\sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{ext}} \\ \sigma\in\overline{K}\cap\partial\Omega}} \mathrm{m}(\sigma)d_\sigma \Big(\frac{g(y_\sigma) - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)\Big)^2 \leq C\mathrm{size}(\mathcal{T})^2. \end{aligned} \tag{9.51}$$

Remark 9.11

1. Inequality (9.49) (resp. (9.50)) yields an estimate of order 1 for the discrete H^1_0 norm (resp. L^2 norm) of the error on the solution. Note also that, since $u \in C^1(\overline{\Omega})$, one deduces, from (9.50), the existence of C only depending on u and Ω such that $\|u - u_\mathcal{T}\|_{L^2(\Omega)} \leq C\mathrm{size}(\mathcal{T})$. Inequality (9.51) may be seen as an estimate of order 1 for the L^2 norm of the flux.

2. In Baranger, Maitre and Oudin [8], finite element tools are used to obtain error estimates of order size$(\mathcal{T})^2$ in the case $d = 2$, $\mathbf{v} = b = g = 0$ and if the elements of $\mathcal{T}$ are triangles of a finite element mesh satisfying the Delaunay condition (see section 12 page 85). Note that this result is quite different of those of the remarks 6.2 page 18 and 9.1 page 35, which are obtained by using a higher order approximation of the flux.

3. The proof of Theorem 9.3 given below is close to that of error estimates for finite element schemes in the sense that it uses the coerciveness of the operator (the discrete Poincaré inequality) instead of the discrete maximum principle of Proposition 9.2 page 43 (which is used for error estimates with finite difference schemes).

Proof of Theorem 9.3

Let $u_{\mathcal{T}} \in X(\mathcal{T})$ be defined a.e. in Ω by $u_{\mathcal{T}}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$, where $(u_K)_{K\in\mathcal{T}}$ is the solution to (9.20)-(9.23). Let us write the flux balance for any $K \in \mathcal{T}$;

$$\sum_{\sigma\in\mathcal{E}_K} \left(\overline{F}_{K,\sigma} + \overline{V}_{K,\sigma}\right) + b\int_K u(x)dx = \int_K f(x)dx, \tag{9.52}$$

where $\overline{F}_{K,\sigma} = -\int_\sigma \nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)$, and $\overline{V}_{K,\sigma} = \int_\sigma u(x)\mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)$ are respectively the diffusion and convection fluxes through σ outward to K.
Let $F^\star_{K,\sigma}$ and $V^\star_{K,\sigma}$ be defined by

$$F^\star_{K,\sigma} = -\tau_{K|L}(u(x_L) - u(x_K)),\ \forall\sigma = K|L \in \mathcal{E}_K \cap \mathcal{E}_{\rm int},\ \forall K \in \mathcal{T},$$

$$F^\star_{K,\sigma} d(x_K,\sigma) = -{\rm m}(\sigma)(u(y_\sigma) - u(x_K)),\ \forall\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm ext},\ \forall K \in \mathcal{T},$$

$$V^\star_{K,\sigma} = v_{K,\sigma}u(x_{\sigma,+}), \forall\sigma \in \mathcal{E}_K,\ \forall K \in \mathcal{T},$$

where $x_{\sigma,+} = x_K$ (resp. x_L) if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$ and $v_{K,\sigma} \geq 0$ (resp. $v_{K,\sigma} \leq 0$) and $x_{\sigma,+} = x_K$ (resp. y_σ) if $\sigma = \mathcal{E}_K \cap \mathcal{E}_{\rm ext}$ and $v_{K,\sigma} \geq 0$ (resp. $v_{K,\sigma} \leq 0$). Then, the consistency error on the diffusion and convection fluxes may be defined as

$$R_{K,\sigma} = \frac{1}{{\rm m}(\sigma)}(\overline{F}_{K,\sigma} - F^\star_{K,\sigma}), \tag{9.53}$$

$$r_{K,\sigma} = \frac{1}{{\rm m}(\sigma)}(\overline{V}_{K,\sigma} - V^\star_{K,\sigma}), \tag{9.54}$$

Thanks to the regularity of u and $\mathbf{v}$, there exists $C_1 \in \mathbb{R}$, only depending on u and $\mathbf{v}$, such that $|R_{K,\sigma}| + |r_{K,\sigma}| \leq C_1{\rm size}(\mathcal{T})$ for any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$. For $K \in \mathcal{T}$, let

$$\rho_K = u(x_K) - (1/{\rm m}(K))\int_K u(x)dx,$$

so that $|\rho_K| \leq C_2{\rm size}(\mathcal{T})$ with some $C_2 \in \mathbb{R}_+$ only depending on u.

Substract (9.20) to (9.52); thanks to (9.53) and (9.54), one has

$$\sum_{\sigma\in\mathcal{E}_K} \left(G_{K,\sigma} + W_{K,\sigma}\right) + b{\rm m}(K)e_K = b{\rm m}(K)\rho_K - \sum_{\sigma\in\mathcal{E}_K} {\rm m}(\sigma)(R_{K,\sigma} + r_{K,\sigma}), \tag{9.55}$$

where
$G_{K,\sigma} = F^\star_{K,\sigma} - F_{K,\sigma}$ is such that

$$G_{K,\sigma} = -\tau_{K|L}(e_L - e_K),\ \forall K \in \mathcal{T},\ \forall\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm int}, \sigma = K|L,$$

$$G_{K,\sigma}d(x_K,\sigma) = \mathrm{m}(\sigma)e_K,\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext},$$

with $e_K = u(x_K) - u_K$, and $W_{K,\sigma} = V^\star_{K,\sigma} - V_{K,\sigma} = v_{K,\sigma}(u(x_{\sigma,+}) - u_{\sigma,+})$
Multiply (9.55) by e_K, sum for $K\in\mathcal{T}$, and note that

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}e_K = \sum_{\sigma\in\mathcal{E}}|D_\sigma e|^2\frac{\mathrm{m}(\sigma)}{d_\sigma} = \|e\|^2_{1,\mathcal{T}}.$$

Hence

$$\|e_\mathcal{T}\|^2_{1,\mathcal{T}} + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}e_{\sigma,+}e_K + b\|e_\mathcal{T}\|^2_{L^2(\Omega)} \le b\sum_{K\in\mathcal{T}}\mathrm{m}(K)\rho_K e_K - \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)(R_{K,\sigma}+r_{K,\sigma})e_K, \quad (9.56)$$

where
$e_\mathcal{T}\in X(\mathcal{T})$, $e_\mathcal{T}(x) = e_K$ for a.e. $x\in K$ and for all $K\in\mathcal{T}$,
$|D_\sigma e| = |e_K - e_L|$, if $\sigma\in\mathcal{E}_{\rm int}$, $\sigma = K|L$, $|D_\sigma e| = |e_K|$, if $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}$,
$e_{\sigma,+} = u(x_{\sigma,+}) - u_{\sigma,+}$.
By Young's inequality, the first term of the left hand side satisfies:

$$|\sum_{K\in\mathcal{T}}\mathrm{m}(K)\rho_K e_K| \le \frac{1}{2}\|e_\mathcal{T}\|^2_{L^2(\Omega)} + \frac{1}{2}C_2^2(\mathrm{size}(\mathcal{T}))^2\mathrm{m}(\Omega). \quad (9.57)$$

Thanks to the assumption div$\mathbf{v}\ge 0$, one obtains, through a computation similar to (9.27)-(9.28) page 43 that

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}e_{\sigma,+}e_K \ge 0.$$

Hence, (9.56) and (9.57) yield that there exists C_3 only depending on u, b and Ω such that

$$\|e_\mathcal{T}\|^2_{1,\mathcal{T}} + \frac{1}{2}b\|e_\mathcal{T}\|^2_{L^2(\Omega)} \le C_3(\mathrm{size}(\mathcal{T}))^2 - \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)(R_{K,\sigma}+r_{K,\sigma})e_K, \quad (9.58)$$

Thanks to the property of conservativity, one has $R_{K,\sigma} = -R_{L,\sigma}$ and $r_{K,\sigma} = -r_{L,\sigma}$ for $\sigma\in\mathcal{E}_{\rm int}$ such that $\sigma = K|L$. Let $R_\sigma = |R_{K,\sigma}|$ and $r_\sigma = |r_{K,\sigma}|$ if $\sigma\in\mathcal{E}_K$. Reordering the summation over the edges and from the Cauchy-Schwarz inequality, one then obtains

$$\begin{aligned}|\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)(R_{K,\sigma}+r_{K,\sigma})e_K| \le \sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)(D_\sigma e)(R_\sigma+r_\sigma) \le \\ \Big(\sum_{\sigma\in\mathcal{E}}\frac{\mathrm{m}(\sigma)}{d_\sigma}(D_\sigma e)^2\Big)^{\frac{1}{2}}\Big(\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)d_\sigma(R_\sigma+r_\sigma)^2\Big)^{\frac{1}{2}}.\end{aligned} \quad (9.59)$$

Now, since $|R_\sigma + r_\sigma| \le C_1\mathrm{size}(\mathcal{T})$ and since $\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)d_\sigma = d\,\mathrm{m}(\Omega)$, (9.58) and (9.59) yield the existence of $C_4\in\mathbb{R}_+$ only depending on $u, \mathbf{v}$ and Ω such that

$$\|e_\mathcal{T}\|^2_{1,\mathcal{T}} + \frac{1}{2}b\|e_\mathcal{T}\|^2_{L^2(\Omega)} \le C_3(\mathrm{size}(\mathcal{T}))^2 + C_4\mathrm{size}(\mathcal{T})\|e\|_{1,\mathcal{T}}.$$

Using again Young's inequality, there exists C_5 only depending on u, $\mathbf{v}$, b and Ω such that

$$\|e_\mathcal{T}\|^2_{1,\mathcal{T}} + b\|e_\mathcal{T}\|^2_{L^2(\Omega)} \le C_5(\mathrm{size}(\mathcal{T}))^2. \quad (9.60)$$

This inequality yields Estimate (9.49) and, in the case $b>0$, Estimate (9.50). In the case where $b=0$, one uses the discrete Poincaré inequality (9.13) and the inequality (9.60) to obtain

$$\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} \leq \mathrm{diam}(\Omega)^2 C_5 (\mathrm{size}(\mathcal{T}))^2,$$

which yields (9.50).

Remark now that (9.49) can be written

$$\begin{aligned} &\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u_L-u_K}{d_\sigma} - \frac{u(x_L)-u(x_K)}{d_\sigma}\Big)^2 + \\ &\sum_{\substack{\sigma\in\mathcal{E}_{\rm ext}\\ \sigma\in\overline{K}\cap\partial\Omega}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{g(y_\sigma)-u_K}{d_\sigma} - \frac{u(y_\sigma)-u(x_K)}{d_\sigma}\Big)^2 \leq (C\mathrm{size}(\mathcal{T}))^2. \end{aligned} \tag{9.61}$$

From Definition (9.53) and the consistency of the fluxes, one has

$$\begin{aligned} &\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u(x_L)-u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\Big)^2 + \\ &\sum_{\substack{\sigma\in\mathcal{E}_{\rm ext}\\ \sigma\in\overline{K}\cap\partial\Omega}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u(y_\sigma)-u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\Big)^2 = \\ &\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma R_\sigma^2 \leq d\mathrm{m}(\Omega) C_1^2 (\mathrm{size}(\mathcal{T}))^2. \end{aligned} \tag{9.62}$$

Then (9.61) and (9.62) give (9.51). ∎

9.6 H^2 error estimate

In Theorem 9.3, the hypothesis $u \in C^2(\overline{\Omega})$ was used. In the following theorem (Theorem 9.4), one obtains Estimates (9.49) and (9.50), in the case $b = \mathbf{v} = 0$ and assuming some additional assumption on the mesh (see Definition 9.3 below), under the weaker assumption $u \in H^2(\Omega)$. This additional assumption on the mesh is not completely necessary (see Remark 9.13 and Gallouët, Herbin and Vignal [72]). It is also possible to obtain Estimates (9.49) and (9.50) in the cases $b \neq 0$ or $\mathbf{v} \neq 0$ assuming $u \in H^2(\Omega)$ (see Remark 9.13 and Gallouët, Herbin and Vignal [72]). Some similar results are also in Lazarov, Mishev and Vassilevski [99] and Coudière, Vila and Villedieu [41].

Definition 9.3 (Restricted admissible meshes) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3. A restricted admissible finite volume mesh of Ω, denoted by $\mathcal{T}$, is an admissible mesh in the sense of Definition 9.1 such that, for some $\zeta > 0$, one has $d_{K,\sigma} \geq \zeta \mathrm{diam}(K)$ for all control volumes K and for all $\sigma \in \mathcal{E}_K$.

Theorem 9.4 (H^2 regularity) *Under Assumption 9.1 page 32 with $b = \mathbf{v} = 0$, let $\mathcal{T}$ be a restricted admissible mesh in the sense of Definition 9.3 and $u_{\mathcal{T}} \in X(\mathcal{T})$ (see Definition 9.2 page 39) be the approximate solution defined in Ω by $u_{\mathcal{T}}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$, where $(u_K)_{K\in\mathcal{T}}$ is the (unique) solution to (9.20)-(9.23) (existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ are given by Lemma 9.2). Assume that the unique solution, u, of (9.3) (with $b = \mathbf{v} = 0$) belongs to $H^2(\Omega)$. For each control volume K, let $e_K = u(x_K) - u_K$, and $e_{\mathcal{T}} \in X(\mathcal{T})$ defined by $e_{\mathcal{T}}(x) = e_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$.*
Then, there exists C, only depending on u, ζ and Ω, such that (9.49), (9.50) and (9.51) hold.

Remark 9.12

1. In Theorem 9.4, the function $e_{\mathcal{T}}$ is still well defined, and so is the quantity "$\nabla u \cdot \mathbf{n}_\sigma$" on σ, for all $\sigma \in \mathcal{E}$. Indeed, since $u \in H^2(\Omega)$ (and $d \leq 3$), one has $u \in C(\overline{\Omega})$ (and then $u(x_K)$ is well defined for all control volumes K) and $\nabla u \cdot \mathbf{n}_\sigma$ belongs to $L^2(\sigma)$ (for the $(d-1)$-dimensional Lebesgue measure on σ) for all $\sigma \in \mathcal{E}$.

2. Note that, under Assumption 9.1 with $b = \mathbf{v} = g = 0$ the (unique) solution of (9.3) is necessarily in $H^2(\Omega)$ provided that Ω is convex.

PROOF of Theorem 9.4

Let K be a control volume and $\sigma \in \mathcal{E}_K$. Define $\mathcal{V}_{K,\sigma} = \{tx_K + (1-t)x,\ x \in \sigma,\ t \in [0,1]\}$. For $\sigma \in \mathcal{E}_{\rm int}$, let $\mathcal{V}_\sigma = \mathcal{V}_{K,\sigma} \cup \mathcal{V}_{L,\sigma}$, if K and L are the control volumes such that $\sigma = K|L$. For $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$, let $\mathcal{V}_\sigma = \mathcal{V}_{K,\sigma}$.

The main part of the proof consists in proving the existence of some C, only depending on the space dimension d and ζ (given in Definition 9.3), such that, for all control volumes K and for all $\sigma \in \mathcal{E}_K$,

$$|R_{K,\sigma}|^2 \le C \frac{(\mathrm{size}(\mathcal{T}))^2}{\mathrm{m}(\sigma) d_\sigma} \int_{\mathcal{V}_\sigma} |H(u)(z)|^2 dz, \tag{9.63}$$

where$H$ is the Hessian matrix of u and

$$|H(u)(z)|^2 = \sum_{i,j=1}^{d} |D_i D_j u(z)|^2,$$

and D_i denotes the (weak) derivative with respect to the component z_i of $z = (z_1, \cdots, z_d)^t \in \mathbb{R}^d$.
Recall that $R_{K,\sigma}$ is the consistency error on the diffusion flux (see (9.53)), that is:

$$R_{K,\sigma} = \frac{u(x_L) - u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x), \text{ if } \sigma \in \mathcal{E}_{\rm int} \text{ and } \sigma = K|L,$$

$$R_{K,\sigma} = \frac{u(y_\sigma) - u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x), \text{ if } \sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K.$$

Note that $R_{K,\sigma}$ is well defined, thanks to $u \in H^2(\Omega)$, see Remark 9.12.

In Step 1, one proves (9.63), and, in Step 2, we conclude the proof of Estimates (9.49) and (9.50).

Step 1. Proof of (9.63).

Let $\sigma \in \mathcal{E}$. Since $u \in H^2(\Omega)$, the restriction of u to $\mathcal{V}_\sigma$ belongs to $H^2(\mathcal{V}_\sigma)$. The space $C^2(\overline{\mathcal{V}_\sigma})$ is dense in $H^2(\mathcal{V}_\sigma)$ (see, for instance, NEČAS [113], this can be proved quite easily be a regularization technique). Then, by a density argument, one needs only to prove (9.63) for $u \in C^2(\overline{\mathcal{V}_\sigma})$. Therefore, in the remainder of Step 1, it is assumed $u \in C^2(\overline{\mathcal{V}_\sigma})$.

First, one proves (9.63) if $\sigma \in \mathcal{E}_{\rm int}$. Let K and L be the 2 control volumes such that $\sigma = K|L$.
It is possible to assume, for simplicity of notations and without loss of generality, that $\sigma = 0 \times \tilde{\sigma}$, with some $\tilde{\sigma} \subset \mathbb{R}^{d-1}$, and $x_K = (-\alpha, 0)^t$, $x_L = (\beta, 0)^t$, with some $\alpha > \zeta \mathrm{diam}(K)$, $\beta > \zeta \mathrm{diam}(L)$ (ζ is defined in Definition 9.3 page 55).
Let $x = (0, \tilde{x})^t \in \sigma$. In order to obtain a suitable integral remainder for the consistency error, as suggested in Remark 6.3, we introduce the function $\varphi : [0,1] \to \mathbb{R}$, defined by $\varphi(t) = u(tx_K + (1-t)x)$, which is twice continuously differentiable and we have:

$$\varphi(1) = x_K, \varphi(0) = u(x), \varphi'(t) = \nabla u(tx_K + (1-t)x) \cdot (x_K - x)$$
$$\text{and } \varphi''(t) = Hu(tx_K + (1-t)x)(x_K - x) \cdot (x_K - x),$$

where $H(u)(z)$ denotes the Hessian matrix of u at point z. Therefore, writing that

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\, dt = \int_0^1 \varphi'(t)\, dt = \varphi'(0) - \int_0^1 (t-1)\varphi''(t)\, dt$$

yields that

$$u(x_K) - u(x) = \int_0^1 H(u)(tx + (1-t)x_K)(x_K - x) \cdot (x_K - x) t dt \text{ for a.e. } x = (0, \tilde{x})^t \in \sigma$$

(for the $(d-1)$-dimensional Lebesgue measure on σ). Similarly, we have

$$u(x_L) - u(x) = \nabla u(x) \cdot (x_L - x) + \int_0^1 H(u)(tx + (1-t)x_L)(x_L - x) \cdot (x_L - x) t dt.$$

Subtracting one equation to the other and integrating over σ yields (note that $x_L - x_K = \mathbf{n}_{K,\sigma} d_\sigma$) $|R_{K,\sigma}| \leq B_{K,\sigma} + B_{L,\sigma}$, with

$$B_{K,\sigma} = \frac{C_1}{\mathrm{m}(\sigma) d_\sigma} \int_\sigma \int_0^1 |H(u)(tx + (1-t)x_K)||x_K - x|^2 t dt d\gamma(x), \tag{9.64}$$

for some C_1 only depending on d, The quantity $B_{L,\sigma}$ is obtained with $B_{K,\sigma}$ by changing K in L.
Let us perform the change of variables

$$\begin{aligned} h :]0,1[\times \sigma &\to \mathcal{V}_{K,\sigma} \\ (t,x) &\mapsto h(t,x) = tx + (1-t)x_K, \end{aligned}$$

in (9.64). let z_1 denote the first component of z and $\bar{z}$ the $d-1$ last components of z; thus $z = (z_1, \overline{z})^t$ and $z_1 = (t-1)\alpha$, so that

$$dz = t^{d-1} \alpha \delta t d\gamma(x).$$

Since $|x_K - x| \leq \mathrm{diam}(K)$ we obtain

$$B_{K,\sigma} \leq \frac{C_1 (\mathrm{diam}(K))^2}{\mathrm{m}(\sigma) d_\sigma} \int_{\mathcal{V}_{K,\sigma}} |H(u)(z)| \frac{\alpha^{d-2}}{\alpha (z_1 + \alpha)^{d-2}} dz.$$

This gives, with the famous Cauchy-Schwarz inequality,

$$B_{K,\sigma} \leq \frac{C_1 \alpha^{d-3} (\mathrm{diam}(K))^2}{\mathrm{m}(\sigma) d_\sigma} \Big(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 dz \Big)^{\frac{1}{2}} \Big(\int_{\mathcal{V}_{K,\sigma}} \frac{1}{(z_1 + \alpha)^{(d-2)2}} dz \Big)^{\frac{1}{2}}.$$

For $d = 2$, (9.6) gives

$$B_{K,\sigma} \leq \frac{C_1 (\mathrm{diam}(K))^2}{\alpha \mathrm{m}(\sigma) d_\sigma} \Big(\frac{\alpha \mathrm{m}(\sigma)}{2} \Big)^{\frac{1}{2}} \Big(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 dz \Big)^{\frac{1}{2}},$$

and therefore

$$B_{K,\sigma} \leq \frac{C_1 (\mathrm{diam}(K))^2}{2^{\frac{1}{2}} (\mathrm{m}(\sigma) d_\sigma)^{\frac{1}{2}} (d_\sigma \alpha)^{\frac{1}{2}}} \Big(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 dz \Big)^{\frac{1}{2}}.$$

A similar estimate holds on $B_{L,\sigma}$ by changing K in L and α in β. Since $\alpha, \beta \geq \zeta \mathrm{diam}(K)$ and $d_\sigma = \alpha + \beta \geq \zeta \mathrm{diam}(K)$, these estimates on $B_{K,\sigma}$ and $B_{L,\sigma}$ yield (9.63) for some C only depending on d and ζ.
For $d = 3$, the computation of the integral $A = \int_{\mathcal{V}_{K,\sigma}} \frac{1}{(z_1+\alpha)^2} dz$ by the following change of variable (see Figure (9.6)):

$$A = \int_{-d}^0 \frac{1}{(z_1 + \alpha)^2} \Big(\int_{\overline{z} \in t\tilde{\sigma}} d\overline{z} \Big) dz_1, \text{ where } t = \frac{z_1 + \alpha}{d_{K,\sigma}}.$$

Now,

$$\int_{\overline{z} \in t\tilde{\sigma}} d\overline{z} = \int_{y \in \tilde{\sigma}} t^2 dy = \frac{(z_1 + \alpha)^2}{\alpha^2} \mathrm{m}(\sigma),$$

and therefore $A = \frac{\mathrm{m}(\sigma)}{\alpha}$, and (9.6) yields that:

$$B_{K,\sigma} \leq \frac{C_3 \, (\mathrm{diam}(K))^2}{(\mathrm{m}(\sigma) \, d_\sigma^2 \, d_{K,\sigma})^{1/2}} \left(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 \, dz \right)^{1/2} \leq \frac{C_3 \, \mathrm{size}(\mathcal{T})}{\sqrt{2} \, \zeta \, (\mathrm{m}(\sigma) \, d_\sigma)^{1/2}} \, \|H(u)\|_{L^2(\mathcal{V}_{K,\sigma})}.$$

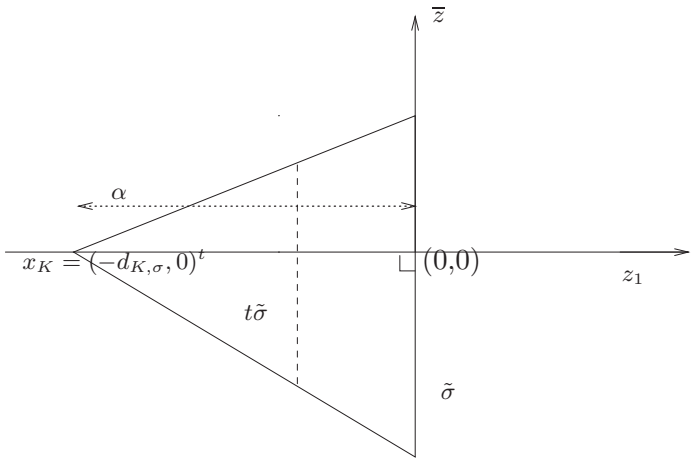


Figure 3.3: Consistency error, $d = 3$

and therefore (9.6) gives:

$$B_{K,\sigma} \leq \frac{C_1(\mathrm{diam}(K))^2}{\mathrm{m}(\sigma)d_\sigma}\Big(\int_{-\alpha}^{0} \frac{\mathrm{m}(\sigma)}{\alpha^2}dz_1\Big)^{\frac{1}{2}}\Big(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 dz\Big)^{\frac{1}{2}},$$

and then

$$B_{K,\sigma} \leq \frac{C_1(\mathrm{diam}(K))^2}{(\mathrm{m}(\sigma)d_\sigma)^{\frac{1}{2}}(d_\sigma\alpha)^{\frac{1}{2}}}\Big(\int_{\mathcal{V}_{K,\sigma}} |H(u)(z)|^2 dz\Big)^{\frac{1}{2}}.$$

With a similar estimate on $B_{L,\sigma}$, this yields (9.63) for some C only depending on d and ζ.

Now, one proves (9.63) if $\sigma \in \mathcal{E}_{\mathrm{ext}}$. Let K be the control volume such that $\sigma \in \mathcal{E}_K$. One can assume, without loss of generality, that $x_K = 0$ and $\sigma = \{2\alpha\} \times \tilde{\sigma}$ with $\tilde{\sigma} \subset \mathbb{R}^{d-1}$ and some $\alpha \geq \frac{1}{2}\zeta\mathrm{diam}(K)$. The above proof gives (see Definition 9.1 page 37 for the definition of y_σ), with some C_2 only depending on d,

$$\Big|\frac{u(y_\sigma) - u(x_K)}{2\alpha} - \frac{1}{\mathrm{m}(\hat{\sigma})}\int_{\hat{\sigma}} \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\Big|^2 \leq C_2 \frac{(\mathrm{size}(\mathcal{T}))^2}{\mathrm{m}(\sigma)d_\sigma}\int_{\mathcal{V}_{\hat{\sigma}}} |H(u)(z)|^2 dz, \tag{9.65}$$

with $\hat{\sigma} = \{(\alpha\,\frac{x}{2}), x \in \tilde{\sigma}\}$, and $\mathcal{V}_{\hat{\sigma}} = \{ty_\sigma + (1-t)x,\ x \in \hat{\sigma},\ t \in [0,1]\} \cup \{tx_K + (1-t)x,\ x \in \hat{\sigma},\ t \in [0,1]\}$. Note that $\mathrm{m}(\hat{\sigma}) = \frac{\mathrm{m}(\sigma)}{2^{d-1}}$ and that $\mathcal{V}_{\hat{\sigma}} \subset \mathcal{V}_\sigma$.
One has now to compare $I_\sigma = \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)$ with $I_{\hat{\sigma}} = \frac{1}{\mathrm{m}(\hat{\sigma})}\int_{\hat{\sigma}} \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)$.
A Taylor expansion gives

$$I_\sigma - I_{\hat{\sigma}} = \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \int_{\frac{1}{2}}^{1} H(u)(x_K + t(x - x_K))(x - x_K)\cdot \mathbf{n}_{K,\sigma} dt d\gamma(x).$$

The change of variables in this last integral $z = x_K + t(x - x_K)$, which gives $dz = 2\alpha t^{d-1} dt d\gamma(x)$, yields, with $E_\sigma = \{tx + (1-t)x_K,\ x \in \sigma,\ t \in [\frac{1}{2}, 1]\}$ and some C_3 only depending on d (note that $t \geq \frac{1}{2}$),

$$|I_\sigma - I_{\hat{\sigma}}| \leq \frac{C_3}{\mathrm{m}(\sigma)\alpha}\int_{E_\sigma} |H(u)(z)||x - x_K| dz.$$

Then, from the Cauchy-Schwarz inequality and since $|x - x_K| \leq \mathrm{diam}(K)$,

$$|I_\sigma - I_{\hat\sigma}|^2 \leq \frac{C_4(\mathrm{diam}(K))^2}{\mathrm{m}(\sigma)d_\sigma}\int_{E_\sigma}|H(u)(z)|^2dz, \tag{9.66}$$

with some C_4 only depending on d and ζ.
Inequalities (9.65) and (9.66) yield (9.63) for some C only depending on d and ζ.
One may therefore choose $C \in \mathbb{R}_+$ such that (9.63) holds for $\sigma \in \mathcal{E}_{\rm int}$ or $\sigma \in \mathcal{E}_{\rm ext}$. This concludes Step 1.

Step 2. Proof of Estimates (9.49), (9.50) and (9.51).
In order to obtain Estimate (9.49) (and therefore (9.50) from the discrete Poincaré inequality (9.13)), one proceeds as in Theorem 9.3. Inequality (9.56) reads here, since $R_{K,\sigma} = -R_{L,\sigma}$, if $\sigma = K|L$,

$$\|e_\mathcal{T}\|_{1,\mathcal{T}}^2 \leq \sum_{\sigma\in\mathcal{E}} R_\sigma |D_\sigma e|\mathrm{m}(\sigma),$$

with $R_\sigma = |R_{K,\sigma}|$, if $\sigma \in \mathcal{E}_K$. Recall also that $|D_\sigma e| = |e_K - e_L|$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$ and $|D_\sigma e| = |e_K|$, if $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$. Cauchy and Schwarz strike again:

$$\|e_\mathcal{T}\|_{1,\mathcal{T}}^2 \leq \big(\sum_{\sigma\in\mathcal{E}} R_\sigma^2 \mathrm{m}(\sigma)d_\sigma\big)^{\frac12}\big(\sum_{\sigma\in\mathcal{E}}|D_\sigma e|^2\frac{\mathrm{m}(\sigma)}{d_\sigma}\big)^{\frac12}.$$

The main consequence of (9.63) is that

$$\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)d_\sigma R_\sigma^2 \leq C(\mathrm{size}(\mathcal{T}))^2\sum_{\sigma\in\mathcal{E}}\int_{\mathcal{V}_\sigma}|H(u)(z)|^2dz = C(\mathrm{size}(\mathcal{T}))^2\int_\Omega |H(u)(z)|^2dz. \tag{9.67}$$

Then, one obtains

$$\|e_\mathcal{T}\|_{1,\mathcal{T}} \leq \sqrt{C}\mathrm{size}(\mathcal{T})\big(\int_\Omega |H(u)(z)|^2dz\big)^{\frac12}.$$

This concludes the proof of (9.49) since $u \in H^2(\Omega)$ implies $\int_\Omega |H(u)(z)|^2dz < \infty$.
Estimate (9.51) follows from (9.67) in a similar manner as in the proof of Theorem 9.3. This concludes the proof of Theorem 9.4. ∎

Remark 9.13 (Generalizations)

1. By developping the method used to bound the consistency error on the flux on the elements of $\mathcal{E}_{\rm ext}$, it is possible to replace, in Theorem 9.4, the hypothesis $d_{K,\sigma} \geq \zeta\mathrm{diam}(K)$ in Definition 9.3 page 55 by the weaker hypothesis $d_\sigma \geq \zeta\mathrm{diam}(\sigma)$ provided that $\mathcal{V}_\sigma$ is convex. Note also that, in this case, the hypothesis $x_K \in K$ is not necessary, it suffices that $x_L - x_K = d_\sigma \mathbf{n}_{K,\sigma}$, for all $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$ (for $\sigma \in \mathcal{E}_{\rm ext}$, one always needs $y_\sigma - x_K = d_\sigma \mathbf{n}_{K,\sigma}$).

2. It is also possible to prove Theorem 9.4 if $b \neq 0$ or $\mathbf{v} \neq 0$ (or, of course, $b \neq 0$ and $\mathbf{v} \neq 0$). Indeed, if the solution, u, to (9.3) is not only in $H^2(\Omega)$ but is also Lipschitz continuous on $\overline{\Omega}$ (this is the case if, for instance, there exists $p > d$ such that $u \in W^{2,p}(\Omega)$), the treatment of the consistency error terms due to the terms involving b and $\mathbf{v}$ are exactly as in Theorem 9.3. If u is not Lipschitz continuous on $\overline{\Omega}$, one has to deal with the consistency error terms due to b and $\mathbf{v}$ similarly as in the proof of Theorem 9.4 (see also Eymard, Gallouët and Herbin [55] or Gallouët, Herbin and Vignal [72]).

It is also possible, essentially under Assumption 9.1 page 32, to obtain an L^q estimate of the error, for $2 \leq q < +\infty$ if $d = 2$, and for $1 \leq q \leq 6$ if $d = 3$, see [39]. The error estimate for the L^q norm is a consequence of the following lemma:

Lemma 9.5 (Discrete Sobolev Inequality) *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$ and $\mathcal{T}$ be a general finite volume mesh of Ω in the sense of definition 10.1 page 63, and let $\zeta > 0$ be such that*

$$\forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K, \qquad d_{K,\sigma} \geq \zeta d_\sigma, \tag{9.68}$$

Let be $u \in X(\mathcal{T})$ (see definition 9.2 page 39), then, there exists $C > 0$ only depending on Ω and ζ, such that for all $q \in [2, +\infty)$, if $d = 2$, and $q \in [2, 6]$, if $d = 3$,

$$\|u\|_{L^q(\Omega)} \leq Cq\|u\|_{1,\mathcal{T}}, \tag{9.69}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in definition 9.2 page 39.

Proof of Lemma 9.5

Let us first prove the two-dimensional case. Assume $d = 2$ and let $q \in [2, +\infty)$. Let $\boldsymbol{d}_1 = (1, 0)^t$ and $\boldsymbol{d}_2 = (0, 1)^t$; for $x \in \Omega$, let $\mathcal{D}_x^1$ and $\mathcal{D}_x^2$ be the straight lines going through x and defined by the vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$.

Let $v \in X(\mathcal{T})$. For all control volume K, one denotes by v_K the value of v on K. For any control volume K and a.e. $x \in K$, one has

$$v_K^2 \leq \sum_{\sigma\in\mathcal{E}} D_\sigma v\, \chi_\sigma^{(1)}(x) \sum_{\sigma\in\mathcal{E}} D_\sigma v\, \chi_\sigma^{(2)}(x), \tag{9.70}$$

where $\chi_\sigma^{(1)}$ and $\chi_\sigma^{(2)}$ are defined by

$$\chi_\sigma^{(i)}(x) = \begin{cases} 1 & \text{if } \sigma \cap \mathcal{D}_x^i \neq \emptyset \\ 0 & \text{if } \sigma \cap \mathcal{D}_x^i = \emptyset \end{cases} \quad \text{for } i = 1, 2.$$

Recall that $D_\sigma v = |v_K - v_L|$, if $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$ and $D_\sigma v = |v_K|$, if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$. Integrating (9.70) over K and summing over $K \in \mathcal{T}$ yields

$$\int_\Omega v^2(x)dx \leq \int_\Omega \Big(\sum_{\sigma\in\mathcal{E}} D_\sigma v\, \chi_\sigma^{(1)}(x) \sum_{\sigma\in\mathcal{E}} D_\sigma v\, \chi_\sigma^{(2)}(x) \Big) dx.$$

Note that $\chi_\sigma^{(1)}$ (resp. $\chi_\sigma^{(2)}$) only depends on the second component x_2 (resp. the first component x_1) of x and that both functions are non zero on a region the width of which is less than $\mathrm{m}(\sigma)$; hence

$$\int_\Omega v^2(x)dx \leq \Big(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) D_\sigma v \Big)^2. \tag{9.71}$$

Applying the inequality (9.71) to $v = |u|^\alpha \mathrm{sign}(u)$, where $u \in X(\mathcal{T})$ and $\alpha > 1$ yields

$$\int_\Omega |u(x)|^{2\alpha} dx \leq \Big(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) D_\sigma v \Big)^2.$$

Now, since $|v_K - v_L| \leq \alpha(|u_K|^{\alpha-1} + |u_L|^{\alpha-1})|u_K - u_L|$, if $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$ and $|v_K| \leq \alpha(|u_K|^{\alpha-1})|u_K|$, if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$,

$$\Big(\int_\Omega |u(x)|^{2\alpha} dx \Big)^{\frac{1}{2}} \leq \alpha \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma) |u_K|^{\alpha-1} D_\sigma u.$$

Using Hölder's inequality with $p, p' \in \mathbb{R}_+$ such that $\frac{1}{p} + \frac{1}{p'} = 1$ yields that

$$\Big(\int_\Omega |u(x)|^{2\alpha} dx \Big)^{\frac{1}{2}} \leq \alpha \Big(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} |u_K|^{p(\alpha-1)} \mathrm{m}(\sigma) d_{K,\sigma} \Big)^{\frac{1}{p}} \Big(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \frac{|D_\sigma u|^{p'}}{{d_{K,\sigma}}^{p'}} \mathrm{m}(\sigma) d_{K,\sigma} \Big)^{\frac{1}{p'}}.$$

Since $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)d_{K,\sigma} = 2\mathrm{m}(K)$, this gives

$$\Big(\int_\Omega |u(x)|^{2\alpha}dx\Big)^{\frac{1}{2}} \le \alpha 2^{\frac{1}{p}}\big(\int_\Omega |u(x)|^{p(\alpha-1)}dx\big)^{\frac{1}{p}}\big(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{|D_\sigma u|^{p'}}{{d_{K,\sigma}}^{p'}}\mathrm{m}(\sigma)d_{K,\sigma}\big)^{\frac{1}{p'}},$$

which yields, choosing p such that $p(\alpha-1)=2\alpha$, i.e. $p=\frac{2\alpha}{\alpha-1}$ and $p'=\frac{2\alpha}{\alpha+1}$,

$$\|u\|_{L^q(\Omega)} = \Big(\int_\Omega |u(x)|^{2\alpha}dx\Big)^{\frac{1}{2\alpha}} \le \alpha 2^{\frac{1}{p}}\big(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{|D_\sigma u|^{p'}}{{d_{K,\sigma}}^{p'}}\mathrm{m}(\sigma)d_{K,\sigma}\big)^{\frac{1}{p'}}, \tag{9.72}$$

where $q=2\alpha$. Let $r=\frac{2}{p'}$ and $r'=\frac{2}{2-p'}$, Hölder's inequality yields

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{|D_\sigma u|^{p'}}{{d_{K,\sigma}}^{p'}}\mathrm{m}(\sigma)d_{K,\sigma} \le \big(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{|D_\sigma u|^{2}}{{d_{K,\sigma}}^{2}}\mathrm{m}(\sigma)d_{K,\sigma}\big)^{\frac{p'}{2}}\big(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)d_{K,\sigma}\big)^{\frac{1}{r'}},$$

replacing in (9.72) gives

$$\|u\|_{L^q(\Omega)} \le \alpha 2^{\frac{1}{p}}\big(\frac{2}{\zeta}\big)^{\frac{1}{2}}(2\mathrm{m}(\Omega))^{\frac{1}{p'r'}}\|u\|_{1,\mathcal{T}}$$

and then (9.69) with, for instance, $C=(\frac{2}{\zeta})^{\frac{1}{2}}((2\mathrm{m}(\Omega))^{\frac{1}{2}}+1)$.

Let us now prove the three-dimensional case. Let $d=3$. Using the same notations as in the two-dimensional case, let $\boldsymbol{d}_1=(1,0,0)^t$, $\boldsymbol{d}_2=(0,1,0)^t$ and $\boldsymbol{d}_3=(0,0,1)^t$; for $x\in\Omega$, let $\mathcal{D}_x^1$, $\mathcal{D}_x^2$ and $\mathcal{D}_x^3$ be the straight lines going through x and defined by the vectors $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ and $\boldsymbol{d}_3$. Let us again define the functions $\chi_\sigma^{(1)}$, $\chi_\sigma^{(2)}$ and $\chi_\sigma^{(3)}$ by

$$\chi_\sigma^{(i)}(x) = \begin{cases} 1 & \text{if } \sigma\cap\mathcal{D}_x^i \neq \emptyset \\ 0 & \text{if } \sigma\cap\mathcal{D}_x^i = \emptyset \end{cases} \quad \text{for } i=1,2,3.$$

Let $v\in X(\mathcal{T})$ and let $A\in\mathbb{R}_+$ such that $\Omega\subset[-A,A]^3$; we also denote by v the function defined on $[-A,A]^3$ which equals v on Ω and 0 on $[-A,A]^3\setminus\Omega$. By the Cauchy-Schwarz inequality, one has:

$$\begin{aligned}&\int_{-A}^A\int_{-A}^A |v(x_1,x_2,x_3)|^{\frac{3}{2}}dx_1dx_2 \\ &\le \Big(\int_{-A}^A\int_{-A}^A |v(x_1,x_2,x_3)|dx_1dx_2\Big)^{\frac{1}{2}}\Big(\int_{-A}^A\int_{-A}^A |v(x_1,x_2,x_3)|^2dx_1dx_2\Big)^{\frac{1}{2}}.\end{aligned} \tag{9.73}$$

Now remark that

$$\int_{-A}^A\int_{-A}^A |v(x_1,x_2,x_3)|dx_1dx_2 \le \sum_{\sigma\in\mathcal{E}} D_\sigma v\int_{-A}^A\int_{-A}^A \chi_\sigma^{(3)}(x)dx_1dx_2 \le \sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)D_\sigma v.$$

Moreover, computations which were already performed in the two-dimensional case give that

$$\int_{-A}^A\int_{-A}^A |v(x_1,x_2,x_3)|^2dx_1dx_2 \le \int_{-A}^A\int_{-A}^A \sum_{\sigma\in\mathcal{E}} D_\sigma v\chi_\sigma^{(1)}(x)\sum_{\sigma\in\mathcal{E}} D_\sigma v\chi_\sigma^{(2)}(x)dx_1dx_2 \le \Big(\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma_{x_3})D_\sigma v\Big)^2,$$

where σ_{x_3} denotes the intersection of σ with the plane which contains the point $(0,0,x_3)$ and is orthogonal to $\boldsymbol{d}_3$. Therefore, integrating (9.73) in the third direction yields:

$$\int_\Omega |v(x)|^{\frac{3}{2}}dx \le \Big(\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)D_\sigma v\Big)^{\frac{3}{2}}. \tag{9.74}$$

Now let $v = |u|^4\mathrm{sign}(u)$, since $|v_K - v_L| \le 4(|u_K|^3 + |u_L|^3)|u_K - u_L|$, Inequality (9.74) yields:

$$\int_\Omega |u(x)|^6 dx \le \Big[4 \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} |u_K|^3 D_\sigma u \mathrm{m}(\sigma)\Big]^{\frac{3}{2}}.$$

By Cauchy-Schwarz' inequality and since $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma) d_{K,\sigma} = 3\mathrm{m}(K)$, this yields

$$\|u\|_{L^6} \le 4\sqrt{3} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} (D_\sigma u)^2 \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}},$$

and since $d_{K,\sigma} \ge \zeta d_\sigma$, this yields (9.69) with, for instance, $C = \frac{4\sqrt{3}}{\sqrt{\zeta}}$. ∎

Remark 9.14 (Discrete Poincaré Inequality) In the above proof, Inequality (9.71) leads to another proof of some discrete Poincaré inequality (as in Lemma 9.1 page 40) in the two-dimensional case. Indeed, let Ω be an open bounded polygonal subset of $\mathbb{R}^2$. Let $\mathcal{T}$ be an admissible finite volume mesh of Ω in the sense of Definition 9.1 page 37 (but more general meshes are possible). Let $v \in X(\mathcal{T})$. Then, (9.71), the Cauchy-Schwarz inequality and the fact that $\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma = 2\,\mathrm{m}(\Omega)$ yield

$$\|v\|^2_{L^2(\Omega)} \le 2\,\mathrm{m}(\Omega) \|v\|^2_{1,\mathcal{T}}.$$

A similar result holds in the three-dimensional case.

Corollary 9.1 (Error estimate) *Under the same assumptions and with the same notations as in Theorem 9.3 page 52, or as in Theorem 9.4 page 55, and assuming that the mesh satisfies, for some $\zeta > 0$, $d_{K,\sigma} \ge \zeta d_\sigma$, for all $\sigma \in \mathcal{E}_K$ and for all control volume K, there exists $C > 0$ only depending on u, ζ and Ω such that*

$$\|e_\mathcal{T}\|_{L^q(\Omega)} \le C q \mathrm{size}(\mathcal{T}); \text{ for any } q \in \begin{cases} [1,6] & \text{if } d = 3, \\ [1,+\infty) & \text{if } d = 2; \end{cases} \tag{9.75}$$

furthermore, there exists $C \in \mathbb{R}_+$ only depending on u, ζ, $\zeta_\mathcal{T} = \min\{\frac{\mathrm{m}(K)}{\mathrm{size}(\mathcal{T})^d}, K \in \mathcal{T}\}$, and Ω, such that

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} \le C \mathrm{size}(\mathcal{T})(|\ln(\mathrm{size}(\mathcal{T}))| + 1), \qquad \text{if } d = 2. \tag{9.76a}$$

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} \le C \mathrm{size}(\mathcal{T})^{2/3}, \qquad \text{if } d = 3. \tag{9.76b}$$

Proof of Corollary 9.1

Estimate (9.49) of Theorem 9.3 (or Theorem 9.4) and Inequality (9.69) of Lemma 9.5 immediately yield Estimate (9.75) in the case $d = 2$. Let us now prove (9.76). Remark that

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} = \max\{|e_K|, K \in \mathcal{T}\} \le \Big(\frac{1}{\zeta_\mathcal{T} \mathrm{size}(\mathcal{T})^2}\Big)^{\frac{1}{q}} \|e_\mathcal{T}\|_{L^q}. \tag{9.77}$$

For $d = 2$, a study of the real function defined, for $q \ge 2$, by $q \mapsto \ln q + (1 - \frac{2}{q}) \ln h$ (with $h = \mathrm{size}(\mathcal{T})$) shows that its minimum is attained for $q = -2 \ln h$, if $\ln h \le -\frac{1}{2}$. Therefore (9.75) and (9.77) yield(9.76). The 3 dimensional case is an immediate consequence of(9.75) with $q = 6$. ∎

10 Neumann boundary conditions

This section is devoted to the proof of convergence of the finite volume scheme when Neumann boundary conditions are imposed. The discretization of a general convection-diffusion equation with Dirichlet, Neumann and Fourier boundary conditions is considered in section 11 below, and the convection term is largely studied in the previous section. Hence we shall limit here the presentation to the pure diffusion operator. Consider the following elliptic problem:

$$-\Delta u(x) = f(x),\ x \in \Omega, \tag{10.1}$$

with Neumann boundary conditions:

$$\nabla u(x) \cdot \boldsymbol{n}(x) = g(x),\ x \in \partial\Omega, \tag{10.2}$$

where $\partial\Omega$ denotes the boundary of Ω and $\boldsymbol{n}$ its unit normal vector outward to Ω.
The following assumptions are made on the data:

Assumption 10.1

1. *Ω is an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3,*
2. *$g \in L^2(\partial\Omega)$, $f \in L^2(\Omega)$ and $\int_{\partial\Omega} g(x)d\gamma(x) + \int_\Omega f(x)dx = 0$.*

Under Assumption 10.1, Problem (10.1), (10.2) has a unique (variational) solution, u, belonging to $H^1(\Omega)$ and such that $\int_\Omega u(x)dx = 0$. It is the unique solution of the following problem:

$$u \in H^1(\Omega),\ \int_\Omega u(x)dx = 0, \tag{10.3}$$

$$\int_\Omega \nabla u(x)\nabla\psi(x) = \int_\Omega f(x)\psi(x)dx + \int_{\partial\Omega} g(x)\overline{\gamma}(\psi)(x)d\gamma(x),\ \forall\psi \in H^1(\Omega). \tag{10.4}$$

Recall that $\overline{\gamma}$ is the "trace" operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$ (or to $H^{\frac{1}{2}}(\partial\Omega)$).

10.1 Meshes and schemes

Admissible meshes

The definition of the scheme in the case of Neumann boundary conditions is easier, since the finite volume scheme naturally introduces the fluxes on the boundaries in its formulation. Hence the class of admissible meshes considered here is somewhat wider than the one considered in Definition 9.1 page 37, thanks to the Neumann boundary conditions and the absence of convection term.

Definition 10.1 (Admissible meshes) Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$, or 3. An admissible finite volume mesh of Ω for the discretization of Problem (10.1), (10.2), denoted by $\mathcal{T}$, is given by a family of "control volumes", which are open disjoint polygonal convex subsets of Ω, a family of subsets of $\overline{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$, denoted by $\mathcal{E}$ (these are the "sides" of the control volumes), with strictly positive $(d-1)$-dimensional Lebesgue measure, and a family of points of Ω denoted by $\mathcal{P}$ satisfying properties (i), (ii), (iii) and (iv) of Definition 9.1 page 37.

The same notations as in Definition 9.1 page 37 are used in the sequel.

One defines the set $X(\mathcal{T})$ of piecewise constant functions on the control volumes of an admissible mesh as in Definition 9.2 page 39.

Definition 10.2 (Discrete H^1 seminorm) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3, and $\mathcal{T}$ an admissible finite volume mesh in the sense of Definition 10.1.
For $u \in X(\mathcal{T})$, the discrete H^1 seminorm of u is defined by

$$|u|_{1,\mathcal{T}} = \Big(\sum_{\sigma \in \mathcal{E}_{\rm int}} \tau_\sigma (D_\sigma u)^2 \Big)^{\frac{1}{2}},$$

where $\tau_\sigma = \frac{\mathrm{m}(\sigma)}{d_\sigma}$ and $\mathcal{E}_{\rm int}$ are defined in Definition 9.1 page 37, u_K is the value of u in the control volume K and $D_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$.

The finite volume scheme

Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 . For $K \in \mathcal{T}$, let us define:

$$f_K = \frac{1}{\mathrm{m}(K)} \int_K f(x) dx, \tag{10.5}$$

$$\begin{aligned} &g_K = \frac{1}{\mathrm{m}(\partial K \cap \partial\Omega)} \int_{\partial K \cap \partial\Omega} g(x) d\gamma(x) \text{ if } \mathrm{m}(\partial K \cap \partial\Omega) \neq 0, \\ &g_K = 0 \text{ if } \mathrm{m}(\partial K \cap \partial\Omega) = 0. \end{aligned} \tag{10.6}$$

Recall that, in formula (10.5), $\mathrm{m}(K)$ denotes the d-dimensional Lebesgue measure of K, and, in (10.6), $\mathrm{m}(\partial K \cap \partial\Omega)$ denotes the $(d-1)$-dimensional Lebesgue measure of $\partial K \cap \partial\Omega$. Note that $g_K = 0$ if the dimension of $\partial K \cap \partial\Omega$ is less than $d-1$. Let $(u_K)_{K\in\mathcal{T}}$ denote the discrete unknowns; the numerical scheme is defined by (9.20)-(9.22) page 42, with $b = 0$ and $\mathbf{v} = 0$. This yields:

$$-\sum_{L \in \mathcal{N}(K)} \tau_{K|L} \Big(u_L - u_K \Big) = \mathrm{m}(K) f_K + \mathrm{m}(\partial K \cap \partial\Omega) g_K, \ \forall K \in \mathcal{T}, \tag{10.7}$$

(see the notations in Definitions 9.1 page 37 and 10.1 page 63). The condition (10.3) is discretized by:

$$\sum_{K \in \mathcal{T}} \mathrm{m}(K) u_K = 0. \tag{10.8}$$

Then, the approximate solution, $u_\mathcal{T}$, belongs to $X(\mathcal{T})$ (see Definition 9.2 page 39) and is defined by

$$u_\mathcal{T}(x) = u_K, \text{ for a.e. } x \in K, \ \forall K \in \mathcal{T}.$$

The following lemma gives existence and uniqueness of the solution of (10.7) and (10.8).

Lemma 10.1 *Under Assumption 10.1. let $\mathcal{T}$ be an admissible mesh (see Definition 10.1) and $\{f_K, K \in \mathcal{T}\}$, $\{g_K, K \in \mathcal{T}\}$ defined by (10.5), (10.6). Then, there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to (10.7)-(10.8).*

Proof of lemma 10.1
Let $N = \mathrm{card}(\mathcal{T})$. The equations (10.7) are a system of N equations with N unknowns, namely $(u_K)_{K\in\mathcal{T}}$. Ordering the unknowns (and the equations), this system can be written under a matrix form with a $N \times N$ matrix A. Using the connexity of Ω, the null space of this matrix is the set of "constant" vectors (that is $u_K = u_L$, for all $K, L \in \mathcal{T}$). Indeed, if $f_K = g_K = 0$ for all $K \in \mathcal{T}$ and $\{u_K, K \in \mathcal{T}\}$ is solution of (10.7), multiplying (10.7) (for $K \in \mathcal{T}$) by u_K and summing over $K \in \mathcal{T}$ yields

$$\sum_{\sigma \in \mathcal{E}_{\rm int}} \tau_\sigma (D_\sigma u)^2 = 0,$$

where $D_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$. This gives, thanks to the positivity of τ_σ and the connexity of Ω, $u_K = u_L$, for all $K, L \in \mathcal{T}$.

For general $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$, a necessary condition, in order that (10.7) has a solution, is that

$$\sum_{K\in\mathcal{T}} (\mathrm{m}(K)f_K + \mathrm{m}(\partial K \cap \partial\Omega)g_K) = 0. \tag{10.9}$$

Since the dimension of the null space of A is one, this condition is also a sufficient condition. Therefore, System (10.7) has a solution if and only if (10.9) holds, and this solution is unique up to an additive constant. Adding condition (10.8) yields uniqueness. Note that (10.9) holds thanks to the second item of Assumption 10.1; this concludes the proof of Lemma 10.1. ∎

10.2 Discrete Poincaré inequality

The proof of an error estimate, under a regularity assumption on the exact solution, and of a convergence result, in the general case (under Assumption 10.1), requires a "discrete Poincaré" inequality as in the case of the Dirichlet problem.

Lemma 10.2 (Discrete mean Poincaré inequality) *Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3. Then, there exists $C \in \mathbb{R}_+$, only depending on Ω, such that for all admissible meshes (in the sense of Definition 10.1 page 63), $\mathcal{T}$, and for all $u \in X(\mathcal{T})$ (see Definition 9.2 page 39), the following inequality holds:*

$$\|u\|^2_{L^2(\Omega)} \le C|u|^2_{1,\mathcal{T}} + 2(\mathrm{m}(\Omega))^{-1}(\int_\Omega u(x)dx)^2, \tag{10.10}$$

where $|\cdot|_{1,\mathcal{T}}$ is the discrete H^1 seminorm defined in Definition 10.2.

Proof of Lemma 10.2

The proof given here is a "direct proof"; another proof, by contradiction, is possible (see Remark 10.2). Let $\mathcal{T}$ be an admissible mesh and $u \in X(\mathcal{T})$. Let $m_\Omega(u)$ be the mean value of u over Ω, that is

$$m_\Omega(u) = \frac{1}{\mathrm{m}(\Omega)} \int_\Omega u(x)dx.$$

Since

$$\|u\|^2_{L^2(\Omega)} \le 2\|u - m_\Omega(u)\|^2_{L^2(\Omega)} + 2(m_\Omega(u))^2\mathrm{m}(\Omega),$$

proving Lemma 10.2 amounts to proving the existence of $D \ge 0$, only depending on Ω, such that

$$\|u - m_\Omega(u)\|^2_{L^2(\Omega)} \le D|u|^2_{1,\mathcal{T}}. \tag{10.11}$$

The proof of (10.11) may be decomposed into three steps (indeed, if Ω is convex, the first step is sufficient).

Step 1 (Estimate on a convex part of Ω)

Let ω be an open convex subset of Ω, $\omega \neq \emptyset$ and $m_\omega(u)$ be the mean value of u on ω. In this step, one proves that there exists C_0, depending only on Ω, such that

$$\|u(x) - m_\omega(u)\|^2_{L^2(\omega)} \le \frac{1}{\mathrm{m}(\omega)} C_0 |u|^2_{1,\mathcal{T}}. \tag{10.12}$$

(Taking $\omega = \Omega$, this proves (10.11) and Lemma 10.2 in the case where Ω is convex.)

Noting that

$$\int_\omega (u(x) - m_\omega(u))^2 dx \le \frac{1}{\mathrm{m}(\omega)} \int_\omega \big(\int_\omega (u(x) - u(y))^2 dy \big) dx,$$

(10.12) is proved provided that there exists $C_0 \in \mathbb{R}_+$, only depending on Ω, such that

$$\int_\omega \int_\omega (u(x)-u(y))^2 dxdy \le C_0 |u|^2_{1,\mathcal{T}}. \tag{10.13}$$

For $\sigma \in \mathcal{E}_{\rm int}$, let the function χ_σ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\{0,1\}$ be defined by

$$\chi_\sigma(x,y) = 1, \text{ if } x,y \in \overline{\Omega}, [x,y] \cap \sigma \neq \emptyset,$$
$$\chi_\sigma(x,y) = 0, \text{ if } x \notin \overline{\Omega} \text{ or } y \notin \overline{\Omega} \text{ or } [x,y] \cap \sigma = \emptyset.$$

(Recall that $[x,y] = \{tx + (1-t)y,\ t \in [0,1]\}$.) For a.e. $x, y \in \omega$, one has, with $D_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$,

$$(u(x)-u(y))^2 \le \Big(\sum_{\sigma \in \mathcal{E}_{\rm int}} |D_\sigma u| \chi_\sigma(x,y) \Big)^2,$$

(note that the convexity of ω is used here) which yields, thanks to the Cauchy-Schwarz inequality,

$$(u(x)-u(y))^2 \le \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_{\sigma,y-x}} \chi_\sigma(x,y) \sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_{\sigma,y-x} \chi_\sigma(x,y), \tag{10.14}$$

with

$$c_{\sigma,y-x} = \Big| \frac{y-x}{|y-x|} \cdot \mathbf{n}_\sigma \Big|,$$

recall that $\mathbf{n}_\sigma$ is a unit normal vector to σ, and that $x_K - x_L = \pm d_\sigma \mathbf{n}_\sigma$ if $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$. For a.e. $x, y \in \omega$, one has

$$\sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_{\sigma,y-x} \chi_\sigma(x,y) = \Big| (x_K - x_L) \cdot \frac{y-x}{|y-x|} \Big|,$$

for some convenient control volumes K and L, depending on x, y and σ (the convexity of ω is used again here). Therefore,

$$\sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_{\sigma,y-x} \chi_\sigma(x,y) \le \mathrm{diam}(\Omega).$$

Thus, integrating (10.14) with respect to x and y in ω,

$$\int_\omega \int_\omega (u(x)-u(y))^2 dxdy \le \mathrm{diam}(\Omega) \int_\omega \int_\omega \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_{\sigma,y-x}} \chi_\sigma(x,y) dxdy,$$

which gives, by a change of variables,

$$\int_\omega \int_\omega (u(x)-u(y))^2 dxdy \le \mathrm{diam}(\Omega) \int_{\mathbb{R}^d} \Big(\sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_{\sigma,z}} \int_\omega \chi_\sigma(x,x+z) dx \Big) dz. \tag{10.15}$$

Noting that, if $|z| > \mathrm{diam}(\Omega)$, $\chi_\sigma(x,x+z) = 0$, for a.e. $x \in \Omega$, and

$$\int_\Omega \chi_\sigma(x,x+z) dx \le \mathrm{m}(\sigma) |z \cdot \mathbf{n}_\sigma| = \mathrm{m}(\sigma) |z| c_{\sigma,z} \text{ for a.e. } z \in \mathbb{R}^d,$$

therefore, with (10.15):

$$\int_\omega \int_\omega (u(x)-u(y))^2 dxdy \le (\mathrm{diam}(\Omega))^2 \mathrm{m}(B_\Omega) \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{\mathrm{m}(\sigma) |D_\sigma u|^2}{d_\sigma},$$

where B_Ω denotes the ball of $\mathbb{R}^d$ of center 0 and radius $\mathrm{diam}(\Omega)$.

This inequality proves (10.13) and then (10.12) with $C_0 = (\mathrm{diam}(\Omega))^2 \mathrm{m}(B_\Omega)$ (which only depends on Ω). Taking $\omega = \Omega$, it concludes the proof of Lemma 10.2 in the case where Ω is convex.

Step 2 (Estimate with respect to the mean value on a part of the boundary)
In this step, one proves the same inequality than (10.12) but with the mean value of u on a (arbitrary) part I of the boundary of ω instead of $m_\omega(u)$ and with a convenient C_1 depending on I, Ω and ω instead of C_0.
More precisely, let ω be a polygonal open convex subset of Ω and let $I \subset \partial\omega$, with $m(I) > 0$ ($m(I)$ is the $(d-1)$-Lebesgue measure of I). Assume that I is included in a hyperplane of $\mathbb{R}^d$. Let $\overline{\gamma}(u)$ be the "trace" of u on the boundary of ω, that is $\overline{\gamma}(u)(x) = u_K$ if $x \in \partial\omega \cap \overline{K}$, for $K \in \mathcal{T}$. (If $x \in \overline{K} \cap \overline{L}$, the choice of $\overline{\gamma}(u)(x)$ between u_K and u_L does not matter). Let $m_I(u)$ be the mean value of $\overline{\gamma}(u)$ on I. This step is devoted to the proof that there exists C_1, only depending on Ω, ω and I, such that

$$\|u - m_I(u)\|^2_{L^2(\omega)} \leq C_1 |u|^2_{1,\mathcal{T}}. \tag{10.16}$$

For the sake of simplicity, only the case $d = 2$ is considered here. Since I is included in a hyperplane, it may be assumed, without loss of generality, that $I = \{0\} \times J$, with $J \subset \mathbb{R}$ and $\omega \subset \mathbb{R}_+ \times \mathbb{R}$ (one uses here the convexity of ω).
Let $\alpha = \max\{x_1,\ x = (x_1, x_2)^t \in \overline{\omega}\}$ and $a = (\alpha, \beta)^t \in \overline{\omega}$. In the following, a is fixed. For a.e. $x = (x_1, x_2)^t \in \omega$ and for a.e. (for the 1-Lebesgue measure) $y = (0, \overline{y})^t \in I$ (with $\overline{y} \in J$), one sets $z(x, y) = ta + (1-t)y$ with $t = x_1/\alpha$. Note that, thanks to the convexity of ω, $z(x, y) = (z_1, z_2)^t \in \overline{\omega}$, with $z_1 = x_1$. The following inequality holds:

$$\pm(u(x) - \overline{\gamma}(u)(y)) \leq |u(x) - u(z(x, y))| + |u(z(x, y) - \overline{\gamma}(u)(y))|.$$

In the following, the notation C_i, $i \in \mathbb{N}^\star$, will be used for quantities only depending on Ω, ω and I.
Let us integrate the above inequality over $y \in I$, take the power 2, from the Cauchy-Schwarz inequality, an integration over $x \in \omega$ leads to

$$\int_\omega (u(x) - m_I(u))^2 dx \leq \frac{2}{m(I)} \int_\omega \int_I (u(x) - u(z(x, y)))^2 d\gamma(y) dx$$
$$+ \frac{2}{m(I)} \int_\omega \int_I (u(z(x, y)) - u(y))^2 d\gamma(y) dx.$$

Then,

$$\int_\omega (u(x) - m_I(u))^2 dx \leq \frac{2}{m(I)}(A + B),$$

with, since ω is convex,

$$A = \int_\omega \int_I \Big(\sum_{\sigma \in \mathcal{E}_{\mathrm{int}}} |D_\sigma u| \chi_\sigma(x, z(x, y)) \Big)^2 d\gamma(y) dx,$$

and

$$B = \int_\omega \int_I \Big(\sum_{\sigma \in \mathcal{E}_{\mathrm{int}}} |D_\sigma u| \chi_\sigma(z(x, y), y) \Big)^2 d\gamma(y) dx.$$

Recall that, for $\xi, \eta \in \overline{\Omega}$, $\chi_\sigma(\xi, \eta) = 1$ if $[\xi, \eta] \cap \sigma \neq \emptyset$ and $\chi_\sigma(\xi, \eta) = 0$ if $[\xi, \eta] \cap \sigma = \emptyset$. Let us now look for some bounds of A and B of the form $C|u|^2_{1,\mathcal{T}}$.
The bound for A is easy. Using the Cauchy-Schwarz inequality and the fact that

$$\sum_{\sigma \in \mathcal{E}_{\mathrm{int}}} c_{\sigma, x - z(x,y)} d_\sigma \chi_\sigma(x, z(x, y)) \leq \mathrm{diam}(\Omega)$$

(recall that $c_{\sigma,\eta} = |\frac{\eta}{|\eta|} \cdot \mathbf{n}_\sigma|$ (for $\eta \in \mathbb{R}^2 \setminus 0$) gives

$$A \leq C_2 \int_\omega \int_I \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2 \chi_\sigma(x, z(x,y))}{c_{\sigma,x-z(x,y)} d_\sigma} dx d\gamma(y).$$

Since $z_1 = x_1$, one has $c_{\sigma,x-z(x,y)} = c_{\sigma,e}$, with $e = (0,1)^t$. Let us perform the integration of the right hand side of the previous inequality, with respect to the first component of x, denoted by x_1, first. The result of the integration with respect to x_1 is bounded by $|u|^2_{1,\mathcal{T}}$. Then, integrating with respect to x_2 and $y \in I$ gives $A \leq C_3 |u|^2_{1,\mathcal{T}}$.
In order to obtain a bound B, one remarks, as for A, that

$$B \leq C_4 \int_\omega \int_I \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2 \chi_\sigma(z(x,y), y)}{c_{\sigma,y-z(x,y)} d_\sigma} dx d\gamma(y).$$

In the right hand side of this inequality, the integration with respect to $y \in I$ is transformed into an integration with respect to $\xi = (\xi_1, \xi_2)^t \in \sigma$, this yields (note that $c_{\sigma,y-z(x,y)} = c_{\sigma,a-y}$)

$$B \leq C_4 \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma} \int_\omega \int_\sigma \frac{\psi_\sigma(x,\xi)}{c_{I,a-y(\xi)}} \frac{|a-y(\xi)|}{|a-\xi|} dx d\gamma(\xi),$$

where $y(\xi) = s\xi + (1-s)a$, with $s\xi_1 + (1-s)\alpha = 0$, and where ψ_σ is defined by

$$\psi_\sigma(x,\xi) = 1, \text{ if } y(\xi) \in I \text{ and } \xi_1 \leq x_1$$
$$\psi_\sigma(x,\xi) = 0, \text{ if } y(\xi) \notin I \text{ or } \xi_1 > x_1.$$

Noting that $c_{I,a-y(\xi)} \geq C_5 > 0$, one deduces that

$$B \leq C_6 \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma} \int_\sigma \Big(\int_\omega \psi_\sigma(x,\xi) \frac{|a-y(\xi)|}{|a-\xi|} dx \Big) d\gamma(\xi) \leq C_7 |u|^2_{1,\mathcal{T}},$$

with, for instance, $C_7 = C_6 (\mathrm{diam}(\omega))^2$. The bounds on A and B yield (10.16).
Step 3 (proof of (10.11))
Let us now prove that there exists $D \in \mathbb{R}_+$, only depending on Ω such that (10.11) hold. Since Ω is a polygonal set ($d = 2$ or 3), there exists a finite number of disjoint convex polygonal sets, denoted by $\{\Omega_1, \ldots, \Omega_n\}$, such that $\overline{\Omega} = \cup_{i=1}^n \overline{\Omega_i}$. Let $I_{i,j} = \overline{\Omega_i} \cap \overline{\Omega_j}$, and B be the set of couples $(i,j) \in \{1,\ldots,n\}^2$ such that $i \neq j$ and the $(d-1)$-dimensional Lebesgue measure of $I_{i,j}$, denoted by $m(I_{i,j})$, is positive.
Let m_i denote the mean value of u on Ω_i, $i \in \{1,\ldots,n\}$, and $m_{i,j}$ denote the mean value of u on $I_{i,j}$, $(i,j) \in B$. (For $\sigma \in \mathcal{E}_{\rm int}$, in order that u be defined on σ, a.e. for the $(d-1)$-dimensional Lebesgue measure, let $K \in \mathcal{T}$ be a control volume such that $\sigma \in \mathcal{E}_K$, one sets $u = u_K$ on σ.) Note that $m_{i,j} = m_{j,i}$ for all $(i,j) \in B$.

Step 1 gives the existence of C_i, $i \in \{1,\ldots,n\}$, only depending on Ω (since the Ω_i only depend on Ω), such that

$$\|u - m_i\|^2_{L^2(\Omega_i)} \leq C_i |u|^2_{1,\mathcal{T}}, \ \forall i \in \{1,\ldots,n\}, \tag{10.17}$$

Step 2 gives the existence of $C_{i,j}$, $i,j \in B$, only depending on Ω, such that

$$\|u - m_{i,j}\|^2_{L^2(\Omega_i)} \leq C_{i,j} |u|^2_{1,\mathcal{T}}, \ \forall (i,j) \in B.$$

Then, one has $(m_i - m_{i,j})^2 \mathrm{m}(\Omega_i) \leq 2(C_i + C_{i,j}) |u|^2_{1,\mathcal{T}}$, for all $(i,j) \in B$. Since Ω is connected, the above inequality yields the existence of M, only depending on Ω, such that $|m_i - m_j| \leq M |u|_{1,\mathcal{T}}$ for all $(i,j) \in \{1,\ldots,n\}^2$, and therefore $|m_\Omega(u) - m_i| \leq M |u|_{1,\mathcal{T}}$ for all $i \in \{1,\ldots,n\}$. Then, (10.17) yields the existence of D, only depending on Ω, such that (10.11) holds. This completes the proof of Lemma 10.2. ∎

An easy consequence of the proof of Lemma 10.2 is the following lemma. Although this lemma is not used in the sequel, it is interesting in its own sake.

Lemma 10.3 (Mean boundary Poincaré inequality) *Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3. Let $I \subset \partial\Omega$ such that the $(d-1)$- Lebesgue measure of I is positive. Then, there exists $C \in \mathbb{R}_+$, only depending on Ω and I, such that for all admissible mesh (in the sense of Definition 10.1 page 63) $\mathcal{T}$ and for all $u \in X(\mathcal{T})$ (see Definition 9.2 page 39), the following inequality holds:*

$$\|u - m_I(u)\|^2_{L^2(\Omega)} \leq C|u|^2_{1,\mathcal{T}}$$

where $|\cdot|_{1,\mathcal{T}}$ is the discrete H^1 seminorm defined in Definition 10.2 and $m_I(u)$ is the mean value of $\overline{\gamma}(u)$ on I with $\overline{\gamma}(u)$ defined a.e. on $\partial\Omega$ by $\overline{\gamma}(u)(x) = u_K$ if $x \in \sigma$, $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$, $K \in \mathcal{T}$.

Note that this last lemma also gives as a by-product a discrete Poincaré inequality in the case of a Dirichlet boundary condition on a part of the boundary if the domain is assumed to be connex, see Remark 9.4.
Finally, let us point out that a continuous version of lemmata 10.2 (known as the Poincaré-Wirtinger inequality) and 10.3 holds and that the proof is similar and rather easier. Let us state this continuous version which can be proved by contradiction or with a technique similar to Lemma 9.4 page 49. The advantage of the latter is that it gives a more explicit bound.

Lemma 10.4 *Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3. Let $I \subset \partial\Omega$ such that the $(d-1)$- Lebesgue measure of I is positive.*
Then, there exists $C \in \mathbb{R}_+$, only depending on Ω, and $\tilde{C} \in \mathbb{R}_+$, only depending on Ω and I, such that, for all $u \in H^1(\Omega)$, the following inequalities hold:

$$\|u\|^2_{L^2(\Omega)} \leq C|u|^2_{H^1(\Omega)} + 2(\mathrm{m}(\Omega))^{-1}(\int_\Omega u(x)dx)^2$$

and

$$\|u - m_I(u)\|^2_{L^2(\Omega)} \leq \tilde{C}|u|^2_{H^1(\Omega)},$$

where $|\cdot|_{H^1(\Omega)}$ is the H^1 seminorm defined by $|v|^2_{H^1(\Omega)} = \|\nabla u\|^2_{(L^2(\Omega))^d} = \int_\Omega |\nabla v(x)|^2 dx$ for all $v \in H^1(\Omega)$, and $m_I(u)$ is the mean value of $\overline{\gamma}(u)$ on I. Recall that $\overline{\gamma}$ is the trace operator from $H^1(\Omega)$ to $H^{1/2}(\partial\Omega)$.

10.3 Error estimate

Under Assumption 10.1, let $\mathcal{T}$ be an admissible mesh (see Definition 10.1) and $\{f_K,\ K \in \mathcal{T}\}$, $\{g_K,\ K \in \mathcal{T}\}$ defined by (10.5), (10.6). By Lemma 10.1, there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to (10.7)-(10.8). Under an additional regularity assumption on the exact solution, the following error estimate holds:

Theorem 10.1 *Under Assumption 10.1 page 63, let $\mathcal{T}$ be an admissible mesh (see Definition 10.1 page 63) and $h = \mathrm{size}(\mathcal{T})$. Let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (10.7) and (10.8) (thanks to (10.5) and (10.6), existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 10.1). Let $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2 page 39) be defined by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Assume that the unique solution, u, to Problem (10.3), (10.4) satisfies $u \in C^2(\overline{\Omega})$.*
Then there exists $C \in \mathbb{R}_+$ which only depends on u and Ω such that

$$\|u_\mathcal{T} - u\|_{L^2(\Omega)} \leq Ch, \tag{10.18}$$

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)d_\sigma(\frac{u_L - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x))^2 \leq Ch^2. \tag{10.19}$$

Recall that, in the above theorem, $K|L$ denotes the element σ of $\mathcal{E}_{\rm int}$ such that $\overline{\sigma} = \partial K \cap \partial L$, with K, $L \in \mathcal{T}$.

Proof of Theorem 10.1

Let $C_{\mathcal{T}} \in \mathbb{R}$ be such that

$$\sum_{K\in\mathcal{T}} \overline{u}(x_K)\mathrm{m}(K) = 0,$$

where $\overline{u} = u + C_{\mathcal{T}}$.

Let, for each $K \in \mathcal{T}$, $e_K = \overline{u}(x_K) - u_K$, and $e_{\mathcal{T}} \in X(\mathcal{T})$ defined by $e_{\mathcal{T}}(x) = e_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Let us first prove the existence of C only depending on u and Ω such that

$$|e_{\mathcal{T}}|_{1,\mathcal{T}} \le Ch \quad \text{and} \quad \|e_{\mathcal{T}}\|_{L^2(\Omega)} \le Ch. \tag{10.20}$$

Integrating (10.1) page 63 over $K \in \mathcal{T}$, and taking (10.2) page 63 into account yields:

$$\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) = \int_K f(x)dx + \int_{\partial K\cap\partial\Omega} g(x)d\gamma(x). \tag{10.21}$$

For $\sigma \in \mathcal{E}_{\rm int}$ such that $\sigma = K|L$, let us define the consistency error on the flux from K through σ by:

$$R_{K,\sigma} = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) - \frac{u(x_L) - u(x_K)}{d_\sigma}. \tag{10.22}$$

Note that the definition of $R_{K,\sigma}$ remains with $\overline{u}$ instead of u in (10.22).

Thanks to the regularity of the solution u, there exists $C_1 \in \mathbb{R}_+$, only depending on u, such that $|R_{K,L}| \le C_1 h$. Using (10.21), (10.22) and (10.7) yields

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(e_L - e_K)^2 \le d\mathrm{m}(\Omega)(C_1 h)^2,$$

which gives the first part of (10.20).

Thanks to the discrete Poincaré inequality (10.10) applied to the function $e_{\mathcal{T}}$, and since

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)e_K = 0$$

(which is the reason why $e_{\mathcal{T}}$ was defined with $\overline{u}$ instead of u) one obtains the second part of (10.20), that is the existence of C_2 only depending on u and Ω such that

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)(e_K)^2 \le C_2 h^2.$$

From (10.20), one deduces (10.18) from the fact that $u \in C^1(\overline{\Omega})$. Indeed, let C_2 be the maximum value of $|\nabla u|$ in Ω. One has $|u(x) - u(y)| \le C_2 h$, for all x, $y \in K$, for all $K \in \mathcal{T}$. Then, from $\int_\Omega u(x)dx = 0$, one deduces $C_{\mathcal{T}} \le C_2 h$. Furthermore, one has

$$\sum_{K\in\mathcal{T}} \int_K (u(x_K) - u(x))^2 dx \le \sum_{K\in\mathcal{T}} \mathrm{m}(K)(C_2 h)^2 = \mathrm{m}(\Omega)(C_2 h)^2.$$

Then, noting that

$$\|u_{\mathcal{T}} - u\|^2_{L^2(\Omega)} = \sum_{K\in\mathcal{T}} \int_K (u_K - u(x))^2 dx$$
$$\le 3\sum_{K\in\mathcal{T}} \mathrm{m}(K)(e_K)^2 + 3(C_{\mathcal{T}})^2\mathrm{m}(\Omega) + 3\sum_{K\in\mathcal{T}} \int_K (u(x_K) - u(x))^2 dx$$

yields (10.18).

The proof of Estimate (10.19) is exactly the same as in the Dirichlet case. This property will be useful in the study of the convergence of finite volume methods in the case of a system consisting of an elliptic equation and a hyperbolic equation (see Section 37.6). ∎

As for the Dirichlet problem, the hypothesis $u \in C^2(\overline{\Omega})$ is not necessary to obtain error estimates. Assuming an additional assumption on the mesh (see Definition 10.3), Estimates (10.20) and (10.19) hold under the weaker assumption $u \in H^2(\Omega)$ (see Theorem 10.2 below). It is therefore also possible to obtain (10.18) under the additional assumption that u is Lipschitz continuous.

Definition 10.3 (Neumann restricted admissible meshes) Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3. A restricted admissible mesh for the Neumann problem, denoted by $\mathcal{T}$, is an admissible mesh in the sense of Definition 10.1 such that, for some $\zeta > 0$, one has $d_{K,\sigma} \geq \zeta \mathrm{diam}(K)$ for all control volume K and for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{int}}$.

Theorem 10.2 (H^2 regularity, Neumann problem) *Under Assumption 10.1 page 63, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.3 and $h = \mathrm{size}(\mathcal{T})$. Let $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2 page 39) be the approximated solution defined in Ω by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$, where $(u_K)_{K\in\mathcal{T}}$ is the (unique) solution to (10.7) and (10.8) (thanks to (10.5) and (10.6), existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 10.1). Assume that the unique solution, u, of (10.3), (10.4) belongs to $H^2(\Omega)$. Let $C_\mathcal{T} \in \mathbb{R}$ be such that*

$$\sum_{K\in\mathcal{T}} \overline{u}(x_K)\mathrm{m}(K) = 0 \text{ where } \overline{u} = u + C_\mathcal{T}.$$

Let, for each control volume $K \in \mathcal{T}$, $e_K = \overline{u}(x_K) - u_K$, and $e_\mathcal{T} \in X(\mathcal{T})$ defined by $e_\mathcal{T}(x) = e_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$.
Then there exists C, only depending on u, ζ and Ω, such that (10.20) and (10.19) hold.

Note that, in Theorem 10.2, the function $e_\mathcal{T}$ is well defined, and the quantity "$\nabla u \cdot \mathbf{n}_\sigma$" is well defined on σ, for all $\sigma \in \mathcal{E}$ (see Remark 9.12).

Proof of Theorem 10.2

The proof is very similar to that of Theorem 9.4 page 55, from which the same notations are used.
There exists some C, depending only on the space dimension (d) and ζ (given in Definition 10.3), such that, for all $\sigma \in \mathcal{E}_{\mathrm{int}}$,

$$|R_\sigma|^2 \leq C\frac{h^2}{\mathrm{m}(\sigma)d_\sigma}\int_{\mathcal{V}_\sigma} |(H(u)(z)|^2 dz, \tag{10.23}$$

and therefore

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \mathrm{m}(\sigma)d_\sigma R_\sigma^2 \leq Ch^2\int_\Omega |H(u)(z)|^2 dz. \tag{10.24}$$

The proof of (10.23) (from which (10.24) is an easy consequence) was already done in the proof of Theorem 9.4 (note that, here, there is no need to consider the case of $\sigma \in \mathcal{E}_{\mathrm{ext}}$). In order to obtain Estimate (10.20), one proceeds as in Theorem 9.4. Recall

$$|e_\mathcal{T}|_{1,\mathcal{T}}^2 \leq \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} R_\sigma |D_\sigma e|\mathrm{m}(\sigma),$$

where $|D_\sigma e| = |e_K - e_L|$ if $\sigma \in \mathcal{E}_{\mathrm{int}}$ is such that $\sigma = K|L$; hence, from the Cauchy-Schwarz inequality, one obtains that

$$|e_{\mathcal{T}}|_{1,\mathcal{T}}^2 \leq \big(\sum_{\sigma\in\mathcal{E}_{\rm int}} R_\sigma^2 {\rm m}(\sigma) d_\sigma\big)^{\frac12} \big(\sum_{\sigma\in\mathcal{E}_{\rm int}} |D_\sigma e|^2 \frac{{\rm m}(\sigma)}{d_\sigma}\big)^{\frac12}.$$

Then, one obtains, with (10.24),

$$|e_{\mathcal{T}}|_{1,\mathcal{T}} \leq \sqrt{C} h \big(\int_\Omega |H(u)(z)|^2 dz\big)^{\frac12}.$$

This concludes the proof of the first part of (10.20). The second part of (10.20) is a consequence of the discrete Poincaré inequality (10.10). Using (10.24) also easily leads (10.19).
Note also that, if u is Lipschitz continuous, Inequality (10.18) follows from the second part of (10.20) and the definition of $\overline{u}$ as in Theorem 10.1.
This concludes the proof of Theorem 10.2. ∎

Some generalizations of Theorem 10.2 are possible, as for the Dirichlet case, see Remark 9.13 page 59.

10.4 Convergence

A convergence result, under Assumption 10.1, may be proved without any regularity assumption on the exact solution.
The proof of convergence uses the following preliminary inequality on the "trace" of an element of $X(\mathcal{T})$ on the boundary:

Lemma 10.5 (Trace inequality) *Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^d$, $d = 2$ or 3 (indeed, the connexity of Ω is not used in this lemma). Let $\mathcal{T}$ be an admissible mesh, in the sense of Definition 10.1 page 63, and $u \in X(\mathcal{T})$ (see Definition 9.2 page 39). Let u_K be the value of u in the control volume K. Let $\overline{\gamma}(u)$ be defined by $\overline{\gamma}(u) = u_K$ a.e. (for the $(d-1)$-dimensional Lebesgue measure) on σ, if $\sigma \in \mathcal{E}_{\rm ext}$ and $\sigma \in \mathcal{E}_K$. Then, there exists C, only depending on Ω, such that*

$$\|\overline{\gamma}(u)\|_{L^2(\partial\Omega)} \leq C(|u|_{1,\mathcal{T}} + \|u\|_{L^2(\Omega)}). \tag{10.25}$$

Remark 10.1 The result stated in this lemma still holds if Ω is not assumed connected. Indeed, one needs only modify (in an obvious way) the definition of admissible meshes (Definition 10.1 page 63) so as to take into account non connected subsets.

Proof of Lemma 10.5

By compactness of the boundary of $\partial\Omega$, there exists a finite number of open hyper-rectangles ($d = 2$ or 3), $\{R_i, i = 1, \ldots, N\}$, and normalized vectors of $\mathbb{R}^d$, $\{\eta_i, i = 1, \ldots, N\}$, such that

$$\begin{cases} \partial\Omega \subset \cup_{i=1}^N R_i, \\ \eta_i \cdot \mathbf{n}(x) \geq \alpha > 0 \text{ for all } x \in R_i \cap \partial\Omega, i \in \{1, \ldots, N\}, \\ \{x + t\eta_i, x \in R_i \cap \partial\Omega, t \in \mathbb{R}_+\} \cap R_i \subset \Omega, \end{cases}$$

where α is some positive number and $\mathbf{n}(x)$ is the normal vector to $\partial\Omega$ at x, inward to Ω. Let $\{\alpha_i, i = 1, \ldots, N\}$ be a family of functions such that $\sum_{i=1}^N \alpha_i(x) = 1$, for all $x \in \partial\Omega$, $\alpha_i \in C_c^\infty(\mathbb{R}^d, \mathbb{R}_+)$ and $\alpha_i = 0$ outside of R_i, for all $i = 1, \ldots, N$. Let $\Gamma_i = R_i \cap \partial\Omega$; let us prove that there exists C_i only depending on α and α_i such that

$$\|\alpha_i \overline{\gamma}(u)\|_{L^2(\Gamma_i)} \leq C_i\big(|u|_{1,\mathcal{T}} + \|u\|_{L^2(\Omega)}\big). \tag{10.26}$$

The existence of C, only depending on Ω, such that (10.25) holds, follows easily (taking $C = \sum_{i=1}^N C_i$, and using $\sum_{i=1}^N \alpha_i(x) = 1$, note that α and α_i depend only on Ω). It remains to prove (10.26).

Let us introduce some notations. For $\sigma \in \mathcal{E}$ and $K \in \mathcal{T}$, define χ_σ and χ_K from $\mathbb{R}^d \times \mathbb{R}^d$ to $\{0,1\}$ by $\chi_\sigma(x,y) = 1$, if $[x,y] \cap \sigma \neq \emptyset$, $\chi_\sigma(x,y) = 0$, if $[x,y] \cap \sigma = \emptyset$, and $\chi_K(x,y) = 1$, if $[x,y] \cap K \neq \emptyset$, $\chi_K(x,y) = 0$, if $[x,y] \cap K = \emptyset$.

Let $i \in \{1,\dots,N\}$ and let $x \in \Gamma_i$. There exists a unique $t > 0$ such that $x + t\eta_i \in \partial R_i$, let $y(x) = x + t\eta_i$. For $\sigma \in \mathcal{E}$, let $z_\sigma(x) = [x,y(x)] \cap \sigma$ if $[x,y(x)] \cap \sigma \neq \emptyset$ and is reduced to one point. For $K \in \mathcal{T}$, let $\xi_K(x), \eta_K(x)$ be such that $[x,y(x)] \cap K = [\xi_K(x), \eta_K(x)]$ if $[x,y(x)] \cap K \neq \emptyset$.
One has, for a.e. (for the $(d-1)$-dimensional Lebesgue measure) $x \in \Gamma_i$,

$$|\alpha_i \overline{\gamma}(u)(x)| \leq \sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} |\alpha_i(z_\sigma(x))(u_K - u_L)|\chi_\sigma(x,y(x)) + \sum_{K \in \mathcal{T}} |(\alpha_i(\xi_K(x) - \alpha_i(\eta_K(x))u_K|\chi_K(x,y(x)),$$

that is,

$$|\alpha_i \overline{\gamma}(u)(x)|^2 \leq A(x) + B(x) \tag{10.27}$$

with

$$A(x) = 2(\sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} |\alpha_i(z_\sigma(x))(u_K - u_L)|\chi_\sigma(x,y(x)))^2,$$

$$B(x) = 2(\sum_{K \in \mathcal{T}} |(\alpha_i(\xi_K(x)) - \alpha_i(\eta_K(x)))u_K|\chi_K(x,y(x)))^2.$$

A bound on $A(x)$ is obtained for a.e. $x \in \Gamma_i$, by remarking that, from the Cauchy-Schwarz inequality:

$$A(x) \leq D_1 \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_\sigma} \chi_\sigma(x,y(x)) \sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_\sigma \chi_\sigma(x,y(x)),$$

where D_1 only depends on α_i and $c_\sigma = |\eta_i \cdot \mathbf{n}_\sigma|$. (Recall that $D_\sigma u = |u_K - u_L|$.) Since

$$\sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_\sigma \chi_\sigma(x,y(x)) \leq \mathrm{diam}(\Omega),$$

this yields:

$$A(x) \leq \mathrm{diam}(\Omega) D_1 \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_\sigma} \chi_\sigma(x,y(x)).$$

Then, since

$$\int_{\Gamma_i} \chi_\sigma(x,y(x)) d\gamma(x) \leq \frac{1}{\alpha} c_\sigma \mathrm{m}(\sigma),$$

there exists D_2, only depending on Ω, such that

$$A = \int_{\Gamma_i} A(x) d\gamma(x) \leq D_2 |u|^2_{1,\mathcal{T}}.$$

A bound $B(x)$ for a.e. $x \in \Gamma_i$ is obtained with the Cauchy-Schwarz inequality:

$$B(x) \leq D_3 \sum_{K \in \mathcal{T}} u_K^2 \chi_K(x,y(x)) |\xi_K(x) - \eta_K(x)| \sum_{K \in \mathcal{T}} |\xi_K(x) - \eta_K(x)| \chi_K(x,y(x)),$$

where D_3 only depends on α_i. Since

$$\sum_{K \in \mathcal{T}} |\xi_K(x) - \eta_K(x)| \chi_K(x,y(x)) \leq \mathrm{diam}(\Omega) \text{ and } \int_{\Gamma_i} \chi_K(x,y(x)) |\xi_K(x) - \eta_K(x)| d\gamma(x) \leq \frac{1}{\alpha} \mathrm{m}(K),$$

there exists D_4, only depending on Ω, such that

$$B = \int_{\Gamma_i} B(x) d\gamma(x) \leq D_4 \|u\|^2_{L^2(\Omega)}.$$

Integrating (10.27) over Γ_i, the bounds on A and B lead (10.26) for some convenient C_i and it concludes the proof of Lemma 10.5. ∎

Remark 10.2 Using this "trace inequality" (10.25) and the Kolmogorov theorem (see Theorem 14.1 page 94, it is possible to prove Lemma 10.2 page 65 (Discrete Poincaré inequality) by way of contradiction. Indeed, assume that there exists a sequence $(u_n)_{n\in\mathbb{N}}$ such that, for all $n \in \mathbb{N}$, $\|u_n\|_{L^2(\Omega)} = 1$, $\int_\Omega u_n(x)dx = 0$, $u_n \in X(\mathcal{T}_n)$ (where $\mathcal{T}_n$ is an admissible mesh in the sense of Definition 10.1) and $|u_n|_{1,\mathcal{T}_n} \leq \frac{1}{n}$. Using the trace inequality, one proves that $(u_n)_{n\in\mathbb{N}}$ is relatively compact in $L^2(\Omega)$, as in Theorem 10.3 page 74. Then, one can assume that $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$. The function u satisfies $\|u\|_{L^2(\Omega)} = 1$, since $\|u_n\|_{L^2(\Omega)} = 1$, and $\int_\Omega u(x)dx = 0$, since $\int_\Omega u_n(x)dx = 0$. Using $|u_n|_{1,\mathcal{T}_n} \leq \frac{1}{n}$, a proof similar to that of Theorem 14.3 page 95, yields that $D_i u = 0$, for all $i \in \{1, \ldots, n\}$ (even if size$(\mathcal{T}_n) \not\to 0$, as $n \to \infty$), where $D_i u$ is the derivative in the distribution sense with respect to x_i of u. Since Ω is connected, one deduces that u is constant on Ω, but this is impossible since $\|u\|_{L^2(\Omega)} = 1$ and $\int_\Omega u(x)dx = 0$.

Let us now prove that the scheme (10.7) and (10.8), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (10.5) and (10.6) is stable: the approximate solution given by the scheme is bounded independently of the mesh, as we proceed to show.

Lemma 10.6 (Estimate for the Neumann problem) *Under Assumption 10.1 page 63, let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 10.1 page 63). Let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (10.7) and (10.8), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (10.5) and (10.6); the existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 10.1. Let $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2) be defined by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Then, there exists $C \in \mathbb{R}_+$, only depending on Ω, g and f, such that*

$$|u_\mathcal{T}|_{1,\mathcal{T}} \leq C, \tag{10.28}$$

where $|\cdot|_{1,\mathcal{T}}$ is defined in Definition 10.2 page 64.

Proof of Lemma 10.6
Multiplying (10.7) by u_K and summing over $K \in \mathcal{T}$ yields

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(u_L - u_K)^2 = \sum_{K\in\mathcal{T}} \mathrm{m}(K) f_K u_K + \sum_{\sigma\in\mathcal{E}_{\rm ext}} u_{K_\sigma} g_{K_\sigma} \mathrm{m}(\sigma), \tag{10.29}$$

where, for $\sigma \in \mathcal{E}_{\rm ext}$, $K_\sigma \in \mathcal{T}$ is such that $\sigma \in \mathcal{E}_{K_\sigma}$.
We get (10.28) from (10.29) using (10.25), (10.10) and the Cauchy-Schwarz inequality. ∎

Using the estimate (10.28) on the approximate solution, a convergence result is given in the following theorem.

Theorem 10.3 (Convergence in the case of the Neumann problem)
Under Assumption 10.1 page 63, let u be the unique solution to (10.3),(10.4). For an admissible mesh (in the sense of Definition 10.1 page 63) $\mathcal{T}$, let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (10.7) and (10.8) (where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (10.5) and (10.6), the existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 10.1) and define $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 9.2) by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Then,

$$u_{\mathcal{T}} \to u \text{ in } L^2(\Omega) \text{ as } \mathrm{size}(\mathcal{T}) \to 0,$$

$$|u_{\mathcal{T}}|^2_{1,\mathcal{T}} \to \int_\Omega |\nabla u(x)|^2 dx \text{ as } \mathrm{size}(\mathcal{T}) \to 0$$

and

$$\overline{\gamma}(u_{\mathcal{T}}) \to \overline{\gamma}(u) \text{ in } L^2(\partial\Omega) \text{ for the weak topology as } \mathrm{size}(\mathcal{T}) \to 0,$$

where the function $\overline{\gamma}(u)$ stands for the trace of u on $\partial\Omega$ in the sense given in Lemma 10.5 when $u \in X(\mathcal{T})$ and in the sense of the classical trace operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$ (or $H^{\frac{1}{2}}(\partial\Omega)$) when $u \in H^1(\Omega)$.

Proof of Theorem 10.3

Step 1 (Compactness)
Denote by Y the set of approximate solutions $u_{\mathcal{T}}$ for all admisible meshes $\mathcal{T}$. Thanks to Lemma 10.6 and to the discrete Poincaré inequality (10.10), the set Y is bounded in $L^2(\Omega)$. Let us prove that Y is relatively compact in $L^2(\Omega)$, and that, if $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is a sequence of admissible meshes such that size$(\mathcal{T}_n)$ tends to 0 and $u_{\mathcal{T}_n}$ tends to u, in $L^2(\Omega)$, as n tends to infinity, then u belongs to $H^1(\Omega)$. Indeed, these results follow from theorems 14.1 and 14.3 page 95, provided that there exists a real positive number C only depending on Ω, f and g such that

$$\|\tilde{u}_{\mathcal{T}}(\cdot+\eta) - \tilde{u}_{\mathcal{T}}\|^2_{L^2(\mathbb{R}^d)} \le C|\eta|, \text{ for any admissible mesh } \mathcal{T} \text{ and for any } \eta \in \mathbb{R}^d,\ |\eta| \le 1, \tag{10.30}$$

and that, for any compact subset $\bar{\omega}$ of Ω,

$$\begin{array}{l}\|u_{\mathcal{T}}(\cdot+\eta) - u_{\mathcal{T}}\|^2_{L^2(\bar{\omega})} \le C|\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})), \text{ for any admissible mesh } \mathcal{T} \\ \text{and for any } \eta \in \mathbb{R}^d \text{ such that } |\eta| < d(\bar{\omega}, \Omega^c).\end{array} \tag{10.31}$$

Recall that $\tilde{u}_{\mathcal{T}}$ is defined by $\tilde{u}_{\mathcal{T}}(x) = u_{\mathcal{T}}(x)$ if $x \in \Omega$ and $\tilde{u}_{\mathcal{T}}(x) = 0$ otherwise. In order to prove (10.30) and (10.31), define χ_σ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\{0,1\}$ by $\chi_\sigma(x,y) = 1$ if $[x,y] \cap \sigma \neq \emptyset$ and $\chi_\sigma(x,y) = 0$ if $[x,y] \cap \sigma = \emptyset$. Let $\eta \in \mathbb{R}^d \setminus \{0\}$. Then:

$$|\tilde{u}(x+\eta) - \tilde{u}(x)| \le \sum_{\sigma\in\mathcal{E}_{\rm int}} \chi_\sigma(x, x+\eta)|D_\sigma u| + \sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi_\sigma(x, x+\eta)|u_\sigma|, \text{ for a.e. } x \in \Omega, \tag{10.32}$$

where, for $\sigma \in \mathcal{E}_{\rm ext}$, $u_\sigma = u_K$, and K is the control volume such that $\sigma \in \mathcal{E}_K$. Recall also that $D_\sigma u = |u_K - u_L|$, if $\sigma = K|L$. Let us first prove Inequality (10.31). Let $\bar{\omega}$ be a compact subset of Ω. If $x \in \bar{\omega}$ and $|\eta| < d(\bar{\omega}, \Omega^c)$, the second term of the right hand side of (10.32) is 0, and the same proof as in Lemma 9.3 page 44 gives, from an integration over $\bar{\omega}$ instead of Ω and from (9.33) with $C = 2$ since $[x, x+\eta] \subset \Omega$ for $x \in \bar{\omega}$,

$$\|u_{\mathcal{T}}(\cdot+\eta) - u_{\mathcal{T}}\|^2_{L^2(\bar{\omega})} \le |u|^2_{1,\mathcal{T}}|\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})). \tag{10.33}$$

In order to prove (10.30), remark that the number of non zero terms in the second term of the right hand side of (10.32) is, for a.e. $x \in \Omega$, bounded by some real positive number, which only depends on Ω, which can be taken, for instance, as the number of sides of Ω, denoted by N. Hence, with $C_1 = (N+1)^2$ (which only depends on Ω. Indeed, if Ω is convex, $N = 2$ is also convenient), one has

$$|\tilde{u}(x+\eta) - \tilde{u}(x)|^2 \le C_1\Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \chi_\sigma(x, x+\eta)|D_\sigma u|\Big)^2 + C_1 \sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi_\sigma(x, x+\eta)u_\sigma^2, \text{ for a.e. } x \in \Omega. \tag{10.34}$$

Let us integrate this inequality over $\mathbb{R}^d$. As seen in the proof of Lemma 9.3 page 44,

$$\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \chi_\sigma(x,x+\eta)|D_\sigma u|\Big)^2 dx \le |u|_{1,\mathcal{T}}^2 |\eta|(|\eta| + 2(N-1)\mathrm{size}(\mathcal{T}));$$

hence, by Lemma 10.6 page 74, there exists a real positive number C_2, only depending on Ω, f and g, such that (if $|\eta| \le 1$)

$$\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \chi_\sigma(x,x+\eta)|D_\sigma u|\Big)^2 dx \le C_2|\eta|.$$

Let us now turn to the second term of the right hand side of (10.34) integrated over $\mathbb{R}^d$;

$$\begin{aligned}\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi_\sigma(x,x+\eta)u_\sigma^2\Big)dx &\le \sum_{\sigma\in\mathcal{E}_{\rm ext}} \mathrm{m}(\sigma)|\eta|u_\sigma^2 \\ &\le \|\overline{\gamma}(u_\mathcal{T})\|_{L^2(\partial\Omega)}^2|\eta|;\end{aligned}$$

therefore, thanks to Lemma 10.5, Lemma 10.6 and to the discrete Poincaré inequality (10.10), there exists a real positive number C_3, only depending on Ω, f and g, such that

$$\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi_\sigma(x,x+\eta)u_\sigma^2\Big)dx \le C_3|\eta|.$$

Hence (10.30) is proved for some real positive number C only depending on Ω, f and g.

Step 2 (Passage to the limit)
In this step, the convergence of $u_\mathcal{T}$ to the solution of (10.3), (10.4) (in $L^2(\Omega)$ as $\mathrm{size}(\mathcal{T}) \to 0$) is first proved.
Since the solution to (10.3), (10.4) is unique, and thanks to the compactness of the set Y described in Step 1, it is sufficient to prove that, if $u_{\mathcal{T}_n} \to u$ in $L^2(\Omega)$ and $\mathrm{size}(\mathcal{T}_n) \to 0$ as $n \to 0$, then u is a solution to (10.3)-(10.4).
Let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a sequence of admissible meshes and $(u_{\mathcal{T}_n})_{n\in\mathbb{N}}$ be the corresponding solutions to (10.7)-(10.8) page 64 with $\mathcal{T} = \mathcal{T}_n$. Assume $u_{\mathcal{T}_n} \to u$ in $L^2(\Omega)$ and $\mathrm{size}(\mathcal{T}_n) \to 0$ as $n \to 0$. By Step 1, one has $u \in H^1(\Omega)$ and since the mean value of $u_{\mathcal{T}_n}$ is zero, one also has $\int_\Omega u(x)dx = 0$. Therefore, u is a solution of (10.3). It remains to show that u satisfies (10.4). Since $(\overline{\gamma}(u_{\mathcal{T}_n}))_{n\in\mathbb{N}}$ is bounded in $L^2(\partial\Omega)$, one may assume (up to a subsequence) that it converges to some v weakly in $L^2(\partial\Omega)$. Let us first prove that

$$\begin{aligned}-\int_\Omega u(x)\Delta\varphi(x)dx + \int_{\partial\Omega} \nabla\varphi(x)\cdot\mathbf{n}(x)v(x)d\gamma(x) &= \int_\Omega f(x)\varphi(x)dx \\ &+ \int_{\partial\Omega} g(x)\varphi(x)d\gamma(x),\ \forall\varphi\in C^2(\overline{\Omega}),\end{aligned} \tag{10.35}$$

and then that u satisfies (10.4).

Let $\mathcal{T}$ be an admissible mesh, $u_\mathcal{T}$ the corresponding approximate solution to the Neumann problem, given by (10.7) and (10.8), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (10.5) and (10.6) and let $\varphi \in C^2(\overline{\Omega})$. Let $\varphi_K = \varphi(x_K)$, define $\varphi_\mathcal{T}$ by $\varphi_\mathcal{T}(x) = \varphi_K$, for a.e. $x \in K$ and for any control volume K, and $\overline{\gamma}(\varphi_\mathcal{T})(x) = \varphi_K$ for a.e. $x \in \sigma$ (for the $(d-1)$-dimensional Lebegue measure), for any $\sigma \in \mathcal{E}_{\rm ext}$ and control volume K such that $\sigma \in \mathcal{E}_K$.
Multiplying (10.7) by φ_K, summing over $K \in \mathcal{T}$ and reordering the terms yields

$$\sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\varphi_L - \varphi_K) = \int_\Omega f(x)\varphi_\mathcal{T}(x)dx + \int_{\partial\Omega} \overline{\gamma}(\varphi_\mathcal{T})(x)g(x)d\gamma(x). \tag{10.36}$$

Using the consistency of the fluxes and the fact that $\varphi \in C^2(\overline{\Omega})$, there exists C only depending on φ such that

$$\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(\varphi_L-\varphi_K)=\int_K\Delta\varphi(x)dx-\int_{\partial\Omega\cap\partial K}\nabla\varphi(x)\cdot\mathbf{n}(x)d\gamma(x)+\sum_{L\in\mathcal{N}(K)}R_{K,L}(\varphi),$$

with $R_{K,L}=-R_{L,K}$, for all $L\in\mathcal{N}(K)$ and $K\in\mathcal{T}$, and $|R_{K,L}|\leq C_4\mathrm{m}(K|L)\mathrm{size}(\mathcal{T})$, where C_4 only depends on φ. Hence (10.36) may be rewritten as

$$\begin{aligned}-\int_\Omega u_\mathcal{T}(x)\Delta\varphi(x)dx+\int_{\partial\Omega}\nabla\varphi(x)\cdot\mathbf{n}(x)\overline{\gamma}(u_\mathcal{T})(x)d\gamma(x)+r(\varphi,\mathcal{T})=\\ \int_\Omega f(x)\varphi_\mathcal{T}(x)dx+\int_{\partial\Omega}\overline{\gamma}(\varphi_\mathcal{T})(x)g(x)d\gamma(x),\end{aligned}\tag{10.37}$$

where

$$\begin{aligned}|r(\varphi,\mathcal{T})| &= C_4\sum_{\sigma\in\mathcal{E}_{\rm int}}|D_\sigma u|\mathrm{m}(\sigma)\mathrm{size}(\mathcal{T})\\ &\leq C_4\Big(\sum_{\sigma\in\mathcal{E}_{\rm int}}|D_\sigma u|^2\frac{\mathrm{m}(\sigma)}{d_\sigma}\Big)^{\frac12}\Big(\sum_{\sigma\in\mathcal{E}_{\rm int}}\mathrm{m}(\sigma)d_\sigma\Big)^{\frac12}\mathrm{size}(\mathcal{T})\\ &\leq C_5\mathrm{size}(\mathcal{T}),\end{aligned}$$

where C_5 is a real positive number only depending on f, g, Ω and φ (thanks to Lemma 10.6). Writing (10.37) with $\mathcal{T}=\mathcal{T}_n$ and passing to the limit as n tends to infinity yields (10.35).

Let us now prove that u satifies (10.4). Since $u\in H^1(\Omega)$, an integration by parts in (10.35) yields

$$\begin{aligned}\int_\Omega\nabla u(x)\cdot\nabla\varphi(x)dx+\int_{\partial\Omega}\nabla\varphi(x)\cdot\mathbf{n}(x)(v(x)-\overline{\gamma}(u)(x))d\gamma(x)\\ =\int_\Omega f(x)\varphi(x)dx+\int_{\partial\Omega}g(x)\varphi(x)d\gamma(x),\forall\varphi\in C^2(\overline{\Omega}),\end{aligned}\tag{10.38}$$

where $\overline{\gamma}(u)$ denotes the trace of u on $\partial\Omega$ (which belongs to $L^2(\partial\Omega)$). In order to prove that u is solution to (10.4) (this will conclude the proof of Theorem 10.3), it is sufficient, thanks to the density of $C^2(\overline{\Omega})$ in $H^1(\Omega)$, to prove that $v=\overline{\gamma}(u)$ a.e. on $\partial\Omega$ (for the $(d-1)$ dimensional Lebesgue measure on $\partial\Omega$). Let us now prove that $v=\overline{\gamma}(u)$ a.e. on $\partial\Omega$ by first remarking that (10.38) yields

$$\int_\Omega\nabla u(x)\cdot\nabla\varphi(x)dx=\int_\Omega f(x)\varphi(x)dx,\ \forall\varphi\in C_c^\infty(\Omega),$$

and therefore, by density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$,

$$\int_\Omega\nabla u(x)\cdot\nabla\varphi(x)dx=\int_\Omega f(x)\varphi(x)dx,\ \forall\varphi\in H_0^1(\Omega).$$

With (10.38), this yields

$$-\int_{\partial\Omega}\nabla\varphi(x)\cdot\mathbf{n}(x)(v(x)-\overline{\gamma}(u)(x))d\gamma(x)=0,\ \forall\varphi\in C^2(\overline{\Omega})\text{ such that }\varphi=0\text{ on }\partial\Omega.\tag{10.39}$$

There remains to show that the wide choice of φ in (10.39) allows to conclude $v=\overline{\gamma}(u)$ a.e. on $\partial\Omega$ (for the $(d-1)$-dimensional Lebesgue measure of $\partial\Omega$). Indeed, let I be a part of the boundary $\partial\Omega$, such that I is included in a hyperplane of $\mathbb{R}^d$. Assume that $I=\{0\}\times J$, where J is an open ball of $\mathbb{R}^{d-1}$ centered on the origin. Let $z=(a,\tilde{z})\in\mathbb{R}^d$ with $a\in\mathbb{R}_+^\star$, $\tilde{z}\in\mathbb{R}^{d-1}$ and $B=\{(t,\frac{a-|t|}{a}y+\frac{|t|}{a}\tilde{z});t\in(-a,a),y\in J\}$; assume that, for a convenient a, one has

$$B\cap\Omega=\{(t,\frac{a-|t|}{a}y+\frac{|t|}{a}\tilde{z});t\in(0,a),y\in J\}.$$

Let $\psi \in C_c^\infty(J)$, and for $x = (x_1, y) \in \mathbb{R} \times J$, define $\varphi_1(x) = -x_1\psi(y)$. Then,

$$\varphi_1 \in C^\infty(\mathbb{R}^d) \text{ and } \frac{\partial \varphi_1}{\partial n} = \psi \text{ on } I.$$

(Recall that $\mathbf{n}$ is the normal unit vector to $\partial\Omega$, outward to Ω.) Let $\varphi_2 \in C_c^\infty(B)$ such that $\varphi_2 = 1$ on a neighborhood of $\{0\} \times \{\psi \neq 0\}$, where $\{\psi \neq 0\} = \{x \in J;\ \psi(x) \neq 0\}$, and set $\varphi = \varphi_1\varphi_2$; φ is an admissible test function in (10.39), and therefore

$$\int_J \psi(y)\big(\overline{\gamma}(u)(0,y) - v(0,y)\big)dy = 0,$$

which yields, since ψ is arbitrary in $C_c^\infty(J)$, $v = \overline{\gamma}(u)$ a.e. on I. Since J is arbitrary, this implies that $v = \overline{\gamma}(u)$ a.e. on $\partial\Omega$.

This conclude the proof of $u_\mathcal{T} \to u$ in $L^2(\Omega)$ as size$(\mathcal{T}) \to 0$, where u is the solution to (10.3),(10.4).

Note also that the above proof gives (by way of contradiction) that $\overline{\gamma}(u_\mathcal{T}) \to \overline{\gamma}(u)$ weakly in $L^2(\partial\Omega)$, as size$(\mathcal{T}) \to 0$.

Then, a passage to the limit in (10.29) together with (10.4) yields

$$|u_\mathcal{T}|_{1,\mathcal{T}}^2 \to |||\nabla u|||_{L^2(\Omega)}^2, \text{ as size}(\mathcal{T}) \to 0.$$

This concludes the proof of Theorem 10.3. ∎

Note that, with some discrete Sobolev inequality (similar to (9.69)), the hypothesis "$f \in L^2(\Omega)$ $g \in L^2(\partial\Omega)$" may be relaxed in some way similar to that of Item 2 of Remark 9.7.

11 General elliptic operators

11.1 Discontinuous matrix diffusion coefficients

Meshes and schemes

Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3. We are interested here in the discretization of an elliptic operator with discontinuous matrix diffusion coefficients, which may appear in real case problems such as electrical or thermal transfer problems or, more generally, diffusion problems in heterogeneous media. In this case, the mesh is adapted to fit the discontinuities of the data. Hence the definition of an admissible mesh given in Definition 9.1 must be adapted. As an illustration, let us consider here the following problem, which was studied in Section 7 page 21 in the one-dimensional case:

$$-\mathrm{div}(\Lambda\nabla u)(x) + \mathrm{div}(\mathbf{v}u)(x) + bu(x) = f(x),\ x \in \Omega, \tag{11.1}$$

$$u(x) = g(x),\ x \in \partial\Omega, \tag{11.2}$$

with the following assumptions on the data (one denotes by $\mathbb{R}^{d\times d}$ the set of $d \times d$ matrices with real coefficients):

Assumption 11.1

1. *Λ is a bounded measurable function from Ω to $\mathbb{R}^{d\times d}$ such that for any $x \in \Omega$, $\Lambda(x)$ is symmetric, and that there exists $\underline{\lambda}$ and $\overline{\lambda} \in \mathbb{R}_+^\star$ such that $\underline{\lambda}\xi\cdot\xi \le \Lambda(x)\xi\cdot\xi \le \overline{\lambda}\xi\cdot\xi$ for any $x \in \Omega$ and any $\xi \in \mathbb{R}^d$.*
2. *$\mathbf{v} \in C^1(\overline{\Omega}, \mathbb{R}^d)$, $\mathrm{div}\mathbf{v} \ge 0$ on Ω, $b \in \mathbb{R}_+$.*
3. *f is a bounded piecewise continuous function from Ω to $\mathbb{R}$.*

4. g is such that there exists $\tilde{g} \in H^1(\Omega)$ such that $\overline{\gamma}(\tilde{g}) = g$ (a.e. on $\partial\Omega$) and is a bounded piecewise continuous function from $\partial\Omega$ to $\mathbb{R}$.

(Recall that $\overline{\gamma}$ denotes the trace operator from $H^1(\Omega)$ into $L^2(\partial\Omega)$.) As in Section 9, under Assumption 11.1, there exists a unique variational solution $u \in H^1(\Omega)$ of Problem (11.1), (11.2). This solution satisfies $u = w + \tilde{g}$, where $\tilde{g} \in H^1(\Omega)$ is such that $\overline{\gamma}(\tilde{g}) = g$, a.e. on $\partial\Omega$, and w is the unique function of $H_0^1(\Omega)$ satisfying

$$\int_\Omega \Big(\Lambda(x)\nabla w(x)\cdot\nabla\psi(x) + \mathrm{div}(\mathbf{v}w)(x)\psi(x) + bw(x)\psi(x)\Big)dx =$$
$$\int_\Omega \Big(-\Lambda(x)\nabla\tilde{g}(x)\cdot\nabla\psi(x) - \mathrm{div}(\mathbf{v}\tilde{g})(x)\psi(x) - b\tilde{g}(x)\psi(x) + f(x)\psi(x)\Big)dx,\ \forall\psi \in H_0^1(\Omega).$$

Let us now define an admissible mesh for the discretization of Problem (11.1)-(11.2).

Definition 11.1 (Admissible mesh for a general diffusion operator) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3. An admissible finite volume mesh for the discretization of Problem (11.1)-(11.2) is an admissible mesh $\mathcal{T}$ of Ω in the sense of Definition 9.1 page 37 where items (iv) and (v) are replaced by the two following conditions:

(iv)' The set $\mathcal{T}$ is such that

the restriction of g to each edge $\sigma \in \mathcal{E}_{\rm ext}$ is continuous.

For any $K \in \mathcal{T}$, let Λ_K denote the mean value of Λ on K, that is

$$\Lambda_K = \frac{1}{\mathrm{m}(K)}\int_K \Lambda(x)dx.$$

There exists a family of points

$$\mathcal{P} = (x_K)_{K\in\mathcal{T}} \text{ such that } x_K = \cap_{\sigma\in\mathcal{E}_K}\mathcal{D}_{K,\sigma} \in \overline{K},$$

where $\mathcal{D}_{K,\sigma}$ is a straigth line perpendicular to σ with respect to the scalar product induced by Λ_K^{-1} such that $\mathcal{D}_{K,\sigma}\cap\sigma = \mathcal{D}_{L,\sigma}\cap\sigma \neq \emptyset$ if $\sigma = K|L$. Furthermore, if $\sigma = K|L$, let $y_\sigma = \mathcal{D}_{K,\sigma}\cap\sigma (= \mathcal{D}_{L,\sigma}\cap\sigma)$ and assume that $x_K \neq x_L$.

(v)' For any $\sigma \in \mathcal{E}_{\rm ext}$, let K be the control volume such that $\sigma \in \mathcal{E}_K$ and let $\mathcal{D}_{K,\sigma}$ be the straight line going through x_K and orthogonal to σ with respect to the scalar product induced by Λ_K^{-1}; then, there exists $y_\sigma \in \sigma\cap\mathcal{D}_{K,\sigma}$; let $g_\sigma = g(y_\sigma)$.

The notations are are the same as those introduced in Definition 9.1 page 37.

We shall now define the discrete unknowns of the numerical scheme, with the same notations as in Section 9.2. As in the case of the Dirichlet problem, the primary unknowns $(u_K)_{K\in\mathcal{T}}$ will be used, which aim to be approximations of the values $u(x_K)$, and some auxiliary unknowns, namely the fluxes $F_{K,\sigma}$, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, and some (expected) approximation of u in σ, say u_σ, for all $\sigma \in \mathcal{E}$. Again, these auxiliary unknowns are helpful to write the scheme, but they can be eliminated locally so that the discrete equations will only be written with respect to the primary unknowns $(u_K)_{K\in\mathcal{T}}$. For any $\sigma \in \mathcal{E}_{\rm ext}$, set $u_\sigma = g(y_\sigma)$. The finite volume scheme for the numerical approximation of the solution to Problem (11.1)-(11.2) is obtained by integrating Equation (11.1) over each control volume K, and approximating the fluxes over each edge σ of K. This yields

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}u_{\sigma,+} + \mathrm{m}(K)bu_K = f_K,\ \forall K \in \mathcal{T}, \tag{11.3}$$

where

$v_{K,\sigma} = \int_\sigma \mathbf{v}(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$ (where $\mathbf{n}_{K,\sigma}$ denotes the normal unit vector to σ outward to K); if $\sigma = K_{\sigma,+}|K_{\sigma,-}$, $u_{\sigma,+} = u_{K_{\sigma,+}}$, where $K_{\sigma,+}$ is the upstream control volume, i.e. $v_{K,\sigma} \geq 0$, with $K = K_{\sigma,+}$; if $\sigma \in \mathcal{E}_{\text{ext}}$, then $u_{\sigma,+} = u_K$ if $v_{K,\sigma} \geq 0$ (i.e. K is upstream to σ with respect to v), and $u_{\sigma,+} = u_\sigma$ otherwise.
$F_{K,\sigma}$ is an approximation of $\int_\sigma -\Lambda_K \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$; the approximation $F_{K,\sigma}$ is written with respect to the discrete unknowns $(u_K)_{K\in\mathcal{T}}$ and $(u_\sigma)_{\sigma\in\mathcal{E}}$. For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, let $\lambda_{K,\sigma} = |\Lambda_K \mathbf{n}_{K,\sigma}|$ (recall that $|\cdot|$ denote the Euclidean norm).

- If $x_K \notin \sigma$, a natural expression for $F_{K,\sigma}$ is then

$$F_{K,\sigma} = -\mathrm{m}(\sigma)\lambda_{K,\sigma}\frac{u_\sigma - u_K}{d_{K,\sigma}}.$$

  Writing the conservativity of the scheme, i.e. $F_{L,\sigma} = -F_{K,\sigma}$ if $\sigma = K|L \subset \Omega$, yields the value of u_σ, if $x_L \notin \sigma$, with respect to $(u_K)_{K\in\mathcal{T}}$;

$$u_\sigma = \frac{1}{\frac{\lambda_{K,\sigma}}{d_{K,\sigma}} + \frac{\lambda_{L,\sigma}}{d_{L,\sigma}}}\Big(\frac{\lambda_{K,\sigma}}{d_{K,\sigma}}u_K + \frac{\lambda_{L,\sigma}}{d_{L,\sigma}}u_L\Big).$$

  Note that this expression is similar to that of (7.3) page 22 in the 1D case.

- If $x_K \in \sigma$, one sets $u_\sigma = u_K$.

Hence the value of $F_{K,\sigma}$;

- internal edges:

$$F_{K,\sigma} = -\tau_\sigma(u_L - u_K), \text{ if } \sigma \in \mathcal{E}_{\text{int}},\ \sigma = K|L, \tag{11.4}$$

  where

$$\tau_\sigma = \mathrm{m}(\sigma)\frac{\lambda_{K,\sigma}\lambda_{L,\sigma}}{\lambda_{K,\sigma}d_{L,\sigma} + \lambda_{L,\sigma}d_{K,\sigma}} \text{ if } y_\sigma \neq x_K \text{ and } y_\sigma \neq x_L$$

  and

$$\tau_\sigma = \mathrm{m}(\sigma)\frac{\lambda_{K,\sigma}}{d_{K,\sigma}} \text{ if } y_\sigma \neq x_K \text{ and } y_\sigma = x_L;$$

- boundary edges:

$$F_{K,\sigma} = -\tau_\sigma(g_\sigma - u_K), \text{ if } \sigma \in \mathcal{E}_{\text{ext}} \text{ and } x_K \notin \sigma, \tag{11.5}$$

  where

$$\tau_\sigma = \mathrm{m}(\sigma)\frac{\lambda_{K,\sigma}}{d_{K,\sigma}};$$

  if $x_K \in \sigma$, then the equation associated to u_K is $u_K = g_\sigma$ (instead of that given by (11.3)) and the numerical flux $F_{K,\sigma}$ is an unknown which may be deduced from (11.3).

Remark 11.1 Note that if $\Lambda = Id$, then the scheme (11.3)-(11.5) is the same scheme than the one described in Section 9.2.

Error estimate

Theorem 11.1
Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3. Under Assumption 11.1, let u be the unique variational solution to Problem (11.1)-(11.2). Let $\mathcal{T}$ be an admissible mesh for the discretization of Problem (11.1)-(11.2), in the sense of Definition 11.1. Let ζ_1 and $\zeta_2 \in \mathbb{R}_+$ such that

$$\begin{array}{l}\zeta_1(\mathrm{size}(\mathcal{T}))^2 \leq \mathrm{m}(K) \leq \zeta_2(\mathrm{size}(\mathcal{T}))^2,\\ \zeta_1\mathrm{size}(\mathcal{T}) \leq \mathrm{m}(\sigma) \leq \zeta_2\mathrm{size}(\mathcal{T}),\\ \zeta_1\mathrm{size}(\mathcal{T}) \leq d_\sigma \leq \zeta_2\mathrm{size}(\mathcal{T}).\end{array}$$

Assuming moreover that
the restriction of f to K belongs to $C(\overline{K})$, for any $K \in \mathcal{T}$;
the restriction of Λ to K belongs to $C^1(\overline{K}, \mathbb{R}^{d\times d})$, for any $K \in \mathcal{T}$;
the restriction of u (unique variational solution of Problem (11.1)-(11.2)) to K belongs to $C^2(\overline{K})$, for any $K \in \mathcal{T}$.
(Recall that $C^m(\overline{K}, \mathbb{R}^N) = \{v_{|_K},\, v \in C^m(\mathbb{R}^d, \mathbb{R}^N)\}$ and $C^m(\cdot) = C^m(\cdot, \mathbb{R})$.)
Then, there exists a unique family $(u_K)_{K\in\mathcal{T}}$ satisfying (11.3)-(11.5); furthermore, denoting by $e_K = u(x_K) - u_K$, there exists $C \in \mathbb{R}_+$ only depending on ζ_1, ζ_2, $\gamma = \sup_{K\in\mathcal{T}}(\|D^2u\|_{L^\infty(K)})$ and $\delta = \sup_{K\in\mathcal{T}}(\|D\Lambda\|_{L^\infty(K)})$ such that

$$\sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma e)^2}{d_\sigma}\mathrm{m}(\sigma) \leq C(\mathrm{size}(\mathcal{T}))^2 \tag{11.6}$$

and

$$\sum_{K\in\mathcal{T}} e_K^2\mathrm{m}(K) \leq C(\mathrm{size}(\mathcal{T}))^2. \tag{11.7}$$

Recall that $D_\sigma e = |e_L - e_K|$ for $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$ and $D_\sigma e = |e_K|$ for $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$.

Proof of Theorem 11.1

First, one may use Taylor expansions and the same technique as in the 1D case (see step 2 of the proof of Theorem 7.1, Section 7) to show that the expressions (11.4) and (11.5) are consistent approximations of th exact diffusion flux $\int_\sigma -\Lambda(x)\nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)$, i.e. there exists C_1 only depending on u and Λ such that, for all $\sigma \in \mathcal{E}$, with $F^\star_{K,\sigma} = \tau_\sigma(u(x_L) - u(x_K))$, if $\sigma = K|L$, and $F^\star_{K,\sigma} = \tau_\sigma(u(y_\sigma) - u(x_K))$, if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$,

$$\begin{array}{c}F^\star_{K,\sigma} - \int_\sigma -\Lambda(x)\nabla u(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x) = R_{K,\sigma},\\ \text{with } |R_{K,\sigma}| \leq C_1\mathrm{size}(\mathcal{T})\mathrm{m}(\sigma).\end{array}$$

There also exists C_2 only depending on u and $\mathbf{v}$ such that, for all $\sigma \in \mathcal{E}$,

$$\begin{array}{c}v_{K,\sigma}u(x_{K_{\sigma,+}}) - \int_\sigma \mathbf{v}\cdot\mathbf{n}_{K,\sigma}u = r_{K,\sigma},\\ \text{with } |r_{K,\sigma}| \leq C_2\mathrm{size}(\mathcal{T})\mathrm{m}(\sigma).\end{array}$$

Let us then integrate Equation (11.1) over each control volume, subtract to (11.3) and use the consistency of the fluxes to obtain the following equation on the error:

$$\left\{\begin{array}{l} -\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}e_{\sigma,+} + \mathrm{m}(K)be_K =\\ \sum_{\sigma\in\mathcal{E}_K}(R_{K,\sigma} + r_{K,\sigma}) + S_K,\ \forall K \in \mathcal{T},\end{array}\right.$$

where $G_{K,\sigma} = \tau_\sigma(e_L - e_K)$, if $\sigma = K|L$, and $G_{K,\sigma} = \tau_\sigma(-e_K)$, if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$, $e_{\sigma,+} = e_{K_{\sigma,+}}$ is the error associated to the upstream control volume to σ and $S_K = b(\mathrm{m}(K)u(x_K) - \int_K u(x)dx)$ is such that

$|S_K| \leq \mathrm{m}(K) C_3 h$, where $C_3 \in \mathbb{R}_+$ only depends on u and b. Then, similarly to the proof of Theorem 9.3 page 52, let us multiply by e_K, sum over $K \in \mathcal{T}$, and use the conservativity of the scheme, which yields that if $\sigma = K|L$ then $R_{K,\sigma} = -R_{L,\sigma}$. A reordering of the summation over $\sigma \in \mathcal{E}$ yields the "discrete H_0^1 estimate" (11.6). Then, following HERBIN [84], one shows the following discrete Poincaré inequality:

$$\sum_{K \in \mathcal{T}} e_K^2 \mathrm{m}(K) \leq C_4 \sum_{\sigma \in \mathcal{E}} \frac{(D_\sigma e)^2}{d_\sigma} \mathrm{m}(\sigma), \tag{11.8}$$

where C_4 only depends on Ω, ζ_1 and ζ_2, which in turn yields the L^2 estimate (11.7). ∎

Remark 11.2 In the case where Λ is constant, or more generally, in the case where $\Lambda(x) = \lambda(x) Id$, where $\lambda(x) > 0$, the proof of Lemma 9.1 is easily extended. However, for a general matrix Λ, the generalization of this proof is not so clear; this is the reason of the dependency of the estimates (11.6) and (11.7) on ζ_1 and ζ_2, which arises when proving (11.8) as in HERBIN [84].

11.2 Other boundary conditions

The finite volume scheme may be used to discretize elliptic problems with Dirichlet or Neumann boundary conditions, as we saw in the previous sections. It is also easily implemented in the case of Fourier (or Robin) and periodic boundary conditions. The case of interface conditions between two geometrical regions is also generally easy to implement; the purpose here is to present the treatment of some of these boundary and interface conditions. One may also refer to ANGOT [3] and references therein, FIARD, HERBIN [66] for the treatment of more complex boundary conditions and coupling terms in a system of elliptic equations.

Let Ω be (for the sake of simplicity) the open rectangular subset of $\mathbb{R}^2$ defined by $\Omega = (0,1) \times (0,2)$, let $\Omega_1 = (0,1) \times (0,1)$, $\Omega_2 = (0,1) \times (1,2)$, $\Gamma_1 = [0,1] \times \{0\}$, $\Gamma_2 = \{1\} \times [0,2]$, $\Gamma_3 = [0,1] \times \{2\}$, $\Gamma_4 = \{0\} \times [0,2]$ and $I = [0,1] \times \{1\}$. Let λ_1 and $\lambda_2 > 0$, $f \in C(\overline{\Omega})$, $\alpha > 0$, $\overline{u} \in \mathbb{R}$, $g \in C(\Gamma_4)$, θ and $\Phi \in C(I)$. Consider here the following problem (with some "natural" notations):

$$-\mathrm{div}(\lambda_i \nabla u)(x) = f(x),\ x \in \Omega_i,\ i = 1,2, \tag{11.9}$$

$$-\lambda_i \nabla u(x) \cdot \mathbf{n}(x) = \alpha(u(x) - \overline{u}),\ x \in \Gamma_1 \cup \Gamma_3, \tag{11.10}$$

$$\nabla u(x) \cdot \mathbf{n}(x) = 0,\ x \in \Gamma_2, \tag{11.11}$$

$$u(x) = g(x),\ x \in \Gamma_4, \tag{11.12}$$

$$(\lambda_2 \nabla u(x) \cdot \mathbf{n}_I(x))_{|2} = (\lambda_1 \nabla u(x) \cdot \mathbf{n}_I(x))_{|1} + \theta(x),\ x \in I, \tag{11.13}$$

$$u_{|2}(x) - u_{|1}(x) = \Phi(x),\ x \in I, \tag{11.14}$$

where $\mathbf{n}$ denotes the unit normal vector to $\partial\Omega$ outward to Ω and $\mathbf{n}_I = (0,1)^t$ (it is a unit normal vector to I).

Let $\mathcal{T}$ be an admissible mesh for the discretization of (11.9)-(11.14) in the sense of Definition 11.1. For the sake of simplicity, let us assume here that $d_{K,\sigma} > 0$ for all $K \in \mathcal{T}, \sigma \in \mathcal{E}_K$. Integrating Equation (11.9) over each control volume K, and approximating the fluxes over each edge σ of K yields the following finite volume scheme:

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = f_K,\ \forall K \in \mathcal{T}, \tag{11.15}$$

where $F_{K,\sigma}$ is an approximation of $\int_\sigma -\lambda_i \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$, with i such that $K \subset \Omega_i$.

Let $N_\mathcal{T} = \mathrm{card}(\mathcal{T})$, $N_\mathcal{E} = \mathrm{card}(\mathcal{E})$, $N_\mathcal{E}^0 = \mathrm{card}(\{\sigma \in \mathcal{E}; \sigma \not\subset \partial\Omega \cup I\})$, $N_\mathcal{E}^i = \mathrm{card}(\{\sigma \in \mathcal{E}; \sigma \subset \Gamma_i\})$, and $N_\mathcal{E}^I = \mathrm{card}(\{\sigma \in \mathcal{E}; \sigma \subset I\})$ (note that $N_\mathcal{E} = N_\mathcal{E}^0 + \sum_{i=1}^4 N_\mathcal{E}^i + N_\mathcal{E}^I$). Introduce the $N_\mathcal{T}$ (primary) discrete unknowns $(u_K)_{K \in \mathcal{T}}$; note that the number of (auxiliary) unknowns of the type $F_{K,\sigma}$ is $2(N_\mathcal{E}^0 + N_\mathcal{E}^I) +$

$\sum_{i=1}^4 N_{\mathcal{E}}^i$; let us introduce the discrete unknowns $(u_\sigma)_{\sigma\in\mathcal{E}}$, which aim to be approximations of u on σ. In order to take into account the jump condition (11.14), two unknowns of this type are necessary on the edges $\sigma \subset I$, namely $u_{\sigma,1}$ and $u_{\sigma,2}$. Hence the number of (auxiliary) unknowns of the type u_σ is $N_{\mathcal{E}}^0 + \sum_{i=1}^4 N_{\mathcal{E}}^i + 2N_{\mathcal{E}}^I$. Therefore, the total number of discrete unknowns is

$$N_{tot} = N_{\mathcal{T}} + 3N_{\mathcal{E}}^0 + 4N_{\mathcal{E}}^I + 2\sum_{i=1}^4 N_{\mathcal{E}}^i.$$

Hence, it is convenient, in order to obtain a well-posed system, to write N_{tot} discrete equations. We already have $N_{\mathcal{T}}$ equations from (11.15). The expression of $F_{K,\sigma}$ with respect to the unknowns u_K and u_σ is

$$F_{K,\sigma} = -\mathrm{m}(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}},\ \forall\, K \in \mathcal{T}; K \subset \Omega_i\ (i = 1, 2),\ \forall\, \sigma \in \mathcal{E}_K; \tag{11.16}$$

which yields $2(N_{\mathcal{E}}^0 + N_{\mathcal{E}}^I) + \sum_{i=1}^4 N_{\mathcal{E}}^i$. (In (11.16), u_σ stands for $u_{\sigma,i}$ if $\sigma \subset I$.)
Let us now take into account the various boundary and interface conditions:

- Fourier boundary conditions. Discretizing condition (11.10) yields

$$F_{K,\sigma} = \alpha \mathrm{m}(\sigma)(u_\sigma - \overline{u}), \forall\, K \in \mathcal{T}, \forall\, \sigma \in \mathcal{E}_K;\ \sigma \subset \Gamma_1 \cup \Gamma_3, \tag{11.17}$$

  that is $N_{\mathcal{E}}^1 + N_{\mathcal{E}}^3$ equations.

- Neumann boundary conditions. Discretizing condition (11.11) yields

$$F_{K,\sigma} = 0, \forall\, K \in \mathcal{T}, \forall\, \sigma \in \mathcal{E}_K;\ \sigma \subset \Gamma_2, \tag{11.18}$$

  that is $N_{\mathcal{E}}^2$ equations.

- Dirichlet boundary conditions. Discretizing condition (11.12) yields

$$u_\sigma = g(y_\sigma),\ \forall\, \sigma \in \mathcal{E}; \sigma \subset \Gamma_4, \tag{11.19}$$

  that is $N_{\mathcal{E}}^4$ equations.

- Conservativity of the flux. Except at interface I, the flux is continuous, and therefore

$$F_{K,\sigma} = -F_{L,\sigma}, \forall\, \sigma \in \mathcal{E}; \sigma \not\subset (\bigcup_{i=1}^4 \Gamma_i \cup I) \text{ and } \sigma = K|L, \tag{11.20}$$

  that is $N_{\mathcal{E}}^0$ equations.

- Jump condition on the flux. At interface I, condition (11.13) is discretized into

$$F_{K,\sigma} + F_{L,\sigma} = \int_\sigma \theta(x)ds, \forall\, \sigma \in \mathcal{E}; \sigma \subset I \text{ and } \sigma = K|L;\ K \subset \Omega_2, \tag{11.21}$$

  that is $N_{\mathcal{E}}^I$ equations.

- Jump condition on the unknown. At interface I, condition (11.14) is discretized into

$$u_{\sigma,2} = u_{\sigma,1} + \Phi(y_\sigma), \forall\, \sigma \in \mathcal{E}; \sigma \subset I \text{ and } \sigma = K|L. \tag{11.22}$$

  that is another $N_{\mathcal{E}}^I$ equations.

Hence the total number of equations from (11.15) to (11.22) is N_{tot}, so that the numerical scheme can be expected to be well posed.

The finite volume scheme for the discretization of equations (11.9)-(11.14) is therefore completely defined by (11.15)-(11.22). Particular cases of this scheme are the schemes (9.20)-(9.23) page 42 (written for Dirichlet boundary conditions) and (10.7)-(10.8) page 64 (written for Neumann boundary conditions and no convection term) which were thoroughly studied in the two previous sections.

12 Dual meshes and unknowns located at vertices

One of the principles of the classical finite volume method is to associate the discrete unknowns to the grid cells. However, it is sometimes useful to associate the discrete unknowns with the vertices of the mesh; for instance, the finite volume method may be used for the discretization of a hyperbolic equation coupled with an elliptic equation (see Chapter 7). Suppose that an existing finite element code is implemented for the elliptic equation and yields the discrete values of the unknown at the vertices of the mesh. One might then want to implement a finite volume method for the hyperbolic equation with the values of the unknowns at the vertices of the mesh. Note also that for some physical problems, e.g. the modelling of two phase flow in porous media, the conservativity principle is easier to respect if the discrete unknowns have the same location. For these various reasons, we introduce here some finite volume methods where the discrete unknowns are located at the vertices of an existing mesh.
For the sake of simplicity, the treatment of the boundary conditions will be omitted here. Recall that the construction of a finite volume method is carried out (in particular) along the following principles:

1. Divide the spatial domain in control volumes,
2. Associate to each control volume and, for time dependent problems, to each discrete time, one discrete unknown,
3. Obtain the discrete equations (at each discrete time) by integration of the equation over the control volume and the definition of one exchange term between two (adjacent) control volumes.

Recall, in particular, that the definition of one (and one only) exchange term between two control volumes is important; this is called the property of conservativity of a finite volume method. The aim here is to present finite volume methods for which the discrete unknowns are located at the vertices of the mesh. Hence, to each vertex must correspond a control volume. Note that these control volumes may be somehow "fictive" (see the next section); the important issue is to respect the principles given above in the construction of the finite volume scheme. In the three following sections, we shall deal with the two dimensional case; the generalization to the three-dimensional case is the purpose of section 12.4.

12.1 The piecewise linear finite element method viewed as a finite volume method

We consider here the Dirichlet problem. Let Ω be a bounded open polygonal subset of $\mathbb{R}^2$, f and g be some "regular" functions (from Ω or $\partial\Omega$ to $\mathbb{R}$). Consider the following problem:

$$\begin{cases} -\Delta u(x) = f(x), & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega. \end{cases} \tag{12.1}$$

Let us show that the "piecewise linear" finite element method for the discretization of (12.1) may be viewed as a kind of finite volume method. Let $\mathcal{M}$ be a finite element mesh of Ω, consisting of triangles (see e.g. Ciarlet [29] for the conditions on the triangles), and let $\mathcal{V} \subset \overline{\Omega}$ be the set of vertices of $\mathcal{M}$. For $K \in \mathcal{V}$ (note that here K denotes a point of $\overline{\Omega}$), let φ_K be the shape function associated to K in the piecewise linear finite element method for the mesh $\mathcal{M}$. We remark that

$$\sum_{K\in\mathcal{V}} \varphi_K(x) = 1,\ \forall x \in \Omega,$$

and therefore

$$\sum_{K\in\mathcal{V}} \int_\Omega \varphi_K(x)dx = \mathrm{m}(\Omega) \tag{12.2}$$

and

$$\sum_{K\in\mathcal{V}} \nabla\varphi_K(x) = 0, \quad \text{for a.e.} x \in \Omega. \tag{12.3}$$

Using the latter equality, the discrete finite element equation associated to the unknown u_K, if $K \in \Omega$, can therefore be written as

$$\sum_{L\in\mathcal{V}} \int_\Omega (u_L - u_K)\nabla\varphi_L(x)\cdot\nabla\varphi_K(x)dx = \int_\Omega f(x)\varphi_K(x)dx.$$

Then the finite element method may be written as

$$\sum_{L\in\mathcal{V}} -\tau_{K|L}(u_L - u_K) = \int_\Omega f(x)\varphi_K(x)dx, \text{ if } K \in \mathcal{V}\cap\Omega,$$

$$u_K = g(K), \text{ if } K \in \mathcal{V}\cap\partial\Omega,$$

with

$$\tau_{K|L} = -\int_\Omega \nabla\varphi_L(x)\cdot\nabla\varphi_K(x)dx.$$

Under this form, the finite element method may be viewed as a finite volume method, except that there are no "real" control volumes associated to the vertices of $\mathcal{M}$. Indeed, thanks to (12.2), the control volume associated to K may be viewed as the support of φ_K "weighted" by φ_K. This interpretation of the finite element method as a finite volume method was also used in Forsyth [67], Forsyth [68] and Eymard and Gallouët [49] in order to design a numerical scheme for a transport equation for which the velocity field is the gradient of the pressure, which is itself the solution to an elliptic equation (see also Herbin and Labergerie [86] for numerical tests). This method is often referred to as the "control volume finite element" method.

In this finite volume interpretation of the finite element scheme, the notion of "consistency of the fluxes" does not appear. This notion of consistency, however, seems to be an interesting tool in the study of the "classical" finite volume schemes.

Note that the (discrete) maximum principle is satisfied with this scheme if only if the transmissibilities $\tau_{K|L}$ are nonnegative (for all $K, L \in \mathcal{V}$ with $K \in \Omega$) ; this is the case under the classical Delaunay condition; this condition states that the (interior of the) circumscribed circle (or sphere in the three dimensional case) of any triangle (tetrahedron in the three dimensional case) of the mesh does not contain any element of $\mathcal{V}$. This is equivalent, in the case of two dimensional triangular meshes, to the fact that the sum of two opposite angles facing a common edge is less or equal π.

12.2 Classical finite volumes on a dual mesh

Let $\mathcal{M}$ be a mesh of Ω ($\mathcal{M}$ may consist of triangles, but it is not necessary) and $\mathcal{V}$ be the set of vertices of $\mathcal{M}$. In order to associate to each vertex (of $\mathcal{M}$) a control volume (such that the whole spatial domain is the "disjoint union" of the control volumes), a possibility is to construct a "dual mesh" which will be denoted by $\mathcal{T}$. In order for this mesh to be admissible in the sense of Definition 9.1 page 37, a

simple way is to use the Voronoï mesh defined with $\mathcal{V}$ (see Example 9.2 page 39). For a description of the Delaunay-Voronoï discretization and its use for covolume methods, we refer to [115] (and references therein). In order to write the "classical" finite volume scheme with this mesh (see (9.20)-(9.23) page 42), a slight modification is necessary at the boundary for some particular $\mathcal{M}$ (see Example 9.2); this method is denoted CFV/DM (classical finite volume on dual mesh); it is conservative, the numerical fluxes are consistent, and the transmissibilities are nonnegative. Hence, the convergence results and error estimates which were studied in previous sections hold (see, in particular, theorems 9.1 page 45 and 9.3 page 52).

A case of particular interest is found when the primal mesh (that is $\mathcal{M}$) consists in triangles with acute angles. One uses, as dual mesh, the Voronoï mesh defined with $\mathcal{V}$. Then, the dual mesh is admissible in the sense of Definition 9.1 page 37 and is constructed with the orthogonal bisectors of the edges of the elements of $\mathcal{M}$, parts of these orthogonal bisectors (and parts of $\partial\Omega$) give the boundaries to the control volumes of the dual mesh. In this case, the CFV/DM scheme is "close" to the piecewise linear finite element scheme on the primal mesh. Let us elaborate on this point.
For $K \in \mathcal{V}$, let K also denote the control volume (of the dual mesh) associated to K (in the sequel, the notation "K", which denotes either the vertex or the control volume will be used in such a context that it does not yield any confusion) and let φ_K be the shape function associated to the vertex K (in the piecewise linear finite element associated to $\mathcal{M}$). The term $\tau_{K|L}$ (ratio between the length of the edge $K|L$ and the distance between vertices), which is used in the finite volume scheme, verifies

$$\tau_{K|L} = -\int_\Omega \nabla\varphi_K(x)\cdot\nabla\varphi_L(x)dx.$$

Idelsohn S., Onate E., Finite volumes and finite elements: Two good friends, Internat. J. Numer. Methods Engrg. 37 (1994), 33233341 Hughes T.J.R., Engel G., Mazzei L., Larson M.G., The continuous Galerkin method is locally conservative, J. Comput. Phys. 163 (2000), 467488
This wellknown fact may be proven by considering two nodes of the mesh, denoted by $K = x_1$ and $L = x_2$ and the two triangles T and $\tilde{T}$ which share the line segment x_1x_2 as a common edge. Let ϕ_1 and ϕ_2 be the two piecewise linear finite element shape functions respectively associated to the vertices x_1 and x_2. Let us compute

$$\int_\Omega \nabla\phi_1(x)\cdot\nabla\phi_2(x)dx = \int_{T\cup\tilde{T}} \nabla\phi_1(x)\cdot\nabla\phi_2(x)dx.$$

Now let θ_1, θ_2 and θ_3 be the angles at vertices x_1, x_2 and x_3 of T (see Figure 3.4). For $i = 1, 2, 3$, let $\boldsymbol{n}_i$ denote the outward unit normal vector to the side opposite to x_i. Since ϕ_1 (resp. ϕ_2) is a linear function on T, its gradient is constant, and since $\phi_i(x_j) = \delta_{i,j}$, we obtain that $\nabla\phi_i(x) = -\frac{1}{h_i}\boldsymbol{n}_i$ for $i = 1, 2$, where h_i is the height of the triangle with respect to the vertex x_i. Since $\boldsymbol{n}_1\cdot\boldsymbol{n}_2 = -\cos\theta$,

$$-\int_T \nabla\phi_1(x)\cdot\nabla\phi_2(x)dx = |T|\frac{1}{h_1h_2}\cos\theta,$$

where θ is the angle of the triangle at its vertex $M = x_3$. Now the area of the triangle T is equal to $|T| = \frac{1}{2}\sin\theta d(L,M)d(M,K) = \dfrac{h_1h_2}{2\sin\theta}$ Thus,

$$-\int_T \nabla\phi_1(x)\cdot\nabla\phi_2(x)dx = \frac{1}{2\tan\theta}.$$

Let x_T denote the circumcenter of T. Since the triangles $x_1x_Tx_3$, $x_2x_Tx_3$ and $x_2x_Tx_2$ are isoceles, one gets that the angle between the line segment x_1x_T is also θ and therefore, $\tan\theta = \frac{d_{K,L}}{2m_{K,T}}$ where $m_{K,T}$ denotes the distance between $K|L$ and the circumcenter of T, and $d_{K,L}$ denotes the distance between K and L, which is also the length of σ.. Therefore,

$$-\int_T \nabla\phi_1(x)\cdot\nabla\phi_2(x)dx = \frac{m_{K,T}}{d_{K,L}}$$

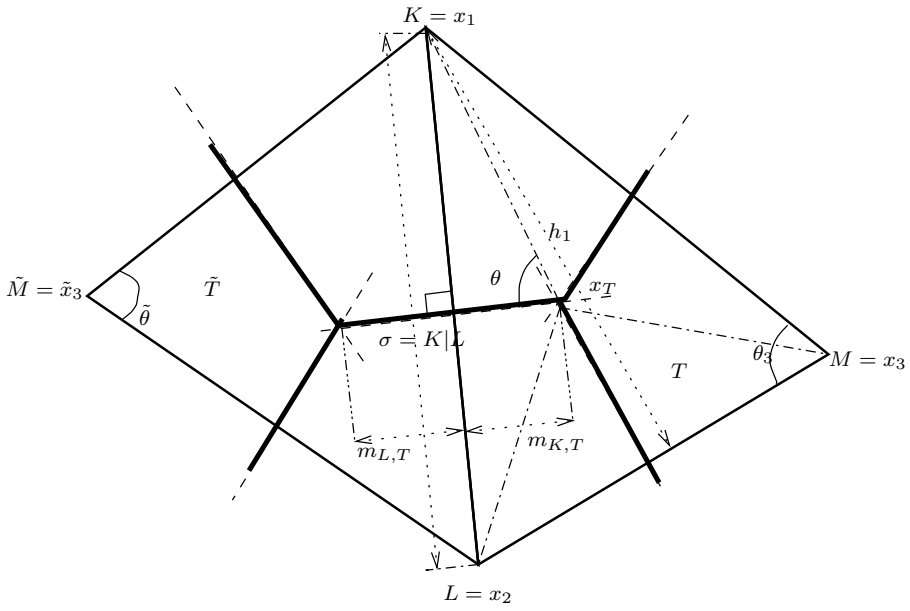


Figure 3.4: Triangular finite element mesh and associate Voronoï cells

The same computation on $\tilde{T}$ yields that:

$$-\int_{\tilde{T}} \nabla\phi_1(x) \cdot \nabla\phi_2(x)dx = \frac{m_{K,\tilde{T}}}{d_{K,L}}$$

if the angles θ and $\tilde{\theta}$ are such that $\theta+\tilde{\theta} < \pi$ (weak Delaunay condition), which, in turn yields the expected result.
The CFV/DM scheme (finite volume scheme on the dual mesh) reads

$$-\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(u_L - u_K) = \int_K f(x)dx, \text{ if } K \in \mathcal{V} \cap \Omega,$$

$$u_K = g(K), \text{ if } K \in \mathcal{V} \cap \partial\Omega,$$

where K stands for an element of $\mathcal{V}$ or for the control volume (of the dual mesh) associated to this point.
The finite element scheme (on the primal mesh) reads

$$-\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(u_L - u_K) = \int_\Omega f(x)\varphi_K(x)dx, \text{ if } K \in \mathcal{V} \cap \Omega,$$

$$u_K = g(K), \text{ if } K \in \mathcal{V} \cap \partial\Omega.$$

Therefore, the only difference between the finite element and the finite volume schemes is in the definition of the right hand sides. Note that these right hand sides may be quite different. Consider for example a node K which is the vertex of four identical triangles featuring an angle of $\frac{\pi}{2}$ at the vertex K, as depicted in Figure 3.5, and denote by a the area of each of these triangles.
Then, for $f \equiv 1$, the right hand side computed for the discrete equation associated to the node K is equal to a in the case of the finite element (piecewise linear finite element) scheme, and equal to $2a$ for the dual mesh finite volume (CFV/DM) scheme. Both schemes may be shown to converge, by using finite volume techniques for the CFV/DM scheme (see previous sections), and finite element techniques for the piecewise linear finite element (see e.g.Ciarlet [29]).

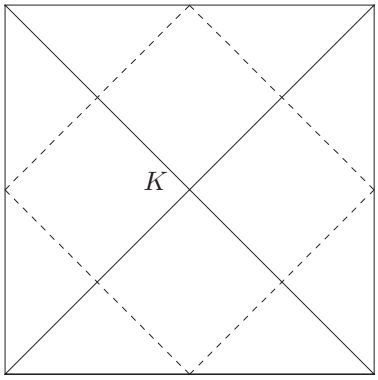
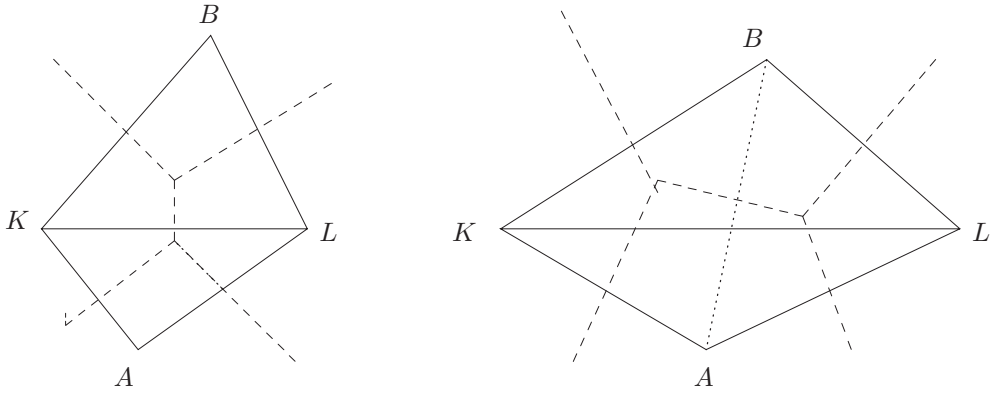


Figure 3.5: An example of a triangular primal mesh (solid line) and a dual Voronoï control volume (dashed line)

Let us now weaken the hypothesis that all angles of the triangles of the primal mesh $\mathcal{M}$ are acute to the so called Delaunay condition and the additional assumption that an angle of an element of $\mathcal{M}$ is less or equal $\pi/2$ if its opposite edge lies on $\partial\Omega$ (see e.g. Vanselow [149]). Under this new assumption the schemes (piecewise linear finite element finite element and CFV/DM with the Voronoï mesh defined with $\mathcal{V}$) still lead to the same transmissibilities and still differ in the definition of the right hand sides.
Recall that the Delaunay condition states that no neighboring element (of $\mathcal{M}$) is included in the circumscribed circle of an arbitrary element of $\mathcal{M}$. This is equivalent to saying that the sum of two opposite angles to an edge is less or equal π. As shown in Figure 3.6, the dual mesh is still admissible in the sense of Definition 9.1 page 37 and is still constructed with the orthogonal bisectors of the edges of the elements of $\mathcal{M}$, parts of these orthogonal bisectors (and parts of $\partial\Omega$) give the boundaries to the control volumes of the dual mesh (see Figure (3.6)) is not the case when $\mathcal{M}$ does not satisfy the Delaunay condition.



Delaunay case Non Delaunay case

Figure 3.6: Construction of the Voronoï dual cells (dashed line) in the case of a triangular primal mesh (solid line) with and without the Delaunay condition

Consider now a primal mesh, $\mathcal{M}$, consisting of triangles, but which does not satisfy the Delaunay condition and let the dual mesh be the Voronoï mesh defined with $\mathcal{V}$. Then, the two schemes, piecewise linear finite element and CFV/DM are quite different. If the Delaunay condition does not hold say between the

angles $\widehat{KAL}$ and $\widehat{KBL}$ (the triplets (K,A,L) and (K,B,L) defining two elements of $\mathcal{M}$), the sum of these two angles is greater than π and the transmissibility $\tau_{K|L} = -\int_\Omega \nabla\varphi_K(x)\cdot\nabla\varphi_L(x)dx$ between the two control volumes associated respectively to K and L becomes negative with the piecewise linear finite element scheme; there is no transmissibility between A and B (since A and B do not belong to a common element of $\mathcal{M}$). Hence the maximum principle is no longer respected for the finite element scheme, while it remains valid for the CFV/DM finite volume scheme. This is due to the fact that the CFV/DM scheme allows an exchange term between A and B, with a positive transmissibility (and leads to no exchange term between K and L), while the finite element scheme does not. Also note also that the common edge to the control volumes (of the dual mesh) associated to A and B is not a part of an orthogonal bisector of an edge of an element of $\mathcal{M}$ (it is a part of the orthogonal bisector of the segment $[A,B]$).

To conclude this section, note that an admissible mesh for the classical finite volume is generally not a dual mesh of a primal triangular mesh consisting of triangles (for instance, the general triangular meshes which are considered in Herbin [84] are not dual meshes to triangular meshes).

12.3 "Finite Volume Finite Element" methods

The "finite volume finite element" method for elliptic problems also uses a dual mesh $\mathcal{T}$ constructed from a finite element primal mesh, such that each cell of $\mathcal{T}$ is associated with a vertex of the primal mesh $\mathcal{M}$. Let $\mathcal{V}$ again denote the set of vertices of $\mathcal{M}$. As in the classical finite volume method, the conservation law is integrated over each cell of the (dual) mesh. Indeed, this integration is performed only if the cell is associated to a vertex (of the primal mesh) belonging to Ω.
Let us consider Problem (12.1). Integrating the conservation law over K_P, where $P\in\mathcal{V}\cap\Omega$ and K_P is the control volume (of the dual mesh) associated to P yields

$$-\int_{\partial K_P}\nabla u(x)\cdot\mathbf{n}_P(x)d\gamma(x) = \int_{K_P} f(x)dx.$$

(Recall that $\mathbf{n}_P$ is the unit normal vector to ∂K_P outward to K_P.) Now, following the idea of finite element methods, the function u is approximated by a Galerkin expansion $\sum_{M\in\mathcal{V}} u_M\varphi_M$, where the functions φ_M are the shape functions of the piecewise linear finite element method. Hence, the discrete unknowns are $\{u_P,\ P\in\mathcal{V}\}$ and the scheme reads

$$-\sum_{M\in\mathcal{V}}\Big(\int_{\partial K_P}\nabla\varphi_M(x)\cdot\mathbf{n}_P(x)d\gamma(x)\Big)u_M = \int_{K_P} f(x)dx,\ \forall P\in\mathcal{V}\cap\Omega, \tag{12.4}$$

$$u_P = g(P),\ \forall P\in\mathcal{V}\cap\partial\Omega.$$

Equations (12.4) may also be written under the conservative form

$$\sum_{Q\in\mathcal{V}} E_{P,Q} = \int_{K_P} f(x)dx,\ \forall P\in\mathcal{V}\cap\Omega, \tag{12.5}$$

$$u_P = g(P),\ \forall P\in\mathcal{V}\cap\partial\Omega, \tag{12.6}$$

where

$$E_{P,Q} = -\sum_{M\in\mathcal{V}}\int_{\partial K_P\cap\partial K_Q}\nabla\varphi_M(x)\cdot\mathbf{n}_P(x)d\gamma(x). \tag{12.7}$$

Note that $E_{Q,P} = -E_{P,Q}$. Unfortunately, the exchange term $E_{P,Q}$ between P and Q is not, in general, a function of the only unknowns u_P and u_Q (this property was used, in the previous sections, to obtain convergence results of finite volume schemes). Another way to write (12.4) is, thanks to (12.3),

$$-\sum_{Q\in\mathcal{V}}\Big(\int_{\partial K_P}\nabla\varphi_Q(x)\cdot\mathbf{n}_P(x)d\gamma(x)\Big)(u_Q-u_P) = \int_{K_P} f(x)dx,\ \forall P\in\mathcal{V}\cap\Omega.$$

Hence a new exchange term from P to Q might be $\bar{E}_{P,Q} = -\Big(\int_{\partial K_P} \nabla\varphi_Q(x)\cdot \mathbf{n}_P(x) d\gamma(x)\Big)(u_Q - u_P)$ and the scheme is therefore conservative if $\bar{E}_{P,Q} = -\bar{E}_{Q,P}$. Unfortunately, this is not the case for a general dual mesh.

There are several ways of constructing a dual mesh from a primal mesh. A common way (see e.g. Fezoui, Lanteri, Larrouturou and Olivier [64])is to take a primal mesh ($\mathcal{M}$) consisting of triangles and to construct the dual mesh with the medians (of the triangles of $\mathcal{M}$), joining the centers of gravity of the triangles to the midpoints of the edges of the primal mesh. The main interest of this way is that the resulting scheme (called FVFE/M below, Finite Volume Finite Element with Medians) is very close to the piecewise linear finite element scheme associated to $\mathcal{M}$. Indeed the FVFE/M scheme is defined by (12.5)-(12.7) while the piecewise linear finite element scheme reads

$$\sum_{Q\in\mathcal{V}} E_{P,Q} = \int_\Omega f(x)\varphi_P(x)dx,\ \forall\, P\in\mathcal{V}\cap\Omega,$$

$$u_P = g(P),\ \forall\, P\in\mathcal{V}\cap\partial\Omega,$$

where $E_{P,Q}$ is defined by (12.7).
These two schemes only differ by the right hand sides and, in fact, these right hand sides are "close" since

$$\mathrm{m}(K_P) = \int_\Omega \varphi_P(x)dx,\ \forall\, P\in\mathcal{V}.$$

This is due to the fact that $\int_T \varphi_P(x)dx = \mathrm{m}(T)/3$ and $\mathrm{m}(K_P\cap T) = \mathrm{m}(T)/3$, for all $T\in\mathcal{M}$ and all vertex P of T.
Thus, convergence properties of the FVFE/M scheme can be proved by using the finite element techniques. Recall however that the piecewise linear finite element scheme (and the FVFE/M scheme) does not satisfy the (discrete) maximum principle if $\mathcal{M}$ does not satisfy the Delaunay condition.

There are other means to construct a dual mesh starting from a primal triangular mesh. One of them is the Voronoï mesh associated to the vertices of the primal mesh, another possibility is to join the centers of gravity; in the latter case, the control volume associated to a vertex, say S, of the primal mesh is then limited by the lines joining the centers of gravity of the neighboring triangles of which S is a vertex (with some convenient modification for the vertices which are on the boundary of Ω). See also Barth [10] for descriptions of dual meshes.

Note that the proof of convergence which we designed for finite volume with admissible meshes does not generalize to any "FVFE" (Finite Volume Finite Element) method for several reasons. In particular, since the exchange term between P and Q (denoted by $E_{P,Q}$) is not, in general, a function of the only unknowns u_P and u_Q (and even if it is the transmissibilities may become negative) and also since, as in the case of the finite element method, the concept of consistency of the fluxes is not clear with the FVFE schemes.

12.4 Generalization to the three dimensional case

The methods described in the three above sections generalize to the three-dimensional case, in particular when the primal mesh is a tetrahedral mesh. With such a mesh, the Delaunay condition no longer ensures the non negativity of the transmissibilities in the case of the piecewise linear finite element method. It is however possible to construct a dual mesh (the "three-dimensional Voronoï" mesh) to a Delaunay triangulation such that the FVFE scheme leads to positive transmissibilities, and therefore such that the maximum principle holds, see Cordes and Putti [38].

Note that the theoretical results (convergence and error estimate) which were shown for the classical finite volume method on an admissible mesh (sections 9.2 page 37 and 10 page 63) still hold for CFV/DM in three-dimensional, since the dual mesh is admissible.

13 Mesh refinement and singularities

Some problems involve singular source terms. In the case of petroleum engineering for instance, one may model (in two space dimensions) the well with a Dirac measure. Other problems may require a better precision of some unknown in certain areas. This section is devoted to the treatment of this kind of problem, either with an adequate treatment of the singularity or by mesh refinement.

13.1 Singular source terms and finite volumes

It is possible to take into account, in the discretization with the finite volume method, the singularities of the solution of an elliptic problem. A common example is the study of wells in petroleum engineering. As a model example we can consider the following problem, which appears, for instance, in the study of a two phase flow in a porous medium. Let B be the ball of $\mathbb{R}^2$ of center 0 and radius r_p (B represents a well of radius r_p). Let $\Omega = (-R, R)^2$ be the whole domain of simulation; r_p is of the order of 10 cm while R can be of the order of 1 km for instance. An approximation to the solution of the following problem is sought:

$$\begin{array}{l} -\mathrm{div}(\nabla u)(x) = 0, \quad x \in \Omega \setminus B, \\ u(x) = P_p, \quad x \in \partial B, \\ \text{“BC”on } \ \partial\Omega, \end{array} \tag{13.1}$$

where “BC” stands for some “smooth” boundary conditions on $\partial\Omega$ (for instance, Dirichlet or Neumann condition). This system is a mathematical model (under convenient assumptions...) of the two phase flow problem, with u representing the pressure of the fluid and P_p an imposed pressure at the well. In order to discretize (13.1) with the finite volume method, a mesh $\mathcal{T}$ of Ω is introduced. For the sake of simplicity, the elements of $\mathcal{T}$ are assumed to be squares of length h (the method is easily generalized to other meshes). It is assumed that the well, represented by B, is located in the middle of one cell, denoted by K_0, so that the origin 0 is the center of K_0. It is also assumed that the mesh size, h, is large with respect to the radius of the well, r_p (which is the case in real applications, where, for instance, h ranges between 10 and 100 m). Following the principle of the finite volume method, one discrete unknown u_K per cell K ($K \in \mathcal{T}$) is introduced in order to discretize the following system:

$$\begin{array}{l} \displaystyle\int_{\partial K} \nabla u(x) \cdot \boldsymbol{n}_K(x) d\gamma(x) = 0, \quad K \in \mathcal{T}, \quad K \neq K_0, \\ \displaystyle\int_{\partial K_0} \nabla u(x) \cdot \boldsymbol{n}_{K_0}(x) d\gamma(x) = \int_{\partial B} \nabla u(x) \cdot \boldsymbol{n}_B(x) d\gamma(x), \end{array} \tag{13.2}$$

where $\boldsymbol{n}_P$ denotes the normal to ∂P, outward to P (with $P = K, K_0$ or B).
Hence, we have to discretize $\nabla u \cdot \boldsymbol{n}_K$ on ∂K (and $\nabla u \cdot \boldsymbol{n}_B$ on ∂B) in terms of $\{u_L, L \in \mathcal{T}\}$ (and “BC” and P_p).
The problems arise in the discretization of $\nabla u \cdot \boldsymbol{n}_{K_0}$ and $\nabla u \cdot n_B$. Indeed, if $\sigma = K|L$ is the common edge to K and L (elements of $\mathcal{T}$), with $K \neq K_0$ and $L \neq K_0$, since the solution of (13.1) is “smooth” enough with respect to the mesh size, except “near” the well, $\nabla u \cdot n_K$ can be discretized by $\frac{1}{h}(u_L - u_K)$ on σ.
In order to discretize ∇u near the well, it is assumed that $\nabla u \cdot n_B$ is constant on ∂B. Let $q(x) = -2\pi r_p \nabla u \cdot n_B$ for $x \in \partial B$ (recall that $\boldsymbol{n}_B$ is the normal to ∂B, outward to B). Then $q \in \mathbb{R}$ is a new unknown, which satisfies

$$\int_{\partial B} -\nabla u \cdot n_B d\gamma(x) = q.$$

Denoting by $|\cdot|$ the euclidian norm in $\mathbb{R}^2$, and u the solution to (13.1), let v be defined by

$$v(x) = \frac{q}{2\pi} \ln(|x|) + u(x), \quad x \in \Omega \setminus B, \tag{13.3}$$

$$v(x) = \frac{q}{2\pi} \ln(r_p) + P_p, \quad x \in B. \tag{13.4}$$

Thanks to the boundary conditions satisfied by u on ∂B, the function v satisfies $-\mathrm{div}(\nabla v) = 0$ on the whole domain Ω, and therefore v is regular on the whole domain Ω. Note that, if we set

$$u(x) = -\frac{q}{2\pi}\ln(|x|) + v(x), \text{ a.e. } x \in \Omega,$$

then

$$-\mathrm{div}(\nabla u) = q\delta_0 \text{ on } \Omega,$$

where δ_0 is the Dirac mass at 0. A discretization of $\nabla u \cdot n_{K_0}$ is now obtained in the following way. Let σ be the common edge to $K_1 \in \mathcal{T}$ and K_0, since v is smooth, it is possible to approximate $\nabla v \cdot \boldsymbol{n}_{K_0}$ on σ by $\frac{1}{h}(v_{K_1} - v_{K_0})$, where v_{K_i} is some approximation of v in K_i (e.g. the value of v at the center of K_i). Then, by (13.4), it is natural to set

$$v_{K_0} = \frac{q}{2\pi}\ln(r_p) + P_p,$$

and by (13.3),

$$v_{K_1} = \frac{q}{2\pi}\ln(h) + u_{K_1}.$$

By (13.3) and from the fact that the integral over σ of $\nabla(\frac{q}{2\pi}\ln(|x|)) \cdot \boldsymbol{n}_{K_0}$ is equal to $\frac{q}{4}$, we find the following approximation for $\int_\sigma \nabla u \cdot n_{K_0} d\gamma$:

$$-\frac{q}{4} + \frac{q}{2\pi}\ln(\frac{h}{r_p}) + u_{K_1} - P_p.$$

The discretization is now complete, there are as many equations as unknowns. The discrete unknowns appearing in the discretized problem are $\{u_K, K \in \mathcal{T}, K \neq K_0\}$ and q. Note that, up to now, the unknown u_{K_0} has not been used. The discrete equations are given by (13.2) where each term of (13.2) is replaced by its approximation in terms of $\{u_K, K \in \mathcal{T}, K \neq K_0\}$ and q. In particular, the discrete equation "associated" to the unknown q is the discretization of the second equation of (13.2), which is

$$\sum_{i=1}^{4}(\frac{q}{2\pi}\ln(\frac{h}{r_p}) + u_{K_i} - P_p) = 0, \tag{13.5}$$

where $\{K_i, i = 1, 2, 3, 4\}$ are the four neighbouring cells to K_0.
It is possible to replace the unknown q by the unknown u_{K_0} (as it is done in petroleum engineering) by setting

$$u_{K_0} = \frac{q}{4} - \frac{q}{2\pi}\ln(\frac{h}{r_p}) + P_p, \tag{13.6}$$

the interest of which is that it yields the usual formula for the discretization of $\nabla u \cdot n_{K_0}$ on σ if σ is the common edge to K_1 and K_0, namely $\frac{1}{h}(u_{K_1} - u_{K_0})$; the discrete equation associated to the unknown u_{K_0} is then (from (13.5))

$$\sum_{i=1}^{4}(u_{K_i} - u_{K_0}) = -q$$

and (13.6) may be written as:

$$q = i_p(P_p - u_{K_0}), \text{ with } i_p = \frac{1}{-\frac{1}{4} + \frac{1}{2\pi}\ln(\frac{h}{r_p})}.$$

This last equation defines i_p, the so called "well-index" in petroleum engineering. With this formula for i_p, the discrete unknowns are now $\{u_K, K \in \mathcal{T}\}$. The discrete equations associated to $\{u_K, K \in \mathcal{T}, K \neq K_0\}$ are given by the first part of (13.2) where each terms of (13.2) is replaced by its approximation in terms of $\{u_K, K \in \mathcal{T}\}$ (using also "BC" on $\partial\Omega$). The discrete equation associated to the unknown u_{K_0} is

$$\sum_{i=1}^{4}(u_{K_i} - u_{K_0}) = -i_p(P_p - u_{K_0}),$$

where $\{K_i, i = 1, 2, 3, 4\}$ are the four neighbouring cells to K_0.
Note that the discrete unknown u_{K_0} is somewhat artificial, it does not really represent the value of u in K_0. In fact, if $x \in K_0$, the "approximate value" of $u(x)$ is $-\frac{q}{2\pi}\ln(\frac{|x|}{r_p}) + P_p$ and $u_{K_0} = \frac{q}{4} - \frac{q}{2\pi}\ln(\frac{h}{r_p}) + P_p$.

13.2 Mesh refinement

Mesh refinement consists in using, in certain areas of the domain, control volumes of smaller size than elsewhere. In the case of triangular grids, a refinement may be performed for instance by dividing each triangle in the refined area into four subtriangles, and those at the boundary of the refined area in two triangles. Then, with some additional technique (e.g. change of diagonal), one may obtain an admissible mesh in the sense of definitions 9.1 page 37, 10.1 page 63 and 11.1 page 79; therefore the error estimates 9.3 page 52, 10.1 page 69 and 11.1 page 81 hold under the same assumptions.

In the case of rectangular grids, the same refining procedure leads to "atypical" nodes and edges, i.e. an edge σ of a given control volume K may be common to two other control volumes, denoted by L and M. This is also true in the triangular case if the triangles of the boundary of the refined area are left untouched.
Let us consider for instance the same problem as in section 9.1 page 33, with the same assumptions and notations, namely the discretization of

$$-\Delta u(x, y) = f(x, y),\ (x, y) \in \Omega = (0, 1) \times (0, 1),$$
$$u(x, y) = 0, (x, y) \in \partial\Omega.$$

It is easily seen that, in this case, if the approximation of the fluxes is performed using differential quotients such as in (9.6) page 34, the fluxes on the "atypical" edge σ cannot be consistent, since the lines joining the centers of K and L and the centers of K and M are not orthogonal to σ. However, the error which results from this lack of consistency can be controlled if the number of atypical edges is not too large.

In the case of rectangular grids (with a refining procedure), denoting by $\mathcal{E}_\dashv$ the set of "atypical" edges of a given mesh $\mathcal{T}$, i.e. edges with separate more than two control volumes, and $\mathcal{T}_\dashv$ the set of "atypical" control volumes, i.e. the control volumes containing an atypical edge in their boundaries; let e_K denote the error between $u(x_K)$ and u_K for each control volume K, and $e_\mathcal{T}$ denote the piecewise constant function defined by $e(x) = e_K$ for any $x \in K$, then one has

$$\|e\|_{L^2(\Omega)} \leq C(\text{size}(\mathcal{T}) + (\sum_{K\in\mathcal{T}_\dashv} \text{m}(K))^{\frac{1}{2}}).$$

The proof is similar to that of Theorem 9.3 page 52. It is detailed in Belmouhoub [11].

14 Compactness results

This section is devoted to some functional analysis results which were used in the previous section. Let Ω be a bounded open set of $\mathbb{R}^d$, $d \geq 1$. Two relative compactness results in $L^2(\Omega)$ for sequences "almost" bounded in $H^1(\Omega)$ which were used in the proof of convergence of the schemes are presented here. Indeed, they are variations of the Rellich theorem (relative compactness in $L^2(\Omega)$ of a bounded sequence in $H^1(\Omega)$ or $H^1_0(\Omega)$). The originality of these results is not the fact that the sequences are relatively compact in $L^2(\Omega)$, which is an immediate consequence of the Kolmogorov theorem (see below), but the fact that the eventual limit, in $L^2(\Omega)$, of the sequence (or of a subsequence) is necessarily in $H^1(\Omega)$ (or in $H^1_0(\Omega)$ for Theorem 14.2), a space which does not contain the elements of the sequence.

We shall make use in this section of the Kolmogorov compactness theorem in $L^2(\Omega)$ which we now recall. The essential part of the proof of this theorem may be found in Brezis [16].

Theorem 14.1 (Kolmogorov compactness lemma) *Let ω be an open bounded set of $\mathbb{R}^N$, $N \geq 1$, $1 \leq q < \infty$ and $A \subset L^q(\omega)$. Then, A is relatively compact in $L^q(\omega)$ if and only if there exists $\{p(u), u \in A\} \subset L^q(\mathbb{R}^N)$ such that*

1. *$p(u) = u$ a.e. on ω, for all $u \in A$,*
2. *$\{p(u), u \in A\}$ is bounded in $L^q(\mathbb{R}^N)$,*
3. *$\|p(u)(\cdot + \eta) - p(u)\|_{L^q(\mathbb{R}^N)} \to 0$ as $\eta \to 0$, uniformly with respect to $u \in A$.*

Let us now state the compactness results used in this chapter.

Theorem 14.2 (Compactness of a bounded sequence and regularity of the limit) *Let Ω be an open bounded set of $\mathbb{R}^d$ with a Lipschitz continuous boundary, $d \geq 1$, and $\{u_n, n \in \mathbb{N}\}$ a bounded sequence of $L^2(\Omega)$. For $n \in \mathbb{N}$, one defines $\tilde{u}_n$ by $\tilde{u}_n = u_n$ a.e. on Ω and $\tilde{u}_n = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$. Assume that there exist $C \in \mathbb{R}$ and $\{h_n, n \in \mathbb{N}\} \subset \mathbb{R}_+$ such that $h_n \to 0$ as $n \to \infty$ and*

$$\|\tilde{u}_n(\cdot + \eta) - \tilde{u}_n\|^2_{L^2(\mathbb{R}^d)} \leq C|\eta|(|\eta| + h_n), \forall n \in \mathbb{N},\ \forall \eta \in \mathbb{R}^d. \tag{14.1}$$

Then, $\{u_n, n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$. Furthermore, if $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$, then $u \in H^1_0(\Omega)$.

Proof of Theorem 14.2

Since $\{h_n, n \in \mathbb{N}\}$ is bounded, the fact that $\{u_n, n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$ is an immediate consequence of Theorem 14.1, taking $N = d$, $\omega = \Omega$, $q = 2$ and $p(u_n) = \tilde{u}_n$. Then, assuming that $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$, it is only necessary to prove that $u \in H^1_0(\Omega)$. Let us first remark that $\tilde{u}_n \to \tilde{u}$ in $L^2(\mathbb{R}^d)$, as $n \to \infty$, with $\tilde{u} = u$ a.e. on Ω and $\tilde{u} = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$.
Then, for $\varphi \in C^\infty_c(\mathbb{R}^d)$, one has, for all $\eta \in \mathbb{R}^d$, $\eta \neq 0$ and $n \in \mathbb{N}$, using the Cauchy-Schwarz inequality and thanks to (14.1),

$$\int_{\mathbb{R}^d} \frac{(\tilde{u}_n(x + \eta) - \tilde{u}_n(x))}{|\eta|} \varphi(x) dx \leq \frac{\sqrt{C|\eta|(|\eta| + h_n)}}{|\eta|} \|\varphi\|_{L^2(\mathbb{R}^d)},$$

which gives, letting $n \to \infty$, since $h_n \to 0$,

$$\int_{\mathbb{R}^d} \frac{(\tilde{u}(x + \eta) - \tilde{u}(x))}{|\eta|} \varphi(x) dx \leq \sqrt{C} \|\varphi\|_{L^2(\mathbb{R}^d)},$$

and therefore, with a trivial change of variables in the integration,

$$\int_{\mathbb{R}^d} \frac{(\varphi(x - \eta) - \varphi(x))}{|\eta|} \tilde{u}(x) dx \leq \sqrt{C} \|\varphi\|_{L^2(\mathbb{R}^d)}. \tag{14.2}$$

Let $\{e_i, i = 1, \ldots, d\}$ be the canonical basis of $\mathbb{R}^d$. For $i \in \{1, \ldots, d\}$ fixed, taking $\eta = he_i$ in (14.2) and letting $h \to 0$ (with $h > 0$, for instance) leads to

$$-\int_{\mathbb{R}^d} \frac{\partial \varphi(x)}{\partial x_i} \tilde{u}(x) dx \leq \sqrt{C} \|\varphi\|_{L^2(\mathbb{R}^d)},$$

for all $\varphi \in C^\infty_c(\mathbb{R}^d)$.
This proves that $D_i \tilde{u}$ (the derivative of $\tilde{u}$ with respect to x_i in the sense of distributions) belongs to $L^2(\mathbb{R}^d)$, and therefore that $\tilde{u} \in H^1(\mathbb{R}^d)$. Since u is the restriction of $\tilde{u}$ on Ω and since $\tilde{u} = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$, therefore $u \in H^1_0(\Omega)$. This completes the proof of Theorem 14.2. ∎

Remark 14.1 (Direct proof of the regularity of the limit) *In fact, the proof that a possible limit is in H_0^1 can be directly drawn from the boundedness of the sequence of approximate solutions in the H^1 discrete norm. Let us detail this point.*
Let Ω be an open bounded set of $\mathbb{R}^d$ with a Lipschitz continuous boundary, $d \geq 1$. Let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a sequence of admissible meshes (in fact, the orthogonality condition is not required for this direct proof) such that size$(\mathcal{T}_n)$ *tends to 0, and let $\{u_n \in X_{\mathcal{T}_n},\ n \in \mathbb{N}\}$ be a sequence of functions of $L^2(\Omega)$ weakly converging to u. Let us assume that there exists a real number C not depending on n such that $\|u_n\|_{1,\mathcal{T}_n} \leq C$. Then $u \in H_0^1(\Omega)$.*
Indeed, let $\boldsymbol{\varphi} \in (C_c^\infty(\mathbb{R}^d))^d$ (note that $\boldsymbol{\varphi}$ does not vanish on the boundary of Ω); as in Step 4 of the proof of Theorem 14.2, let us define for $n \in \mathbb{N}$, $\tilde{u}_n$ by $\tilde{u}_n = u_n$ a.e. on Ω and $\tilde{u}_n = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$.
Then

$$\begin{aligned}\int_{\mathbb{R}^d} \tilde{u}_n \mathrm{div}\boldsymbol{\varphi} dx &= \sum_{K\in\mathcal{T}} u_K \int_{\partial K} \boldsymbol{\varphi}\cdot\boldsymbol{n} d\gamma \\ &\leq \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)|u_K - u_L||\varphi_\sigma| + \sum_{\substack{\sigma\in\mathcal{E}_{\rm ext}\\ \sigma\in\mathcal{E}_K}} \mathrm{m}(\sigma)|u_K||\varphi_\sigma|\end{aligned}$$

where φ_σ is the mean value of $\boldsymbol{\varphi}\cdot\boldsymbol{n}$ over σ. By the Cauchy-Schwarz inequality, we obtain

$$\int_{\mathbb{R}^d} \tilde{u}_n \mathrm{div}\boldsymbol{\varphi} dx \leq \|u\|_{1,\mathcal{T}_n} \left(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma \varphi_\sigma^2\right)^{1/2}.$$

Defining by $\widetilde{\varphi}_\sigma$ the mean value of $\boldsymbol{\varphi}\cdot\boldsymbol{n}$ over D_σ, we get

$$|\widetilde{\varphi}_\sigma - \varphi_\sigma| \leq \mathrm{size}(\mathcal{T}_n)\|\boldsymbol{\varphi}\|_{1,\infty},$$

where $\|\boldsymbol{\varphi}\|_{1,\infty}$ is a bound of the first derivatives of $\boldsymbol{\varphi}$. Therefore we get

$$\left(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma \varphi_\sigma^2\right)^{1/2} \leq \left(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma \widetilde{\varphi}_\sigma^2\right)^{1/2} + \sqrt{d\mathrm{m}(\Omega)}\mathrm{size}(\mathcal{T}_n)\|\boldsymbol{\varphi}\|_{1,\infty}.$$

From the above results, we may write

$$\int_{\mathbb{R}^d} \tilde{u}_n \mathrm{div}\boldsymbol{\varphi} dx \leq \sqrt{d}\ \|u\|_{1,\mathcal{T}_n} |||\boldsymbol{\varphi}|||_{L^2(\Omega)} + \sqrt{d\mathrm{m}(\Omega)}\mathrm{size}(\mathcal{T}_n)\|\boldsymbol{\varphi}\|_{1,\infty}.$$

Passing to the limit, we then get that

$$\lim_{n\to+\infty} \int_{\mathbb{R}^d} \tilde{u}_n \mathrm{div}\boldsymbol{\varphi} dx \leq C|||\boldsymbol{\varphi}|||_{L^2(\Omega)}.$$

This in turn shows that that $\nabla u \in (L^2(\mathbb{R}^d))^d$ and $u \in H_0^1(\Omega)$.

Theorem 14.3 *Let Ω be an open bounded set of $\mathbb{R}^d$, $d \geq 1$, and $\{u_n,\ n \in \mathbb{N}\}$ a bounded sequence of $L^2(\Omega)$. For $n \in \mathbb{N}$, one defines $\tilde{u}_n$ by $\tilde{u}_n = u_n$ a.e. on Ω and $\tilde{u}_n = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$. Assume that there exist $C \in \mathbb{R}$ and $\{h_n,\ n \in \mathbb{N}\} \subset \mathbb{R}_+$ such that $h_n \to 0$ as $n \to \infty$ and such that*

$$\|\tilde{u}_n(\cdot+\eta) - \tilde{u}_n\|^2_{L^2(\mathbb{R}^d)} \leq C|\eta|, \forall n \in \mathbb{N},\ \forall \eta \in \mathbb{R}^d, \tag{14.3}$$

and, for all compact $\bar{\omega} \subset \Omega$,

$$\|u_n(\cdot+\eta) - u_n\|^2_{L^2(\bar{\omega})} \leq C|\eta|(|\eta| + h_n), \forall n \in \mathbb{N},\ \forall \eta \in \mathbb{R}^d,\ |\eta| < d(\bar{\omega}, \Omega^c). \tag{14.4}$$

(The distance between $\bar{\omega}$ and $\mathbb{R}^d \setminus \Omega$ is denoted by $d(\bar{\omega}, \Omega^c)$.)
Then $\{u_n,\ n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$. Furthermore, if $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$, then $u \in H^1(\Omega)$.

PROOF of Theorem 14.3

The proof is very similar to that of Theorem 14.2. Using assumption 14.3, Theorem 14.1 yields that $\{u_n, n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$. Assuming now that $u_n \to u$ in $L^2(\Omega)$, as $n \to \infty$, one has to prove that $u \in H^1(\Omega)$.

Let $\varphi \in C_c^\infty(\Omega)$ and $\varepsilon > 0$ such that $\varphi(x) = 0$ if the distance from x to $\mathbb{R}^d \setminus \Omega$ is less than ε. Assumption 14.4 yields

$$\int_\Omega \frac{(u_n(x+\eta) - u_n(x))}{|\eta|}\varphi(x)dx \leq \frac{\sqrt{C|\eta|(|\eta| + h_n)}}{|\eta|}\|\varphi\|_{L^2(\Omega)},$$

for all $\eta \in \mathbb{R}^d$ such that $0 < |\eta| < \varepsilon$.

From this inequality, it may be proved, as in the proof of Theorem 14.2 (letting $n \to \infty$ and using a change of variables in the integration),

$$\int_\Omega \frac{(\varphi(x-\eta) - \varphi(x))}{|\eta|}u(x)dx \leq \sqrt{C}\|\varphi\|_{L^2(\Omega)},$$

for all $\eta \in \mathbb{R}^d$ such that $0 < |\eta| < \varepsilon$.

Then, taking $\eta = he_i$ and letting $h \to 0$ (with $h > 0$, for instance) one obtains, for all $i \in \{1, \ldots, d\}$,

$$-\int_\Omega \frac{\partial\varphi(x)}{\partial x_i}u(x)dx \leq \sqrt{C}\|\varphi\|_{L^2(\Omega)},$$

for all $\varphi \in C_c^\infty(\Omega)$.

This proves that $D_i u$ (the derivative of u with respect to x_i in the sense of distributions) belongs to $L^2(\Omega)$, and therefore that $u \in H^1(\Omega)$. This completes the proof of Theorem 14.3. ∎

Chapter 4

Parabolic equations

15 Introduction

The aim of this chapter is the study of finite volume schemes applied to a class of linear or nonlinear parabolic problems. We consider the following transient diffusion-convection equation:

$$u_t(x,t) - \Delta\varphi(u)(x,t) + \mathrm{div}(\mathbf{v}u)(x,t) + bu(x,t) = f(x,t),\ x \in \Omega,\ t \in (0,T), \tag{15.1}$$

where Ω is an open polygonal bounded subset of $\mathbb{R}^d$, with $d = 2$ or $d = 3$, $T > 0$, $b \geq 0$, $\mathbf{v} \in \mathbb{R}^d$ is, for the sake of simplicity, a constant velocity field, f is a function defined on $\Omega \times \mathbb{R}_+$ which represents a volumetric source term. The function φ is a nondecreasing Lipschitz continuous function, which arises in the modelling of general diffusion processes. A simplified version of Stefan's problem may be expressed with the formulation (15.1) where φ is a continuous piecewise linear function, which is constant on an interval. The porous medium equation is also included in equation (15.1), with $\varphi(u) = u^m$, $m > 1$. However, the linear case, i.e. $\varphi(u) = u$, is of full interest and the error estimate of section 17 will be given in such a case. In section 18 page 104, we study the convergence of the explicit and of the implicit Euler scheme for the nonlinear case with $\mathbf{v} = 0$ and $b = 0$.

Remark 15.1 One could also consider a nonlinear convection term of the form $\mathrm{div}(\mathbf{v}\psi(u))(x,t)$ where $\psi \in C^1(\mathbb{R},\mathbb{R})$. Such a nonlinear convection term will be largely studied in the framework of nonlinear hyperbolic equations (chapters 5 and 6) and we restrain here to a linear convection term for the sake of simplicity.

An initial condition is given by

$$u(x,0) = u_0(x),\ x \in \Omega. \tag{15.2}$$

Let $\partial\Omega$ denote the boundary of Ω, and let $\partial\Omega_d \subset \partial\Omega$ and $\partial\Omega_n \subset \partial\Omega$ such that $\partial\Omega_d \cup \partial\Omega_n = \partial\Omega$ and $\partial\Omega_d \cap \partial\Omega_n = \emptyset$. A Dirichlet boundary condition is specified on $\partial\Omega_d \subset \partial\Omega$. Let g be a real function defined on $\partial\Omega_d \times \mathbb{R}_+$, the Dirichlet boundary condition states that

$$u(x,t) = g(x,t),\ x \in \partial\Omega_d,\ t \in (0,T). \tag{15.3}$$

A Neumann boundary condition is given with a function $\tilde{g}$ defined on $\partial\Omega_n \times \mathbb{R}_+$:

$$-\nabla\varphi(u)(x,t) \cdot \mathbf{n}(x) = \tilde{g}(x,t),\ x \in \partial\Omega_n,\ t \in (0,T), \tag{15.4}$$

where $\mathbf{n}$ is the unit normal vector to $\partial\Omega$, outward to Ω.

Remark 15.2 Note that, formally, $\Delta\varphi(u) = \mathrm{div}(\varphi'(u)\nabla u)$. Then, if $\varphi'(u)(x,t) = 0$ for some $(x,t) \in \Omega \times (0,T)$, the diffusion coefficient vanishes, so that Equation (15.1) is a "degenerate" parabolic equation. In this case of degeneracy, the choice of the boundary conditions is important in order for the problem to be well-posed. In the case where φ' is positive, the problem is always parabolic.

In the next section, a finite volume scheme for the discretization of (15.1)-(15.4) is presented. An error estimate in the linear case (that is $\varphi(u) = u$) is given in section 17. Finally, a nonlinear (and degenerate) case is studied in section 18; a convergence result is given for subsequences of sequences of approximate solutions, and, when the weak solution is unique, for the whole set of approximate solutions. A uniqueness result is therefore proved for the case of a smooth boundary.

16 Meshes and schemes

In order to perform a finite volume discretization of system (15.1)-(15.4), admissible meshes are used in a similar way to the elliptic cases. Let $\mathcal{T}$ be an admissible mesh of Ω in the sense of Definition 9.1 page 37 with the additional assumption that any $\sigma \in \mathcal{E}_{\rm ext}$ is included in the closure of $\partial\Omega_d$ or included in the closure of $\partial\Omega_n$. The time discretization may be performed with a variable time step; in order to simplify the notations, we shall choose a constant time step $k \in (0,T)$. Let $N_k \in \mathbb{N}^\star$ such that $N_k = \max\{n \in \mathbb{N}, nk < T\}$, and we shall denote $t_n = nk$, for $n \in \{0,\ldots,N_k+1\}$. Note that with a variable time step, error estimates and convergence results similar to that which are given in the next sections hold.

Denote by $\{u_K^n,\ K \in \mathcal{T},\ n \in \{0,\ldots,N_k+1\}\}$ the discrete unknowns; the value u_K^n is an expected approximation of $u(x_K, nk)$.
In order to obtain the numerical scheme, let us integrate formally Equation (15.1) over each control volume K of $\mathcal{T}$, and time interval $(nk,(n+1)k)$, for $n \in \{0,\ldots,N_k\}$:

$$\begin{array}{l}\displaystyle\int_K (u(x,t_{n+1}) - u(x,t_n))dx - \int_{nk}^{(n+1)k}\int_{\partial K} \nabla\varphi(u)(x,t)\cdot\mathbf{n}_K(x)d\gamma(x)dt + \\ \displaystyle\int_{nk}^{(n+1)k}\int_{\partial K} \mathbf{v}\cdot\mathbf{n}_K(x)u(x,t)d\gamma(x)dt + b\int_{nk}^{(n+1)k}\int_K u(x,t)dxdt = \int_{nk}^{(n+1)k}\int_K f(x,t)dxdt.\end{array} \tag{16.1}$$

where $\mathbf{n}_K$ is the unit normal vector to ∂K, outward to K.

Recall that, as usual, the stability condition for an explicit discretization of a parabolic equation requires the time step to be limited by a power two of the space step, which is generally too strong a condition in terms of computational cost. Hence the choice of an implicit formulation in the left hand side of (16.1) which yields

$$\begin{array}{l}\displaystyle\frac{1}{k}\int_K (u(x,t_{n+1}) - u(x,t_n))dx - \int_{\partial K} \nabla\varphi(u)(x,t_{n+1})\cdot\mathbf{n}_K(x)d\gamma(x) + \\ \displaystyle\int_{\partial K} \mathbf{v}\cdot\mathbf{n}_K(x)u(x,t_{n+1})d\gamma(x) + b\int_K u(x,t_{n+1})dxdt = \frac{1}{k}\int_{nk}^{(n+1)k}\int_K f(x,t)dxdt,\end{array} \tag{16.2}$$

There now remains to replace in Equation (16.1) each term by its approximation with respect to the discrete unknowns (and the data). Before doing so, let us remark that another way to obtain (16.2) is to integrate (in space) formally Equation (15.1) over each control volume K of $\mathcal{T}$, at time $t \in (0,T)$. This gives

$$\begin{array}{l}\displaystyle\int_K u_t(x,t)dx - \int_{\partial K} \nabla\varphi(u)(x,t)\cdot\mathbf{n}_K(x)d\gamma(x) + \\ \displaystyle\int_{\partial K} \mathbf{v}\cdot\mathbf{n}_K(x)u(x,t)d\gamma(x) + b\int_K u(x,t)dx = \int_K f(x,t)dx.\end{array} \tag{16.3}$$

An implicit time discretization is then obtained by taking $t = t_{n+1}$ in the left hand side of (16.3), and replacing $u_t(x, t_{n+1})$ by $(u(x, t_{n+1}) - u(x, t_n))/k$. For the right hand side of (16.3) a mean value of f between t_n and t_{n+1} may be used. This gives (16.2). It is also possible to take $f(x, t_{n+1})$ in the right hand side of (16.3). This latter choice is simpler for the proof of some error estimates (see Section 17).

Writing the approximation of the various terms in Equation (16.2) with respect to the discrete unknowns (namely, $\{u_K^n,\, K \in \mathcal{T},\, n \in \{0, \ldots, N_k + 1\}\}$) and taking into account the initial and boundary conditions yields the following implicit finite volume scheme for the discretization of (15.1)-(15.4), using the same notations and introducing some auxiliary unknowns as in Chapter 3 (see equations (9.20)-(9.23) page 42):

$$\begin{array}{l} \mathrm{m}(K)\dfrac{u_K^{n+1} - u_K^n}{k} + \displaystyle\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}^{n+1} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+}^{n+1} + \mathrm{m}(K) b u_K^{n+1} = \mathrm{m}(K) f_K^n, \\ \forall K \in \mathcal{T},\, \forall n \in \{0, \ldots, N_k\}, \end{array} \tag{16.4}$$

with

$$d_{K,\sigma} F_{K,\sigma}^n = -\mathrm{m}(\sigma)\Big(\varphi(u_\sigma^n) - \varphi(u_K^n)\Big) \text{ for } \sigma \in \mathcal{E}_K, \text{ for } n \in \{1, \ldots, N_k + 1\}, \tag{16.5}$$

$$F_{K,\sigma}^n = -F_{L,\sigma}^n \text{ for all } \sigma \in \mathcal{E}_{\mathrm{int}} \text{ such that } \sigma = K|L, \text{ for } n \in \{1, \ldots, N_k + 1\}, \tag{16.6}$$

$$F_{K,\sigma}^n = \frac{1}{k} \int_{(n-1)k}^{nk} \int_\sigma \tilde{g}(x,t) d\gamma(x) dt \quad \text{for } \sigma \in \mathcal{E}_K \text{ such that } \sigma \subset \partial\Omega_n, \text{ for } n \in \{1, \ldots, N_k + 1\}, \tag{16.7}$$

and

$$u_\sigma^n = g(y_\sigma, nk) \quad \text{for } \sigma \subset \partial\Omega_d, \text{ for } n \in \{1, \ldots, N_k + 1\}, \tag{16.8}$$

The upstream choice for the convection term is performed as in the elliptic case (see page 41, recall that $v_{K,\sigma} = \mathrm{m}(\sigma)\mathbf{v}.\mathbf{n}_{K,\sigma}$),

$$u_{\sigma,+}^n = \begin{cases} u_K^n, \text{ if } \mathbf{v} \cdot \mathbf{n}_{K,\sigma} \geq 0, \\ u_L^n, \text{ if } \mathbf{v} \cdot \mathbf{n}_{K,\sigma} < 0, \end{cases} \quad \text{for all } \sigma \in \mathcal{E}_{\mathrm{int}} \text{ such that } \sigma = K|L, \tag{16.9}$$

$$u_{\sigma,+}^n = \begin{cases} u_K^n, & \text{if } \mathbf{v} \cdot \mathbf{n}_{K,\sigma} \geq 0, \\ u_\sigma^n & \text{if } \mathbf{v} \cdot \mathbf{n}_{K,\sigma} < 0, \end{cases} \quad \text{for all } \sigma \in \mathcal{E}_K \text{ such that } \sigma \subset \partial\Omega. \tag{16.10}$$

Note that, in the same way as in the elliptic case, the unknowns u_σ^{n+1} may be eliminated using (16.5)-(16.8). There remains to define the right hand side, which may be defined by:

$$f_K^n = \frac{1}{k\ \mathrm{m}(K)} \int_{nk}^{(n+1)k} \int_K f(x,t) dx dt, \forall K \in \mathcal{T},\, \forall n \in \{0, \ldots, N_k\}, \tag{16.11}$$

or by:

$$f_K^n = \frac{1}{\mathrm{m}(K)} \int_K f(x, t_{n+1}) dx, \forall K \in \mathcal{T},\, \forall n \in \{0, \ldots, N_k\}. \tag{16.12}$$

Initial conditions can be taken into account by different ways, depending on the regularity of the data u_0. For example, it is possible to take

$$u_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x) dx,\ K \in \mathcal{T}, \tag{16.13}$$

or

$$u_K^0 = u_0(x_K), K \in \mathcal{T}. \tag{16.14}$$

Remark 16.1 It is not obvious to prove that the implicit finite volume scheme (16.4)-(16.10) (with (16.11) or (16.12) and (16.13) or (16.14)) has a solution. Once the unknowns $F_{K,\sigma}^{n+1}$ are eliminated, a nonlinear system of equations has to be solved. A proof of the existence and uniqueness of a solution to this system is proved in the next section for the linear case, and is sketched in Remark 18.4 for the nonlinear case.

Remark 16.2 (Comparison with finite differences) Let us consider the case of the heat equation, that is the case where $\mathbf{v} = 0$, $b = 0$, $\varphi(s) = s$ for all $s \in \mathbb{R}$, with Dirichlet condition on the whole boundary ($\partial\Omega_d = \partial\Omega$). If the mesh consists in rectangular control volumes with constant space step in each direction, then the discretization obtained with the finite volume method gives (as in the case of the Laplace operator), the same scheme than the one obtained with the finite difference method (for which the discretization points are the centers of the elements of $\mathcal{T}$) except at the boundary. In the general nonlinear case, finite difference methods have been used in Attey [6], Kamenomostskaja, S.L. [89] and Meyer [108], for example.

Remark 16.3 (Comparison with mass-lumped finite element) Finite element methods are classically used for elliptic or parabolic problems, see for instance Amiez and Gremaud [2] or Ciavaldini [31]. Let $\mathcal{M}$ be a finite element mesh of Ω, consisting of triangles (see e.g. Ciarlet [29] for the conditions on the triangles), with N internal nodes. A finite element formulation for (15.1), with the implicit Euler scheme in time, yields

$$\frac{1}{k}\Big(\int_\Omega (u^{n+1}(x) - u^n(x))\phi_i(x)dx\Big) + \int_\Omega \nabla u^{n+1}(x)\cdot\nabla\phi_i(x)dx = \int_\Omega f(x,t_{n+1})\phi_i(x)dx,$$

where ϕ_i is the shape function of the finite element basis, associated with node i, for $i = 1,\dots N$. Let us approximate u^n by the following Galerkin expansion:

$$u^{n+1} = \sum_{j=1}^{\overline{N}} u_j^{n+1}\phi_j \text{ and } u^n = \sum_{j=1}^{\overline{N}} u_j^n\phi_j,$$

where $\overline{N}$ is the total number of nodes, and u_j^n is expected to be an approximation of u at time t_n and node j, for all j and n; replacing in the above equation, this yields:

$$\frac{1}{k}\sum_{j=1}^{\overline{N}}\int_\Omega (u_j^{n+1} - u_j^n)\phi_j(x)\phi_i(x)dx + \sum_{j=1}^{\overline{N}}\int_\Omega u_j^{n+1}\nabla\phi_j(x)\cdot\nabla\phi_i(x)dx = \int_\Omega f(x,t_{n+1})\phi_i(x)dx. \quad (16.15)$$

Hence, the finite element formulation yields, at each time step, a linear system of the form $CU^{n+1} + AU^{n+1} = B$ (where $U^{n+1} = (u_1,\dots,u_N)^t$, and A and C are $N\times N$ matrices); this scheme, however, is generally used after a mass-lumping, i.e. by assigning to the diagonal term of C the sum of the coefficients of the corresponding line and setting the extra-diagonal terms to 0, thereby transforming C into a diagonal matrix; we already saw in section 12.1 that the part AU^{n+1} may be seen as a linear system derived from a finite volume formulation over the associated Voronoï cells. With the mass lumping technique, the term $C_{i,i}$ corresponding to the i-th node is in fact equal to the integral $\int_K \phi_i$ of the i-th shape function ϕ_i; since for an element K whose vertices contain the i-th node, one has $\int_K \phi_i = \frac{1}{3}|K|$, therefore the integral $\int_K \phi_i$ is also equal to the area of the so-called Donald dual cell , which is the dual cell around the i-th node obtained by joining the barycenter of each cell around the node to the middle of its edges. The term CU^{n+1} may thus be interpreted as a discretization by a finite volume scheme over this Donald dual cell.

17 Error estimate for the linear case

We consider, in this section, the linear case, $\varphi(s) = s$ for all $s \in \mathbb{R}$, and assume $\partial\Omega_d = \partial\Omega$, i.e. that a Dirichlet boundary condition is given on the whole boundary, in which case Problem (15.1)-(15.4)

becomes

$$u_t(x,t) - \Delta u(x,t) + \mathrm{div}(\mathbf{v}u)(x,t) + bu(x,t) = f(x,t),\ x \in \Omega,\ t \in (0,T),$$

$$u(x,0) = u_0(x),\ x \in \Omega,$$

$$u(x,t) = g(x,t),\ x \in \partial\Omega,\ t \in (0,T);$$

the finite volume scheme (16.4)-(16.10) then becomes, assuming, for the sake of simplicity, that $x_K \in K$ for all $K \in \mathcal{T}$,

$$\mathrm{m}(K)\frac{u_K^{n+1} - u_K^n}{k} + \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}^{n+1} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+}^{n+1} + \mathrm{m}(K) b u_K^{n+1} = \mathrm{m}(K) f_K^n,$$
$$\forall K \in \mathcal{T},\ \forall n \in \{0,\ldots,N_k\}, \quad (17.1)$$

with

$$F_{K,\sigma}^n = -\tau_{K|L}(u_L^n - u_K^n) \text{ for all } \sigma \in \mathcal{E}_{\mathrm{int}} \text{ such that } \sigma = K|L,\ \text{for } n \in \{1,\ldots,N_k+1\}, \quad (17.2)$$

$$F_{K,\sigma}^n = -\tau_\sigma(g(y_\sigma, nk) - u_K^n) \text{ for all } \sigma \in \mathcal{E}_K \text{ such that } \sigma \subset \partial\Omega,\ \text{for } n \in \{1,\ldots,N_k+1\}, \quad (17.3)$$

and

$$\begin{cases} u_{\sigma,+}^n = u_K^n, \text{ if } \mathbf{v}\cdot\mathbf{n}_{K,\sigma} \geq 0, \\ u_{\sigma,+}^n = u_L^n, \text{ if } \mathbf{v}\cdot\mathbf{n}_{K,\sigma} < 0, \end{cases} \text{ for all } \sigma \in \mathcal{E}_{\mathrm{int}} \text{ such that } \sigma = K|L, \quad (17.4)$$

$$\begin{cases} u_{\sigma,+}^n = u_K^n, & \text{if } \mathbf{v}\cdot\mathbf{n}_{K,\sigma} \geq 0, \\ u_{\sigma,+}^n = g(y_\sigma, nk), & \text{if } \mathbf{v}\cdot\mathbf{n}_{K,\sigma} < 0, \end{cases} \text{ for all } \sigma \in \mathcal{E}_K \text{ such that } \sigma \subset \partial\Omega. \quad (17.5)$$

The source term and initial condition f and u_0 are discretized by (16.12) and (16.14).

A convergence analysis of a one-dimensional vertex-centered scheme was performed in Guo and Stynes [79] by writing the scheme in a finite element framework. Here we shall use direct finite volume techniques which also handle the multi-dimensional case.

The following theorem gives an L^∞ estimate (on the approximate solution) and an error estimate. Some easy generalizations are possible (for instance, the same theorem holds with $b < 0$, the only difference is that in the L^∞ estimate (17.6) the bound c also depends on b).

Theorem 17.1 *Let Ω be an open polygonal bounded subset of $\mathbb{R}^d$, $T > 0$, $u \in C^2(\overline{\Omega} \times \mathbb{R}_+, \mathbb{R})$, $b \geq 0$ and $\mathbf{v} \in \mathbb{R}^d$. Let $u_0 \in C^2(\overline{\Omega}, \mathbb{R})$ be defined by $u_0 = u(\cdot, 0)$, let $f \in C^0(\overline{\Omega} \times \mathbb{R}_+, \mathbb{R})$ be defined by $f = u_t - \mathrm{div}(\nabla u) + \mathrm{div}(\mathbf{v}u) + bu$ and $g \in C^0(\partial\Omega \times \mathbb{R}_+, \mathbb{R})$ defined by $g = u$ on $\partial\Omega \times \mathbb{R}_+$. Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 9.1 page 37 and $k \in (0,T)$. Then there exists a unique vector $(u_K)_{K\in\mathcal{T}}$ satisfying (17.1)-(17.5) (or (16.4)-(16.10)) with (16.12) and (16.14). There exists c only depending on u_0, T, f and g such that*

$$\sup\{|u_K^n|, K \in \mathcal{T}, n \in \{1,\ldots,N_k+1\}\} \leq c. \quad (17.6)$$

Furthermore, let $e_K^n = u(x_K, t_n) - u_K^n$, for $K \in \mathcal{T}$ and $n \in \{1,\ldots,N_k+1\}$, and $h = \mathrm{size}(\mathcal{T})$. Then there exists $C \in \mathbb{R}_+$ only depending on b, u, $\mathbf{v}$, Ω and T such that

$$\Big(\sum_{K\in\mathcal{T}} (e_K^n)^2 \mathrm{m}(K)\Big)^{\frac{1}{2}} \leq C(h+k),\ \forall\, n \in \{1,\ldots,N_k+1\}. \quad (17.7)$$

PROOF of Theorem 17.1

For simplicity, let us assume that $x_K \in K$ for all $K \in \mathcal{T}$. Generalization without this condition is straightforward.

(i) Existence, uniqueness, and L^∞ estimate

For a given $n \in \{0, \ldots, N_k\}$, set $f_K^n = 0$ and $u_K^n = 0$ in (17.1), and $g(y_\sigma, (n+1)k) = 0$ for all $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega$. Multiplying (17.1) by u_K^{n+1} and using the same technique as in the proof of Lemma 9.2 page 42 yields that $u_K^{n+1} = 0$ for all $K \in \mathcal{T}$. This yields the uniqueness of the solution $\{u_K^{n+1},\, K \in \mathcal{T}\}$ to (17.1)-(17.5) for given $\{u_K^n,\, K \in \mathcal{T}\}$, $\{f_K^n,\, K \in \mathcal{T}\}$ and $\{g(y_\sigma, (n+1)k),\, \sigma \in \mathcal{E}, \sigma \subset \partial\Omega_d\}$. The existence follows immediately, since (17.1)-(17.5) is a finite dimensional linear system with respect to the unknown $\{u_K^{n+1},\, K \in \mathcal{T}\}$ (with as many unknowns as equations).

Let us now prove the estimate (17.6). Let $m_f = \min\{f(x,t),\, x \in \overline{\Omega},\, t \in [0, 2T]\}$ and $m_g = \min\{g(x,t),\, x \in \partial\Omega,\, t \in [0, 2T]\}$. Let $n \in \{0, \ldots, N_k\}$. Then, we claim that

$$\min\{u_K^{n+1},\, K \in \mathcal{T}\} \geq \min\{\min\{u_K^n,\, K \in \mathcal{T}\} + km_f, 0, m_g\}. \tag{17.8}$$

Indeed, if $\min\{u_K^{n+1},\, K \in \mathcal{T}\} < \min\{0, m_g\}$, let $K_0 \in \mathcal{T}$ such that $u_{K_0}^{n+1} = \min\{u_K^{n+1},\, K \in \mathcal{T}\}$. Let us write (17.1) with $K = K_0$. Since $u_{K_0}^{n+1} < 0$ and $u_{K_0}^{n+1} < m_g$, we get that $F_{K_0,\sigma}^{n+1} \leq 0$. Moreover, since $\boldsymbol{v}$ is constant, we have $\displaystyle\sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} = 0$, so that

$$\sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+}^{n+1} = \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} \left(u_{\sigma,+}^{n+1} - u_{K_0}^{n+1}\right) \leq 0;$$

therefore

$$u_{K_0}^{n+1} \geq u_{K_0}^n + k f_{K_0}^n \geq \min\{u_K^n,\, K \in \mathcal{T}\} + km_f,$$

this proves (17.8), which yields, by induction, that:

$$\min\{u_K^n,\, K \in \mathcal{T}\} \geq \min\{\min\{u_K^0,\, K \in \mathcal{T}\}, 0, m_g\} + nk \min\{m_f, 0\}, \forall\, n \in \{0, \ldots, N_k + 1\}.$$

Similarly,

$$\max\{u_K^n,\, K \in \mathcal{T}\} \leq \max\{\max\{u_K^0,\, K \in \mathcal{T}\}, 0, M_g\} + nk \max\{M_f, 0\}, \forall\, n \in \{0, \ldots, N_k + 1\},$$

with $M_f = \max\{f(x,t),\, x \in \overline{\Omega},\, t \in [0, 2T]\}$ and $M_g = \max\{g(x,t),\, x \in \partial\Omega,\, t \in [0, 2T]\}$.
This proves (17.6) with $c = \|u_0\|_{L^\infty(\Omega)} + \|g\|_{L^\infty(\partial\Omega \times (0,2T))} + 2T\|f\|_{L^\infty(\Omega \times (0,2T))}$.

(ii) Error estimate

As in the stationary case (see the proof of Theorem 9.3 page 52), one uses the regularity of the data and the solution to write an equation for the error $e_K^n = u(x_K, t_n) - u_K^n$, defined for $K \in \mathcal{T}$ and $n \in \{0, \ldots, N_k + 1\}$. Note that $e_K^0 = 0$ for $K \in \mathcal{T}$. Let $n \in \{0, \ldots, N_k\}$. Integrating (in space) Equation (15.1) over each control volume K of $\mathcal{T}$, at time $t = t_{n+1}$, gives, thanks to the choice of f_K^n (see (16.12)),

$$\int_K u_t(x, t_{n+1})dx - \int_{\partial K} \Big(\nabla u(x, t_{n+1}) - \mathbf{v}u(x, t_{n+1})\Big) \cdot \mathbf{n}_K(x) d\gamma(x) + b \int_K u(x, t_{n+1})dx = \mathrm{m}(K) f_K^n. \tag{17.9}$$

Note that, for all $x \in K$ and all $K \in \mathcal{T}$, a Taylor expansion yields, thanks to the regularity of u:

$$u_t(x, t_{n+1}) = (1/k)(u(x_K, t_{n+1}) - u(x_K, t_n)) + s_K^n(x) \text{ with } |s_K^n(x)| \leq C_1(h + k)$$

with some C_1 only depending on u and T. Therefore, defining $S_K^n = \displaystyle\int_K s_K^n(x)dx$, one has: $|S_K^n| \leq C_1 \mathrm{m}(K)(h + k)$.

One follows now the lines of the proof of Theorem 9.3 page 52, adding the terms due to the time derivative u_t. Substracting (17.1) to (17.9) yields

$$\begin{array}{l} \mathrm{m}(K)\dfrac{e_K^{n+1}-e_K^n}{k}+\displaystyle\sum_{\sigma\in\mathcal{E}_K}\Big(G_{K,\sigma}^{n+1}+W_{K,\sigma}^{n+1}\Big)+b\mathrm{m}(K)e_K^{n+1}= \\ b\mathrm{m}(K)\rho_K^n-\displaystyle\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)(R_{K,\sigma}^n+r_{K,\sigma}^n)-S_K^n,\ \forall K\in\mathcal{T}, \end{array} \tag{17.10}$$

where (with the notations of Definition 9.1 page 37),

$$\begin{array}{l} G_{K,\sigma}^{n+1}=-\tau_\sigma(e_L^{n+1}-e_K^{n+1}),\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm int},\ \sigma=K|L, \\ G_{K,\sigma}^{n+1}=\tau_\sigma e_K^{n+1},\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}, \\ W_{K,\sigma}^{n+1}=\mathrm{m}(\sigma)\mathbf{v}\cdot\mathbf{n}_{K,\sigma}(u(x_{\sigma,+},t_{n+1})-u_{\sigma,+}^{n+1}), \end{array}$$

where $x_{\sigma,+}=x_K$ (resp. x_L) if $\sigma\in\mathcal{E}_{\rm int}$, $\sigma=K|L$ and $\mathbf{v}\cdot\mathbf{n}_{K,\sigma}\geq 0$ (resp. $\mathbf{v}\cdot\mathbf{n}_{K,\sigma}<0$) and $x_{\sigma,+}=x_K$ (resp. y_σ) if $\sigma=\mathcal{E}_K\cap\mathcal{E}_{\rm ext}$ and $\mathbf{v}\cdot\mathbf{n}_{K,\sigma}\geq 0$ (resp. $\mathbf{v}\cdot\mathbf{n}_{K,\sigma}<0$),

$$\rho_K^n=u(x_K,t^{n+1})-\frac{1}{\mathrm{m}(K)}\int_K u(x,t_{n+1})dx,$$

$$\mathrm{m}(\sigma)R_{K,\sigma}^n=\tau_\sigma(u(x_K,t^{n+1})-u(x_L,t^{n+1}))+\int_\sigma\nabla u(x,t_n).\mathbf{n}_{K,\sigma}d\gamma(x)\text{ if }\sigma=K|L\in\mathcal{E}_{\rm int},$$

$$\mathrm{m}(\sigma)R_{K,\sigma}^n=\tau_\sigma(u(x_K,t^{n+1})-g(y_\sigma,t^{n+1})+\int_\sigma\nabla u(x,t_n).\mathbf{n}_{K,\sigma}d\gamma(x)\text{ if }\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm int},$$

and

$$\mathrm{m}(\sigma)r_{K,\sigma}^n=\mathbf{v}\cdot\mathbf{n}_{K,\sigma}(\mathrm{m}(\sigma)u(x_{\sigma,+},t_{n+1})-\int_{\rm m}(\sigma)u(x,t_{n+1})d\gamma(x),\ \text{for any }\sigma\in\mathcal{E}.$$

As in Theorem 9.3, thanks to the regularity of u, there exists C_2, only depending on u, $\mathbf{v}$ and T, such that $|R_{K,\sigma}^n|+|r_{K,\sigma}^n|\leq C_2h$ and $|\rho_K^n|\leq C_2h$, for any $K\in\mathcal{T}$ and $\sigma\in\mathcal{E}_K$.

Multiplying (17.10) by e_K^{n+1}, summing for $K\in\mathcal{T}$, and performing the same computations as in the proof of Theorem 9.3 between (9.56) to (9.60) page 54 yields, with some C_3 only depending on u, $\mathbf{v}$, b, Ω and T,

$$\begin{array}{l} \dfrac{1}{k}\displaystyle\sum_{K\in\mathcal{T}}\mathrm{m}(K)(e_K^{n+1})^2+\frac{1}{2}\|e_{\mathcal{T}}^{n+1}\|_{1,\mathcal{T}}^2+\frac{1}{2}b\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2\leq \\ C_3h^2+C_1(h+k)\displaystyle\sum_{K\in\mathcal{T}}\mathrm{m}(K)|e_K^{n+1}|+\frac{1}{k}\sum_{K\in\mathcal{T}}\mathrm{m}(K)e_K^{n+1}e_K^n, \end{array} \tag{17.11}$$

where the second term of the right hand side is due to the bound on S_K^n and where $e_{\mathcal{T}}^{n+1}$ is a piecewise constant function defined by

$$e_{\mathcal{T}}^{n+1}(x)=e_K^{n+1},\ \text{for }x\in K,K\in\mathcal{T}.$$

Inequality (17.11) yields

$$\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2\leq 2kC_3h^2+2kC_1\mathrm{m}(\Omega)(k+h)\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}+\|e_{\mathcal{T}}^n\|_{L^2(\Omega)}^2,$$

which gives

$$\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2\leq\|e_{\mathcal{T}}^n\|_{L^2(\Omega)}^2+C_4\big(kh^2+k(k+h)\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}\big), \tag{17.12}$$

where $C_4\in\mathbb{R}_+$ only depends on u, $\mathbf{v}$, b, Ω and T. Remarking that for $\varepsilon>0$, the following inequality holds:

$$C_4k(k+h)\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)} \le \varepsilon^2\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2 + (1/\varepsilon^2)C_4^2k^2(k+h)^2,$$

taking $\varepsilon^2 = k/(k+1)$, (17.12) yields

$$\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2 \le (1+k)\|e_{\mathcal{T}}^{n}\|_{L^2(\Omega)}^2 + C_4kh^2(1+k) + (1+k)^2C_4^2k(k+h)^2. \tag{17.13}$$

Then, if $\|e_{\mathcal{T}}^{n}\|_{L^2(\Omega)}^2 \le c_n(h+k)^2$, with $c_n \in \mathbb{R}_+$, one deduces from (17.13), using $h \le h+k$ and $k < T$, that

$$\|e_{\mathcal{T}}^{n+1}\|_{L^2(\Omega)}^2 \le c_{n+1}(h+k)^2 \text{ with } c_{n+1} = (1+k)c_n + C_5k \text{ and } C_5 = C_4(1+T) + C_4^2(1+T)^2.$$

(Note that C_5 only depends on u, $\mathbf{v}$, b, Ω and T).
Choosing $c_0 = 0$ (since $\|e_{\mathcal{T}}^{0}\|_{L^2(\Omega)} = 0$), the relation between c_n and c_{n+1} yields (by induction) $c_n \le C_5e^{2kn}$. Estimate (17.7) follows with $C^2 = C_5e^{4T}$. ∎

Remark 17.1 The error estimate given in Theorem 17.1 may be generalized to the case of discontinuous coefficients. The admissibility of the mesh is then redefined so that the data and the solution are piecewise regular on the control volumes as in Definition 11.1 page 79, see also Herbin [85].

18 Convergence in the nonlinear case

18.1 Solutions to the continuous problem

We consider Problem (15.1)-(15.4) with $\mathbf{v} = 0$, $b = 0$, $\partial\Omega_n = \partial\Omega$ and $\tilde{g} = 0$, that is a homogeneous Neumann condition on the whole boundary, in which case the problem becomes

$$u_t(x,t) - \Delta\varphi(u)(x,t) = f(x,t), \quad \text{for } (x,t) \in \Omega \times (0,T), \tag{18.1}$$

with

$$\nabla\varphi(u)(x,t)\cdot\boldsymbol{n}(x) = 0, \quad \text{for } (x,t) \in \partial\Omega \times (0,T), \tag{18.2}$$

and the initial condition

$$u(x,0) = u_0(x), \quad \text{for all } x \in \Omega. \tag{18.3}$$

We suppose that the following hypotheses are satisfied:

Assumption 18.1

(i) Ω *is an open bounded polygonal subset of* $\mathbb{R}^d$ *and* $T > 0$.

(ii) The function $\varphi \in C(\mathbb{R},\mathbb{R})$ *is a nondecreasing locally Lipschitz continuous function.*

(iii) The initial data u_0 *satisfies* $u_0 \in L^\infty(\Omega)$.

(iv) The right hand side f *satisfies* $f \in L^\infty(\Omega \times \mathbb{R}_+^\star)$.

Equation (18.1) is a degenerate parabolic equation. Formally, $\Delta\varphi(u) = \mathrm{div}(\varphi'(u)\nabla u)$, so that, if $\varphi'(u) = 0$, the diffusion coefficient vanishes. Let us give a definition of a weak solution u to Problem (18.1)-(18.3) (the proof of the existence of such a solution is given in Kamenomostskaja, S.L. [89], Ladyženskaja, Solonnikov and Ural'ceva [97], Meirmanov [107], Oleinik [119]).

Definition 18.1 Under Assumption 18.1, a measurable function u is a weak solution of (18.1)-(18.3) if

$$\begin{array}{l} u \in L^\infty(\Omega \times (0,T)), \\ \displaystyle\int_0^T \int_\Omega \Big(u(x,t)\psi_t(x,t) + \varphi(u(x,t))\Delta\psi(x,t) + f(x,t)\psi(x,t) \Big)\, dx\, dt + \\ \displaystyle\int_\Omega u_0(x)\psi(x,0)dx = 0, \text{ for all } \psi \in \mathcal{A}_T, \end{array} \tag{18.4}$$

where $\mathcal{A}_T = \{\psi \in C^{2,1}(\overline{\Omega} \times [0,T]),\ \nabla\psi \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \times [0,T], \text{ and } \psi(\cdot,T) = 0\}$, and $C^{2,1}(\overline{\Omega} \times [0,T])$ denotes the set of functions which are restrictions on $\overline{\Omega} \times [0,T]$ of functions from $\mathbb{R}^d \times \mathbb{R}$ into $\mathbb{R}$ which are twice (resp. once) continuously differentiable with respect to the first (resp. second) variable. (Recall that, as usual, $\mathbf{n}$ is the unit normal vector to $\partial\Omega$, outward to Ω.)

Remark 18.1 It is possible to use a solution in a stronger sense, using only one integration by parts for the space term. It then leads to a larger test function space than $\mathcal{A}_T$.

Remark 18.2 Note that the function u formally satisfies the conservation law

$$\int_\Omega u(x,t)dx = \int_\Omega u_0(x)dx + \int_0^t \int_\Omega f(x,t)dxdt, \tag{18.5}$$

for all $t \in [0,T]$. This property is also satisfied by the finite volume approximation.

18.2 Definition of the finite volume approximate solutions

As in sections 9.2 page 37 and 10.1 page 63, an admissible mesh of Ω is defined, with respect to which a functional space is introduced: this space contains the approximate solutions obtained from the finite volume discretization over the admissible mesh.

Definition 18.2 Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63, $T > 0$, $k \in (0,T)$ and $N_k = \max\{n \in \mathbb{N}; nk < T\}$. Let $X(\mathcal{T},k)$ be the set of functions u from $\Omega \times (0,(N_k+1)k)$ to $\mathbb{R}$ such that there exists a family of real values $\{u_K^n,\ K \in \mathcal{T},\ n \in \{0,\ldots,N_k\}\}$, with $u(x,t) = u_K^n$ for a.e. $x \in K$, $K \in \mathcal{T}$ and for a.e. $t \in [nk,(n+1)k)$, $n \in \{0,\ldots,N_k\}$.

Since we only consider, for the sake of simplicity, a Neumann boundary condition, we can easily eliminate the unknowns $F_{K,\sigma}^n$ located at the edges in equation (16.4) using the equations (16.5), (16.6), and (16.7). An explicit version of the scheme can then be written in the following way:

$$\begin{array}{l} \mathrm{m}(K)\dfrac{u_K^{n+1} - u_K^n}{k} - \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big) = \mathrm{m}(K) f_K^n, \\ \forall K \in \mathcal{T},\ \forall n \in \{0,\ldots,N_k\}. \end{array} \tag{18.6}$$

$$u_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x)dx,\ \forall K \in \mathcal{T}, \tag{18.7}$$

$$f_K^n = \frac{1}{k\ \mathrm{m}(K)} \int_{nk}^{(n+1)k} \int_K f(x,t)dxdt, \forall K \in \mathcal{T},\ \forall n \in \{0,\ldots,N_k\}. \tag{18.8}$$

(Recall that $\tau_{K|L} = \dfrac{\mathrm{m}(K|L)}{d_{K|L}}$, see Definition 10.1 page 63.)

Remark 18.3 The definition using the mean value in (18.7) is motivated by the lack of regularity assumed on the data u_0.

The scheme (18.6)-(18.8) is then used to build an approximate solution, $u_{\mathcal{T},k} \in X(\mathcal{T},k)$ by

$$u_{\mathcal{T},k}(x,t) = u_K^n, \forall x \in K, \forall t \in [nk,(n+1)k), \forall K \in \mathcal{T}, \forall n \in \{0,\ldots,N_k\}. \tag{18.9}$$

Remark 18.4 The implicit finite volume scheme is defined by

$$\begin{aligned} &\mathrm{m}(K)\frac{u_K^{n+1}-u_K^n}{k} - \sum_{L\in\mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^{n+1}) - \varphi(u_K^{n+1})\Big) = \mathrm{m}(K) f_K^n, \\ &\forall K \in \mathcal{T},\ \forall n \in \{0,\ldots,N_k\}. \end{aligned} \tag{18.10}$$

The proof of the existence of u_K^{n+1}, for any $n \in \{0,\ldots,N_k\}$, can be obtained using the following fixed point method:

$$u_K^{n+1,0} = u_K^n, \quad \text{for all } K \in \mathcal{T}, \tag{18.11}$$

and

$$\begin{aligned} &\mathrm{m}(K)\frac{u_K^{n+1,m+1}-u_K^n}{k} - \sum_{L\in\mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^{n+1,m}) - \varphi(u_K^{n+1,m+1})\Big) = \mathrm{m}(K) f_K^n, \\ &\forall K \in \mathcal{T},\ \forall m \in \mathbb{N}. \end{aligned} \tag{18.12}$$

Equation (18.12) gives a contraction property, which leads first to prove that for all $K \in \mathcal{T}$, the sequence $(\varphi(u_K^{n+1,m}))_{m\in\mathbb{N}}$ converges. Then we deduce that $(u_K^{n+1,m})_{m\in\mathbb{N}}$ also converges.
We shall see further that all results obtained for the explicit scheme are also true, with convenient adaptations, for the implicit scheme. The function $u_{\mathcal{T},k}$ is then defined by $u_{\mathcal{T},k}(x,t) = u_K^{n+1}$, for all $x \in K$, for all $t \in [nk,(n+1)k)$.

The mathematical problem is to study, under Assumption 18.1 and with a mesh in the sense of Definition 10.1, the convergence of $u_{\mathcal{T},k}$ to a weak solution of Problem (18.1)-(18.3), when $h = \mathrm{size}(\mathcal{T}) \to 0$ and $k \to 0$. Exactly in the same manner as for the elliptic case, we shall use estimates on the approximate solutions which are discrete versions of the estimates which hold on the solution of the continous problem and which ensure the stability of the scheme. We present the proofs in the case of the explicit scheme and show in several remarks how they can be extended to the case of the implicit scheme (which is significantly easier to study). The proof of convergence of the scheme uses a weak-$\star$ convergence property, as in Ciavaldini [31], which is proved in a general setting in section 18.5 page 116. For the sake of completeness, the proof of uniqueness of the weak solution of Problem (18.1)-(18.3) is given for the case of a regular boundary; this allows to prove that the whole sequence of approximate solutions converges to the weak solution of problem (18.1)-(18.3), in which case an admissible mesh for a smooth domain can easily be defined (see Definition 18.4 page 116).

18.3 Estimates on the approximate solution

Maximum principle

Lemma 18.1 *Under Assumption 18.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63 and $k \in (0,T)$. Let $U = \|u_0\|_{L^\infty(\Omega)} + T\|f\|_{L^\infty(\Omega\times(0,T))}$, $B = \sup_{-U\le x<y\le U} \dfrac{\varphi(x)-\varphi(y)}{x-y}$. Assume that the condition*

$$k \le \frac{\mathrm{m}(K)}{B\displaystyle\sum_{L\in\mathcal{N}(K)} \tau_{K|L}}, \quad \textit{for all } K \in \mathcal{T}, \tag{18.13}$$

is satisfied. Then the function $u_{\mathcal{T},k}$ defined by (18.6)-(18.9) verifies

$$\|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times(0,T))} \le U. \tag{18.14}$$

Proof of Lemma 18.1

Let $n \in \{0,\dots,N_k-1\}$ and assume $u_K^n \in [-U,+U]$ for all $K \in \mathcal{T}$.
Let $K \in \mathcal{T}$, Equation (18.6) can be written as

$$\begin{aligned} u_K^{n+1} = \ & \Big(1 - \frac{k}{\mathrm{m}(K)} \sum_{L\in\mathcal{N}(K)} \tau_{K|L} \frac{\varphi(u_L^n)-\varphi(u_K^n)}{u_L^n - u_K^n}\Big) u_K^n + \\ & \frac{k}{\mathrm{m}(K)} \sum_{L\in\mathcal{N}(K)} \Big(\tau_{K|L} \frac{\varphi(u_L^n)-\varphi(u_K^n)}{u_L^n - u_K^n}\Big) u_L^n + k f_K^n, \end{aligned}$$

with the convention that $\dfrac{\varphi(u_L^n)-\varphi(u_K^n)}{u_L^n-u_K^n} = 0$ if $u_L^n - u_K^n = 0$.
Thanks to the condition (18.13) and since φ is nondecreasing, the following inequality can be deduced:

$$|u_K^{n+1}| \le \sup_{L\in\mathcal{T}} |u_L^n| + k\|f\|_{L^\infty(\Omega\times(0,T))}.$$

Then, since K is arbitrary in $\mathcal{T}$,

$$\sup_{K\in\mathcal{T}} |u_K^{n+1}| \le \sup_{L\in\mathcal{T}} |u_L^n| + k\|f\|_{L^\infty(\Omega\times(0,T))}. \tag{18.15}$$

Using (18.15), an induction on n yields, for $n \in \{0,\dots,N_k\}$, $\sup_{K\in\mathcal{T}} |u_K^n| \le \|u_0\|_{L^\infty(\Omega)} + nk\|f\|_{L^\infty(\Omega\times(0,T))}$, which leads to Inequality (18.14) since $N_k k \le T$. ∎

Remark 18.5 Assume that there exist α, β, $\gamma \in \mathbb{R}_+^\star$ such that $\mathrm{m}(K) \ge \alpha h^d$, $\mathrm{m}(\partial K) \le \beta h^{d-1}$, for all $K \in \mathcal{T}$, and $d_{K|L} \ge \gamma h$, for all $K|L \in \mathcal{E}_{\rm int}$ (recall that $h = \mathrm{size}(\mathcal{T})$). Then, $k \le Ch^2$ with $C = (\alpha\gamma)/(B\beta)$ yields (18.13).

Remark 18.6 Let $(\mathcal{T}_n, k_n)_{n\in\mathbb{N}}$ be a sequence of admissible meshes and time steps, and $(u_{\mathcal{T}_n,k_n})_{n\in\mathbb{N}}$ the associated sequence of approximate finite volume solutions; then , thanks to (18.14), there exists a function $u \in L^\infty(\Omega\times(0,T))$ and a subsequence of $(u_{\mathcal{T}_n,k_n})_{n\in\mathbb{N}}$ which converges to u for the weak-$\star$ topology of $L^\infty(\Omega\times(0,T))$.

Remark 18.7 Estimate (18.14) is also true, with $U = \|u_0\|_{L^\infty(\Omega)} + 2T\|f\|_{L^\infty(\Omega\times(0,2T))}$, for the implicit scheme, because the fixed point method guarantees (18.15) (with $\|f\|_{L^\infty(\Omega\times(0,2T))}$ instead of $\|f\|_{L^\infty(\Omega\times(0,T))}$ and until $n = N_k$), without any condition on k.

Space translates of approximate solutions

Let us now define a seminorm, which is the discrete version of the seminorm in the space $L^2(0,T;H^1(\Omega))$.

Definition 18.3 (Discrete $L^2(0,T;H^1(\Omega))$ seminorm) Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $\mathcal{T}$ an admissible finite volume mesh in the sense of Definition 10.1 page 63, $T > 0$, $k \in (0,T)$ and $N_k = \max\{n \in \mathbb{N}; nk < T\}$. For $u \in X(\mathcal{T},k)$, let the following seminorms be defined by:

$$|u(\cdot,t)|_{1,\mathcal{T}}^2 = \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(u_L^n - u_K^n)^2, \text{ for a.e. } t \in (0,T) \text{ and } n = \max\{n \in \mathbb{N}; nk \le t\}, \tag{18.16}$$

and

$$|u|_{1,\mathcal{T},k}^2 = \sum_{n=0}^{N_k} k \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(u_L^n - u_K^n)^2. \tag{18.17}$$

Let us now state some preliminary lemmata to the use of Kolmogorov's theorem (compactness properties in $L^2(\Omega \times (0,T))$) in the proof of convergence of the approximate solutions.

Lemma 18.2 *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $\mathcal{T}$ an admissible mesh in the sense of Definition 10.1 page 63, $T > 0$, $k \in (0,T)$ and $u \in X(\mathcal{T},k)$. For all $\eta \in \mathbb{R}^d$, let Ω_η be defined by $\Omega_\eta = \{x \in \Omega,\ [x, x+\eta] \subset \Omega\}$. Then:*

$$\|u(\cdot+\eta,\cdot) - u(\cdot,\cdot)\|^2_{L^2(\Omega_\eta \times (0,T))} \leq |u|^2_{1,\mathcal{T},k} |\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})), \forall \eta \in \mathbb{R}^d, \tag{18.18}$$

Proof of Lemma 18.2

Reproducing the proof of Lemma 9.3 page 44 (see also the proof of (10.31) page 75), we get, for a.e. $t \in (0,T)$:

$$\|u(\cdot+\eta,t) - u(\cdot,t)\|^2_{L^2(\Omega_\eta)} \leq |u(\cdot,t)|^2_{1,\mathcal{T}} |\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})), \forall \eta \in \mathbb{R}^d. \tag{18.19}$$

Integrating (18.19) on $t \in (0,T)$ gives (18.18). ∎

The set Ω_η defined in Lemma 18.2 verifies $\Omega \setminus \Omega_\eta \subset \cup_{\sigma \in \mathcal{E}_{\rm ext}} \omega^-_{\eta,\sigma}$, with $\omega_{\eta,\sigma} = \{y - t\eta,\ y \in \sigma,\ t \in [0,1]\}$. Then, $\mathrm{m}(\Omega \setminus \Omega_\eta) \leq |\eta|\ \mathrm{m}(\partial\Omega)$, since $\mathrm{m}(\bar{\omega}_\eta) \leq \eta \mathrm{m}(\sigma)$. Then, an immediate corollary of Lemma 18.2 is the following:

Lemma 18.3 *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $\mathcal{T}$ an admissible mesh in the sense of Definition 10.1 page 63, $T > 0$, $k \in (0,T)$ and $u \in X(\mathcal{T},k)$. Let $\tilde{u}$ be defined by $\tilde{u} = u$ a.e. on $\Omega \times (0,T)$, and $\tilde{u} = 0$ a.e. on $\mathbb{R}^{d+1} \setminus \Omega \times (0,T)$. Then:*

$$\left\{ \begin{array}{l} \|\tilde{u}(\cdot+\eta,\cdot) - \tilde{u}(\cdot,\cdot)\|^2_{L^2(\mathbb{R}^{d+1})} \leq |\eta| \Big(|u|^2_{1,\mathcal{T},k} (|\eta| + 2\,\mathrm{size}(\mathcal{T})) + 2\mathrm{m}(\partial\Omega) \|u\|^2_{L^\infty(\Omega\times(0,T))} \Big), \\ \forall \eta \in \mathbb{R}^d. \end{array} \right. \tag{18.20}$$

Remark 18.8 Estimate (18.20) makes use of the $L^\infty(\Omega\times(0,T))$-norm of $u \in X(\mathcal{T},k)$. A similar estimate may be proved with the $L^2(\Omega \times (0,T))$-norm of u (instead of the $L^\infty(\Omega \times (0,T))$-norm). Indeed, the right hand side of (18.20) may be replaced by $C\eta(|u|^2_{1,\mathcal{T},k} + \|u\|^2_{L^2(\Omega\times(0,T))})$, where C only depends on Ω. This estimate is proved in Theorem 10.3 page 74 where it is used for the convergence of numerical schemes for the Neumann problem (for which no L^∞ estimate on the approximate solutions is available). The key to its proof is the "trace lemma" 10.5 page 72.

Let us now state the following lemma, which gives an estimate of the discrete $L^2(0,T;H^1(\Omega))$ seminorm of the nonlinearity.

Lemma 18.4 *Under Assumption 18.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63. Let $\xi \in (0,1)$ and $k \in (0,T)$ such that*

$$k \leq (1-\xi) \frac{\mathrm{m}(K)}{B \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K|L}}, \quad \textit{for all } K \in \mathcal{T}. \tag{18.21}$$

Let $u_{\mathcal{T},k} \in X(\mathcal{T},k)$ be given by (18.6)-(18.9).
Let $U = \|u_0\|_{L^\infty(\Omega)} + T\|f\|_{L^\infty(\Omega\times(0,T))}$ and B be the Lipschitz constant of φ on $[-U,U]$. Then there exists $F_1 \geq 0$, which only depends on Ω, T, φ, u_0, f and ξ such that

$$|\varphi(u_{\mathcal{T},k})|^2_{1,\mathcal{T},k} \leq F_1. \tag{18.22}$$

Proof of lemma 18.4

Let us first remark that the condition (18.21) is stronger than (18.13). Therefore, the result of lemma 18.1 holds, i.e. $|u_K^n| \leq U$, for all $K \in \mathcal{T}$, $n \in \{0, \ldots, N_k\}$. Multiplying equation (18.6) by ku_K^n, and summing the result over $n \in \{0, \ldots, N_k\}$ and $K \in \mathcal{T}$ yields:

$$\begin{aligned} &\sum_{n=0}^{N_k} \sum_{K \in \mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)u_K^n - \\ &\sum_{n=0}^{N_k} k \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big)u_K^n \quad = \sum_{n=0}^{N_k} k \sum_{K \in \mathcal{T}} \mathrm{m}(K) u_K^n f_K^n. \end{aligned} \tag{18.23}$$

In order to obtain a lower bound on the first term on the left hand side of (18.23), let us first remark that:

$$(u_K^{n+1} - u_K^n)u_K^n = \frac{1}{2}(u_K^{n+1})^2 - \frac{1}{2}(u_K^n)^2 - \frac{1}{2}(u_K^{n+1} - u_K^n)^2. \tag{18.24}$$

Now, applying (18.6), using Young's inequality, the following inequality is obtained:

$$(u_K^{n+1} - u_K^n)^2 \leq k^2(1+\xi)\Big[\Big(\frac{1}{\mathrm{m}(K)} \sum_{L \in \mathcal{N}(K)} \tau_{K|L}(\varphi(u_L^n) - \varphi(u_K^n))\Big)^2 + \frac{(f_K^n)^2}{\xi}\Big]. \tag{18.25}$$

which yields in turn, using the Cauchy-Schwarz inequality:

$$\begin{aligned} (u_K^{n+1} - u_K^n)^2 \leq \frac{k^2}{\mathrm{m}(K)^2}(1+\xi)\Big[\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big]\Big[\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big)^2\Big] \\ + \frac{(1+\xi)(k\ f_K^n)^2}{\xi}. \end{aligned} \tag{18.26}$$

Taking condition (18.21) into account gives:

$$(u_K^{n+1} - u_K^n)^2 \leq (1-\xi^2)\frac{k}{B\mathrm{m}(K)}\Big[\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big)^2\Big] + \frac{(1+\xi)(k\ f_K^n)^2}{\xi}. \tag{18.27}$$

Using (18.24) and (18.27) leads to the following lower bound on the first term of the left hand side of (18.23):

$$\begin{aligned} \sum_{n=0}^{N_k} \sum_{K \in \mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)u_K^n \geq \quad & \frac{1}{2}\sum_{K \in \mathcal{T}} \mathrm{m}(K)\Big((u_K^{N_k+1})^2 - (u_K^0)^2\Big) \\ & - \frac{1-\xi^2}{2B}\sum_{n=0}^{N_k} k \sum_{K \in \mathcal{T}}\Big[\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big)^2\Big] \\ & - \frac{k(1+\xi)}{2\xi}\sum_{n=0}^{N_k} k \sum_{K \in \mathcal{T}} \mathrm{m}(K)(f_K^n)^2. \end{aligned} \tag{18.28}$$

Let us now handle the second term on the left hand side of (18.23). Let $\phi \in C(\mathbb{R}, \mathbb{R})$ be defined by $\phi(x) = x\varphi(x) - \int_{x_0}^{x} \varphi(y)dy$, where $x_0 \in \mathbb{R}$ is an arbitrary given real value. Then the following equality holds:

$$\phi(u_L^n) - \phi(u_K^n) = u_K^n(\varphi(u_L^n) - \varphi(u_K^n)) - \int_{u_K^n}^{u_L^n} (\varphi(x) - \varphi(u_L^n))dx. \tag{18.29}$$

The following technical lemma is used here and several times in the sequel:

Lemma 18.5 *Let $g : \mathbb{R} \to \mathbb{R}$ be a monotone Lipschitz continuous function, with a Lipschitz constant $G > 0$. Then:*

$$|\int_c^d (g(x) - g(c))dx| \geq \frac{1}{2G}(g(d) - g(c))^2, \quad \forall c, d \in \mathbb{R}. \tag{18.30}$$

Proof of Lemma 18.5

In order to prove Lemma 18.5, we assume, for instance, that g is nondecreasing and $c < d$ (the other cases are similar). Then, one has $g(s) \geq h(s)$, for all $s \in [c, d]$, where $h(s) = g(c)$ for $s \in [c, d - l]$ and $h(s) = g(c) + (s - d + l)G$ for $s \in [d - l, d]$, with $lG = g(d) - g(c)$, and therefore:

$$\int_c^d (g(s) - g(c))ds \geq \int_c^d (h(s) - g(c))ds = \frac{l}{2}(g(d) - g(c)) = \frac{1}{2G}(g(d) - g(c))^2,$$

this completes the proof of Lemma 18.5.

It is interesting to notice that, for this proof, the fact that g is Lipschitz continuous is not necessary. We only use the fact that $g(s) \geq g(c)$ and $g(d) - g(s) \leq G(d - s)$ for all $s \in [c, d]$ (indeed we use $g(s) \geq g(c)$ for $s \in [c, d - l]$) and $g(d) - g(s) \leq G(d - s)$ for $s \in [d - l, d]$). ∎

Using Lemma 18.5, (18.29) and the equality $\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\phi(u_L^n) - \phi(u_K^n)) = 0$ yields:

$$-\sum_{n=0}^{N_k} k \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L}\Big(\varphi(u_L^n) - \varphi(u_K^n)\Big)u_K^n \geq \frac{1}{2B}\sum_{n=0}^{N_k} k \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\varphi(u_L^n) - \varphi(u_K^n))^2. \tag{18.31}$$

Since $k < T$ we deduce from (18.14) that the right hand side of equation (18.23) satisfies

$$|\sum_{n=0}^{N_k} k \sum_{K\in\mathcal{T}} \mathrm{m}(K)u_K^n f_K^n| \leq 2T\mathrm{m}(\Omega)U\|f\|_{L^\infty(\Omega\times(0,2T))}. \tag{18.32}$$

Relations $k < T$, (18.23), (18.28), (18.31) and (18.32) lead to

$$\begin{gathered}\frac{\xi^2}{2B}\sum_{n=0}^{N_k} k \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\varphi(u_L^n) - \varphi(u_K^n))^2 \leq \\ 2T\mathrm{m}(\Omega)\|f\|_{L^\infty(\Omega\times(0,2T))}\Big(U + \frac{1+\xi}{2\xi}\|f\|_{L^\infty(\Omega\times(0,2T))}T\Big) + \frac{1}{2}\mathrm{m}(\Omega)\|u_0\|_{L^\infty(\Omega)}^2\end{gathered} \tag{18.33}$$

which concludes the proof of the lemma. ∎

Remark 18.9 Estimate (18.22) also holds for the implicit scheme , without any condition on k. One multiplies (18.10) by u_K^{n+1}: the last term on the right hand side of (18.24) appears with the opposite sign, which considerably simplifies the previous proof.

Time translates of approximate solutions

In order to fulfill the hypotheses of Kolmogorov's theorem, the study of time translates must now be performed. The following estimate holds:

Lemma 18.6 *Under Assumption 18.1 page 104, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63 and $k \in (0, T)$. Let $u_{\mathcal{T},k} \in X(\mathcal{T}, k)$ be given by (18.6)-(18.9). Let $U = \|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times(0,T))}$ and B be the Lipschitz constant of φ on $[-U, U]$. Then:*

$$\left\{\begin{array}{l}\|\varphi(u_{\mathcal{T},k}(\cdot, \cdot + \tau)) - \varphi(u_{\mathcal{T},k}(\cdot, \cdot))\|_{L^2(\Omega\times(0,T-\tau))}^2 \leq \\ 2B\tau\Big(|\varphi(u_{\mathcal{T},k})|_{1,\mathcal{T},k}^2 + BT\mathrm{m}(\Omega)U\|f\|_{L^\infty(\Omega\times(0,T))}\Big), \forall \tau \in (0, T).\end{array}\right. \tag{18.34}$$

Proof of Lemma 18.6
Let $\tau \in (0,T)$. Since B is the Lipschitz constant of φ on $[-U,U]$, $U = \|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times(0,T))}$ and φ is nondecreasing, the following inequality holds:

$$\int_{\Omega\times(0,T-\tau)} \Big(\varphi(u_{\mathcal{T},k}(x,t+\tau)) - \varphi(u_{\mathcal{T},k}(x,t))\Big)^2 dxdt \le B\int_0^{T-\tau} A(t)dt, \tag{18.35}$$

where, for almost every $t \in (0,T-\tau)$,

$$A(t) = \int_\Omega \Big(\varphi(u_{\mathcal{T},k}(x,t+\tau)) - \varphi(u_{\mathcal{T},k}(x,t))\Big)\Big(u_{\mathcal{T},k}(x,t+\tau) - u_{\mathcal{T},k}(x,t)\Big)dx.$$

Let $t \in (0,T-\tau)$. Using the definition of $u_{\mathcal{T},k}$ (18.9), this may also be written:

$$A(t) = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\Big(\varphi(u_K^{n_1(t)}) - \varphi(u_K^{n_0(t)})\Big)\Big(u_K^{n_1(t)} - u_K^{n_0(t)}\Big), \tag{18.36}$$

with $n_0(t)$, $n_1(t) \in \{0,\dots,N_k\}$ such that $n_0(t)k \le t < (n_0(t)+1)k$ and $n_1(t)k \le t+\tau < (n_1(t)+1)k$. Equality (18.36) may be written as

$$A(t) = \sum_{K\in\mathcal{T}} (\varphi(u_K^{n_1(t)}) - \varphi(u_K^{n_0(t)}))\Big(\sum_{n=n_0(t)+1}^{n_1(t)} \mathrm{m}(K)(u_K^n - u_K^{n-1})\Big),$$

which also reads

$$A(t) = \sum_{K\in\mathcal{T}} (\varphi(u_K^{n_1(t)}) - \varphi(u_K^{n_0(t)}))\Big(\sum_{n=1}^{N_k} \chi_n(t,t+\tau)\mathrm{m}(K)(u_K^n - u_K^{n-1})\Big), \tag{18.37}$$

with $\chi_n(t,t+\tau) = 1$ if $nk \in (t,t+\tau]$ and $\chi_n(t,t+\tau) = 0$ if $nk \notin (t,t+\tau]$.
In (18.37), the order of summation between n and K is changed and the scheme (18.6) is used. Hence,

$$\begin{aligned} A(t) = k\sum_{n=1}^{N_k} \chi_n(t,t+\tau)\Big[\sum_{K\in\mathcal{T}} (\varphi(u_K^{n_1(t)}) - \varphi(u_K^{n_0(t)})) \\ \Big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\varphi(u_L^{n-1}) - \varphi(u_K^{n-1})) + \mathrm{m}(K)f_K^{n-1}\Big)\Big]. \end{aligned}$$

Gathering by edges, this yields:

$$\begin{aligned} A(t) = k\sum_{n=1}^{N_k}\Big[\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(\varphi(u_K^{n_1(t)}) - \varphi(u_L^{n_1(t)}) - \varphi(u_K^{n_0(t)}) + \varphi(u_L^{n_0(t)})) \\ (\varphi(u_L^{n-1}) - \varphi(u_K^{n-1})) + \sum_{K\in\mathcal{T}} (\varphi(u_K^{n_1(t)}) - \varphi(u_K^{n_0(t)}))\mathrm{m}(K)f_K^{n-1}\Big]\chi_n(t,t+\tau). \end{aligned}$$

Using the inequality $2ab \le a^2 + b^2$, this yields:

$$A(t) \le \frac{1}{2}A_0(t) + \frac{1}{2}A_1(t) + A_2(t) + A_3(t), \tag{18.38}$$

with

$$A_0(t) = k\sum_{n=1}^{N_k} \chi_n(t,t+\tau)\Big(\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(\varphi(u_L^{n_0(t)}) - \varphi(u_K^{n_0(t)}))^2\Big),$$

$$A_1(t) = k\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\big(\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}(\varphi(u_L^{n_1(t)})-\varphi(u_K^{n_1(t)}))^2\big),$$

$$A_2(t) = k\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\big(\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}(\varphi(u_L^{n-1})-\varphi(u_K^{n-1}))^2\big),$$

and

$$A_3(t) = k\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\big(\sum_{K\in\mathcal{T}}(\varphi(u_K^{n_1(t)})-\varphi(u_K^{n_0(t)}))\mathrm{m}(K)f_K^{n-1}\big).$$

Note that, since $t\in(0,T-\tau)$, $n_0(t)\in\{0,\dots,N_k\}$, and, for $m\in\{0,\dots,N_k\}$, $n_0(t)=m$ if and only if $t\in[mk,(m+1)k)$. Therefore,

$$\int_0^{T-\tau}A_0(t)dt \le \sum_{m=0}^{N_k}\int_{mk}^{(m+1)k}k\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\big(\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}(\varphi(u_L^m)-\varphi(u_K^m))^2\big)dt,$$

which also reads

$$\int_0^{T-\tau}A_0(t)dt \le \sum_{m=0}^{N_k}k\int_{mk}^{(m+1)k}\Big(\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\Big)dt\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}(\varphi(u_L^m)-\varphi(u_K^m))^2. \tag{18.39}$$

The change of variable $t=s+(n-m)k$ yields

$$\int_{mk}^{(m+1)k}\chi_n(t,t+\tau)dt = \int_{2mk-nk}^{2mk-nk+k}\chi_n(s+(n-m)k,s+(n-m)k+\tau)ds = \int_{2mk-nk}^{2mk-nk+k}\chi_m(s,s+\tau)ds,$$

then, for all $m\in\{0,\dots,N_k\}$,

$$\int_{mk}^{(m+1)k}\Big(\sum_{n=1}^{N_k}\chi_n(t,t+\tau)\Big)dt \le \int_{\mathbb{R}}\chi_m(s,s+\tau)ds = \tau,$$

since $\chi_m(s,s+\tau)=1$ if and only if $mk\in(s,s+\tau]$ which is equivalent to $s\in[mk-\tau,mk)$.
Therefore (18.39) yields

$$\int_0^{T-\tau}A_0(t)dt \le \tau|\varphi(u_{\mathcal{T},k})|^2_{1,\mathcal{T},k}. \tag{18.40}$$

Similarly:

$$\int_0^{T-\tau}A_1(t)dt \le \tau|\varphi(u_{\mathcal{T},k})|^2_{1,\mathcal{T},k}. \tag{18.41}$$

Let us now study the term $\int_0^{T-\tau}A_2(t)dt$:

$$\int_0^{T-\tau}A_2(t)dt \le \sum_{n=1}^{N_k}k\sum_{K|L\in\mathcal{E}_{\rm int}}\tau_{K|L}(\varphi(u_L^{n-1})-\varphi(u_K^{n-1}))^2\int_0^{T-\tau}\chi_n(t,t+\tau)dt. \tag{18.42}$$

Since $\int_0^{T-\tau}\chi_n(t,t+\tau)\le\tau$ (recall that $\chi_n(t,t+\tau)=1$ if and only if $t\in[nk-\tau,nk)$), the following inequality holds:

$$\int_0^{T-\tau} A_2(t)dt \le \tau|\varphi(u_{\mathcal{T},k})|^2_{1,\mathcal{T},k}. \tag{18.43}$$

In the same way:

$$\begin{aligned}\textstyle\int_0^{T-\tau} A_3(t)dt &\le \sum_{n=1}^{N_k} k\big(\sum_{K\in\mathcal{T}} \mathrm{m}(K)2BU\|f\|_{L^\infty(\Omega\times(0,T))}\big)\int_0^{T-\tau}\chi_n(t,t+\tau)dt\\ &\le \tau T\mathrm{m}(\Omega)2BU\|f\|_{L^\infty(\Omega\times(0,T))}.\end{aligned} \tag{18.44}$$

Using inequalities (18.35), (18.38) and (18.40)-(18.44), (18.34) is proved. ∎

Remark 18.10 Estimate (18.34) is again true for the implicit scheme , with $\|f\|_{L^\infty(\Omega\times(0,2T))}$ instead of $\|f\|_{L^\infty(\Omega\times(0,T))}$.

An immediate corollary of Lemma 18.6 is the following.

Lemma 18.7 *Under Assumption 18.1 page 104, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63 and $k\in(0,T)$. Let $u_{\mathcal{T},k}\in X(\mathcal{T},k)$ be given by (18.6)-(18.9). Let $U=\|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times(0,T)}$ and B be the Lipschitz constant of φ on $[-U,U]$. One defines $\tilde u$ by $\tilde u = u_{\mathcal{T},k}$ a.e. on $\Omega\times(0,T)$, and $\tilde u = 0$ a.e. on $\mathbb{R}^{d+1}\setminus\Omega\times(0,T)$. Then:*

$$\begin{aligned}\|\varphi(\tilde u(\cdot,\cdot+\tau))-\varphi(\tilde u(\cdot,\cdot))\|^2_{L^2(\mathbb{R}^{d+1})} \le 2|\tau|B\Big(\quad &|\varphi(u_{\mathcal{T},k})|^2_{1,\mathcal{T},k}+\\ &BT\mathrm{m}(\Omega)U\|f\|_{L^\infty(\Omega\times(0,T))}+B\mathrm{m}(\Omega)U^2\Big),\\ \forall\tau\in\mathbb{R}.\end{aligned}$$

18.4 Convergence

Theorem 18.1 *Under Assumption 18.1 page 104, let $U=\|u_0\|_{L^\infty(\Omega)}+T\|f\|_{L^\infty(\Omega\times(0,T))}$ and*

$$B=\sup_{-U\le x<y\le U}\frac{\varphi(x)-\varphi(y)}{x-y}.$$

Let $\xi\in(0,1)$ be a given real value. For $m\in\mathbb{N}$, let $\mathcal{T}_m$ be an admissible mesh in the sense of Definition 10.1 page 63 and $k_m\in(0,T)$ satisfying the condition (18.21) with $\mathcal{T}=\mathcal{T}_m$ and $k=k_m$. Let $u_{\mathcal{T}_m,k_m}$ be given by (18.6)-(18.9) with $\mathcal{T}=\mathcal{T}_m$ and $k=k_m$. Assume that $\mathrm{size}(\mathcal{T}_m)\to 0$ *as $m\to\infty$.*
Then, there exists a subsequence of the sequence of approximate solutions, still denoted by $(u_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$, which converges to a weak solution u of Problem (18.1)-(18.3), as $m\to\infty$, in the following sense:
(i) $u_{\mathcal{T}_m,k_m}$ converges to u in $L^\infty(\Omega\times(0,T))$, for the weak-$\star$ topology as m tends to $+\infty$,
(ii) $(\varphi(u_{\mathcal{T}_m,k_m}))$ converges to $\varphi(u)$ in $L^1(\Omega\times(0,T))$ as m tends to $+\infty$,
where $u_{\mathcal{T}_m,k_m}$ and $\varphi(u_{\mathcal{T}_m,k_m})$ also denote the restrictions of these functions to $\Omega\times(0,T)$.

Proof of Theorem 18.1

Let us set $u_m=u_{\mathcal{T}_m,k_m}$ and assume, without loss of generality, that $\varphi(0)=0$. First remark that, by (18.21), $k_m\to 0$ as $m\to 0$. Thanks to Lemma 18.1 page 106, the sequence $(u_m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(\Omega\times(0,T))$. Then, there exists a subsequence, still denoted by $(u_m)_{m\in\mathbb{N}}$, such that u_m converges, as $m\to\infty$, to u in $L^\infty(\Omega\times(0,T))$, for the weak-$\star$ topology.
For the study of the sequence $(\varphi(u_m))_{m\in\mathbb{N}}$, we shall apply Theorem 14.1 page 94 with $N=d+1$, $q=2$, $\omega=\Omega\times(0,T)$ and $p(v)=\tilde v$ with $\tilde v$ defined, as usual, by $\tilde v=v$ on $\Omega\times(0,T)$ and $\tilde v=0$ on $\mathbb{R}^{d+1}\setminus\Omega\times(0,T)$. The first and second items of Theorem 14.1 are clearly satisfied; let us prove hereafter that the third is also satisfied. By Lemma 18.4, the sequence $(|\varphi(u_m)|_{1,\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$ is bounded. Let $\eta\in\mathbb{R}^d$ and $\tau\in\mathbb{R}$, since

$$\|\varphi(\tilde{u}_m(\cdot+\eta,\cdot+\tau))-\varphi(\tilde{u}_m(\cdot,\cdot))\|_{L^2(\mathbb{R}^{d+1})}\leq$$
$$\|\varphi(\tilde{u}_m(\cdot+\eta,\cdot))-\varphi(\tilde{u}_m(\cdot,\cdot))\|_{L^2(\mathbb{R}^{d+1})}+\|\varphi(\tilde{u}_m(\cdot,\cdot+\tau))-\varphi(\tilde{u}_m(\cdot,\cdot))\|_{L^2(\mathbb{R}^{d+1})},$$

lemmata 18.3 and 18.7 give the third item of Theorem 14.1 and this yields the compactness of the sequence $(\varphi(u_m))_{m\in\mathbb{N}}$ in $L^2(\Omega\times(0,T))$.
Therefore, there exists a subsequence, still denoted by $(\varphi(u_m))_{m\in\mathbb{N}}$, and there exists $\chi\in L^2(\Omega\times(0,T))$ such that $\varphi(u_{\mathcal{T}_m,k_m})$ converges, as $m\to\infty$, to χ in $L^2(\Omega\times(0,T))$. Indeed, since $(\varphi(u_m))_{m\in\mathbb{N}}$ is bounded in $L^\infty(\Omega\times(0,T))$, this convergence holds in $L^q(\Omega\times(0,T))$ for all $1\leq q<\infty$. Furthermore, since φ is nondecreasing, Theorem 18.2 page 116 gives that $\chi=\varphi(u)$.

Up to now, the following properties have been shown to be satisfied by a convenient subsequence:

(i) $(u_m)_{m\in\mathbb{N}}$ converges to u, as $m\to\infty$, in $L^\infty(\Omega\times(0,T))$ for the weak-$\star$ topology,

(ii) $(\varphi(u_m))_{m\in\mathbb{N}}$ converges to $\varphi(u)$ in $L^1(\Omega\times(0,T))$ (and even in $L^p(\Omega\times(0,T))$ for all $p\in[1,\infty)$).

There remains to show that u is a weak solution of Problem (18.1)-(18.3), which concludes the proof of Theorem 18.1.
Let $m\in\mathbb{N}$. For the sake of simplicity, we shall use the notations $\mathcal{T}=\mathcal{T}_m$, $h=\text{size}(\mathcal{T})$ and $k=k_m$. Let $\psi\in\mathcal{A}_T$. We multiply (18.6) page 105 by $k\psi(x_K,nk)$, and sum the result on $n\in\{0,\ldots,N_k\}$ and $K\in\mathcal{T}$. We obtain

$$T_{1m}+T_{2m}=T_{3m},\tag{18.45}$$

with

$$T_{1m}=\sum_{n=0}^{N_k}\sum_{K\in\mathcal{T}}\mathrm{m}(K)(u_K^{n+1}-u_K^n)\psi(x_K,nk),$$

$$T_{2m}=-\sum_{n=0}^{N_k}k\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}\Big(\varphi(u_L^n)-\varphi(u_K^n)\Big)\psi(x_K,nk),$$

and

$$T_{3m}=\sum_{n=0}^{N_k}k\sum_{K\in\mathcal{T}}\psi(x_K,nk)\mathrm{m}(K)f_K^n.$$

We first consider T_{1m}.

$$\begin{aligned}T_{1m}=\ &\sum_{n=1}^{N_k}\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^n\Big(\psi(x_K,(n-1)k)-\psi(x_K,nk)\Big)\quad+\\ &\sum_{K\in\mathcal{T}}\mathrm{m}(K)\Big(u_K^{N_k+1}\psi(x_K,kN_k)-u_K^0\psi(x_K,0)\Big).\end{aligned}$$

Performing one more step of the induction in Lemma 18.1, it is clear that $|u_K^{N_k+1}|<U+2T\|f\|_{L^\infty(\Omega\times(0,2T))}$, for all $K\in\mathcal{T}$.
Since $0<T-N_kk\leq k$, there exists $C_{1,\psi}$ which only depends on ψ, T and Ω, such that $|\psi(x_K,N_kk)|\leq kC_{1,\psi}$. Hence,

$$\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^{N_k+1}\psi(x_K,kN_k)\to 0\text{ as }m\to\infty.$$

Since

$$\|\sum_{K\in\mathcal{T}} u_K^0 1_K - u_0\|_{L^1(\Omega)} \to 0, \text{ as } m\to\infty,$$

(where $1_K(x)=1$ if $x\in K$, 0 otherwise), one has

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^0 \psi(x_K,0) \to \int_\Omega u_0(x)\psi(x,0)dx \text{ as } m\to\infty.$$

Since $(u_m)_{m\in\mathbb{N}}$ converges, as $m\to+\infty$, to u in $L^\infty(\Omega\times(0,T))$, for the weak-$\star$ topology, and since $|u_K^{N_k}| < U + T\|f\|_{L^\infty(\Omega\times(0,T))}$, for all $K\in\mathcal{T}$, the following property also holds:

$$\sum_{n=1}^{N_k}\sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^n \Big(\psi(x_K,(n-1)k) - \psi(x_K,nk)\Big) \to -\int_0^T\int_\Omega u(x,t)\psi_t(x,t)dxdt \text{ as } m\to\infty.$$

Therefore,

$$T_{1m} \to -\int_0^T\int_\Omega u(x,t)\psi_t(x,t)dxdt - \int_\Omega u_0(x)\psi(x,0)dx, \text{ as } m\to\infty.$$

We now study T_{2m}. This term can be rewritten as

$$T_{2m} = \quad -\sum_{n=0}^{N_k} k \sum_{K|L\in\mathcal{E}_{\rm int}} m(K|L)(\varphi(u_L^n)-\varphi(u_K^n))\frac{\psi(x_K,nk)-\psi(x_L,nk)}{d_{K|L}}.$$

It is useful to introduce the following expression:

$$\begin{aligned} T'_{2m} &= \sum_{n=0}^{N_k}\int_{nk}^{(n+1)k}\int_\Omega \varphi(u_{\mathcal{T},k}(x,t))\Delta\psi(x,nk)dxdt \\ &= \sum_{n=0}^{N_k} k\sum_{K\in\mathcal{T}}\varphi(u_K^n)\int_K \Delta\psi(x,nk)dx \\ &= \sum_{n=0}^{N_k} k\sum_{K|L\in\mathcal{E}_{\rm int}} (\varphi(u_K^n)-\varphi(u_L^n))\int_{K|L}\nabla\psi(x,nk)\cdot\boldsymbol{n}_{K,L}d\gamma(x). \end{aligned}$$

The sequence $(\varphi(u_m))_{m\in\mathbb{N}}$ converges to $\varphi(u)$ in $L^1(\Omega\times(0,T))$; furthermore, it is bounded in L^∞ so that the integral between T and $(N_k+1)k$ tends to 0. Therefore:

$$T'_{2m} \to \int_0^T\int_\Omega \varphi(u(x,t))\Delta\psi(x,t)dxdt, \text{ as } m\to\infty.$$

The term $T_{2m}+T'_{2m}$ can be written as

$$T_{2m}+T'_{2m} \quad = \sum_{n=0}^{N_k} k\sum_{K|L\in\mathcal{E}} m(K|L)(\varphi(u_K^n)-\varphi(u_L^n))R_{K,L}^n,$$

with

$$R_{K,L}^n = \frac{1}{m(K|L)}\int_{K|L}\nabla\psi(x,nk)\cdot\boldsymbol{n}_{K,L}d\gamma(x) - \frac{\psi(x_L,nk)-\psi(x_K,nk)}{d_{K|L}}.$$

Thanks to the regularity properties of ψ there exists C_ψ, which only depends on ψ, such that $|R_{K,L}^n|\le C_\psi h$. Then, using the estimate (18.22), we conclude that $T_{2m}+T'_{2m}\to 0$ as $m\to\infty$. Therefore,

$$T_{2m} \to -\int_0^T \int_\Omega \varphi(u(x,t))\Delta\psi(x,t)dxdt, \text{ as } m \to \infty.$$

Let us now study T_{3m}.
Define $f_{\mathcal{T},k} \in X(\mathcal{T},k)$ by $f_{\mathcal{T},k}(x,t) = f_K^n$ if $(x,t) \in K \times (nk, nk+k)$. Since $f_{\mathcal{T},k} \to f$ in $L^1(\Omega \times (0,T)$ and since $f \in L^\infty(\Omega \times (0,2T)$,

$$T_{3m} \to \int_\Omega \int_0^T f(x,t)\psi(x,t)dtdx, \text{ as } m \to \infty.$$

Passing to the limit in Equation (18.45) gives that u is a weak solution of Problem (18.1)-(18.3). This concludes the proof of Theorem 18.1. ∎

Remark 18.11 This convergence proof is quite similar in the case of the implicit scheme, with the additional condition that $(k_m)_{m\in\mathbb{N}}$ converges to zero, since condition (18.21) does not have to be satisfied.

Remark 18.12 The above convergence result was shown for a subsequence only. A convergence theorem is obtained for the full set of approximate solutions, if a uniqueness result is valid. Such a result can be easily obtained in the case of a smooth boundary and is given in section 18.6 below. For this case, an extension to the definition 10.1 page 63 of admissible meshes is given hereafter.

Definition 18.4 (Admissible meshes for regular domains) Let Ω be an open bounded connected subset of $\mathbb{R}^d$, $d = 2$ or 3 with a C^2 boundary $\partial\Omega$. An admissible finite volume mesh of Ω is given by an open bounded polygonal set Ω' containing Ω, and an admissible mesh $\mathcal{T}'$ of Ω' in the sense of Definition 10.1 page 63. The set of control volumes of the mesh of Ω are $\{K' \cap \Omega,\ K' \in \mathcal{T}' \text{ such that } m_d(K' \cap \Omega) > 0\}$ and the set of edges of the mesh is $\mathcal{E} = \{\sigma \cap \overline{\Omega},\ \sigma \in \mathcal{E}' \text{ such that } m_{d-1}(\sigma \cap \overline{\Omega}) > 0\}$, where $\mathcal{E}'$ denotes the set of edges of $\mathcal{T}'$ and m_N denotes the N-dimensional Lebesgue measure.

Remark 18.13 For smooth domains Ω, the set of edges $\mathcal{E}$ of an admissible mesh of Ω does not contain the parts of the boundaries of the control volumes which are included in the boundary $\partial\Omega$ of Ω.

18.5 Weak convergence and nonlinearities

We show here a property which was used in the proof of Theorem 18.1.

Theorem 18.2 *Let $U > 0$ and $\varphi \in C([-U,U])$ be a nondecreasing function. Let ω be an open bounded subset of $\mathbb{R}^N$, $N \geq 1$. Let $(u_n)_{n\in\mathbb{N}} \subset L^\infty(\omega)$ such that*
(i) $-U \leq u_n \leq U$ a.e. in ω, for all $n \in \mathbb{N}$;
(ii) there exists $u \in L^\infty(\omega)$ such that $(u_n)_{n\in\mathbb{N}}$ converges to u in $L^\infty(\omega)$ for the weak-$\star$ topology;
(iii) there exists a function $\chi \in L^1(\omega)$ such that $(\varphi(u_n))_{n\in\mathbb{N}}$ converges to χ in $L^1(\omega)$.
Then $\chi(x) = \varphi(u(x))$, for a.e. $x \in \omega$.

Proof of Theorem 18.2
First we extend the definition of φ by $\varphi(v) = \varphi(-U) + v + U$ for all $v < -U$ and $\varphi(v) = \varphi(U) + v - U$ for all $v > U$, and denote again by φ this extension of φ which now maps $\mathbb{R}$ into $\mathbb{R}$, is continuous and nondecreasing. Let us define $\alpha_\pm$ from $\mathbb{R}$ to $\mathbb{R}$ by $\alpha_-(t) = \inf\{v \in \mathbb{R}, \varphi(v) = t\}$ and $\alpha_+(t) = \sup\{v \in \mathbb{R}, \varphi(v) = t\}$, for all $t \in \mathbb{R}$.
Note that the functions $\alpha_\pm$ are increasing and that
(i) α_- is left continuous and therefore lower semi-continuous, that is

$$t = \lim_{n\to\infty} t_n \Longrightarrow \alpha_-(t) \leq \liminf_{n\to\infty} \alpha_-(t_n),$$

(ii) α_+ is right continuous and therefore upper semi-continuous, that is

$$t = \lim_{n\to\infty} t_n \Longrightarrow \alpha_+(t) \geq \limsup_{n\to\infty} \alpha_+(t_n).$$

Thus, since we may assume, up to a subsequence, that $\varphi(u_n) \to \chi$ a.e. in ω,

$$\alpha_-(\chi(x)) \leq \liminf_{n\to\infty} \alpha_-\Big(\varphi(u_n(x))\Big) \leq \limsup_{n\to\infty} \alpha_+\Big(\varphi(u_n(x))\Big) \leq \alpha_+(\chi(x)), \tag{18.46}$$

for a.e. $x \in \omega$.

A direct application of the definition of the functions α_- and α_+ gives

$$\alpha_-\Big(\varphi(u_n(x))\Big) \leq u_n(x) \leq \alpha_+\Big(\varphi(u_n(x))\Big). \tag{18.47}$$

Let $L^1_+ = \{\psi \in L^1(\omega),\ \psi \geq 0 \text{ a.e.}\}$. Let $\psi \in L^1_+$. We multiply (18.47) by $\psi(x)$ and integrate over ω, it yields

$$\int_\omega \alpha_-\Big(\varphi(u_n(x))\Big)\psi(x)dx \leq \int_\omega u_n(x)\psi(x)dx \leq \int_\omega \alpha_+\Big(\varphi(u_n(x))\Big)\psi(x)dx. \tag{18.48}$$

Applying Fatou's lemma to the sequences of L^1 positive functions $\alpha_-(\varphi(u_n))\psi - \alpha_-(\varphi(-U))\psi$ and $\alpha_+(\varphi(U))\psi - \alpha_+(\varphi(u_n))\psi$ yields, with (18.46),

$$\int_\omega \alpha_-(\chi(x))\psi(x)dx \leq \liminf_{n\to\infty} \int_\omega \alpha_-\Big(\varphi(u_n(x))\Big)\psi(x)dx,$$

and

$$\limsup_{n\to\infty} \int_\omega \alpha_+\Big(\varphi(u_n(x))\Big)\psi(x)dx \leq \int_\omega \alpha_+(\chi(x))\psi(x)dx.$$

Then, passing to the lim inf and lim sup in (18.48) and using the convergence of $(u_n)_{n\in\mathbb{N}}$ to u in $L^\infty(\omega)$ for the weak-$\star$ topology gives

$$\int_\omega \alpha_-(\chi(x))\psi(x)dx \leq \int_\omega u(x)\psi(x)dx \leq \int_\omega \alpha_+(\chi(x))\psi(x)dx.$$

Thus, since ψ is arbitrary in L^1_+, the following inequality holds for a.e. $x \in \omega$:

$$\alpha_-(\chi(x)) \leq u(x) \leq \alpha_+(\chi(x)),$$

which implies in turn that $\chi(x) = \varphi(u(x))$ for a.e. $x \in \omega$. This completes the proof of Theorem 18.2. ∎

Remark 18.14 Another proof of Theorem 18.2 is possible by passing to the limit in the inequality

$$0 \leq \int_\omega (\varphi(u_n)(x) - \varphi(v(x)))(u_n(x) - v(x))dx, \forall v \in L^\infty(\omega),$$

which leads to

$$0 \leq \int_\omega (\chi(x) - \varphi(v(x)))(u(x) - v(x))dx, \forall v \in L^\infty(\omega).$$

From this inequality, one deduces that $\chi = \varphi(u)$ a.e. on ω.

A third proof is possible by using the concept of nonlinear weak-$\star$ convergence, see Definition 32.1 page 200.

18.6 A uniqueness result for nonlinear diffusion equations

The uniqueness of the weak solution to variations of Problem (18.1)-(18.3) has been proved by several authors. For precise references we refer to MEIRMANOV [107]. Also rather similar proofs have been given in BERTSCH, KERSNER and PELETIER [13] and GUEDDA, HILHORST and PELETIER [78]. Recall that this uniqueness result allows to obtain a convergence result on the whole set of finite volume approximate solutions to Problem (15.1)-(15.4) (see Remark 18.12).

The uniqueness of the weak solution to Problem (18.1)-(18.3) immediately results from the following property.

Theorem 18.3 *Let Ω be an open bounded subset of $\mathbb{R}^d$ with a C^2 boundary, and suppose that items (ii), (iii) and (iv) of Assumption 18.1 are satisfied. Let u_1 and u_2 be two solutions of Problem (18.1)-(18.3) in the sense of Definition 18.1 page 105, with initial conditions $u_{0,1}$ and $u_{0,2}$ and source terms v_1 and v_2 respectively, that is, for u_1 (resp. u_2), $u_0 = u_{0,1}$ (resp. $u_0 = u_{0,2}$) in (18.3) and $f = v_1$ (resp. v_2) in (18.1).*
Then for all $T > 0$,

$$\int_0^T \int_\Omega |u_1(x,t) - u_2(x,t)| dxdt \leq T \int_\Omega |u_{0,1}(x) - u_{0,2}(x)| dx + \int_0^T \int_\Omega (T-t)\ |v_1(x,t) - v_2(x,t)| dxdt.$$

Before proving Theorem 18.3, let us first show the following auxiliary result.

The existence of regular solutions to the adjoint problem

Lemma 18.8 *Let Ω be an open bounded subset of $\mathbb{R}^d$ with a C^2 boundary, and suppose that φ is a nondecreasing locally Lipschitz-continuous function. Let $T > 0$, $w \in C_c^\infty(\Omega \times (0,T))$ such that $|w| \leq 1$, and $g \in C^\infty(\overline{\Omega} \times [0,T])$ such that there exists $r \in \mathbb{R}$ with $0 < r \leq g(x,t)$, for all $(x,t) \in \Omega \times (0,T)$.*

Then there exists a unique function $\psi \in C^{2,1}(\overline{\Omega} \times [0,T])$ such that

$$\psi_t(x,t) + g(x,t)\Delta\psi(x,t) = w(x,t), \qquad \text{for all } (x,t) \in \Omega \times (0,T), \tag{18.49}$$

$$\nabla\psi \cdot \mathbf{n}(x,t) = 0, \quad \text{for all } (x,t) \in \partial\Omega \times (0,T), \tag{18.50}$$

$$\psi(x,T) = 0, \quad \text{for all } x \in \Omega. \tag{18.51}$$

Moreover the function ψ satisfies

$$|\psi(x,t)| \leq T - t, \quad \text{for all } (x,t) \in \Omega \times (0,T), \tag{18.52}$$

and

$$\int_0^T \int_\Omega g(x,t)\Big(\Delta\psi(x,t)\Big)^2 dxdt \leq 4T \int_0^T \int_\Omega |\nabla w(x,t)|^2 dxdt. \tag{18.53}$$

PROOF of Lemma 18.8

It will be useful in the following to point out that the right hand side of (18.53) does not depend on g. Since the function g is bounded away from zero, equations (18.49)-(18.51) define a boundary value problem for a usual heat equation with an initial condition, in which the time variable is reversed. Since Ω, g and w are sufficiently smooth, this problem has a unique solution $\psi \in \mathcal{A}_T$, see LADYŽENSKAJA, SOLONNIKOV and URAL'CEVA [97]. Since $|w| \leq 1$, the functions $T-t$ and $-(T-t)$ are respectively upper and lower solutions of Problem (18.49)-(18.50). Hence we get (18.52) (see LADYŽENSKAJA, SOLONNIKOV and URAL'CEVA [97]).

In order to show (18.53), multiply (18.49) by $\Delta\psi(x,t)$, integrate by parts on $\Omega \times (0,\tau)$, for $\tau \in (0,T]$. This gives

$$\begin{aligned}&\frac{1}{2}\int_\Omega |\nabla\psi(x,0)|^2dx \;-\; \frac{1}{2}\int_\Omega |\nabla\psi(x,\tau)|^2dx \;+\; \int_0^\tau\int_\Omega g(x,t)\Big(\Delta\psi(x,t)\Big)^2dxdt = \\ &-\int_0^\tau\int_\Omega \nabla w(x,t)\cdot\nabla\psi(x,t)dxdt.\end{aligned} \tag{18.54}$$

Since $\nabla\psi(\cdot,T)=0$, letting $\tau=T$ in (18.54) leads to

$$\begin{aligned}&\frac{1}{2}\int_\Omega |\nabla\psi(x,0)|^2dx + \int_0^T\int_\Omega g(x,t)\Big(\Delta\psi(x,t)\Big)^2dxdt = \\ &-\int_0^T\int_\Omega \nabla w(x,t)\cdot\nabla\psi(x,t)dxdt.\end{aligned} \tag{18.55}$$

Integrating (18.54) with respect to $\tau\in(0,T)$ leads to

$$\begin{aligned}\frac{1}{2}\int_0^T\int_\Omega |\nabla\psi(x,\tau)|^2dxd\tau \le\;& \frac{T}{2}\int_\Omega |\nabla\psi(x,0)|^2dx &+\\ & T\int_0^T\int_\Omega g(x,t)\Big(\Delta\psi(x,t)\Big)^2dxdt &+\\ & T\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt.\end{aligned} \tag{18.56}$$

Using (18.55) and (18.56), we get

$$\frac{1}{2}\int_0^T\int_\Omega |\nabla\psi(x,\tau)|^2dxd\tau \le 2T\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt. \tag{18.57}$$

Thanks to the Cauchy-Schwarz inequality, the right hand side of (18.57) may be estimated as follows:

$$\begin{aligned}\Big[\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt\Big]^2 \le\;& \int_0^T\int_\Omega |\nabla\psi(x,t)|^2dxdt \\ &\times\int_0^T\int_\Omega |\nabla w(x,t)|^2dxdt.\end{aligned}$$

With (18.57), this implies

$$\begin{aligned}\Big[\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt\Big]^2 \le\;& 4T\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt \\ &\times\int_0^T\int_\Omega |\nabla w(x,t)|^2dxdt.\end{aligned}$$

Therefore,

$$\int_0^T\int_\Omega |\nabla w(x,t)\cdot\nabla\psi(x,t)|dxdt \le 4T\int_0^T\int_\Omega |\nabla w(x,t)|^2dxdt,$$

which, together with (18.55), yields (18.53). ∎

Proof of the uniqueness theorem

Let u_1 and u_2 be two solutions of Problem (18.4), with initial conditions $u_{0,1}$ and $u_{0,2}$ and source terms v_1 and v_2 respectively. We set $u_d=u_1-u_2$, $v_d=v_1-v_2$ and $u_{0,d}=u_{0,1}-u_{0,2}$. Let us also define, for all $(x,t)\in\Omega\times\mathbb{R}_+^\star$, $q(x,t)=\dfrac{\varphi(u_1(x,t))-\varphi(u_2(x,t))}{u_1(x,t)-u_2(x,t)}$ if $u_1(x,t)\neq u_2(x,t)$, else $q(x,t)=0$. For all $T\in\mathbb{R}_+^\star$ and for all $\psi\in\ \mathcal{A}_T$, we deduce from (18.4) that

$$\begin{array}{ll}\int_0^T \int_\Omega \Big[u_d(x,t)\Big(\psi_t(x,t)+q(x,t)\Delta\psi(x,t)\Big)+v_d(x,t)\psi(x,t)\Big]dxdt & + \\ \int_\Omega u_{0,d}(x)\psi(x,0)dx & = 0.\end{array} \tag{18.58}$$

Let $w \in C_c^\infty(\Omega\times(0,T))$, such that $|w| \leq 1$. Since φ is locally Lipschitz continuous, we can define its Lipschitz constant, say B_M, on $[-M,M]$, where $M = \max\{\|u_1\|_{L^\infty(\Omega\times(0,T))}, \|u_2\|_{L^\infty(\Omega\times(0,T))}\}$ so that $0 \leq q \leq B_M$ a.e. on $\Omega\times(0,T)$.

Using mollifiers, functions $q_{1,n} \in C_c^\infty(\Omega\times(0,T))$ may be constructed such that $\|q_{1,n}-q\|_{L^2(\Omega\times(0,T))} \leq \frac{1}{n}$ and $0 \leq q_{1,n} \leq B_M$, for $n \in \mathbb{N}^\star$. Let $q_n = q_{1,n} + \frac{1}{n}$. Then

$$\frac{1}{n} \leq q_n(x,t) \leq B_M + \frac{1}{n}, \text{ for all } (x,t) \in \Omega\times(0,T),$$

and

$$\begin{array}{ll}\int_0^T \int_\Omega \frac{(q_n(x,t)-q(x,t))^2}{q_n(x,t)}dxdt \leq 2\Big(& \int_0^T \int_\Omega \frac{(q_n(x,t)-q_{1,n}(x,t))^2}{q_n(x,t)}dxdt + \\ & \int_0^T \int_\Omega \frac{(q_{1,n}(x,t)-q(x,t))^2}{q_n(x,t)}dxdt\Big),\end{array}$$

which shows that

$$\int_0^T \int_\Omega \frac{(q_n(x,t)-q(x,t))^2}{q_n(x,t)}dxdt \leq 2n\Big(\frac{T\mathrm{m}(\Omega)}{n^2}+\frac{1}{n^2}\Big).$$

It leads to

$$\|\frac{q_n-q}{\sqrt{q_n}}\|_{L^2(\Omega\times(0,T))} \to 0 \ \text{ as } \ n\to\infty. \tag{18.59}$$

Let $\psi_n \in \mathcal{A}_T$ be given by lemma 18.8, with $g = q_n$. Substituting ψ by ψ_n in (18.58), using (with $g = q_n$ and $\psi = \psi_n$) (18.49) and (18.52) give

$$\begin{array}{ll}|\int_0^T \int_\Omega u_d(x,t)\Big(w(x,t)+(q(x,t)-q_n(x,t))\Delta\psi_n(x,t)\Big)dxdt| & \leq \\ \int_0^T \int_\Omega |v_d(x,t)|(T-t)dxdt + T\int_\Omega |u_{0,d}(x)|dx. & \end{array} \tag{18.60}$$

The Cauchy-Schwarz inequality yields

$$\begin{array}{l}\Big[\int_0^T \int_\Omega |u_d(x,t)||(q(x,t)-q_n(x,t))\Delta\psi_n(x,t)|dxdt\Big]^2 \leq 4M^2 \\ \int_0^T \int_\Omega \Big(\frac{q(x,t)-q_n(x,t)}{\sqrt{q_n(x,t)}}\Big)^2 dxdt \int_0^T \int_\Omega q_n(x,t)\Big(\Delta\psi_n(x,t)\Big)^2 dxdt.\end{array} \tag{18.61}$$

We deduce from (18.53) and (18.59) that the right hand side of (18.61) tends to zero as $n\to\infty$. Hence the left hand side of (18.61) also tends to zero as $n\to\infty$. Therefore letting $n\to\infty$ in (18.60) gives

$$\begin{array}{ll}|\int_0^T \int_\Omega u_d(x,t)w(x,t)dxdt| \leq & \int_0^T \int_\Omega |v_d(x,t)|(T-t)dxdt + \\ & T\int_\Omega |u_{0,d}(x)|dx.\end{array} \tag{18.62}$$

Inequality (18.62) holds for any function $w \in C_c^\infty(\Omega\times(0,T))$, with $|w| \leq 1$. Let us take as functions w the elements of a sequence $(w_m)_{m\in\mathbb{N}}$ such that $w_m \in C_c^\infty(\Omega\times(0,T))$ and $|w_m| \leq 1$ for all $m \in \mathbb{N}$, and the sequence $(w_m)_{m\in\mathbb{N}}$ converges to $\mathrm{sign}(u_d(\cdot,\cdot))$ in $L^1(\Omega\times(0,T))$. Letting $m\to\infty$ yields

$$\int_0^T \int_\Omega |u_d(x,t)| dxdt \leq \int_0^T \int_\Omega |v_d(x,t)|(T-t)dxdt + T\int_\Omega |u_{0,d}(x)|dx,$$

which concludes the proof of Theorem 18.3. ∎

Chapter 5

Hyperbolic equations in the one dimensional case

This chapter is devoted to the numerical schemes for one-dimensional hyperbolic conservation laws. Some basics on the solution to linear or nonlinear hyperbolic equations with initial data and without boundary conditions are first recalled. We refer to Godlewski and Raviart [75], Godlewski and Raviart [76], Kröner [91], LeVeque [100] and Serre [135] for extensive studies of theoreticaland/or numerical aspects; we shall highlight here the finite volume point of view for several well known schemes, comparing them with finite difference schemes, either for the linear and the nonlinear case. Convergence results for numerical schemes are presented, using a "weak BV inequality" which is not really necessary in the 1D case (at least for BV initial data), but is crucial in the multidimensional case. We also recall the classical proof of convergence, which uses a "strong BV estimate" and the Lax-Wendroff theorem, and does not seem to extend to the mutidimentional case. Error estimates which also hold are not adressed in this chapter: they are given in the chapter in the multidimensional case (Chapter 6).

Throughout this chapter, we shall focus on explicit schemes. However, all the results which are presented here can be extended to implicit schemes; this requires a bit of work and is detailed in the multidimensional case (see (25.6) page 158 for the scheme).

19 The continuous problem

Consider the nonlinear hyperbolic equation with initial data:

$$\left\{ \begin{array}{lll} u_t(x,t) + (f(u))_x(x,t) & = 0 & x \in \mathbb{R}, \quad t \in \mathbb{R}_+, \\ u(x,0) & = u_0(x), & x \in \mathbb{R}, \end{array} \right. \tag{19.1}$$

where f is a given function from $\mathbb{R}$ to $\mathbb{R}$, of class C^1, $u_0 \in L^\infty(\mathbb{R})$ and where the partial derivatives of u with respect to time and space are denoted by u_t and u_x.

Example 19.1 (Bürgers equation) A simple flow model was introduced by Bürgers and yields the following equation:

$$u_t(x,t) + u(x,t)u_x(x,t) - \varepsilon u_{xx}(x,t) = 0 \tag{19.2}$$

Bürgers studied the limit case which is obtained when ε tends to 0; the resulting equation is (19.1) with $f(s) = \dfrac{s^2}{2}$, i.e.

$$u_t(x,t) + \frac{1}{2}(u^2)_x(x,t) = 0$$

Definition 19.1 (Classical solution) Let $f \in C^1(\mathbb{R},\mathbb{R})$ and $u_0 \in C^1(\mathbb{R},\mathbb{R})$; a classical solution to Problem (19.1) is a function $u \in C^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$ such that

$$\left\{\begin{array}{rll} u_t(x,t)+f'(u(x,t))u_x(x,t) & = 0, & \forall\, x\in\mathbb{R},\, \forall\, t\in\mathbb{R}_+,\\ u(x,0) & = u_0(x), & \forall\, x\in\mathbb{R}. \end{array}\right.$$

Recall that in the linear case, i.e. $f(s)=cs$ for all $s\in\mathbb{R}$, for some $c\in\mathbb{R}$, there exists (for $u_0\in C^1(\mathbb{R},\mathbb{R})$) a unique classical solution. It is $u(x,t)=u_0(x-ct)$, for all $x\in\mathbb{R}$ and for all $t\in\mathbb{R}_+$. In the nonlinear case, the existence of such a solution depends on the initial data u_0; in fact, the following result holds:

Proposition 19.1 *Let $f\in C^1(\mathbb{R},\mathbb{R})$ be a nonlinear function, i.e. such that there exist s_1, $s_2\in\mathbb{R}$ with $f'(s_1)\neq f'(s_2)$; then there exists $u_0\in C_c^\infty(\mathbb{R},\mathbb{R})$ such that Problem (19.1) has no classical solution.*

Proposition 19.1 is an easy consequence of the following remark.

Remark 19.1 If u is a classical solution to (19.1), then u is constant along the characteristic lines which are defined by

$$x(t)=f'(u_0(x_0))t+x_0,\ t\in\mathbb{R}_+,$$

where $x_0\in\mathbb{R}$ is the origin of the characteristic. This is the equation of a straight line issued from the point $(x_0,0)$ (in the (x,t) coordinates). Note that if f depends on x and u (rather than only on u), the characteristics are no longer straight lines.

The concept of weak solution is introduced in order to define solutions of (19.1) when classical solutions do not exist.

Definition 19.2 (Weak solution) Let $f\in C^1(\mathbb{R},\mathbb{R})$ and $u_0\in L^\infty(\mathbb{R})$; a weak solution to Problem (19.1) is a function u such that

$$\left\{\begin{array}{l} u\in L^\infty(\mathbb{R}\times\mathbb{R}_+^\star),\\ \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx+\int_{\mathbb{R}}\int_{\mathbb{R}_+} f(u(x,t))\varphi_x(x,t)dtdx+\int_{\mathbb{R}} u_0(x)\varphi(x,0)dx=0,\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\varphi\in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}). \end{array}\right. \tag{19.3}$$

Remark 19.2
1. If $u\in C^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})\cap L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$ then u is a weak solution if and only if u is a classical solution.
2. Note that in the above definition, we require the test function φ to belong to $C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$, so that φ may be non zero at time $t=0$.

One may show that there exists at least one weak solution to (19.1). In the linear case, i.e. $f(s)=cs$, for all $s\in\mathbb{R}$, for some $c\in\mathbb{R}$, this solution is unique (it is $u(x,t)=u_0(x-ct)$ for a.e. $(x,t)\in\mathbb{R}\times\mathbb{R}_+$). However, the uniqueness of this weak solution in the general nonlinear case is no longer true. Hence the concept of entropy weak solution, for which an existence and uniqueness result is known.

Definition 19.3 (Entropy weak solution) Let $f\in C^1(\mathbb{R},\mathbb{R})$ and $u_0\in L^\infty(\mathbb{R})$; the entropy weak solution to Problem (19.1) is a function u such that

$$\left\{\begin{array}{l} u\in L^\infty(\mathbb{R}\times\mathbb{R}_+^\star),\\ \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}_+}\eta(u(x,t))\varphi_t(x,t)dtdx+\int_{\mathbb{R}}\int_{\mathbb{R}_+}\Phi(u(x,t))\varphi_x(x,t)dtdx+\int_{\mathbb{R}}\eta(u_0(x))\varphi(x,0)dx\geq 0,\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\varphi\in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+),\\ \qquad\qquad \text{for all convex function } \eta\in C^1(\mathbb{R},\mathbb{R}) \text{ and } \Phi\in C^1(\mathbb{R},\mathbb{R}) \text{ such that } \Phi'=\eta' f'. \end{array}\right. \tag{19.4}$$

Remark 19.3 The solutions of (19.4) are necessarily solutions of (19.3). This can be shown by taking in (19.4) $\eta(s) = s$ for all $s \in \mathbb{R}$, $\eta(s) = -s$, for all $s \in \mathbb{R}$, and regularizations of the positive and negative parts of the test functions of the weak formulation.

Theorem 19.1 *Let $f \in C^1(\mathbb{R}, \mathbb{R})$, $u_0 \in L^\infty(\mathbb{R})$, then there exists a unique entropy weak solution to Problem (19.1).*

The proof of this result was first given by Vol'pert in Vol'pert [156], introducing the space $BV(\mathbb{R})$ which is defined hereafter and assuming $u_0 \in BV(\mathbb{R})$, see also Oleinik [120] for the convex case. In Krushkov [94], Krushkov proved the theorem of existence and uniqueness in the general case $u_0 \in L^\infty(\mathbb{R})$, using a regularization of u_0 in $BV(\mathbb{R})$, under the slightly stronger assumption $f \in C^3(\mathbb{R}, \mathbb{R})$. Krushkov also proved that the solution is in the space $C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$. Krushkov's proof uses particular entropies, namely the functions $|\cdot - \kappa|$ for all $\kappa \in \mathbb{R}$, which are generally referred to as "Krushkov's entropies". The "entropy flux" associated to $|\cdot - \kappa|$ may be taken as $f(\cdot \top \kappa) - f(\cdot \bot \kappa)$, where $a \top b$ denotes the maximum of a and b and $a \bot b$ denotes the minimum of a and b, for all real values a, b (recall that $f(a \top b) - f(a \bot b) = \text{sign}(a - b)(f(a) - f(b))$).

Definition 19.4 ($BV(\mathbb{R})$) A function $v \in L^1_{loc}(\mathbb{R})$ is of bounded variation, that is $v \in BV(\mathbb{R})$, if

$$|v|_{BV(\mathbb{R})} = \sup\{\int_{\mathbb{R}} v(x)\varphi_x(x)dx, \varphi \in C^1_c(\mathbb{R}, \mathbb{R}), |\varphi(x)| \leq 1 \ \forall x \in \mathbb{R}\} < +\infty. \tag{19.5}$$

Remark 19.4

1. If $v : \mathbb{R} \to \mathbb{R}$ is piecewise constant, that is if there exists an increasing sequence $(x_i)_{i \in \mathbb{Z}}$ with $\mathbb{R} = \cup_{i \in \mathbb{Z}} [x_i, x_{i+1}]$ and a sequence $(v_i)_{i \in \mathbb{Z}}$ such that $v|_{(x_i, x_{i+1})} = v_i$, then $|v|_{BV(\mathbb{R})} = \sum_{i \in \mathbb{Z}} |v_{i+1} - v_i|$.
2. If $v \in C^1(\mathbb{R}, \mathbb{R})$ then $|v|_{BV(\mathbb{R})} = \|v_x\|_{L^1(\mathbb{R})}$.
3. The space $BV(\mathbb{R})$ is included in the space $L^\infty(\mathbb{R})$; furthermore, if $u \in BV(\mathbb{R}) \cap L^1(\mathbb{R})$ then $\|u\|_{L^\infty(\mathbb{R})} \leq |u|_{BV(\mathbb{R})}$.
4. Let $u \in BV(\mathbb{R})$ and let $(x_{i+1/2})_{i \in \mathbb{Z}}$ be an increasing sequence of real values such that $\mathbb{R} = \cup_{i \in \mathbb{Z}} [x_{i-1/2}, x_{i+1/2}]$. For $i \in \mathbb{Z}$, let $K_i = (x_{i-1/2}, x_{i+1/2})$ and u_i be the mean value of u over K_i. Then, choosing conveniently φ in the definition of $|u|_{BV(\mathbb{R})}$, it is easy to show that

$$\sum_{i \in \mathbb{Z}} |u_{i+1} - u_i| \leq |u|_{BV(\mathbb{R})}. \tag{19.6}$$

   Inequality (19.6) is used for the classical proof of "BV estimates" for the approximate solutions given by finite volume schemes (see Lemma 21.5 page 142 and Corollary 21.1 page 142).

   Note that (19.6) is also true when u_i is the mean value of u over a subinterval of K_i instead of the mean value of u over K_i.

Krushkov used a characterization of entropy weak solutions which is given in the following proposition.

Proposition 19.2 (Entropy weak solution using "Krushkov's entropies") *Let $f \in C^1(\mathbb{R}, \mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$, u is the unique entropy weak solution to Problem (19.1) if and only if u is such that*

$$\left\{ \begin{array}{l} u \in L^\infty(\mathbb{R} \times \mathbb{R}^\star_+), \\ \displaystyle\int_{\mathbb{R}} \int_{\mathbb{R}_+} |u(x,t) - \kappa| \varphi_t(x,t) dt dx + \\ \displaystyle\int_{\mathbb{R}} \int_{\mathbb{R}_+} \Big(f(u(x,t) \top \kappa) - f(u(x,t) \bot \kappa) \Big) \varphi_x(x,t) dt dx + \int_{\mathbb{R}} |u_0(x) - \kappa| \varphi(x,0) dx \geq 0, \\ \qquad\qquad \forall \varphi \in C^1_c(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+), \forall \kappa \in \mathbb{R}. \end{array} \right. \tag{19.7}$$

The result of existence of an entropy weak solution defined by (19.4) was already proved by passing to the limit on the solutions of an appropriate numerical scheme, see e.g. Oleinik [120], and may also be obtained by passing to the limit on finite volume approximations of the solution (see Theorem 21.1 page 139 in the one-dimensional case and Theorem 29.2 page 187 in the multidimensional case).

Remark 19.5 An entropy weak solution is sometimes defined as a function u satisfying:

$$\left\{\begin{array}{l} \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}}\int_{\mathbb{R}_+} f(u(x,t))\varphi_x(x,t)dtdx + \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0, \\ \qquad\qquad \forall \varphi \in C^1_c(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}). \\ \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}_+} \eta(u(x,t))\varphi_t(x,t)dtdx + \int_{\mathbb{R}}\int_{\mathbb{R}_+} \Phi(u(x,t))\varphi_x(x,t)dtdx \geq 0, \\ \qquad\qquad \forall \varphi \in C^1_c(\mathbb{R}\times\mathbb{R}^\star_+,\mathbb{R}_+), \\ \qquad \text{for all convex function } \eta \in C^1(\mathbb{R},\mathbb{R}) \text{ and } \Phi \in C^1(\mathbb{R},\mathbb{R}) \text{ such that } \Phi' = \eta' f'. \end{array}\right. \tag{19.8}$$

The uniqueness of an entropy weak solution thus defined depends on the functional space to which u is chosen to belong. Indeed, the uniqueness result given in Theorem 19.1 is no longer true with u defined by (19.8) such that

$$u,\ f(u) \in L^1_{loc}(\mathbb{R}\times\mathbb{R}_+),\ u \in L^\infty(\mathbb{R}\times(\varepsilon,\infty)),\ \forall \varepsilon \in \mathbb{R}^\star_+. \tag{19.9}$$

Under Assumption (19.9), every term in (19.8) makes sense. Note that (19.9)-(19.8) is weaker than (19.4). An easy counterexample to a uniqueness result of the solution to (19.8)-(19.9) is obtained with $f(s) = s^2$ for all $s \in \mathbb{R}$ and $u_0(x) = 0$ for a.e. $x \in \mathbb{R}$. In this case, a first solution to (19.8)-(19.9) is $u(x,t) = 0$ for a.e. $(x,t) \in \mathbb{R}\times\mathbb{R}_+$ (it is the entropy weak solution). A second solution to (19.8)-(19.9) is defined for a.e. $(x,t) \in \mathbb{R}\times\mathbb{R}_+$ by

$$\begin{array}{l} u(x,t) = 0, \text{ if } x < -\sqrt{t} \text{ or } x > \sqrt{t}, \\ u(x,t) = \frac{x}{2t}, \text{ if } -\sqrt{t} < x < \sqrt{t}. \end{array}$$

This second solution is not an entropy weak solution: it does not satisfy (19.4). Also note that this second solution is not in the space $C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$ nor in the space $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ (it belongs to $L^\infty(\mathbb{R}_+, L^1(\mathbb{R}))$). Indeed, under the assumption $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+) \cap C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$, the solution of (19.8) is unique.

The entropy weak solution to (19.1) satisfies the following L^∞ and BV stability properties:

Proposition 19.3 *Let $f \in C^1(\mathbb{R},\mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$. Let u be the entropy weak solution to (19.1). Then, $u \in C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$; furthermore, the following estimates hold:*

1. *$\|u(\cdot,t)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}$, for all $t \in \mathbb{R}_+$.*
2. *If $u_0 \in BV(\mathbb{R})$, then $|u(\cdot,t)|_{BV(\mathbb{R})} \leq |u_0|_{BV(\mathbb{R})}$, for all $t \in \mathbb{R}_+$.*

20 Numerical schemes in the linear case

We shall first introduce the numerical schemes in the linear case $f(u) = u$ in (19.1). The problem considered in this section is therefore

$$\left\{\begin{array}{rll} u_t(x,t) + u_x(x,t) &= 0 & x \in \mathbb{R},\ \ t \in \mathbb{R}_+, \\ u(x,0) &= u_0(x), & x \in \mathbb{R}. \end{array}\right. \tag{20.1}$$

Assume that $u_0 \in C^1(\mathbb{R},\mathbb{R})$; Problem (20.1) has a unique classical solution, as defined in Definition 19.1, which is $u(x,t) = u_0(x-t)$ for all $(x,t) \in \mathbb{R}\times\mathbb{R}_+$. If $u_0 \in L^\infty(\mathbb{R})$, then Problem (20.1) has a unique weak solution, as defined in Definition 19.2, which is again $u(x,t) = u_0(x-t)$ for a.e. $(x,t) \in \mathbb{R}\times\mathbb{R}_+$. Therefore, if $u_0 \geq 0$, the solution u is also nonnegative. Hence, it is advisable for many problems that the solution given by the numerical scheme should preserve the nonnegativity of the solution.

20.1 The centered finite difference scheme

Assume $u_0 \in C(\mathbb{R}, \mathbb{R})$. Let $h \in \mathbb{R}_+^\star$ and $x_i = ih$ for all $i \in \mathbb{Z}$. Let $k \in \mathbb{R}_+^\star$ be the time step. With the explicit Euler scheme for the time discretization (the implicit Euler scheme could also be used), the centered finite difference scheme associated to points x_i and k is

$$\left\{ \begin{array}{rl} \dfrac{u_i^{n+1} - u_i^n}{k} + \dfrac{u_{i+1}^n - u_{i-1}^n}{2h} & = 0, \quad \forall n \in \mathbb{N}, \quad \forall i \in \mathbb{Z}, \\ u_i^0 & = u_0(x_i), \quad \forall i \in \mathbb{Z}. \end{array} \right. \tag{20.2}$$

The discrete unknown u_i^n is expected to be an approximation of $u(x_i, nk)$ where u is the solution to (20.1).
It is well known that this scheme should be avoided. In particular, for the following reasons:

1. it does not preserve positivity, i.e. $u_i^0 \geq 0$ for all $i \in \mathbb{Z}$ does not imply $u_i^1 \geq 0$ for all $i \in \mathbb{Z}$; take for instance $u_i^0 = 0$ for $i \leq 0$ and $u_i^0 = 1$ for $i > 0$, then $u_0^1 = -k/(2h) < 0$;

2. it is not "L^∞-diminishing", i.e. $\max\{|u_i^0|,\, i \in \mathbb{Z}\} = 1$ does not imply that $\max\{|u_i^1|,\, i \in \mathbb{Z}\} \leq 1$; take for instance $u_i^0 = 1$ for $i \leq 0$ and $u_i^0 = 0$ for $i > 0$, then $\max\{|u_i^0|,\, i \in \mathbb{Z}\} = 1$ and $\max\{|u_i^1|,\, i \in \mathbb{Z}\} = 1 + k/(2h)$;

3. it is not "L^2-diminishing", i.e. $\sum_{i\in\mathbb{Z}} (u_i^0)^2 = 1$ does not imply that $\sum_{i\in\mathbb{Z}} (u_i^1)^2 \leq 1$; take for instance $u_i^0 = 0$ for $i \neq 0$ and $u_i^0 = 1$ for $i = 0$, then $u_0^1 = 1, u_1^1 = k/(2h), u_{-1}^1 = -k/(2h)$, so that $\sum_{i\in\mathbb{Z}} (u_i^1)^2 = 1 + k^2/(2h^2) > 1$;

4. it is unstable in the von Neumann sense: if the initial condition is taken under the form $u_0(x) = \exp(ipx)$, where p is given in $\mathbb{Z}$, then $u(x,t) = \exp(-ipt)\exp(ipx)$ (i is, here, the usual complex number, u_0 and u take values in $\mathbb{C}$). Hence $\exp(-ipt)$ can be seen as an amplification factor, and its modulus is 1. The numerical scheme is stable in the von Neumann sense if the amplification factor for the discrete solution is less than or equal to 1. For the scheme (20.2), we have $u_j^1 = u_j^0 - (u_{j+1}^0 - u_{j-1}^0)k/(2h) = \exp(ipjh)\xi_{p,h,k}$, with $\xi_{p,h,k} = 1 - (\exp(iph) - \exp(-iph))k/(2h)$. Hence $|\xi_{p,h,k}|^2 = 1 + (k^2/h^2)\sin^2 ph > 1$ if $ph \neq q\pi$ for any q in $\mathbb{Z}$.

In fact, one can also show that there exists $u_0 \in C_c^1(\mathbb{R}, \mathbb{R})$ such that the solution given by the numerical scheme does not tend to the solution of the continuous problem when h and k tend to 0 (whatever the relation between h and k).

Remark 20.1 The scheme (20.2) is also a finite volume scheme with the (spatial) mesh $\mathcal{T}$ given by $x_{i+1/2} = (i + 1/2)h$ in Definition 20.1 below and with a centered choice for the approximation of $u(x_{i+1/2}, nk)$: the value of $u(x_{i+1/2}, nk)$ is approximated by $(u_i^n + u_{i+1}^n)/2$, see (20.6) where an upstream choice for $u(x_{i+1/2}, nk)$ is performed. In fact, the choice of u_i^0 is different in (20.6) and in (20.2) but this does not change the unstability of the centered scheme.

20.2 The upstream finite difference scheme

Consider now a nonuniform distribution of points x_i, i.e. an increasing sequence of real values $(x_i)_{i\in\mathbb{Z}}$ such that $\lim_{i\to\pm\infty} x_i = \pm\infty$. For all $i \in \mathbb{Z}$, we set $h_{i-1/2} = x_i - x_{i-1}$. The time discretization is performed with the explicit Euler scheme with time step $k > 0$. Still assuming $u_0 \in C(\mathbb{R}, \mathbb{R})$, consider the upwind (or upstream) finite difference scheme defined by

$$\left\{ \begin{array}{rl} \dfrac{u_i^{n+1} - u_i^n}{k} + \dfrac{u_i^n - u_{i-1}^n}{h_{i-\frac{1}{2}}} & = 0, \quad \forall n \in \mathbb{N}, \quad \forall i \in \mathbb{Z}, \\ u_i^0 & = u_0(x_i), \quad \forall i \in \mathbb{Z}. \end{array} \right. \tag{20.3}$$

Rewriting the scheme as

$$u_i^{n+1} = (1 - \frac{k}{h_{i-\frac{1}{2}}})u_i^n + \frac{k}{h_{i-\frac{1}{2}}}u_{i-1}^n,$$

it appears that if $\inf_{i\in\mathbb{Z}} h_{i-1/2} > 0$ and if k is such that

$$k \leq \inf_{i\in\mathbb{Z}} h_{i-1/2} \tag{20.4}$$

then u_i^{n+1} is a convex combination of u_i^n and u_{i-1}^n. By induction, this proves that the scheme (20.3) is stable, in the sense that if u_0 is such that $U_m \leq u_0(x) \leq U_M$ for a.e. $x \in \mathbb{R}$, where U_m, $U_M \in \mathbb{R}$, then $U_m \leq u_i^n \leq U_M$ for any $i \in \mathbb{Z}$ and $n \in \mathbb{N}$.
Moreover, if $u_0 \in C^2(\mathbb{R},\mathbb{R})\cap L^\infty(\mathbb{R})$ and u_0' and u_0'' belong to $L^\infty(\mathbb{R})$, it is easily shown that the scheme is consistent in the finite difference sense; indeed, the consistency error defined by

$$R_i^n = \frac{u(x_i,(n+1)k) - u(x_i,nk)}{k} + \frac{u(x_i,nk) - u(x_{i-1},nk)}{h_{i-\frac{1}{2}}} \tag{20.5}$$

is such that, if the CFL condition (20.4) holds, $|R_i^n| \leq Ch$, where $h = \sup_{i\in\mathbb{Z}} h_i$ and $C \geq 0$ only depends on u_0 (recall that u is the solution to problem (20.1)). Hence the following error estimate holds.

Proposition 20.1 (Error estimate for the upwind finite difference scheme)
Let $u_0 \in C^2(\mathbb{R},\mathbb{R})\cap L^\infty(\mathbb{R})$, such that u_0' and $u_0'' \in L^\infty(\mathbb{R})$. Let $(x_i)_{i\in\mathbb{Z}}$ be an increasing sequence of real values such that $\lim_{i\to\pm\infty} x_i = \pm\infty$. Let $h = \sup_{i\in\mathbb{Z}} h_{i-\frac{1}{2}}$, and assume that $h < \infty$ and $\inf_{i\in\mathbb{Z}} h_{i-1/2} > 0$. Let $k > 0$ such that $k \leq \inf_{i\in\mathbb{Z}} h_{i-1/2}$. Let u denote the unique solution to (20.1) and $\{u_i^n,\ i \in \mathbb{Z},\ n \in \mathbb{N}\}$ be given by (20.3); let $e_i^n = u(x_i,nk) - u_i^n$, for any $n \in \mathbb{N}$ and $i \in \mathbb{Z}$, and let $T \in]0,+\infty[$ (note that $u(x_i,nk)$ is well defined since $u \in C^2(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$).
Then there exists $C \in \mathbb{R}_+$, only depending on u_0, such that $|e_i^n| \leq ChT$, for any $n \in \mathbb{N}$ such that $nk \leq T$, and for any $i \in \mathbb{Z}$.

Proof of Proposition 20.1
Let $i \in \mathbb{Z}$ and $n \in \mathbb{N}$. By definition of the consistency error R_i^n in (20.5), the error e_i^n satisfies

$$\frac{e_i^{n+1} - e_i^n}{k} + \frac{e_i^n - e_{i-1}^n}{h_{i-\frac{1}{2}}} = R_i^n.$$

Hence

$$e_i^{n+1} = e_i^n(1 - \frac{k}{h_{i-\frac{1}{2}}}) + \frac{k}{h_{i-\frac{1}{2}}}e_{i-1}^n + kR_i^n.$$

Using $|R_i^n| \leq Ch$ (for some C only depending on u_0) and the assumption $k \leq \inf_{i\in\mathbb{Z}} h_{i-1/2}$, this yields

$$|e_i^{n+1}| \leq \sup_{j\in\mathbb{Z}} |e_j^n| + Ckh.$$

Since $e_i^0 = 0$ for any $i \in \mathbb{Z}$, an induction yields

$$\sup_{i\in\mathbb{Z}} |e_i^n| \leq Cnkh$$

and the result follows. ∎

Note that in the above proof, the linearity of the equation and the regularity of u_0 are used. The next questions to arise are what to do in the case of a nonlinear equation and in the case $u_0 \in L^\infty(\mathbb{R})$.

20.3 The upwind finite volume scheme

Let us first give a definition of the admissible meshes for the finite volume schemes.

Definition 20.1 (One-dimensional admissible mesh) An admissible mesh $\mathcal{T}$ of $\mathbb{R}$ is given by an increasing sequence of real values $(x_{i+1/2})_{i\in\mathbb{Z}}$, such that $\mathbb{R} = \cup_{i\in\mathbb{Z}}[x_{i-1/2}, x_{i+1/2}]$. The mesh $\mathcal{T}$ is the set $\mathcal{T} = \{K_i,\ i \in \mathbb{Z}\}$ of subsets of $\mathbb{R}$ defined by $K_i = (x_{i-1/2}, x_{i+1/2})$ for all $i \in \mathbb{Z}$. The length of K_i is denoted by h_i, so that $h_i = x_{i+1/2} - x_{i-1/2}$ for all $i \in \mathbb{Z}$. It is assumed that $h = \text{size}(\mathcal{T}) = \sup\{h_i,\ i \in \mathbb{Z}\} < +\infty$ and that, for some $\alpha \in \mathbb{R}_+^\star$, $\alpha h \le \inf\{h_i,\ i \in \mathbb{Z}\}$.

Consider an admissible mesh in the sense of Definition 20.1. Let $k \in \mathbb{R}_+^\star$ be the time step. Assume $u_0 \in L^\infty(\mathbb{R})$ (this is a natural hypothesis for the finite volume framework). Integrating (20.1) on each control volume of the mesh, approximating the time derivatives by differential quotients and using an upwind choice for $u(x_{i+\frac{1}{2}}, nk)$ yields the following (time explicit) scheme:

$$\begin{cases} h_i \dfrac{u_i^{n+1} - u_i^n}{k} + u_i^n - u_{i-1}^n &= 0, \quad \forall n \in \mathbb{N}, \quad \forall i \in \mathbb{Z}, \\ u_i^0 &= \dfrac{1}{h_i}\displaystyle\int_{K_i} u_0(x)dx, \quad \forall i \in \mathbb{Z}. \end{cases} \tag{20.6}$$

The value u_i^n is expected to be an approximation of u (solution to (20.1)) in K_i at time nk. It is easily shown that this scheme is not consistent in the finite difference sense if u_i^n is considered to be an approximation of $u(x_i, nk)$ with, for instance, $x_i = (x_{i-1/2} + x_{i+1/2})/2$ for all $i \in \mathbb{Z}$. Even if $u_0 \in C_c^\infty(\mathbb{R}, \mathbb{R})$, the quantity R_i^n defined by (20.5) does not satisfy (except in particular cases) $|R_i^n| \le Ch$, with some C only depending on u_0.

It is however possible to interpret this scheme as another expression of the upwind finite difference scheme (20.3) (except for the minor modification of u_i^0, $i \in \mathbb{Z}$). One simply needs to consider u_i^n as an approximation of $u(x_{i+1/2}, nk)$ which leads to a consistency property in the finite difference sense. Indeed, taking $x_j = x_{j+1/2}$ (for $j = i$ and $i-1$) in the definition (20.5) of R_i^n yields $|R_i^n| \le Ch$, where C only depends on u_0. Therefore, a convergence result for this scheme is given by the proposition 20.1. This analogy cannot be extended to the general case of "monotone flux schemes" (see Definition 21.1 page 134 below) for a nonlinear equation for which there may be no value of x_i (independent of u) leading to such a consistency property, see Remark 21.1 page 134 for a counterexample (the analogy holds however for the scheme (21.8), convenient for a nondecreasing function f, see Remark 21.3).

The approximate finite volume solution $u_{\mathcal{T},k}$ may be defined on $\mathbb{R}\times\mathbb{R}_+$ from the discrete unknowns u_i^n, $i \in \mathbb{Z}$, $n \in \mathbb{N}$ which are computed in (20.6):

$$u_{\mathcal{T},k}(x,t) = u_i^n \text{ for } x \in K_i \text{ and } t \in [nk, (n+1)k). \tag{20.7}$$

The following L^∞ estimate holds:

Lemma 20.1 (L^∞ estimate in the linear case) *Let $u_0 \in L^\infty(\mathbb{R})$ and U_m, $U_M \in \mathbb{R}$ such that $U_m \le u_0(x) \le U_M$ for a.e. $x \in \mathbb{R}$. Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 and let $k \in \mathbb{R}_+^\star$ satisfying the Courant-Friedrichs-Levy (CFL) condition*

$$k \le \inf_{i\in\mathbb{Z}} h_i.$$

(note that taking $k \le \alpha h$ implies the above condition). Let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (20.6) and (20.7).
Then,

$$U_m \le u_{\mathcal{T},k}(x,t) \le U_M \ \text{ for a.e. } x \in \mathbb{R} \text{ and a.e. } t \in \mathbb{R}_+.$$

Proof of Lemma 20.1

The proof that $U_m \leq u_i^n \leq U_M$, for all $i \in \mathbb{Z}$ and $n \in \mathbb{N}$, as in the case of the upwind finite difference scheme (see (20.3) page 126), consists in remarking that equation (20.6) gives, under the CFL condition, an expression of u_i^{n+1} as a linear convex combination of u_i^n and u_{i-1}^n, for all $i \in \mathbb{Z}$ and $n \in \mathbb{N}$. ∎

The following inequality will be crucial for the proof of convergence.

Lemma 20.2 (Weak BV estimate, linear case) *Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ satisfying the CFL condition*

$$k \leq (1-\xi) \inf_{i \in \mathbb{Z}} h_i, \tag{20.8}$$

for some $\xi \in (0,1)$ (taking $k \leq (1-\xi)\alpha h$ implies this condition).
Let $\{u_i^n,\ i \in \mathbb{Z},\ n \in \mathbb{N}\}$ be given by the finite volume scheme (20.6). Let $R \in \mathbb{R}_+^\star$ and $T \in \mathbb{R}_+^\star$ and assume $h = \mathrm{size}(\mathcal{T}) < R$, $k < T$. Let $i_0 \in \mathbb{Z}$, $i_1 \in \mathbb{Z}$ and $N \in \mathbb{N}$ be such that $-R \in \overline{K}_{i_0}$, $R \in \overline{K}_{i_1}$ and $T \in (Nk, (N+1)k]$ (note that $i_0 < i_1$).
Then there exists $C \in \mathbb{R}_+^\star$, only depending on R, T, u_0, α and ξ, such that

$$\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|u_i^n - u_{i-1}^n| \leq Ch^{-1/2}. \tag{20.9}$$

Proof of Lemma 20.2

Multiplying the first equation of (20.6) by ku_i^n and summing on $i = i_0, \ldots, i_1$ and $n = 0, \ldots N$ yields $A + B = 0$ with

$$A = \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} h_i(u_i^{n+1} - u_i^n)u_i^n$$

and

$$B = \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)u_i^n.$$

Noting that

$$A = -\frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} h_i(u_i^{n+1} - u_i^n)^2 + \frac{1}{2} \sum_{i=i_0}^{i_1} h_i[(u_i^{N+1})^2 - (u_i^0)^2]$$

and using the scheme (20.6) gives

$$A = -\frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} \frac{k^2}{h_i}(u_i^n - u_{i-1}^n)^2 + \frac{1}{2} \sum_{i=i_0}^{i_1} h_i[(u_i^{N+1})^2 - (u_i^0)^2];$$

therefore, using the CFL condition (20.8),

$$A \geq -(1-\xi)\frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)^2 - \frac{1}{2} \sum_{i=i_0}^{i_1} h_i(u_i^0)^2.$$

We now study the term B, which may be rewritten as

$$B = \frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)^2 + \frac{1}{2} \sum_{n=0}^{N} k[(u_{i_1}^n)^2 - (u_{i_0-1}^n)^2].$$

Thanks to the L^∞ estimate of Lemma 20.1 page 128, this last equality implies that

$$B \geq \frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)^2 - T \max\{-U_m, U_M\}^2.$$

Therefore, since $A + B = 0$ and $\sum_{i=i_0}^{i_1} h_i \leq 4R$, the following inequality holds:

$$0 \geq \xi \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)^2 - (4R + 2T) \max\{-U_m, U_M\}^2,$$

which, in turn, gives the existence of $C_1 \in \mathbb{R}_+^\star$, only depending on R, T, u_0 and ξ such that

$$\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k(u_i^n - u_{i-1}^n)^2 \leq C_1. \tag{20.10}$$

Finally, using

$$\sum_{i=i_0}^{i_1} 1 \leq \sum_{i=i_0}^{i_1} \frac{h_i}{\alpha h} \leq \frac{4R}{\alpha h},$$

the Cauchy-Schwarz inequality leads to

$$[\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|u_i^n - u_{i-1}^n|]^2 \leq C_1 2T \frac{4R}{\alpha h},$$

which concludes the proof of the lemma. ∎

Contrary to the discrete H_0^1 estimates which were obtained on the approximate finite volume solutions of elliptic equations, see e.g. (9.24), the weak BV estimate (20.9) is not related to an *a priori* estimate on the solution to the continuous problem (20.1). It does not give any compactness property in the space $L^1_{loc}(\mathbb{R})$ (there are some counterexamples); such a compactness property is obtained thanks to a "strong BV estimate" (with, for instance, an L^∞ estimate) as it is recalled below (see Lemma 21.4). In the one-dimensional case which is studied here such a "strong BV estimate" can be obtained if $u_0 \in BV(\mathbb{R})$, see Corollary 21.1; this is no longer true in the multidimensional case with general meshes, for which only the above weak BV estimate is available.

Remark 20.2 The weak BV estimate is a crucial point for the proof of convergence. Indeed, the property which is used in the proof of convergence (see Proposition 20.2 below) is, with the notations of Lemma 20.2,

$$h \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|u_i^n - u_{i-1}^n| \to 0, \text{ as } h \to 0, \tag{20.11}$$

for R, T, u_0, α and ξ fixed.

If a piecewise constant function $u_{\mathcal{T},k}$, such as given by (20.7) (with some u_i^n in $\mathbb{R}$, not necessarily given by (20.6)), is bounded in (for instance) $L^\infty(\mathbb{R} \times \mathbb{R}_+)$ and converges in $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$ as $h \to 0$ and $k \to 0$ (with a possible relation between k and h) then (20.11) holds. This proves that the hypothesis (20.11) is included in the hypotheses of the classical Lax-Wendroff theorem of convergence (see Theorem 21.2 page 143); note that (20.11) is implied by (20.9) and that it is weaker than (20.9)).

We show in the following remark how the " weak" and " strong" BV estimates may "formally" be obtained on the "continuous equation"; this gives a hint of the reason why this estimate may be obtained even if the exact solution does not belong to the space $BV(\mathbb{R} \times \mathbb{R}_+)$. A similar remark also holds in the nonlinear case (i.e. for Problem (19.1)).

Remark 20.3 (Formal derivations of the strong and weak BV estimates) When approximating the solution to (20.1) by the finite volume scheme (20.6) (with $h_i = h$ for all i, for the sake of simplicity), the equation to which an approximation of a solution is sought is "close" to the equation

$$u_t + u_x - \varepsilon u_{xx} = 0 \tag{20.12}$$

where $\varepsilon = \frac{h-k}{2}$ is positive under the CFL condition (20.8), which ensures that the scheme is diffusive. We assume that u is regular enough, with null limits for $u(x,t)$ and its derivatives as $x \to \pm\infty$.

(i) "Strong" BV estimate.
Derivating the equation (20.12) with respect to the variable x, multiplying by $\mathrm{sign}_r(u_x(x,t))$, where sign_r denotes a nondecreasing regularization of the function sign, and integrating over $\mathbb{R}$ yields

$$\Big(\int_{\mathbb{R}} \phi_r(u_x(x,t))dx\Big)_t + \int_{\mathbb{R}} u_{xx}(x,t)\mathrm{sign}_r(u_x(x,t))dx = -\varepsilon \int_{\mathbb{R}} \mathrm{sign}_r'(u_x(x,t))(u_{xx}(x,t))^2 dx \le 0,$$

where $\phi_r' = \mathrm{sign}_r$ and $\phi_r(0) = 0$. Since

$$\int_{\mathbb{R}} u_{xx}(x,t)\mathrm{sign}_r(u_x(x,t))dx = \int_{\mathbb{R}} (\phi_r(u_x(x,t)))_x dx = 0,$$

this yields, passing to the limit on the regularization, that $\|u_x(\cdot,t)\|_{L^1(\mathbb{R})}$ is nonincreasing with respect to t. Copying this formal proof on the numerical scheme yields a strong BV estimate, which is an *a priori* estimate giving compactness properties in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$, see Lemma 21.5, Corollary 21.1 and Lemma 21.4 page 141.

(ii) "Weak" BV estimate
Multiplying (20.12) by u and summing over $\mathbb{R}\times(0,T)$ yields

$$\frac{1}{2}\int_{\mathbb{R}} u^2(x,T)dx - \frac{1}{2}\int_{\mathbb{R}} u^2(x,0)dx + \int_0^T \int_{\mathbb{R}} \varepsilon u_x^2(x,t)dxdt = 0,$$

which yields in turn

$$\varepsilon \int_0^T \int_{\mathbb{R}} u_x^2(x,t)dxdt \le \frac{1}{2}\|u_0\|^2_{L^2(\mathbb{R})}.$$

This is the continuous analogous of (20.10). Hence if $h - k = \varepsilon \ge \xi h$ (this is Condition (20.8), note that this condition is more restrictive than the usual CFL condition required for the L^∞ stability), the discrete equivalent of this formal proof yields (20.10) (and then (20.9)).

In the first case, we derivate the equation and we use some regularity on u_0 (namely $u_0 \in BV(\mathbb{R})$). In the second case, it is sufficient to have $u_0 \in L^\infty(\mathbb{R})$ but we need the diffusion term to be large enough in order to obtain the estimate which, by the way, does not yield any estimate on the solution of (20.12) with $\varepsilon = 0$. This formal derivation may be carried out similarly in the nonlinear case.

Let us now give a convergence result for the scheme (20.6) in $L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology. Recall that a sequence $(v_n)_{n\in\mathbb{N}} \subset L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$ converges to $v \in L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$ in $L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology if

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}} (v_n(x,t) - v(x,t))\varphi(x,t)dxdt \to 0 \text{ as } n\to\infty,\ \forall \varphi \in L^1(\mathbb{R}\times\mathbb{R}_+^\star).$$

A stronger convergence result is available, and comes from the nonlinear study given in Section 21.

Proposition 20.2 (Convergence in the linear case) *Let $u_0 \in L^\infty(\mathbb{R})$ and u be the unique weak solution to Problem (20.1) page 125 in the sense of Definition 19.2 page 123, with $f(s) = s$ for all $s \in \mathbb{R}$. Let $\xi \in (0,1)$ and $\alpha > 0$ be given. Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ satisfying the CFL condition (20.8) page 129 (taking $k \leq (1-\xi)\alpha h$ implies this condition, note that ξ and α do not depend on $\mathcal{T}$).*
Let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (20.6) and (20.7). Then $u_{\mathcal{T},k} \to u$ in $L^\infty(\mathbb{R} \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology as $h = \mathrm{size}(\mathcal{T}) \to 0$.

Proof of Proposition 20.2

Let $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$ be a sequence of meshes and time steps satisfying the hypotheses of Proposition 20.2 and such that $\mathrm{size}(\mathcal{T}_m) \to 0$ as $m \to \infty$.
Lemma 20.1 gives the existence of a subsequence, still denoted by $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$, and of a function $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+^\star)$ such that $u_{\mathcal{T}_m,k_m} \to u$ in $L^\infty(\mathbb{R} \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology, as $m \to +\infty$. There remains to show that u is the solution of (19.3) (with $f(s) = s$ for all $s \in \mathbb{R}$). The uniqueness of the weak solution to Problem (20.1) will then imply that the full sequence converges to u.

Let $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$. Let $m \in \mathbb{N}$ and $\mathcal{T} = \mathcal{T}_m$, $k = k_m$ and $h = \mathrm{size}(\mathcal{T})$. Let us multiply the first equation of (20.6) by $(k/h_i)\varphi(x, nk)$, integrate over $x \in K_i$ and sum for all $i \in \mathbb{Z}$ and $n \in \mathbb{N}$. This yields

$$A_m + B_m = 0$$

with

$$A_m = \sum_{i\in\mathbb{Z}} \sum_{n\in\mathbb{N}} (u_i^{n+1} - u_i^n) \int_{K_i} \varphi(x, nk)dx$$

and

$$B_m = \sum_{i\in\mathbb{Z}} \sum_{n\in\mathbb{N}} k(u_i^n - u_{i-1}^n)\frac{1}{h_i} \int_{K_i} \varphi(x, nk)dx.$$

Let us remark that $A_m = A_{1,m} - A'_{1,m}$ with

$$A_{1,m} = -\int_k^\infty \int_{\mathbb{R}} u_{\mathcal{T},k}(x,t)\varphi_t(x, t-k)dxdt - \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx$$

and

$$A'_{1,m} = \sum_{i\in\mathbb{Z}} u_i^0 \int_{K_i} \varphi(x,0)dx - \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx.$$

Using the fact that $\sum_{i\in\mathbb{Z}} u_i^0 1_{K_i} \to u_0$ in $L^1_{loc}(\mathbb{R})$ as $m \to \infty$, we get that $A'_{1,m} \to 0$ as $m \to \infty$. (Recall that $1_{K_i}(x) = 1$ if $x \in K_i$ and $1_{K_i}(x) = 0$ if $x \notin K_i$.)
Therefore, since $u_{\mathcal{T},k} \to u$ in $L^\infty(\mathbb{R} \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology as $m \to \infty$, and $\varphi_t(\cdot, \cdot - k)1_{\mathbb{R}\times(k,\infty)} \to \varphi_t$ in $L^1(\mathbb{R} \times \mathbb{R}_+^\star)$ (note that $k \to 0$ thanks to (20.8)),

$$\lim_{m\to+\infty} A_m = \lim_{m\to+\infty} A_{1,m} = -\int_{\mathbb{R}_+} \int_{\mathbb{R}} u(x,t)\varphi_t(x,t)dxdt - \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx.$$

Let us now turn to the study of B_m. We compare B_m with

$$B_{1,m} = -\sum_{n\in\mathbb{N}} \int_{nk}^{(n+1)k} \int_{\mathbb{R}} u_{\mathcal{T},k}(x,t)\varphi_x(x, nk)dxdt,$$

which tends to $-\int_{\mathbb{R}_+} \int_{\mathbb{R}} u(x,t)\varphi_x(x,t)dxdt$ as $m \to \infty$. The term $B_{1,m}$ can be rewritten as

$$B_{1,m} = \sum_{i\in\mathbb{Z}} \sum_{n\in\mathbb{N}} k(u_i^n - u_{i-1}^n)\varphi(x_{i-\frac{1}{2}}, nk).$$

Let $R > 0$ and $T > 0$ be such that $\varphi(x,t) = 0$ if $|x| \geq R$ or $t \geq T$. Then, there exists $C \in \mathbb{R}_+^\star$, only depending on φ, such that, if $h < R$ and $k < T$ (which is true for h small enough, thanks to (20.8)),

$$|B_m - B_{1,m}| \leq Ch \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|u_i^n - u_{i-1}^n|, \tag{20.13}$$

where $i_0 \in \mathbb{Z}$, $i_1 \in \mathbb{Z}$ and $N \in \mathbb{N}$ are such that $-R \in \overline{K}_{i_0}$, $R \in \overline{K}_{i_1}$ and $T \in (Nk, (N+1)k]$. Using (20.13) and Lemma 20.2, we get that $B_m - B_{1,m} \to 0$ and then

$$B_m \to -\int_{\mathbb{R}_+}\int_{\mathbb{R}} u(x,t)\varphi_x(x,t)dxdt \text{ as } m \to \infty,$$

which completes the proof that u is the weak solution to Problem (20.1) page 125 (note that here the useful consequence of lemma 20.2 is (20.11)). ∎

Remark 20.4 In Proposition 20.2, a simpler proof of convergence could be achieved, with $\xi = 0$, using a multiplication of the first equation of (20.6) by $(k/h_i)\varphi(x_{i-1/2}, nk)$. However, this proof does not generalize to the general case of nonlinear hyperbolic problems.

Remark 20.5 Proving the convergence of the finite difference method (with the scheme (20.3)) with $u_0 \in L^\infty(\mathbb{R})$ can be done using the same technique as the proof of the finite volume method (that is considering the finite difference scheme as a finite volume scheme on a convenient mesh).

21 The nonlinear case

In this section, finite volume schemes for the discretization of Problem (19.1) are presented and a theorem of convergence is given (Theorem 21.1) which will be generalized to the multidimensional case in the next chapter. We also recall the classical proof of convergence which uses a "strong BV estimate" and the Lax-Wendroff theorem. This proof, however, does not seem to extend to the multidimensional case for general meshes. The following properties are assumed to be satisfied by the data of problem (19.1).

Assumption 21.1 *The flux function f belongs to $C^1(\mathbb{R},\mathbb{R})$, the initial data u_0 belongs to $L^\infty(\mathbb{R})$ and U_m, $U_M \in \mathbb{R}$ are such that $U_m \leq u_0 \leq U_M$ a.e. on $\mathbb{R}$.*

21.1 Meshes and schemes

Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128 and $k \in \mathbb{R}_+^\star$ be the time step. In the general nonlinear case, the finite volume scheme for the discretization of Problem (19.1) page 122 reads

$$\begin{cases} \dfrac{h_i}{k}(u_i^{n+1} - u_i^n) + f_{i+1/2}^n - f_{i-1/2}^n &= 0, \quad \forall n \in \mathbb{N}, \quad \forall i \in \mathbb{Z}, \\ u_i^0 &= \dfrac{1}{h_i}\displaystyle\int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x)dx, \quad \forall i \in \mathbb{Z}, \end{cases} \tag{21.1}$$

where u_i^n is expected to be an approximation of u at time $t_n = nk$ in cell K_i. The quantity $f_{i+1/2}^n$ is often called the numerical flux at point $x_{i+1/2}$ and time t_n (it is expected to be an approximation of $f(u)$ at point $x_{i+1/2}$ and time t_n). Note that a common expression of $f_{i+1/2}^n$ is used for both equations i and $i+1$ in (21.1); therefore the scheme (21.1) satisfies the property of conservativity, common to all finite

volume schemes. In the case of a so-called $2p+1$ point scheme with ($p \in \mathbb{N}^\star$), the numerical flux may be written

$$f^n_{i+1/2} = g^n_{i+\frac{1}{2}}(u^n_{i-p+1}, \ldots, u^n_{i+p}), \tag{21.2}$$

where $g^n_{i+\frac{1}{2}}$ is the numerical flux function at point $x_{i+\frac{1}{2}}$ and time t_n, which determines the scheme. Note that the numerical flux may thus depend on the interface and the time. This is important in applications, for instance in the case of boundary faces (see section 19) or interfaces coupling different domains. For $p = 1$, the flux reads

$$f^n_{i+1/2} = g^n_{i+\frac{1}{2}}(u^n_i, \ldots, u^n_{i+1}), \tag{21.3}$$

and yields a 3-point scheme.
As in the linear case (20.7) page 128, the approximate finite volume solution is defined by

$$u_{\mathcal{T},k}(x,t) = u^n_i \text{ for } x \in K_i \text{ and } t \in [nk, (n+1)k). \tag{21.4}$$

The property of consistency for the finite volume scheme (21.1), (21.2) with $2p+1$ points, is ensured by writing the following condition:

$$g(s, \ldots, s) = f(s), \quad \forall s \in \mathbb{R}. \tag{21.5}$$

This condition is equivalent to writing the consistency of the approximation of the flux (as in the elliptic and parabolic cases, which were described in the previous chapters, see e.g. Section 5).

Remark 21.1 (Finite volumes and finite differences) We can remark that, as in the elliptic case, the condition (21.5) does not generally give the consistency of the scheme (21.1) when it is considered as a finite difference scheme. For instance, assume $f(s) = s^2$ for all $s \in \mathbb{R}$, $p = 1$ and $g(a,b) = f_1(a) + f_2(b)$ for all a, $b \in \mathbb{R}$ with $f_1(s) = \max\{s, 0\}^2$, $f_2(s) = \min\{s, 0\}^2$ (which is shown below to be a good choice, see Example 21.1). Assume also $h_{2i} = h$ and $h_{2i+1} = h/2$ for all $i \in \mathbb{Z}$. In this case, there is no choice of points $x_i \in \mathbb{R}$ such that the quantity $(f^n_{i+1/2} - f^n_{i-1/2})/h_i$ is an approximation of order 1 of $(f(u))_x(x_i, nk)$, for any regular function u, when $u^n_i = u(x_i, nk)$ for all $i \in \mathbb{Z}$. Indeed, up to second order terms, this property of consistency is achieved if and only if $f_2'(a)|x_{i+1} - x_i| + f_1'(a)|x_{i-1} - x_i| = f'(a)h_i$ for all $i \in \mathbb{Z}$ and for all $a \in \mathbb{R}$. Choosing $a > 0$ and $a < 0$, this condition leads to $|x_{i+1} - x_i| = h_i$ and $|x_{i+1} - x_i| = h_{i+1}$ for all $i \in \mathbb{Z}$, which is impossible.

Examples of convenient choices for the function g will now be given. An interesting class of schemes is the class of 3-points schemes with a monotone flux, which we now define.

Definition 21.1 (Monotone flux schemes) Let $p = 1$. Under Assumption 21.1, the finite volume scheme (21.1)-(21.3) is said to be a "monotone flux scheme" if the function g, only depending on f, U_m and U_M, satisfies the following assumptions:

- g is locally Lipschitz continuous from $\mathbb{R}^2$ to $\mathbb{R}$,
- $g(s,s) = f(s)$, for all $s \in [U_m, U_M]$,
- $(a,b) \mapsto g(a,b)$, from $[U_m, U_M]^2$ to $\mathbb{R}$, is nondecreasing with respect to a and nonincreasing with respect to b.

The monotone flux property seems to be remarkable; indeed, monotone flux schemes are consistent in the finite volume sense, they are L^∞-stable under a condition (the so called Courant-Friedrichs-Levy condition) of the type $k \leq C_1 h$, where C_1 only depends on g and u_0 (see Section 21.2 page 136 below), and they are "consistent with the entropy inequalities" also under a condition of the type $k \leq C_2 h$, where C_2 only depends on g and u_0 (but C_2 may be different of C_1, see Section 21.3 page 136).

Remark 21.2 A monotone flux scheme is a monotone scheme, under a Courant-Friedrichs-Levy condition, which means that the scheme can be written under the form

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

with H nondecreasing with respect to its three arguments.

Example 21.1 (Examples of monotone flux schemes) (see also GODLEWSKI and RAVIART [76], LEVEQUE [100] and references therein). Under Assumption 21.1, here are some numerical flux functions g for which the finite volume scheme (21.1)-(21.2) is a monotone flux scheme (in the sense of Definition 21.1):

- the flux splitting scheme: assume $f = f_1 + f_2$, with $f_1,\ f_2 \in C^1(\mathbb{R}, \mathbb{R})$, $f_1'(s) \geq 0$ and $f_2'(s) \leq 0$ for all $s \in [U_m, U_M]$ (such a decomposition for f is always possible, see the modified Lax-Friedrichs scheme below), and take

  $$g(a, b) = f_1(a) + f_2(b).$$

  Note that if $f' \geq 0$, taking $f_1 = f$ and $f_2 = 0$, the flux splitting scheme boils down to the upwind scheme, i.e. $g(a, b) = f(a)$.

- the Godunov scheme: the Godunov scheme, which was introduced in GODUNOV [77], may be summarized by the following expression.

  $$g(a, b) = \begin{cases} \min\{f(\xi), \xi \in [a, b]\} \text{ if } a \leq b, \\ \max\{f(\xi), \xi \in [b, a]\} \text{ if } b \leq a. \end{cases} \tag{21.6}$$

- the modified Lax-Friedrichs scheme : take

  $$g(a, b) = \frac{f(a) + f(b)}{2} + D(a - b), \tag{21.7}$$

  with $D \in \mathbb{R}$ such that $2D \geq \max\{|f'(s)|, s \in [U_m, U_M]\}$. Note that in this modified version of the Lax-Friedrichs scheme, the coefficient D only depends on f, U_m and U_M, while the original Lax-Friedrichs scheme consists in taking $D = h/(2k)$, in the case $h_i = h$ for all $i \in \mathbb{N}$, and therefore satisfies the three items of Definition 21.1 under the condition $h/k \geq \max\{|f'(s)|, s \in [U_m, U_M]\}$. However, an inverse CFL condition appears to be necessary for the convergence of the original Lax-Friedrichs scheme (see remark 30.1 page 189); such a condition is not necessary for the modified version.

  Note also that the modified Lax-Friedrichs scheme consists in a particular flux splitting scheme with $f_1(s) = (1/2)f(s) + Ds$ and $f_2(s) = (1/2)f(s) - Ds$ for $s \in [U_m, U_M]$.

Remark 21.3 In the case of a nondecreasing (resp. nonincreasing) function f, the Godunov monotone flux scheme (21.6) reduces to $g(a, b) = f(a)$ (resp. $f(b)$). Then, in the case of a nondecreasing function f, the scheme (21.1), (21.2) reduces to

$$h_i \frac{u_i^{n+1} - u_i^n}{k} + f(u_i^n) - f(u_{i-1}^n) = 0, \tag{21.8}$$

i.e. the upstream (or upwind) finite volume scheme. The scheme (21.8) is sometimes called "upstream finite difference" scheme. In that particular case (f monotone and 1D) it is possible to find points x_i in order to obtain a consistent scheme in the finite difference sense (if f is nondecreasing, take $x_i = x_{i+1/2}$ as for the scheme (20.6) page 128).

21.2 L^∞-stability for monotone flux schemes

Lemma 21.1 (L^∞ estimate in the nonlinear case) *Under Assumption 21.1, let $\mathcal{T}$ be an admissible mesh in the sense of definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ be the time step.*
Let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (21.1)-(21.4) and assume that the scheme is a monotone flux scheme in the sense of definition 21.1 page 134. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ with respect to its two arguments.
Under the Courant-Friedrichs-Levy (CFL) condition

$$k \leq \frac{\inf_{i\in\mathbb{Z}} h_i}{g_1 + g_2}, \tag{21.9}$$

(note that taking $k \leq \alpha h/(g_1+g_2)$ implies (21.9)),
the approximate solution $u_{\mathcal{T},k}$ satisfies

$$U_m \leq u_{\mathcal{T},k}(x,t) \leq U_M \ \text{ for a.e. } x \in \mathbb{R} \text{ and a.e. } t \in \mathbb{R}_+.$$

Proof of Lemma 21.1
Let us prove that

$$U_m \leq u_i^n \leq U_M, \ \forall i \in \mathbb{Z}, \ \forall n \in \mathbb{N}, \tag{21.10}$$

by induction on n, which proves the lemma. Assertion (21.10) holds for $n = 0$ thanks to the definition of u_i^0 in (21.1) page 133. Suppose that it holds for $n \in \mathbb{N}$.
For all $i \in \mathbb{Z}$, scheme (21.1), (21.2) (with $p = 1$) gives

$$u_i^{n+1} = (1 - b_{i+\frac{1}{2}}^n - a_{i-\frac{1}{2}}^n)u_i^n + b_{i+\frac{1}{2}}^n u_{i+1}^n + a_{i-\frac{1}{2}}^n u_{i-1}^n, \tag{21.11}$$

with

$$b_{i+\frac{1}{2}}^n = \begin{cases} \dfrac{k}{h_i}\dfrac{g(u_i^n, u_{i+1}^n) - f(u_i^n)}{u_i^n - u_{i+1}^n} \text{ if } u_i^n \neq u_{i+1}^n, \\ 0 \text{ if } u_i^n = u_{i+1}^n, \end{cases}$$

and

$$a_{i-\frac{1}{2}}^n = \begin{cases} \dfrac{k}{h_i}\dfrac{g(u_{i-1}^n, u_i^n) - f(u_i^n)}{u_{i-1}^n - u_i^n} \text{ if } u_i^n \neq u_{i-1}^n, \\ 0 \text{ if } u_i^n = u_{i-1}^n. \end{cases}$$

Since $f(u_i^n) = g(u_i^n, u_i^n)$ and thanks to the monotonicity of g, $0 \leq b_{i+\frac{1}{2}}^n \leq g_2 k/h_i$ and $0 \leq a_{i-\frac{1}{2}}^n \leq g_1 k/h_i$, for all $i \in \mathbb{Z}$. Therefore, under condition (21.9), the value u_i^{n+1} may be written as a convex linear combination of the values u_i^n and u_{i-1}^n. Assertion (21.10) is thus proved for $n+1$, which concludes the proof of the lemma. ∎

21.3 Discrete entropy inequalities

Lemma 21.2 (Discrete entropy inequalities) *Under Assumption 21.1, let $\mathcal{T}$ be an admissible mesh in the sense of definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ be the time step.*
Let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (21.1)-(21.4) and assume that the scheme is a monotone flux scheme in the sense of definition 21.1 page 134. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ with respect to its two arguments. Under the CFL condition (21.9), the following inequation holds:

$$\begin{aligned} &\frac{h_i}{k}\Big(|u_i^{n+1} - \kappa| - |u_i^n - \kappa|\Big) + \\ &g(u_i^n \top \kappa, u_{i+1}^n \top \kappa) - g(u_i^n \perp \kappa, u_{i+1}^n \perp \kappa) - g(u_{i-1}^n \top \kappa, u_i^n \top \kappa) + g(u_{i-1}^n \perp \kappa, u_i^n \perp \kappa) \leq 0, \\ &\qquad\qquad \forall\ n \in \mathbb{N},\ \forall\ i \in \mathbb{Z},\ \forall\ \kappa \in \mathbb{R}. \end{aligned} \tag{21.12}$$

Recall that $a \top b$ (resp. $a \bot b$) denotes the maximum (resp. the minimum) of the two real numbers a and b.

Proof of Lemma 21.2
Thanks to the monotonicity properties of g and to the condition (21.9) (see remark 21.2),

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n),\ \forall i \in \mathbb{Z}\!\!\mathbb{Z}\,,\ \forall n \in \mathbb{N},$$

where H is a function from $\mathbb{R}^3$ to $\mathbb{R}$ which is nondecreasing with respect to all its arguments and such that $\kappa = H(\kappa, \kappa, \kappa)$ for all $\kappa \in \mathbb{R}$.
Hence, for all $\kappa \in \mathbb{R}$,

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

and

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

which yields

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

In the same manner, we get

$$u_i^{n+1} \bot \kappa \geq H(u_{i-1}^n \bot \kappa, u_i^n \bot \kappa, u_{i+1}^n \bot \kappa),$$

and therefore, by substracting the last two equations,

$$|u_i^{n+1} - \kappa| \leq H(u_{i-1}^n \top \kappa, u_i^n \top \kappa, u_{i+1}^n \top \kappa) - H(u_{i-1}^n \bot \kappa, u_i^n \bot \kappa, u_{i+1}^n \bot \kappa),$$

that is (21.12). ∎

In the two next sections, we study the convergence of the schemes defined by (21.1), (21.2) with $p = 1$ (see the remarks 21.4 and 21.5 and Section 22 for the schemes with $2p + 1$ points).
We first develop a proof of convergence for the monotone flux schemes; this proof is based on a weak BV estimate similar to (20.9) like the proof of proposition 20.2 page 132 in the linear case. It will be generalized in the multidimensional case studied in Chapter 6. We then briefly describe the BV framework which gave the first convergence results; its generalization to the multidimensional case is not so easy, except in the case of Cartesian meshes.

21.4 Convergence of the upstream scheme in the general case

A proof of convergence similar to the proof of convergence given in the linear case can be developed. For the sake of simplicity, we shall consider only the case of a nondecreasing function f and of the classical upstream scheme (the general case for f and for the monotone flux schemes being handled in Chapter 6). We shall first prove a "weak BV" estimate.

Lemma 21.3 (Weak BV estimate for the nonlinear case) *Under Assumption 21.1, assume that f is nondecreasing. Let $\xi \in (0,1)$ be a given value. Let $\mathcal{T}$ be an admissible mesh in the sense of definition 20.1 page 128, let M be the Lipschitz constant of f in $[U_m, U_M]$ and let $k \in \mathbb{R}_+^\star$ satisfying the CFL condition*

$$k \leq (1-\xi)\frac{\inf_{i \in \mathbb{Z}\!\!\mathbb{Z}}\, h_i}{M}. \tag{21.13}$$

(The condition $k \leq (1-\xi)\alpha h/M$ implies the above condition.) Let $\{u_i^n,\ i \in \mathbb{Z}\!\!\mathbb{Z}\,,\ n \in \mathbb{N}\}$ be given by the finite volume scheme (21.1), (21.2) with $p = 1$ and $g(a,b) = f(a)$. Let $R \in \mathbb{R}_+^\star$ and $T \in \mathbb{R}_+^\star$ and

assume $h < R$ and $k < T$. Let $i_0 \in \mathbb{Z}$, $i_1 \in \mathbb{Z}$ and $N \in \mathbb{N}$ be such that $-R \in \overline{K}_{i_0}$, $R \in \overline{K}_{i_1}$,and $T \in (Nk, (N+1)k]$. Then there exists $C \in \mathbb{R}_+^\star$, only depending on R, T, u_0, α, f and ξ, such that

$$\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|f(u_i^n) - f(u_{i-1}^n)| \leq Ch^{-1/2}. \tag{21.14}$$

Proof of Lemma 21.3

We multiply the first equation of (21.1) by ku_i^n, and we sum on $i = i_0, \ldots, i_1$ and $n = 0, \ldots, N$. We get $A + B = 0$, with

$$A = \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} h_i(u_i^{n+1} - u_i^n)u_i^n,$$

and

$$B = \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)u_i^n.$$

We have

$$A = -\frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} h_i(u_i^{n+1} - u_i^n)^2 + \frac{1}{2} \sum_{i=i_0}^{i_1} h_i[(u_i^{N+1})^2 - (u_i^0)^2].$$

Using the scheme (21.1), we get

$$A = -\frac{1}{2} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} \frac{k^2}{h_i}\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 + \frac{1}{2} \sum_{i=i_0}^{i_1} h_i[(u_i^{N+1})^2 - (u_i^0)^2],$$

and therefore, using the CFL condition (21.13),

$$A \geq -\frac{1}{2M}(1-\xi) \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 - \frac{1}{2} \sum_{i=i_0}^{i_1} h_i(u_i^0)^2. \tag{21.15}$$

We now study the term B.

Denoting by Φ the function $\Phi(a) = \int_{U_m}^{a} sf'(s)ds$, for all $a \in \mathbb{R}$, an integration by parts yields, for all $(a, b) \in \mathbb{R}^2$,

$$\Phi(b) - \Phi(a) = b(f(b) - f(a)) - \int_a^b (f(s) - f(a))dx.$$

Using the technical lemma 18.5 page 110 which states $\int_a^b (f(s) - f(a))dx \geq \frac{1}{2M}(f(b) - f(a))^2$, we obtain

$$b(f(b) - f(a)) \geq \frac{1}{2M}(f(b) - f(a))^2 + \Phi(b) - \Phi(a).$$

The above inequality with $a = u_{i-1}^n$ and $b = u_i^n$ yields

$$B \geq \frac{1}{2M} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 + \sum_{n=0}^{N} k[\Phi(u_{i_1}^n) - \Phi(u_{i_0-1}^n)].$$

Thanks to the L^∞ estimate of Lemma 20.1 page 128, there exists $C_1 > 0$, only depending on u_0 and f such that

$$B \geq \frac{1}{2M} \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 - TC_1.$$

Therefore, since $A + B = 0$ and $\sum_{i=i_0}^{i_1} h_i \leq 4R$, the following inequality holds:

$$0 \geq \xi \sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 - 4RM \max\{-U_m, U_M\}^2 - 2MTC_1,$$

which gives the existence of $C_2 \in \mathbb{R}_+^\star$, only depending on R, T, u_0, f and ξ such that

$$\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k\Big(f(u_i^n) - f(u_{i-1}^n)\Big)^2 \leq C_2.$$

The Cauchy-Schwarz inequality yields

$$\Big[\sum_{i=i_0}^{i_1} \sum_{n=0}^{N} k|f(u_i^n) - f(u_{i-1}^n)|\Big]^2 \leq C_2 2T \frac{4R}{\alpha h},$$

which concludes the proof of the lemma. ∎

We can now state the convergence theorem.

Theorem 21.1 (Convergence in the nonlinear case) *Assume Assumption 21.1 and f nondecreasing. Let $\xi \in (0,1)$ and $\alpha > 0$ be given. Let M be the Lipschitz constant of f in $[U_m, U_M]$. For an admissible mesh $\mathcal{T}$ in the sense of Definition 20.1 page 128 and for a time step $k \in \mathbb{R}_+^\star$ satisfying the CFL condition (21.13) (taking $k \leq (1-\xi)\alpha h/M$ is a sufficient condition, note that ξ and α do not depend of $\mathcal{T}$), let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (21.1)-(21.4) with $p = 1$ and $g(a,b) = f(a)$.*
Then the function $u_{\mathcal{T},k}$ converges to the unique entropy weak solution u of (19.1) page 122 in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ as $\mathrm{size}(\mathcal{T})$ tends to 0.

Proof

Let Y be the set of approximate solutions, that is the set of $u_{\mathcal{T},k}$, defined by (21.1)-(21.4) with $p = 1$ and $g(a,b) = f(a)$, for all $(\mathcal{T}, k)$ where $\mathcal{T}$ is an admissible mesh in the sense of Definition 20.1 page 128 and $k \in \mathbb{R}_+^\star$ satisfies the CFL condition (21.13). Thanks to Lemma 21.1, the set Y is bounded in $L^\infty(\mathbb{R} \times \mathbb{R}_+)$.

The proof of Theorem 21.1 is performed in three steps. In the first step, a compactness result is given for Y, only using the boundeness of Y in $L^\infty(\mathbb{R} \times \mathbb{R}_+)$. In the second step, it is proved that the eventual limit (in a convenient sense) of a sequence of approximate solutions is a solution (in a convenient sense) of problem (19.1). In the third step a uniqueness result yields the conclusion. For steps 1 and 3, we refer to chapter 6 for a complete proof.

Step 1 (compactness result)
Let us first use a compactness result in $L^\infty(\mathbb{R} \times \mathbb{R}_+)$ which is stated in Proposition 32.1 page 201. Since Y is bounded in $L^\infty(\mathbb{R} \times \mathbb{R}_+)$, for any sequence $(u_m)_{m\in\mathbb{N}}$ of Y there exists a subsequence, still denoted by $(u_m)_{m\in\mathbb{N}}$, and there exists $\mu \in L^\infty(\mathbb{R} \times \mathbb{R}_+ \times (0,1))$ such that $(u_m)_{m\in\mathbb{N}}$ converges to μ in the "nonlinear weak-$\star$ sense", that is

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} \theta(u_m(x,t))\varphi(x,t)dtdx \to \int_{\mathbb{R}} \int_{\mathbb{R}_+} \int_0^1 \theta(\mu(x,t,\alpha))\varphi(x,t)d\alpha dtdx, \text{ as } m \to \infty,$$

for all $\varphi \in L^1(\mathbb{R} \times \mathbb{R}_+)$ and all $\theta \in C(\mathbb{R}, \mathbb{R})$. In other words, for any $\theta \in C(\mathbb{R}, \mathbb{R})$,

$$\theta(u_m) \to \mu_\theta \text{ in } L^\infty(\mathbb{R} \times \mathbb{R}_+) \text{ for the weak-}\star \text{ topology as } m \to \infty, \tag{21.16}$$

where μ_θ is defined by

$$\mu_\theta(x,t) = \int_0^1 \theta(\mu(x,t,\alpha))d\alpha, \text{ for a.e. } (x,t) \in \mathbb{R} \times \mathbb{R}_+.$$

Step 2 (passage to the limit)
Let $(u_m)_{m\in\mathbb{N}}$ be a sequence of Y. Assume that $(u_m)_{m\in\mathbb{N}}$ converges to μ in the nonlinear weak-$\star$ sense and that $u_m = u_{\mathcal{T}_m,k_m}$ (for all $m \in \mathbb{N}$) with size$(\mathcal{T}_m) \to 0$ as $m \to \infty$ (note that $k_m \to 0$ as $m \to \infty$, thanks to (21.13)).
Let us prove that μ is a "solution" to problem (19.1) in the following sense (we shall say that μ is "an entropy process solution" to problem (19.1)):

$$\left\{\begin{array}{l} \mu \in L^\infty(\mathbb{R} \times \mathbb{R}_+ \times (0,1)), \\ \displaystyle\int_{\mathbb{R}} \int_{\mathbb{R}_+} \int_0^1 \Big(|\mu(x,t,\alpha) - \kappa| \varphi_t(x,t) + (f(\mu(x,t,\alpha) \top \kappa) - f(\mu(x,t,\alpha) \bot \kappa)) \varphi_x(x,t) \Big) d\alpha dt dx \\ \qquad\qquad + \displaystyle\int_{\mathbb{R}} |u_0(x) - \kappa| \varphi(x,0) dx \geq 0, \, \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+), \forall \kappa \in \mathbb{R}. \end{array}\right. \tag{21.17}$$

Let $\kappa \in \mathbb{R}$. Setting

$$v(x,t) = \int_0^1 |\mu(x,t,\alpha) - \kappa| d\alpha, \text{ for a.e. } (x,t) \in \mathbb{R} \times \mathbb{R}_+$$

and

$$w(x,t) = \int_0^1 \Big(f(\mu(x,t,\alpha) \top \kappa) - f(\mu(x,t,\alpha) \bot \kappa) \Big) d\alpha, \text{ for a.e. } (x,t) \in \mathbb{R} \times \mathbb{R}_+,$$

the inequality in (21.17) reads

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}} \int_{\mathbb{R}_+} [v(x,t)\varphi_t(x,t) + w(x,t)\varphi_x(x,t)] dt dx + \int_{\mathbb{R}} |u_0(x) - \kappa| \varphi(x,0) dx \geq 0, \\ \qquad\qquad \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+). \end{array} \tag{21.18}$$

Let us prove that (21.18) holds; for $m \in \mathbb{N}$ we shall denote by $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$. We use the result of Lemma 21.2, which reads in the present particular case $f' \geq 0$,

$$h_i \frac{v_i^{n+1} - v_i^n}{k} + w_i^n - w_{i-1}^n \leq 0, \, \forall i \in \mathbb{Z}, \, \forall n \in \mathbb{N},$$

where $v_i^n = |u_i^n - \kappa|$ and $w_i^n = f(u_i^n \top \kappa) - f(u_i^n \bot \kappa) = |f(u_i^n) - f(\kappa)|$.
The functions $v_{\mathcal{T}_m,k_m}$ and $w_{\mathcal{T}_m,k_m}$ are defined in the same way as the function $u_{\mathcal{T}_m,k_m}$, i. e. with constant values v_i^n and w_i^n in each control volume K_i during each time step $(nk,(n+1)k)$. Choosing θ equal to the continuous functions $|\cdot - \kappa|$ and $|f(\cdot) - f(\kappa)|$ in (21.16) yields that the sequences $(v_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$ and $(w_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$ converge to v and w in $L^\infty(\mathbb{R} \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology.
Applying the method which was used in the proof of Proposition 20.2 page 132, taking v_i^l instead of u_i^l in the definition of A_m (for $l = n$ and $n+1$) and w_j^n instead of u_j^n in the definition of B_m (for $j = i$ and $i-1$), we conclude that (21.18) holds.
Indeed, a weak BV inequality holds on the values w_i^n (that is (20.9) page 129 holds with w_j^n instead of u_j^n for $j = i$ and $i-1$), thanks to Lemma 21.3 page 137 and the relation

$$\big| |f(u_i^n) - \kappa| - |f(u_{i-1}^n) - \kappa| \big| \leq |f(u_i^n) - f(u_{i-1}^n)|, \, \forall i \in \mathbb{Z}, \, \forall n \in \mathbb{N}.$$

(Note that here, as in the linear case, the useful consequence of the weak BV inequality, is (20.11) page 130 with w_j^n instead of u_j^n for $j = i$ and $i-1$.)
This concludes Step 2.

Step 3 (uniqueness result for (21.17) and conclusion)
Theorem 29.1 page 183 states that there exists at most one solution to (21.17) and that there exists $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ such that μ solution to (21.17) implies $\mu(x,t,\alpha) = u(x,t)$ for a.e. $(x,t,\alpha) \in \mathbb{R}\times\mathbb{R}_+\times(0,1)$. Then, u is necessarily the entropy weak solution to (19.1).
Furthermore, if $(u_m)_{m\in\mathbb{N}}$ converges to u in the nonlinear weak-$\star$ sense, an easy argument shows that $(u_m)_{m\in\mathbb{N}}$ converges to u in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ (and even in $L^p_{loc}(\mathbb{R}\times\mathbb{R}_+)$ for all $1 \le p < \infty$), see Remark (32.2) page 203.
Then, the conclusion of Theorem 21.1 follows easily from Step 2 and Step 1 by way of contradiction (in order to prove the convergence of a sequence $u_{\mathcal{T}_m,k_m} \subset Y$ to u, if size$(\mathcal{T}_m) \to 0$ as $m \to \infty$, without any extraction of a "subsequence"). ∎

Remark 21.4 In Theorem 21.1, we only consider the case $f' \ge 0$ and the so called "upstream scheme". It is quite easy to generalize the result for any $f \in C^1(\mathbb{R},\mathbb{R})$ and any monotone flux scheme (see the following chapter). It is also possible to consider other schemes (for instance, some 5-points schemes, as in Section 22). For a given scheme, the proof of convergence of the approximate solution towards the entropy weak solution contains 2 steps:

1. prove an L^∞ estimate on the approximate solutions, which allows to use the compactness result of Step 1 of the proof of Theorem 21.1,
2. prove a "weak BV" estimate and some "discrete entropy inequality" in order to have the following property:

   If $(u_m)_{m\in\mathbb{N}}$ is a sequence of approximate solutions which converges in the nonlinear weak-$\star$ sense, then

$$\lim_{m\to\mathbb{N}} \int_{\mathbb{R}}\int_{\mathbb{R}_+} \Big(|u_m(x,t)-\kappa|\varphi_t(x,t) + (f(u_m(x,t)\top\kappa) - f(u_m(x,t)\bot\kappa))\varphi_x(x,t)\Big)dtdx$$
$$+ \int_{\mathbb{R}} |u_0(x)-\kappa|\varphi(x,0)dx \ge 0,\ \forall\varphi \in C^1_c(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+),\ \forall\kappa\in\mathbb{R}.$$

21.5 Convergence proof using BV

We now give the details of the classical proof of convergence (considering only 3 points schemes), which requires regularizations of u_0 in $BV(\mathbb{R})$. It consists in using Helly's compactness theorem (which may also be used in the linear case to obtain a strong convergence of $u_{\mathcal{T},k}$ to u in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$). This theorem is a direct consequence of Kolmogorov's theorem (theorem 14.1 page 94). We give below the definition of $BV(\Omega)$ where Ω is an open subset of $\mathbb{R}^p(\Omega)$, $p \ge 1$ (already given in Definition 19.5 page 124 for $\Omega = \mathbb{R}$) and we give a straightforward consequence of Helly's theorem for the case of interest here.

Definition 21.2 ($BV(\Omega)$) Let $p \in \mathbb{N}^\star$ and let Ω be an open subset of $\mathbb{R}^p$. A function $v \in L^1_{loc}(\Omega)$ has a bounded variation, that is $v \in BV(\Omega)$, if $|v|_{BV(\Omega)} < \infty$ where

$$|v|_{BV(\Omega)} = \sup\{\int_\Omega v(x)\mathrm{div}\varphi(x)dx, \varphi \in C^1_c(\Omega,\mathbb{R}^p), |\varphi(x)| \le 1,\ \forall x\in\Omega\}. \tag{21.19}$$

Lemma 21.4 (Consequence of Helly's theorem) *Let $\mathcal{A} \subset L^\infty(\mathbb{R}^2)$. Assume that there exists $C \in \mathbb{R}_+$ and, for all $T > 0$, there exists $C_T \in \mathbb{R}_+$ such that*

$$\|v\|_{L^\infty(\mathbb{R}^2)} \le C,\ \forall v \in \mathcal{A},$$

and

$$|v|_{BV(\mathbb{R}\times(-T,T))} \leq C_T,\ \forall v \in \mathcal{A},\ \forall T > 0.$$

Then for any sequence $(v_n)_{n\in\mathbb{N}}$ of elements of $\mathcal{A}$, there exists a subsequence, still denoted by $(v_n)_{n\in\mathbb{N}}$, and there exists $v \in L^\infty(\mathbb{R}^2)$, with $\|v\|_{L^\infty(\mathbb{R}^2)} \leq C$ and $|v|_{BV(\mathbb{R}\times(-T,T))} \leq C_T$ for all $T > 0$, such that $v_n \to v$ in $L^1_{loc}(\mathbb{R}^2)$ as $n \to \infty$, that is $\int_{\bar\omega} |v_n(x) - v(x)|dx \to 0$, as $n \to \infty$ for any compact set $\bar\omega$ of $\mathbb{R}^2$.

In order to use Lemma 21.4, one first shows the following BV stability estimate for the approximate solution:

Lemma 21.5 (Discrete space BV estimate) *Under Assumption 21.1, assume that $u_0 \in BV(\mathbb{R})$; let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ be the time step. Let $\{u_i^n,\ i \in \mathbb{Z},\ n \in \mathbb{N}\}$ be given by (21.1), (21.2) and assume that the scheme is a monotone flux scheme in the sense of Definition 21.1 page 134. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ with respect to its two arguments. Then, under the CFL condition (21.9), the following inequality holds:*

$$\sum_{i\in\mathbb{Z}} |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i\in\mathbb{Z}} |u_{i+1}^n - u_i^n|,\ \forall\ n \in \mathbb{N}. \tag{21.20}$$

Proof of Lemma 21.5

First remark that, for $n = 0$, $\sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0| \leq |u_0|_{BV(\mathbb{R})}$ (see Remark 19.4 page 124).
For all $i \in \mathbb{Z}$, the scheme (21.1), (21.2) (with $p = 1$) leads to

$$u_i^{n+1} = u_i^n + b_{i+\frac12}^n(u_{i+1}^n - u_i^n) + a_{i-\frac12}^n(u_{i-1}^n - u_i^n),$$

and

$$u_{i+1}^{n+1} = u_{i+1}^n + b_{i+\frac32}^n(u_{i+2}^n - u_{i+1}^n) + a_{i+\frac12}^n(u_i^n - u_{i+1}^n),$$

where $a_{i+1/2}$ and $b_{i+1/2}$ are defined (for all $i \in \mathbb{Z}$) in Lemma 21.1 page 136. Substracting one equation to the other leads to

$$u_{i+1}^{n+1} - u_i^{n+1} = (u_{i+1}^n - u_i^n)(1 - b_{i+\frac12}^n - a_{i+\frac12}^n) + b_{i+\frac32}^n(u_{i+2}^n - u_{i+1}^n) + a_{i-\frac12}^n(u_i^n - u_{i-1}^n).$$

Under the condition (21.9), we get

$$|u_{i+1}^{n+1} - u_i^{n+1}| \leq |u_{i+1}^n - u_i^n|(1 - b_{i+\frac12}^n - a_{i+\frac12}^n) + b_{i+\frac32}^n|u_{i+2}^n - u_{i+1}^n| + a_{i-\frac12}^n|u_i^n - u_{i-1}^n|.$$

Summing the previous equation over $i \in \mathbb{Z}$ gives (21.20). ∎

Corollary 21.1 (Discrete BV estimate) *Under assumption 21.1, let $u_0 \in BV(\mathbb{R})$; let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128 and let $k \in \mathbb{R}_+^\star$ be the time step. Let $u_{\mathcal{T},k}$ be the finite volume approximate solution defined by (21.1)-(21.4) and assume that the scheme is a monotone flux scheme in the sense of Definition 21.1 page 134. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ with respect to its two arguments and assume that k satisfies the CFL condition (21.9). Let $u_{\mathcal{T},k}(x,t) = u_i^0$ for a.e. $(x,t) \in K_i \times \mathbb{R}_-$, for all $i \in \mathbb{Z}$ (hence $u_{\mathcal{T},k}$ is defined a.e. on $\mathbb{R}^2$). Then, for any $T > 0$, there exists $C \in \mathbb{R}_+^\star$, only depending on u_0, g and T such that:*

$$|u_{\mathcal{T},k}|_{BV(\mathbb{R}\times(-T,T))} \leq C. \tag{21.21}$$

Proof of Corollary 21.1

As in Lemma 21.5, remark that $\sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0| \leq |u_0|_{BV(\mathbb{R})}$.
Let us first assume that $T \leq k$. Then, the BV semi-norm of $u_{\mathcal{T},k}$ satisfies

$$|u_{\mathcal{T},k}|_{BV(\mathbb{R}\times(-T,T))} \leq 2T \sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0|.$$

Hence the estimate (21.21) is true for $C = 2T|u_0|_{BV(\mathbb{R})}$.

Let us now assume that $k < T$. Let $N \in \mathbb{N}^\star$ such that $Nk < T \leq (N+1)k$. The definition of $|\cdot|_{BV(\mathbb{R}\times(-T,T))}$ yields

$$\begin{aligned} &|u_{\mathcal{T},k}|_{BV(\mathbb{R}\times(-T,T))} \leq T\textstyle\sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0| + \\ &\sum_{n=0}^{N-1} \sum_{i\in\mathbb{Z}} k|u_{i+1}^n - u_i^n| + (T-Nk) \sum_{i\in\mathbb{Z}} |u_{i+1}^N - u_i^N| + \sum_{n=0}^{N-1} \sum_{i\in\mathbb{Z}} h_i |u_i^{n+1} - u_i^n|. \end{aligned} \tag{21.22}$$

Lemma 21.5 gives $\sum_{i\in\mathbb{Z}} |u_{i+1}^n - u_i^n| \leq |u_0|_{BV(\mathbb{R})}$ for all $n \in \mathbb{N}$, and therefore,

$$\sum_{n=0}^{N-1} \sum_{i\in\mathbb{Z}} k|u_{i+1}^n - u_i^n| + (T-Nk) \sum_{i\in\mathbb{Z}} |u_{i+1}^N - u_i^N| \leq T|u_0|_{BV(\mathbb{R})}. \tag{21.23}$$

In order to bound the last term of (21.22), using the scheme (21.1) yields, for all $i \in \mathbb{Z}$ and all $n \in \mathbb{N}$,

$$|u_i^{n+1} - u_i^n| \leq \frac{k}{h_i} g_1 |u_i^n - u_{i-1}^n| + \frac{k}{h_i} g_2 |u_i^n - u_{i+1}^n|.$$

Therefore,

$$\sum_{i\in\mathbb{Z}} h_i |u_i^{n+1} - u_i^n| \leq k(g_1+g_2) \sum_{i\in\mathbb{Z}} |u_i^n - u_{i+1}^n|, \text{ for all } n \in \mathbb{N},$$

which yields, since $Nk < T$,

$$\sum_{n=0}^{N-1} \sum_{i\in\mathbb{Z}} h_i |u_i^{n+1} - u_i^n| \leq T(g_1+g_2)|u_0|_{BV(\mathbb{R})}. \tag{21.24}$$

Therefore Inequality (21.21) follows from (21.22), (21.23) and (21.24) with $C = T(2+g_1+g_2)|u_0|_{BV(\mathbb{R})}$. ∎

Consider a sequence of admissible meshes and time steps verifying the CFL condition, and the associated sequence of approximate solutions (prolonged on $\mathbb{R}\times\mathbb{R}_-$ as in Corollary 21.1). By Lemma 21.1 page 136 and Corollary 21.1, the sequence of approximate solutions satisfies the hypotheses of Lemma 21.4 page 141. It is therefore possible to extract a subsequence which converges in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ to a function $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+^\star)$. It must still be shown that the function u is the unique weak entropy solution of Problem (19.1). This may be proven by using the discrete entropy inequalities (21.12) and the strong BV estimate (21.20) or the classical Lax-Wendroff theorem recalled below.

Theorem 21.2 (Lax-Wendroff) *Under Assumption 21.1, let $\alpha > 0$ be given and let $(\mathcal{T}_m)_{m\in\mathbb{N}}$ be a sequence of admissible meshes in the sense of Definition 20.1 page 128 (note that, for all $m \in \mathbb{N}$, the mesh $\mathcal{T}_m$ satisfies the hypotheses of Definition 20.1 where $\mathcal{T} = \mathcal{T}_m$ and α is independent of m). Let $(k_m)_{m\in\mathbb{N}}$ be a sequence of (positive) time steps. Assume that* $\mathrm{size}(\mathcal{T}_m) \to 0$ *and $k_m \to 0$ as $m \to \infty$.*

For $m \in \mathbb{N}$, setting $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, let $u_m = u_{\mathcal{T},k}$ be the solution of (21.1)-(21.4) with $p = 1$ and some g from $\mathbb{R}^2$ to $\mathbb{R}$, only depending on f and u_0, locally Lipschitz continuous and such that $g(s,s) = f(s)$ for all $s \in \mathbb{R}$.

Assume that $(u_m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ and that $u_m \to u$ a.e. on $\mathbb{R}\times\mathbb{R}_+$. Then, u is a weak solution to problem (19.1) (that is u satisfies (19.3)).

Furthermore, assume that for any $\kappa \in \mathbb{R}$ there exists some locally Lipschitz continuous function G_κ from $\mathbb{R}^2$ to $\mathbb{R}$, only depending on f, u_0 and κ, such that $G_\kappa(s,s) = f(s\top\kappa) - f(s\bot\kappa)$ for all $s \in \mathbb{R}$ and such that for all $m \in \mathbb{N}$

$$\frac{1}{k}(|u_i^{n+1}-\kappa|-|u_i^n-\kappa|)+\frac{1}{h_i}(G_\kappa(u_i^n,u_{i+1}^n)-G_\kappa(u_{i-1}^n,u_i^n))\le 0,\ \forall i\in\mathbb{Z},\ \forall n\in\mathbb{N}, \tag{21.25}$$

where $\{u_i^n,\ i\in\mathbb{Z},\ n\in\mathbb{N}\}$ *is the solution to (21.1)-(21.2) for* $\mathcal{T}=\mathcal{T}_m$ *and* $k=k_m$*. Then,* u *is the entropy weak solution to Problem (19.1) (that is* u *is the unique solution of (19.4)).*

Proof of Theorem 21.2

Since $(u_m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ and $u_m\to u$ a.e. on $\mathbb{R}\times\mathbb{R}_+$, the sequence $(u_m)_{m\in\mathbb{N}}$ converges to u in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$. This implies in particular (from Kolmogorov's theorem, see Theorem 14.1) that, for all $R>0$ and all $T>0$,

$$\sup_{m\in\mathbb{N}}\int_0^{2T}\int_{-2R}^{2R}|u_m(x,t)-u_m(x-\eta,t)|dxdt\to 0 \text{ as } \eta\to 0.$$

Then, taking $\eta=\alpha\,\mathrm{size}(\mathcal{T}_m)$ (for $m\in\mathbb{N}$) and letting $m\to\infty$ yields, in particular,

$$\int_0^{2T}\int_{-2R}^{2R}|u_m(x,t)-u_m(x-\alpha\,\mathrm{size}(\mathcal{T}_m),t)|dxdt\to 0 \text{ as } m\to\infty. \tag{21.26}$$

For $m\in\mathbb{N}$, let $\{u_i^n,\ i\in\mathbb{Z},\ n\in\mathbb{N}\}$ be the solution to (21.1)-(21.2) for $\mathcal{T}=\mathcal{T}_m$ and $k=k_m$ (note that u_i^n depends on m, even though this dependency is not so clear in the notation). We also set $k_m=k$ and $\mathrm{size}(\mathcal{T}_m)=h$, so that k and h depend on m (but recall that α does not depend on m).

Let $R>0$ and $T>0$. Let $i_0\in\mathbb{Z}$, $i_1\in\mathbb{Z}$ and $N\in\mathbb{N}$ be such that $-R\in\overline{K}_{i_0}$, $R\in\overline{K}_{i_1}$ and $T\in(Nk,(N+1)k]$. Then, for $h<R$ and $k<T$ (which is true for m large enough),

$$\alpha h\sum_{i=i_0}^{i_1}\sum_{n=0}^{N}k|u_i^n-u_{i-1}^n|\le\int_0^{2T}\int_{-2R}^{2R}|u_m(x,t)-u_m(x-\alpha h,t)|dxdt.$$

Therefore, Inequality (21.26) leads to (20.11), that is

$$h\sum_{i=i_0}^{i_1}\sum_{n=0}^{N}k|u_i^n-u_{i-1}^n|\to 0 \text{ as } m\to\infty. \tag{21.27}$$

Using (21.27), the remainder of the proof of Theorem 21.2 is very similar to the proof of Proposition 20.2 page 132 and to Step 2 in the proof of Theorem 21.1 page 139 (Inequality (21.27) replaces the weak BV inequality).

In order to prove that u is solution to (19.3), let us multiply the first equation of (21.1) by $(k/h_i)\varphi(x,nk)$, integrate over $x\in K_i$ and sum for all $i\in\mathbb{Z}$ and $n\in\mathbb{N}$. This yields

$$A_m+B_m=0$$

with

$$A_m=\sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}}(u_i^{n+1}-u_i^n)\int_{K_i}\varphi(x,nk)dx$$

and

$$B_m=\sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}}k(g(u_i^n,u_{i+1}^n)-g(u_{i-1}^n,u_i^n))\frac{1}{h_i}\int_{K_i}\varphi(x,nk)dx.$$

As in the proof of Proposition 20.2, one has

$$\lim_{m\to+\infty}A_m=-\int_{\mathbb{R}_+}\int_{\mathbb{R}}u(x,t)\varphi_t(x,t)dxdt-\int_{\mathbb{R}}u_0(x)\varphi(x,0)dx.$$

Let us now turn to the study of B_m. We compare B_m with

$$B_{1,m} = -\sum_{n\in\mathbb{N}} \int_{nk}^{(n+1)k} \int_{\mathbb{R}} f(u_{\mathcal{T},k}(x,t))\varphi_x(x,nk)dxdt,$$

which tends to $-\int_{\mathbb{R}_+}\int_{\mathbb{R}} f(u(x,t))\varphi_x(x,t)dxdt$ as $m\to\infty$ since $f(u_{\mathcal{T},k})\to f(u)$ in $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ as $m\to\infty$.
The term $B_{1,m}$ can be rewritten as

$$B_{1,m} = \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(f(u_i^n) - f(u_{i-1}^n))\varphi(x_{i-1/2},nk),$$

which yields, introducing $g(u_{i-1}^n, u_i^n)$,

$$\begin{aligned} B_{1,m} = &\sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(f(u_i^n) - g(u_{i-1}^n,u_i^n))\varphi(x_{i-1/2},nk) \\ &+ \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(g(u_{i-1}^n,u_i^n) - f(u_{i-1}^n))\varphi(x_{i-1/2},nk). \end{aligned}$$

Similarly, introducing $f(u_i^n)$ in B_m,

$$\begin{aligned} B_m = &\sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(f(u_i^n) - g(u_{i-1}^n,u_i^n))\frac{1}{h_i}\int_{K_i}\varphi(x,nk)dx \\ &+ \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(g(u_i^n,u_{i+1}^n) - f(u_i^n))\frac{1}{h_i}\int_{K_i}\varphi(x,nk)dx. \end{aligned}$$

In order to compare B_m and $B_{1,m}$, let $R>0$ and $T>0$ be such that $\varphi(x,t)=0$ if $|x|\geq R$ or $t\geq T$. Let $A>0$ be such that $\|u_m\|_{L^\infty(\mathbb{R}\times\mathbb{R}_+)}\leq A$ for all $m\in\mathbb{N}$. Then there exists $C>0$, only depending on φ and the Lipschitz constants on g on $[-A,A]^2$, such that, if $h<R$ and $k<T$ (which is true for m large enough),

$$|B_m - B_{1,m}| \leq Ch\sum_{i=i_0}^{i_1}\sum_{n=0}^{N} k|u_i^n - u_{i-1}^n|, \tag{21.28}$$

where $i_0\in\mathbb{Z}$, $i_1\in\mathbb{Z}$ and $N\in\mathbb{N}$ are such that $-R\in\overline{K}_{i_0}$, $R\in\overline{K}_{i_1}$ and $T\in(Nk,(N+1)k]$.
Using (21.28) and (21.27), we get $|B_m - B_{1,m}|\to 0$ and then

$$B_m \to -\int_{\mathbb{R}}\int_{\mathbb{R}_+} f(u(x,t))\varphi_x(x,t)dtdx \text{ as } m\to\infty,$$

which completes the proof that u is a solution to problem (19.3).

Under the additional assumption that u_m satisfies (21.25), one proves that u satisfies (19.7) page 124 (and therefore that u satisfies (19.4)) and is the entropy weak solution to Problem (19.1) by a similar method.
Indeed, let $\kappa\in\mathbb{R}$. One replaces u_i^l by $|u_i^l-\kappa|$ in A_m (for $l=n$ and $n+1$) and one replaces g by G_κ in B_m. Then, passing to the limit in $A_m+B_m\leq 0$ (which is a consequence of the inequation (21.25)) leads the desired result.
This concludes the proof of Theorem 21.2 ∎

Remark 21.5 Theorem 21.2 still holds with $(2p+1)$-points schemes ($p>1$). The generalization of the first part of Theorem 21.2 (the proof that u is a solution to (19.3)) is quite easy. For the second part of Theorem 21.2 (entropy inequalities) the discrete entropy inequalities may be replaced by some weaker ones (in order to handle interesting schemes such as those which are described in the following section).

However, the use of Theorem 21.2 needs a compactness property of sequences of approximate solutions in the space $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$. Such a compactness property is generally achieved with a "strong BV estimate" (similar to (21.20)). Hence an extensive literature on "TVD schemes" (see HARTEN [80]), "ENO schemes"... (see GODLEWSKI and RAVIART [75], GODLEWSKI and RAVIART [76] and references therein). The generalization of this method in the multidimensional case (studied in the following chapter) does not seem so clear except in the case of Cartesian meshes.

22 Higher order schemes

Consider a monotone flux scheme in the sense of Definition 21.1 page 134. By definition, the considered scheme is a 3 points scheme; recall that the numerical flux function is denoted by g. The approximate solution obtained with this scheme converges to the entropy weak solution of Problem (19.1) page 122 as the mesh size tends to 0 and under a so called CFL condition (it is proved in Theorem 21.1 for a particular case and in the next chapter for the general case). However, 3-points schemes are known to be diffusive, so that the approximate solution is not very precise near the discontinuities. An idea to reduce the diffusion is to go to a 5-points scheme by introducing "slopes" on each discretization cell and limiting the slopes in order for the scheme to remain stable. A classical way to do this is the "MUSCL" (Monotonic Upwind Scheme for Conservation Laws, see VAN LEER [148]) technique .
Reconstructing a slope on each cell enables to compute interface values on each side of an interface $x_{i+\frac{1}{2}}$. These values are then used in the computation of the fluxes

We briefly describe, with the notations of Section 21.1, an example of such a scheme, see e.g. GODLEWSKI and RAVIART [75] and GODLEWSKI and RAVIART [76] for further details. Let $n \in \mathbb{N}$.

- Computation of the slopes

$$\tilde{p}_i^n = \frac{u_{i+1}^n - u_{i-1}^n}{h_i + \frac{h_{i-1}}{2} + \frac{h_{i+1}}{2}}, i \in \mathbb{Z}\,.$$

- Limitation of the slopes

  $p_i^n = \alpha_i^n \tilde{p}_i^n$, $i \in \mathbb{Z}$, where α_i^n is the largest number in $[0,1]$ such that

$$u_i^n + \frac{h_i}{2}\alpha_i^n \tilde{p}_i^n \in [u_i^n \perp u_{i+1}^n, u_i^n \top u_{i+1}^n] \text{ and } u_i^n - \frac{h_i}{2}\alpha_i^n \tilde{p}_i^n \in [u_i^n \perp u_{i-1}^n, u_i^n \top u_{i-1}^n].$$

  In practice, other formulas giving smaller values of α_i^n are sometimes needed for stability reasons.

- Computation of u_i^{n+1} for $i \in \mathbb{Z}$ One replaces $g(u_i^n, u_{i+1}^n)$ in (21.2) by :

$$\overline{g}(u_{i-1}^n, u_i^n, u_{i+1}^n, u_{i+2}^n) = g(u_i^n + \frac{h_i}{2}p_i^n, u_{i+1}^n - \frac{h_{i+1}}{2}p_{i+1}^n).$$

The scheme thus constructed is less diffusive than the original one and it remains stable thanks to the limitation of the slope. Indeed, if the limitation of the slopes is not active (that is $\alpha_i^n = 1$), the space diffusion term disappears from this new scheme, while the time "antidiffusion" term remains. Hence it seems appropriate to use a higher order scheme for the time discretization. This may be done by using, for instance, an RK2 (Runge Kutta order 2, or Heun) method for the discretization of the time derivative. The MUSCL scheme may be written as

$$\frac{U^{n+1} - U^n}{k} = \overline{H}(U^n) \text{ for } n \in \mathbb{N},$$

where $U^n = (u_i^n)_{i \in \mathbb{Z}}$; hence it may be seen as the explicit Euler discretization of

$$U_t = \overline{H}(U);$$

therefore, the RK2 time discretization yields to the following scheme:

$$\frac{U^{n+1} - U^n}{k} = \frac{1}{2}\overline{H}(U^n) + \frac{1}{2}\Big(\overline{H}(U^n + k\overline{H}(U^n))\Big) \text{ for } n \in \mathbb{N}.$$

Going to a second order discretization in time allows larger time steps, without loss of stability.

Results of convergence are possible with these new schemes (with eventually some adaptation of the slope limitations to obtain convenient discrete entropy inequalities, see Vila [154]. It is also possible to obtain error estimates in the spirit of those given in the following chapter, in the multidimensional case, see e.g. Chainais-Hillairet [22], Noëlle [117], Kröner, Noelle and Rokyta [93]. However these error estimates are somewhat unsatisfactory since they are of a similar order to that of the original 3-points scheme (although these schemes are numerically more precise that the original 3-points schemes).

The higher order schemes are nonlinear even if Problem (19.1) page 122 is linear, because of the limitation of the slopes.

Implicit versions of these higher order schemes are more or less straightforward. However, the numerical implementation of these implicit versions requires the solution of nonlinear systems. In many cases, the solutions to these nonlinear systems seem impossible to reach for large k; in fact, the existence of the solutions is not so clear, see Pfertzel [125]. Since the advantage of implicit schemes is essentially the possibility to use large values of k, the above flaw considerably reduces the opportunity of their use. Therefore, although implicit 3-points schemes are very diffusive, they remain the basic schemes in several industrial environments. See also Section 35.3 page 213 for some clues on implicit schemes applied to complex industrial applications.

23 Boundary conditions

A general convergence result is presented here in the case of a scalar equation. Then, this result will be applied to understand the sense of the boundary condition, described at $x = 1$ in the previous section, in a simplified scalar case.

23.1 A general convergence result

The unknown is now a function $u : (0,1) \times \mathbf{R}+ \to \mathbf{R}$. The flux is a function $f \in C^1(\mathbf{R}, \mathbf{R})$ (or $f : \mathbf{R} \to \mathbf{R}$ Lipschitz continuous) and the initial datum is $u_0 \in L^\infty((0,1))$. Let $A, B \in \mathbf{R}$ be such that $A \leq u_0 \leq B$ a.e.. The problem to solve is:

$$u_t + (f(u))_x = 0,\ x \in (0,1),\ t \in \mathbf{R}_+, \tag{23.1}$$

with the initial condition :

$$u(x,0) = u_0(x),\ x \in (0,1), \tag{23.2}$$

and some boundary conditions which will be prescribed later.

As in the previous section, let $h = \frac{1}{N}$ (with $N \in \mathbf{N}^\star$) be the mesh size and $k > 0$ be the time step (assumed to be constant, for the sake of simplicity). The discrete unknowns are now the values $u_i^n \in \mathbf{R}$ for $i \in \{1, \ldots, N\}$ and $n \in \mathbf{N}$. In order to define the approximate solution a.e. in $(0,1) \times \mathbf{R}$, one sets $u_{h,k}(x,t) = u_i^n$ for $x \in ((i-1)h, ih)$, $t \in (nk, (n+1)k)$, $i \in \{1, \ldots, N\}$, $n \in \mathbf{N}$.

The discretization of the initial condition leads to

$$u_i^0 = \frac{1}{h}\int_{(i-1)h}^{ih} u_0(x)dx,\ i \in \{1,\dots,N\}. \tag{23.3}$$

For the computation of u_i^n for $n > 0$, one uses, as before, an explicit, 3-points scheme:

$$\frac{h}{k}(u_i^{n+1} - u_i^n) + f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n = 0,\ i \in \{1,\dots,N\},\ n \in \mathbf{N}. \tag{23.4}$$

For $i \in 1,\dots,N-1$, one takes

$$f_{i+\frac{1}{2}}^n = g(u_i^n, u_{i+1}^n), \tag{23.5}$$

where g is the numerical flux. Sufficient conditions on $g : [A,B]^2 \to \mathbf{R}$, in order to have a convergent scheme if $x \in \mathbf{R}$ instead of $(0,1)$, are:

C1: g is non decreasing with respect to its first argument and nonincreasing with respect to its second argument,

C2: $g(s,s) = f(s)$, for all $s \in [A,B]$,

C3: g is Lipschitz continuous.

Let L be a Lipschitz constant for g (on $[A,B]^2$) and $\zeta > 0$. If $(0,1)$ is replaced by $\mathbf{R}$, It is well known (see e.g. [53]) that, if $k \le (1-\zeta)\frac{h}{L}$, the approximate solution $u_{h,k}$, that is the solution defined by (23.3)-(23.5) (with $i \in \mathbf{Z}$), takes its values in $[A,B]$ and converges towards the unique entropy weak solution of (23.1)-(23.2) in $L^p_{loc}(\mathbf{R}\times\mathbf{R}_+)$ as $h \to 0$.

In the case $x \in (0,1)$ instead of $x \in \mathbf{R}$, one assumes the same conditions on g, namely (C1)-(C3). In order to complete the scheme, one has to define $f_{\frac{1}{2}}^n$ and $f_{N+\frac{1}{2}}^n$.

Let $\overline{u}, \overline{\overline{u}} \in L^\infty(\mathbf{R}_+)$ be such that $A \le \overline{u}, \overline{\overline{u}} \le B$, a.e. on $\mathbf{R}_+$, let $g_0, g_1 : [A,B]^2 \to \mathbf{R}$, satisfying (C1)-(C3), and define:

$$\begin{array}{l} f_{\frac{1}{2}}^n = g_0(\overline{u}^n, u_1^n);\ \ \overline{u}^n = \frac{1}{k}\int_{nk}^{(n+1)k} \overline{u}(t)dt \\ f_{N+\frac{1}{2}}^n = g_1(u_N^n, \overline{\overline{u}}^n,);\ \ \overline{\overline{u}} = \frac{1}{k}\int_{nk}^{(n+1)k} \overline{\overline{u}}(t)dt, \end{array} \tag{23.6}$$

Then, a convergence theorem can be proven as in the case $x \in \mathbf{R}$, see [157]:

Theorem 23.1 *Let $f \in C^1(\mathbf{R},\mathbf{R})$ (or $f : \mathbf{R} \to \mathbf{R}$ Lipschitz continuous). Let $u_0 \in L^\infty((0,1))$, $\overline{u}, \overline{\overline{u}} \in L^\infty(\mathbf{R}_+)$ and $A, B \in \mathbf{R}$ be such that $A \le u_0 \le B$ a.e. on $(0,1)$, $A \le \overline{u}, \overline{\overline{u}} \le B$ a.e. on $\mathbf{R}_+$. Let $g_0, g_1 : [A,B]^2 \to \mathbf{R}$, satisfying (C1)-(C3). Let L be a common Lipschitz constant for g, g_0 and g_1 (on $[A,B]^2$) and let $\zeta > 0$. Then, if $k \le (1-\zeta)\frac{h}{L}$, the equations (23.3)-(23.6) define an approximate solution $u_{h,k}$ which takes its values in $[A,B]$ and converges towards the unique solution of (23.7) in $L^p_{loc}([0,1]\times\mathbf{R}_+)$ for any $1 \le p < \infty$, as $h \to 0$:*

$$\begin{array}{l} u \in L^\infty((0,1)\times(0,\infty)), \\ \displaystyle\int_0^\infty \int_0^1 [(u-\kappa)^\pm \varphi_t + sign_\pm(u-\kappa)(f(u)-f(\kappa))\varphi_x]dxdt \\ \displaystyle +M\int_0^\infty (\overline{u}(t)-\kappa)^\pm \varphi(0,t)dt + M\int_0^\infty (\overline{\overline{u}}(t)-\kappa)^\pm \varphi(1,t)dt \\ \displaystyle + \int_0^1 (u_0-\kappa)^\pm \varphi(x,0)dx \ge 0, \\ \forall \kappa \in [A,B],\ \forall \varphi \in C_c^1([0,1]\times[0,\infty),\mathbf{R}_+). \end{array} \tag{23.7}$$

In (23.7), M is any bound for $|f'|$ on $[A,B]$ (and the solution of (23.7) does not depends on the choice of M). The definition of $sign_\pm$ is: $sign_+(s) = 1$ if $s > 0$, $sign_+(s) = 0$ if $s < 0$, $sign_-(s) = 0$ if $s > 0$, $sign_-(s) = -1$ if $s < 0$.

Remark 23.1

1. *It is interesting to remark that this convergence result is also true if the function g depends on i and n, provided that L is a common Lipschitz constant for all these functions.*

2. *The definition (23.7) of solution of (23.1)-(23.2) with the "weak" boundary conditions $\overline{u}$ and $\overline{\overline{u}}$ at $x = 0$ and $x = 1$ is essentially due to F. Otto, see [122].*

3. *It is interesting also to remark that if one replaces, in (23.7), the two entropies $(u-\kappa)^{\pm}$ by the sole entropy $|u-\kappa|$, one has an existence result (since $|u-\kappa| = (u-\kappa)^+ + (u-\kappa)^-$) but no uniqueness result, see [157] for a counter-example to uniqueness.*

4. *This convergence result can be generalized to the multidimensional case, see Sect. 31 and [157].*

If u, solution of (23.7), is regular enough (say $u \in C^1([0,1] \times \mathbf{R}_+)$, for instance), u satisfies $u(0,t) = \overline{u}(t)$ and $u(1,t) = \overline{\overline{u}}(t)$ in the weak sense given in [9]. This condition is very simple if f is monotone:
If $f' > 0$, then $u(0,\cdot) = \overline{u}$ and u does not depend on $\overline{\overline{u}}$.
If $f' < 0$, then $u(1,\cdot) = \overline{\overline{u}}$ and u does not depend on $\overline{u}$.

23.2 A very simple example

One considers here Equation (23.1), with initial condition (23.2) and weak boundary condition $\overline{u}$ and $\overline{\overline{u}}$ at $x = 0$ and $x = 1$, that is in the sense of (23.7), in the particular case $f' > 0$. In this case, the main example of numerical flux is $g = g_0 = g_1$, $g(a,b) = f(a)$, which leads to the well known upstream scheme. With this choice of g_0 and g_1, using the notations of Sect. 23.1, the boundary conditions are taken into account in the form:

$$f^n_{\frac{1}{2}} = f(\overline{u}^n),\ f^n_{N+\frac{1}{2}} = f(u^n_N), \tag{23.8}$$

with $\overline{u}^n = \frac{1}{k}\int_{nk}^{(n+1)k} \overline{u}(t)dt$. One may apply the general convergence theorem. The approximate solutions converge (as $h \to 0$) towards the solution of (23.7). In this case, the approximate solutions, as well as the solution of (23.7), do not depends on $\overline{\overline{u}}$.

In the case $f' < 0$ the main example is $g = g_0 = g_1$, $g(a,b) = f(b)$, which also leads to the upstream scheme. The boundary conditions are taken into account in the following way:

$$f^n_{\frac{1}{2}} = f(u^n_1),\ f^n_{N+\frac{1}{2}} = f(\overline{\overline{u}}^n), \tag{23.9}$$

with $\overline{\overline{u}}^n = \frac{1}{k}\int_{nk}^{(n+1)k} \overline{\overline{u}}(t)dt$.

These simple cases suggest the following scheme for any f, which is the scalar version of the scheme described in Sect. 38.1 (note that $f'(u)$ is the Jacobian matrix at point $u \in \mathbf{R}$):

- Boundary condition at $x = 0$:

$$\begin{cases} f^n_{\frac{1}{2}} = f(\overline{u}^n), & \text{if} \quad f'(u^n_1) > 0, \\ f^n_{\frac{1}{2}} = f(u^n_1), & \text{if} \quad f'(u^n_1) < 0. \end{cases} \tag{23.10}$$

- Boundary condition at $x = 1$:

$$\begin{cases} f^n_{N+\frac{1}{2}} = f(\overline{\overline{u}}^n), & \text{if} \quad f'(u^n_N) < 0, \\ f^n_{N+\frac{1}{2}} = f(u^n_N), & \text{if} \quad f'(u^n_N) > 0. \end{cases} \tag{23.11}$$

This solution is not always satisfactory as can be shown on the following simple example with the Bürgers equation:
Let $f(s) = s^2$, $u_0 = 1$ a.e. on $(0,1)$, $\overline{u} = 1$ a.e. on $\mathbf{R}_+$ and $\overline{\overline{u}} = -2$ a.e. on $\mathbf{R}_+$. The exact solution which has to be approached by the numerical scheme is the unique solution of (23.7) with these values of f, u_0, $\overline{u}$ and $\overline{\overline{u}}$. Computing the approximate solution with (23.3)-(23.5), the function g satisfying $(C2)$, and with (23.10)-(23.11), leads to an approximate solution which is constant and equal to 1 for any h and k. Then, it does not converge (as h and k go to 0) towards the exact solution which is not constant and equal to 1 since, for the exact solution, a shock wave with a negative speed starts from the point $x = 1$ at time $t = 0$. Indeed, one can also remark that this approximate solution is the exact solution of (23.7) with the same values of f, u_0, $\overline{u}$ and with any $\overline{\overline{u}}$ satisfying $\overline{\overline{u}} \geq -1$ a.e. on $\mathbf{R}_+$. In order to obtain a convergent approximation of the exact solution corresponding to $\overline{\overline{u}} = -2$, a good choice is, instead of (23.11), $f^n_{N+\frac{1}{2}} = g_1(u^n_N, -2)$ with g_1 satisfying (C1)-(C3).

23.3 A simplifed model for two phase flows in pipelines

It is now possible to understand the treatment of the boundary described in Sect. 38.1 on a simplified model. This simplifed model for two phase flows in pipelines is given in [124]. For this model, the densities are constant so that there are no longer pressure waves but only the void fraction wave, corresponding to the second eigenvalue of the original system (38.1). It is also easy to see that for this model, the total flux (that is the sum of the fluxes of the two phases) is constant in space. One also assumes that this total flux is constant in time (and positive). System (38.1) is then reduced to a scalar equation, Equation (23.1), where the unknown, $u : (0,1) \times \mathbf{R} \to \mathbf{R}$, is the gas fraction which takes its values between 0 and 1.

The function f can be taken as $f(s) = as - bs^2$, where $a, b \in \mathbf{R}$ are given and such that $0 < b < a < 2b$. In (23.1), the quantity $f(u)$ is the flux of gas. Then, $f(1) - f(u)$ is the flux of liquid. The function f is increasing between 0 and $u_M = a/(2b)$ and decreasing between u_M and 1. An important value is $u_m \in [0, u_M]$ such that $f(u_m) = f(1)$.

One takes $u_0 = 0$ a.e. on $[0,1]$ as an initial condition. At $x = 0$, the gas flux is given (as in the complete model, see Sect. 38.1), one takes $f(u(0,\cdot)) = \overline{f}$ with $\overline{f}(t) = c$ for $t \leq T$ and $\overline{f}(t) = 0$ for $t \geq T$, where c and T are given with $c > f(1)$ and T large enough so that f' changes sign at $x = 1$ during the simulation. Indeed, in this simplified model, it is also necessary to take T not too large in order to avoid a problem at $x = 0$ (for T too large, f' will also changes sign at $x = 0$). The boundary condition at $x = 1$ will be described on the discrete problem below.

The discretization of the problem is performed as before with (23.3)-(23.5), with g satisfying (C1)-(C3).

For the discretization of the boundary condition at $x = 0$, the method described in Sect. 38.1 leads here to

$$f^n_{\frac{1}{2}} = \overline{f}(nk), \tag{23.12}$$

which is indeed in accordance with the fact that $f'(u^n_1) > 0$ for all n, at least if T is not too large.

For the discretization of the boundary condition at $x = 1$, the first method described in Sect. 38.1 and given in (23.11), using the sign of $f'(u^n_N)$ leads to

$$\begin{cases} f^n_{N+\frac{1}{2}} = f(u^n_N), & \text{if} \quad u^n_N < u_M, \\ f^n_{N+\frac{1}{2}} = f(1), & \text{if} \quad u^n_N > u_M, \end{cases} \tag{23.13}$$

and does not lead to the desired results. Note also that $f^n_{N+\frac{1}{2}}$, given by (23.13), is a discontinuous function of u^n_N.

The second method, described in Sect. 38.1, uses the fact that the liquid flux cannot be negative at

$x = 1$. Since the liquid flux at $x = 1$ is $f(1) - f_{N+\frac{1}{2}}$ and since $f(u_m) = f(1)$, this method leads to

$$\left\{ \begin{array}{lll} f^n_{N+\frac{1}{2}} = f(u^n_N), & \text{if} & u^n_N \leq u_m, \\ f^n_{N+\frac{1}{2}} = f(u_m), & \text{if} & u^n_N > u_m, \end{array} \right. \tag{23.14}$$

Note that $f^n_{N+\frac{1}{2}}$, given by (23.14), is a continuous function of u^n_N. We shall apply the convergence theorem, Theorem 23.1 given in Sect. 23.1, for the boundary conditions (23.12) and (23.14), and understand the boundary conditions satisfied by the limit of the approximate solutions. In order to do so, we need to find g_0 and g_1, satisfying (C1)-(C3), and $\overline{u}, \overline{\overline{u}} \in L^\infty(\mathbf{R}_+)$ such that $f^n_{\frac{1}{2}}$ and $f^n_{N+\frac{1}{2}}$, respectively defined by (23.12) and (23.14), satisfy (23.6). Indeed, it is shown in [56] that both boundary fluxes $f^n_{\frac{1}{2}}$ and $f^n_{N+\frac{1}{2}}$ may be expressed with the Godunov flux in the following way:

- Boundary flux at $x = 1$. One takes $\overline{\overline{u}} = 1$ a.e. on $\mathbf{R}_+$ and g_0 equal to the Godunov flux, that is $g_0 = g_G$ with

$$g_G(\alpha, \beta) = \left\{ \begin{array}{ll} \min\{f(s),\ s \in [\alpha, \beta]\} & \text{if} \quad \alpha \leq \beta, \\ \max\{f(s),\ s \in [\beta, \alpha]\} & \text{if} \quad \alpha > \beta. \end{array} \right.$$

  The formula (23.14) reads

$$f^n_{N+\frac{1}{2}} = g_G(u^n_N, 1) = \left\{ \begin{array}{lll} f(u^n_N) & \text{if} & u^n_N \leq u_m, \\ f(1) & \text{if} & u^n_N > u_m. \end{array} \right. \tag{23.15}$$

- Boundary flux at $x = 0$. One assumes (for simplicity) that $\frac{T}{k} \in \mathbf{N}$. let $\alpha, \beta \in (0, 1)$ such that $\alpha < \beta$ and $f(\alpha) = f(\beta) = c$. One takes

$$\overline{u}(t) = \left\{ \begin{array}{lll} \alpha & \text{if} & t < T, \\ 0 & \text{if} & t > T, \end{array} \right. \tag{23.16}$$

  so that, recalling that $\overline{u}^n = \frac{1}{k} \int_{nk}^{(n+1)k} \overline{u}(t)dt$,

$$f(\overline{u}^n) = \left\{ \begin{array}{lll} c & \text{if} & nk < T, \\ 0 & \text{if} & nk \geq T, \end{array} \right.$$

  Then, if $u^n_1 \leq \beta$, the formula (23.12) reads

$$f^n_{\frac{1}{2}} = g_G(\overline{u}^n, u^n_1), \tag{23.17}$$

  since, in this case, $g_G(\overline{u}^n, u^n_1) = f(\overline{u}^n)$. The fact that $u^n_1 \leq \beta$ is true for all n if T is not too large. If T is too large, the convergence result can be applied with (23.17) instead of (23.12).

It is now possible to apply Theorem 23.1. Let L be a common Lipschitz constant for g and g_G (on $[0, 1]^2$) and let $\zeta > 0$. If $k \leq (1-\zeta)\frac{h}{L}$, the approximate solution $u_{h,k}$, that is the solution defined by (23.3)-(23.5), with the boundary fluxes (23.15)-(23.17) (and $u_0 = 0$, $\overline{\overline{u}} = 1$ and $\overline{u}$ given by (23.16)), takes its values in $[0, 1]$ and converges towards the unique solution of (23.18) in $L^p_{loc}([0, 1] \times \mathbf{R}_+)$ for any $1 \leq p < \infty$, as $h \to 0$:

$$\begin{array}{l} u \in L^\infty((0, 1) \times (0, \infty)), \\ \displaystyle\int_0^\infty \int_0^1 [(u - \kappa)^\pm \varphi_t + \text{sign}_\pm(u - \kappa)(f(u) - f(\kappa))\varphi_x]dxdt \\ \displaystyle +M \int_0^\infty (\overline{u}(t) - \kappa)^\pm \varphi(0, t)dt + M \int_0^\infty (1 - \kappa)^\pm \varphi(1, t)dt \\ \displaystyle + \int_0^1 (0 - \kappa)^\pm \varphi(x, 0)dx \geq 0, \\ \forall \kappa \in [0, 1],\ \forall \varphi \in C^1_c([0, 1] \times [0, \infty), \mathbf{R}_+). \end{array} \tag{23.18}$$

If u, solution of (23.18), is regular enough on $[0,1]\times(0,T)$, then, it is possible to prove that u satisfies the boundary conditions, for $0<t<T$, in the following sense (see [157] and [56]):

- Boundary condition at $x=0$ (recall that $\overline{u}$ is given by (23.16)): $u(0,t)=\alpha$ or $u(0,t)\geq\beta$. In fact, if T is not too large, one has $u(0,t)=\alpha$.
- Boundary condition at $x=1$: $u(1,t)\leq u_m$ or $u(1,t)=1$,

Thanks to Theorem 23.1, it is possible to give other choices for $f^n_{N+\frac{1}{2}}$ for which the approximate solutions obtained with this new choice of $f^n_{N+\frac{1}{2}}$ converge towards the same function u, which is the unique solution of (23.18). Indeed, let $h : [0,1]\to\mathbf{R}$ be a nondecreasing function such that $h\leq f$ and $h(1)=f(1)$ and take:

$$f^n_{N+\frac{1}{2}}=h(u^n_N). \tag{23.19}$$

One may construct a function g_1 satisfying (C1)-(C3) such that $h(s)=g_1(s,1)$, for all $s\in[0,1]$, and then use Theorem 23.1. Let L be a common Lipschitz constant for g and g_G and g_1 (on $[0,1]^2$) and let $\zeta>0$. If $k\leq(1-\zeta)\frac{h}{L}$, the approximate solution $u_{h,k}$, that is the solution defined by (23.3)-(23.5), with the boundary fluxes (23.19) and (23.17) (and $u_0=0$, $\overline{\overline{u}}=1$ and $\overline{u}$ given by (23.16)), takes its values in $[0,1]$ and converges towards the unique solution of (23.18) in $L^p_{loc}([0,1]\times\mathbf{R}_+)$ for any $1\leq p<\infty$, as $h\to 0$.

Turning back to the complete system described in Sect. 38.1, the analysis of this simplified model for two phase flows in pipelines may also suggest another way to take into account the boundary condition at $x=1$ (with a given numerical flux g_1):

1. Compute $DF(w^n_N)$, its eigenvalues $\{\lambda_1,\lambda_2,\lambda_3\}$ and a basis of $\mathbf{R}^3$, $\{\varphi_1,\varphi_2,\varphi_3\}$, such that $DF(w^n_N)\varphi_i=\lambda_i\varphi_i$, $i=1,2,3$,
2. write w^n_N on the basis $\{\varphi_1,\varphi_2,\varphi_3\}$, namely $w^n_N=\alpha_1\varphi_1+\alpha_2\varphi_2+\alpha_3\varphi_3$,
3. Since $\lambda_3<0$ and since one wants $Q_l\geq 0$, compute $w^n_{N+1}=\beta_1\varphi_1+\beta_2\varphi_2+\beta_3\varphi_3$ and $F^n_{N+\frac{1}{2}}=g_1(w^n_N,w^n_{N+1})$ with the following 3 conditions on the components of w^n_{N+1}: usual condition on the pressure, $\beta_3=\alpha_3$ and $R^n_{N+1}=1$ where R^n_{N+1} is the gas fraction computed with w^n_{N+1}.

Chapter 6

Multidimensional nonlinear hyperbolic equations

The aim of this chapter is to define and study finite volume schemes for the approximation of the solution to a nonlinear scalar hyperbolic problem in several space dimensions. Explicit and implicit time discretizations are considered. We prove the convergence of the approximate solution towards the entropy weak solution of the problem and give an error estimate between the approximate solution and the entropy weak solution with respect to the discretization mesh size.

24 The continuous problem

We consider here the following nonlinear hyperbolic equation in d space dimensions ($d \geq 1$), with initial condition

$$u_t(x,t) + \mathrm{div}(\mathbf{v} f(u))(x,t) = 0,\ x \in \mathbb{R}^d,\ t \in \mathbb{R}_+, \tag{24.1}$$

$$u(x,0) = u_0(x),\ x \in \mathbb{R}^d, \tag{24.2}$$

where u_t denotes the time derivative of u ($t \in \mathbb{R}_+$), and div the divergence operator with respect to the space variable (which belongs to $\mathbb{R}^d$). Recall that $|x|$ denotes the euclidean norm of x in $\mathbb{R}^d$, and $x \cdot y$ the usual scalar product of x and y in $\mathbb{R}^d$.

The following hypotheses are made on the data:

Assumption 24.1

(i) $u_0 \in L^\infty(\mathbb{R}^d)$, $U_m, U_M \in \mathbb{R}$, $U_m \leq u_0 \leq U_M$ *a.e.*,
(ii) $\mathbf{v} \in C^1(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}^d)$,
(iii) $\mathrm{div}\mathbf{v}(x,t) = 0$, $\forall (x,t) \in \mathbb{R}^d \times \mathbb{R}_+$,
(iv) $\exists V < \infty$ *such that* $|\mathbf{v}(x,t)| \leq V$, $\forall (x,t) \in \mathbb{R}^d \times \mathbb{R}_+$,
(v) $f \in C^1(\mathbb{R}, \mathbb{R})$.

Remark 24.1 Note that part (iv) of Assumption 24.1 is crucial. It ensures the property of "propagation in finite time" which is needed for the uniqueness of the solution of (24.3) and for the stability (under a "Courant-Friedrichs-Levy" (CFL) condition) of the time explicit numerical scheme. Part (iii) of Assumption 24.1, on the other hand, is only considered for the sake of simplicity; the results of existence and uniqueness of the entropy weak solution and convergence (including error estimates as in the theorems 30.1 page 188 and 30.2 page 189) of the numerical schemes presented below may be extended to the case $\mathrm{div}\mathbf{v} \neq 0$. However, part (iii) of Assumption 24.1 is natural in many "applications" and avoids several

technical complications. Note, in particular, that, for instance, if div$\mathbf{v} \neq 0$, the L^∞-bound on the solution of (24.3) and the L^∞ estimate (in Lemma 26.1 and Proposition 27.1) on the approximate solution depend on $\mathbf{v}$ and T. The case $F(x,t,u)$ instead of $\mathbf{v}(x,t)f(u)$ is also feasible, but somewhat more technical, see Chainais-Hillairet [22] and Chainais-Hillairet [23].

Problem (24.1)-(24.2) has a unique entropy weak solution, which is the solution to the following equation (which is the multidimensional extension of the one-dimensional definition 19.3 page 123).

$$\left\{\begin{array}{l} u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star), \\ \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} \Big[\eta(u(x,t))\varphi_t(x,t) + \Phi(u(x,t))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big]dxdt + \\ \displaystyle\int_{\mathbb{R}^d} \eta(u_0(x))\varphi(x,0)dx \geq 0,\ \forall \varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+), \\ \forall \eta \in C^1(\mathbb{R},\mathbb{R}),\ \text{convex function, and } \Phi \in C^1(\mathbb{R},\mathbb{R}) \text{ such that } \Phi' = f'\eta', \end{array}\right.$$

where $\nabla\varphi$ denotes the gradient of the function φ with respect to the space variable (which belongs to $\mathbb{R}^d$). Recall that $C_c^m(E,F)$ denotes the set of functions C^m from E to F, with compact support in E. The characterization of the entropy weak solution by the Krushkov entropies (proposition 19.2 page 124) still holds in the multidimensional case. Let us define again, for all $\kappa \in \mathbb{R}$, the Krushkov entropies $(|\cdot - \kappa|)$ for which the entropy flux is $f(\cdot \top \kappa) - f(\cdot \bot \kappa)$ (for any pair of real values a, b, we denote again by $a \top b$ the maximum of a and b, and by $a \bot b$ the minimum of a and b). The unique entropy weak solution is also the unique solution to the following problem:

$$\left\{\begin{array}{l} u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star), \\ \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} \Big[|u(x,t)-\kappa|\varphi_t(x,t) + \Big(f(u(x,t)\top\kappa) - f(u(x,t)\bot\kappa)\Big)\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big]dxdt + \\ \displaystyle\int_{\mathbb{R}^d} |u_0(x)-\kappa|\varphi(x,0)dx \geq 0,\ \forall \kappa \in \mathbb{R},\ \forall \varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+). \end{array}\right. \tag{24.3}$$

As in the one-dimensional case (Theorem 19.1 page 124), existence and uniqueness results are also known for the entropy weak solution to Problem (24.1)-(24.2) under assumptions which differ slightly from assumption 24.1 (see e.g. Krushkov [94], Vol'pert [156]). In particular, these results are obtained with a nonlinearity F (in our case $F = \mathbf{v}f$) of class C^3. We recall that the methods which were used in Krushkov [94] require a regularization in $BV(\mathbb{R}^d)$ of the function u_0, in order to take advantage, for any $T > 0$, of compactness properties which are similar to those given in Lemma 21.4 page 141 for the case $d = 1$. Recall that the space $BV(\Omega)$ where Ω is an open subset of $\mathbb{R}^p$, $p \geq 1$, was defined in Definition 21.2 page 141; it will be used later with $\Omega = \mathbb{R}^d$ or $\Omega = \mathbb{R}^d \times (-T, T)$.

The existence of solutions to similar problems to (24.1)-(24.2) was already proved by passing to the limit on solutions of an appropriate numerical scheme, see Conway and Smoller [36]. The work of Conway and Smoller [36] uses a finite difference scheme on a uniform rectangular grid, in two space dimensions, and requires that the initial condition u_0 belongs to $BV(\mathbb{R}^d)$ (and thus, the solution to Problem (24.1)-(24.2) also has a locally bounded variation). These assumptions (on meshes and on u_0) yield, as in Lemma 21.4 page 141, a (strong) compactness property in $L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ on a family of approximate solutions. In the following, however, we shall only require that $u_0 \in L^\infty(\mathbb{R}^d)$ and we shall be able to deal with more general meshes. We may use, for instance, a triangular mesh in the case of two space dimensions. For each of these reasons, the BV framework may not be used and a (strong) compactness property in L^1_{loc} on a family of approximate solutions is not easy to obtain (although this compactness property does hold and results from this chapter). In order to prove the existence of a solution to (24.1)-(24.2) by passing to the limit on the approximate solutions given by finite volume schemes on general meshes (in the sense used below) in two or three space dimensions, we shall work with some "weak" compactness result in L^∞, namely Proposition 32.1, which yields the "nonlinear weak-$\star$ convergence" (see Definition 32.1 page 200) of a family of approximate solutions. When doing so, passing to the limit with the approximate solutions

will give the existence of an "entropy process solution" to Problem (24.1)-(24.2), see Definition 29.1 page 181. A uniqueness result for the entropy process solution to Problem (24.1)-(24.2) is then proven. This uniqueness result proves that the entropy process solution is indeed the entropy weak solution, hence the existence and uniqueness of the entropy weak solution. This uniqueness result also allows us to conclude to the convergence of the approximate solution given by the numerical scheme (that is (25.4), (25.2)) towards the entropy weak solution to (24.1)-(24.2) (this convergence holds in $L^p_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ for any $1 \leq p < \infty$).

Note that uniqueness results for "generalized" solutions (namely measure valued solutions) to (24.1)-(24.2) have recently been proved (see DiPerna [46], Szepessy [140], Gallouët and Herbin [71]). The proofs of these results rely on the one hand on the concept of measure valued solutions and on the other hand on the existence of an entropy weak solution. The direct proof of the uniqueness of a measure valued solution (i.e. without assuming any existence result of entropy weak solutions) leads to a difficult problem involving the application of the theorem of continuity in mean. This difficulty is easier to deal within the framework of entropy process solutions (but in fact, measure valued solutions and entropy process solutions are two presentations of the same concept).

Developing the above analysis gives a (strong) convergence result of approximate solutions towards the entropy weak solution. But moreover, we also derive some error estimates depending on the regularity of u_0.

In the case of a Cartesian grid, the convergence and error analysis reduces essentially to a one-dimensional discretization problem for which results were proved some time ago, see e.g. Kuznetsov [96], Crandall and Majda [43], Sanders [133]. In the case of general meshes, the numerical schemes are not generally "TVD" (Total Variation Diminushing) and therefore the classical framework of the 1D case (see Section 21.5 page 141) may not be used. More recent works deal with several convergence results and error estimates for time explicit finite volume schemes, see e.g. Cockburn, Coquel and LeFloch [32], Champier, Gallouët and Herbin [25], Vila [155], Kröner and Rokyta [92], Kröner, Noelle and Rokyta [93], Kröner [91]: following Szepessy's work on the convergence of the streamline diffusion method (see Szepessy [140]), most of these works use DiPerna's uniqueness theorem, see DiPerna [46] (or an adaptation of it, see Gallouët and Herbin [71] and Eymard, Gallouët and Herbin [54]), and the error estimates generalize the work by Kuznetsov [96]. Here we use the framework of Champier, Gallouët and Herbin [25], Eymard, Gallouët, Ghilani and Herbin [52]; we prove directly that any monotone flux scheme (defined below) satisfies a "weak BV" estimate (see lemmata 26.2 page 161 and 27.1 page 167). This inequality appears to be a key for the proof of convergence and for the error estimate. Some convergence results and error estimates are also possible with some so called "higher order schemes" which are not monotone flux schemes (briefly presented for the 1D case in section 22 page 146). These results are not presented here, see Noëlle [117] and Chainais-Hillairet [22] for some of them.

Note that the nonlinearity considered here is of the form $\mathbf{v}(x,t)f(u)$. This kind of flux is often encountered in porous medium modelling, where the hyperbolic equation may then be coupled with an elliptic or parabolic equation (see e.g. Eymard and Gallouët [49], Vignal [151], Vignal [152], Herbin and Labergerie [86]). It adds an extra difficulty to the case $F(u)$ because of the dependency on x and t. Note again (see Remark 24.1) that the method which we present here for a nonlinearity of the form $\mathbf{v}(x,t)f(u)$ also yields the same results in the case of a nonlinearity of the form $F(x,t,u)$, see the recent work of Chainais-Hillairet [23].

The time implicit discretization adds the extra difficulties of proving the existence of the approximate solution (see Lemma 27.1 page 165) and proving a so called "strong time BV estimate" (see Lemma 27.3 page 170) in order to show that the error estimate for the implicit scheme may still be of order $h^{1/4}$ even if the time step k is of order $\sqrt{h}$, at least in particular cases.

We first describe in section 25 finite volume schemes using a "general" mesh for the discretization of

(24.1)-(24.2). In sections 26 and 27 some estimates on the approximate solution given by the numerical schemes are shown and in Section 28 some entropy inequalities are proven. We then prove in section 29 the convergence of convenient subsequences of sequences of approximate solutions towards an entropy process solution, by passing to the limit when the mesh size and the time step go to 0. A byproduct of this result is the existence of an entropy process solution to (24.1)-(24.2) (see Definition 29.1 page 181). The uniqueness of the entropy process solution to problem (24.1)-(24.2) is then proved; we can therefore conclude to the existence and uniqueness of the entropy weak solution and also to the L^p_{loc} convergence for any finite p of the approximate solution towards the entropy weak solution (Section 29). Using the existence of the entropy weak solution, an error estimate result is given in Section 30 (which also yields the convergence result). Therefore the main interest of this convergence result is precisely to prove the existence of the entropy weak solution to (24.1)-(24.2) without any regularity assumption on the initial data. Section 32 describes the notion of nonlinear weak-$\star$ convergence, which is widely used in the proof of convergence of section 29.

Section 33 is not related to the previous sections. It describes a finite volume approach which may be used to stabilize finite element schemes for the discretization of a hyperbolic equation (or system).

25 Meshes and schemes

Let us first define an admissible mesh of $\mathbb{R}^d$ as a generalization of the notion of admissible mesh of $\mathbb{R}$ as defined in definition 20.1 page 128.

Definition 25.1 (Admissible meshes) An admissible finite volume mesh of $\mathbb{R}^d$, with $d = 1, 2$ or 3 (for the discretization of Problem (24.1)-(24.2)), denoted by $\mathcal{T}$, is given by a family of disjoint polygonal connected subsets of $\mathbb{R}^d$ such that $\mathbb{R}^d$ is the union of the closure of the elements of $\mathcal{T}$ (which are called control volumes in the following) and such that the common "interface" of any two control volumes is included in a hyperplane of $\mathbb{R}^d$ (this is not necessary but is introduced to simplify the formulation). Denoting by $h = \text{size}(\mathcal{T}) = \sup\{\text{diam}(K), K \in \mathcal{T}\}$, it is assumed that $h < +\infty$ and that, for some $\alpha > 0$,

$$\begin{array}{l} \alpha h^d \leq \text{m}(K), \\ \text{m}(\partial K) \leq \frac{1}{\alpha} h^{d-1}, \ \forall K \in \mathcal{T}, \end{array} \tag{25.1}$$

where $\text{m}(K)$ denotes the d-dimensional Lebesgue measure of K, $\text{m}(\partial K)$ denotes the $(d-1)$-dimensional Lebesgue measure of ∂K (∂K is the boundary of K) and $\mathcal{N}(K)$ denotes the set of neighbours of the control volume K; for $L \in \mathcal{N}(K)$, we denote by $K|L$ the common interface between K and L, and by $\boldsymbol{n}_{K,L}$ the unit normal vector to $K|L$ oriented from K to L. The set of all the interfaces is denoted by $\mathcal{E}$.

Note that, in this definition, the terminology is "mixed". For $d = 3$, "polygonal" stands for "polyhedral" and, for $d = 2$, "interface" stands for "edge". For $d = 1$ definition 25.1 is equivalent to definition 20.1 page 128.

In order to define the numerical flux, we consider functions $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfying the following assumptions:

Assumption 25.1 *Under Assumption 24.1 the function g, only depending on f, $\mathbf{v}$, U_m and U_M, satisfies*

- *g is locally Lipschitz continuous from $\mathbb{R}^2$ to $\mathbb{R}$,*
- *$g(s, s) = f(s)$, for all $s \in [U_m, U_M]$,*
- *$(a, b) \mapsto g(a, b)$, from $[U_m, U_M]^2$ to $\mathbb{R}$, is nondecreasing with respect to a and nonincreasing with respect to b.*

Let us denote by g_1 and g_2 the Lipschitz constants of g on $[U_m, U_M]^2$ with respect to its two arguments.

The hypotheses on g are the same as those presented for monotone flux schemes in the one-dimensional case (see definition 21.1 page 134); the function g allows the construction of a numerical flux, see Remark 25.2 below.

Remark 25.1 In Assumption 25.1, the third item will ensure some stability properties of the schemes defined below. In particular, in the case of the "explicit scheme" (see (25.4)), it yields the monotonicity of the scheme under a CFL condition (namely, condition (25.3) with $\xi = 0$). The second item is essential since it ensures the consistency of the fluxes. All the examples of functions g given in Examples 21.1 page 135 satisfy these assumptions. We again give the important example of the "generalized 1D Godunov scheme" obtained with a one-dimensional Godunov scheme for each interface (see e.g., for the explicit scheme, see Cockburn, Coquel and LeFloch [32], Vila [155]),

$$g(a,b) = \begin{cases} \max\{f(s),\, b \le s \le a\} \text{ if } b \le a \\ \min\{f(s),\, a \le s \le b\} \text{ if } a \le b, \end{cases}$$

and also the framework of some "flux splitting" schemes:

$$g(a,b) = f_1(a) + f_2(b),$$

with $f_1,\, f_2 \in C^1(\mathbb{R}, \mathbb{R})$, $f = f_1 + f_2$, f_1 nondecreasing and f_2 nonincreasing (this framework is considerably more simple that the general framework, because it reduces the study to the particular case of two monotone nonlinearities).

Besides, it is possible to replace Assumption 25.1 on g by some slightly more general assumption, in order to handle, in particular, the case of some "Lax-Friedrichs type" schemes (see Remark 30.1 below).

In order to describe the numerical schemes considered here, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$ be the time step. The discrete unknowns are u_K^n, $n \in \mathbb{N}^\star$, $K \in \mathcal{T}$. The set $\{u_K^0,\, K \in \mathcal{T}\}$ is given by the initial condition,

$$u_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x)dx, \forall\, K \in \mathcal{T}. \tag{25.2}$$

The equations satisfied by the discrete unknowns, u_K^n, $n \in \mathbb{N}^\star$, $K \in \mathcal{T}$, are obtained by discretizing equation (24.1). We now describe the explicit and implicit schemes.

25.1 Explicit schemes

We present here the "explicit scheme" associated to a function g satisfying Assumption 25.1. In this case, for stability reasons (see lemmata 26.1 and 26.2), the time step $k \in \mathbb{R}_+^\star$ is chosen such that

$$k \le (1-\xi)\frac{\alpha^2 h}{V(g_1+g_2)}, \tag{25.3}$$

where $\xi \in (0,1)$ is a given real value; recall that g_1 and g_2 are the Lipschitz constants of g with respect to the first and second variables on $[U_m, U_M]^2$ and that $U_m \le u_0 \le U_M$ a.e. and $|\mathbf{v}(x,t)| \le V < +\infty$, for all $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$. Consider the following explicit numerical scheme:

$$\mathrm{m}(K)\frac{u_K^{n+1} - u_K^n}{k} + \sum_{L \in \mathcal{N}(K)} \left(v_{K,L}^n\, g(u_K^n, u_L^n) - v_{L,K}^n\, g(u_L^n, u_K^n) \right) = 0,\ \forall K \in \mathcal{T},\ \forall n \in \mathbb{N}, \tag{25.4}$$

where

$$v_{K,L}^n = \frac{1}{k} \int_{nk}^{(n+1)k} \int_{K|L} (\mathbf{v}(x,t) \cdot \boldsymbol{n}_{K,L})^+ d\gamma(x)dt$$

and

$$v_{L,K}^n = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K|L}(\mathbf{v}(x,t)\cdot\boldsymbol{n}_{L,K})^+ d\gamma(x)dt$$
$$= \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K|L}(\mathbf{v}(x,t)\cdot\boldsymbol{n}_{K,L})^- d\gamma(x)dt.$$

Recall that $a^+ = a\top 0$ and $a^- = -(a\perp 0)$ for all $a \in \mathbb{R}$ and that $d\gamma$ is the integration symbol for the $(d-1)$-dimensional Lebesgue measure on the considered hyperplane.

Remark 25.2 (Numerical fluxes) The numerical flux at the interface between the control volume K and the control volume $L \in \mathcal{N}(K)$ is then equal to $v_{K,L}^n\ g(u_K^n, u_L^n) - v_{L,K}^n\ g(u_L^n, u_K^n)$; this expression yields a monotone flux such as defined in definition 21.1 page 134, given in the one-dimensional case. However, in the multidimensional case, the expression of the numerical flux depends on the considered interface; this was not so in the one-dimensional case for which the numerical flux is completely defined by the function g.

The approximate solution, denoted by $u_{\mathcal{T},k}$, is defined a.e. from $\mathbb{R}^d \times \mathbb{R}_+$ to $\mathbb{R}$ by

$$u_{\mathcal{T},k}(x,t) = u_K^n, \text{ if } x \in K,\ t \in [nk,(n+1)k),\ K \in \mathcal{T},\ n \in \mathbb{N}. \tag{25.5}$$

25.2 Implicit schemes

The use of implicit schemes is steadily increasing in industrial codes for reasons such as robustness and computational cost. Hence we consider in our analysis the following implicit numerical scheme (for which condition (25.3) is no longer needed) associated to a function g satisfying Assumption 25.1:

$$\mathrm{m}(K)\frac{u_K^{n+1}-u_K^n}{k} + \sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(u_K^{n+1},u_L^{n+1}) - v_{L,K}^n\ g(u_L^{n+1},u_K^{n+1})) = 0,\ \forall K \in \mathcal{T},\ \forall n \in \mathbb{N}. \tag{25.6}$$

where $\{u_K^0,\ K \in \mathcal{T}\}$ is still determined by (25.2). The implicit approximate solution $u_{\mathcal{T},k}$, is defined now a.e. from $\mathbb{R}^d \times \mathbb{R}_+$ to $\mathbb{R}$ by

$$u_{\mathcal{T},k}(x,t) = u_K^{n+1}, \text{ if } x \in K,\ t \in (nk,(n+1)k],\ K \in \mathcal{T},\ n \in \mathbb{N}. \tag{25.7}$$

25.3 Passing to the limit

We show in section 29 page 181 the convergence of the approximate solutions $u_{\mathcal{T},k}$ (given by the numerical schemes above described) towards the unique entropy weak solution u to (24.1)-(24.2) in an adequate sense, when size$(\mathcal{T}) \to 0$ and $k \to 0$ (with, possibly, a stability condition). In order to describe the general line of thought leading to this convergence result, we shall simply consider the explicit scheme (that is (25.2), (25.4) and (25.5)) (the implicit scheme will also be fully investigated later).
First, in section 26, by writing u_K^{n+1} as a convex combination of u_K^n and $(u_L^n)_{L\in\mathcal{N}(K)}$, the L^∞ stability is easily shown under the CFL condition (25.3) ($u_{\mathcal{T},k}$ is proved to be bounded in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$, independently of size$(\mathcal{T})$ and k).
Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and let k satisfy (25.3)); by a classical argument, if any possible limit of a family of approximate solutions $u_{\mathcal{T},k}$ is the entropy weak solution to problem (24.1)-(24.2) then $u_{\mathcal{T},k}$ converges (in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology, for instance), as $h = \text{size}(\mathcal{T}) \to 0$ (and k satisfies (25.3)), towards the unique entropy weak solution to problem (24.1)-(24.2). Unfortunately, the L^∞ estimate of section 26 does not yield that any possible limit of a family

of approximate solutions is solution to problem (24.1)-(24.2), even in the linear case ($f(u) = u$) (see the proofs of convergence of Chapter 5). The "BV stability" can be used (combined with the L^∞ stability) to show the convergence in the case of one space dimension (see section 21.5 page 141) and in the case of Cartesian meshes in two or three space dimensions. Indeed, in the case of Cartesian meshes, assuming $u_0 \in BV(\mathbb{R}^d)$ and assuming (for simplicity) $\mathbf{v}$ to be constant (a generalization is possiblefor $\mathbf{v}$ regular enough), the following estimate holds, for all $T \geq k$:

$$k \sum_{n=0}^{N_{T,k}} \sum_{K|L \in \mathcal{E}} \mathrm{m}(K|L)|u_K^n - u_L^n| \leq T|u_0|_{BV(\mathbb{R}^d)},$$

where $N_{T,k} \in \mathbb{N}$ is such that $(N_{T,k}+1)k \leq T < (N_{T,k}+2)k$, and the values u_K^n are given by (25.2) and (25.4). Such an estimate is wrong in the general case of admissible meshes in the sense of Definition 25.1 page 156, as it can be shown with easy counterexamples. It is, however, not necessary for the proof of convergence. A weaker inequality, which is called "weak BV" as in the one-dimensional case (see lemma 21.3 page 137) will be shown in the multidimensional case for both explicit and implicit schemes (see lemmata 26.2 page 161 and 27.1 page 167); the weak BV estimate yields the convergence of the scheme in the general case. As an illustration, consider the case $f' \geq 0$; using an upwind scheme, i.e. $g(a,b) = f(a)$, the weak BV inequality (26.4) page 161, which is very close to that of the 1D case (lemma 21.3 page 137), reads

$$\sum_{n=0}^{N_{T,k}} k \sum_{(K,L) \in \mathcal{E}_R^n} (v_{K,L}^n + v_{L,K}^n)|f(u_K^n) - f(u_L^n)| \leq \frac{C}{\sqrt{h}}, \tag{25.8}$$

where $\mathcal{E}_R^n = \{(K,L) \in \mathcal{T}^2, L \in \mathcal{N}(K), K|L \subset B(0,R) \text{ and } u_K^n > u_L^n\}$ and C only depends on $\mathbf{v}$, g, u_0, α, ξ, R and T (see Lemma 26.2).
We say that Inequality (25.8) is "weak", but it is in fact "three times weak" for the following reasons:

1. the inequality is of order $\frac{1}{\sqrt{h}}$, and not of order 1.
2. In the left hand side of (25.8), the quantity which is associated to the $K|L \in \mathcal{E}_R^n$ interface is zero if f is constant on the interval to which the values u_K^n and u_L^n belong; variations of the discrete unknowns in this interval are therefore not taken into account.
3. The left hand side of (25.8) involves terms $(v_{K,L}^n + v_{L,K}^n)$ which are not uniformly bounded from below by $C\ \mathrm{m}(K|L)$ with some $C > 0$ only depending on the data (that is $\mathbf{v}$, u_0 and g) and not on $\mathcal{T}$ (note that, for instance, $v_{K,L}^n = v_{L,K}^n = 0$ if $\mathbf{v} \cdot \mathbf{n}_{K,L} = 0$).

For the convergence result (namely Theorem 29.2 page 187) the useful consequence of (25.8) is

$$h \sum_{n=0}^{N_{T,k}} k \sum_{(K,L) \in \mathcal{E}_R^n} (v_{K,L}^n + v_{L,K}^n)|f(u_K^n) - f(u_L^n)| \to 0 \text{ as } h \to 0,$$

as in the 1D case, see Theorem 21.1 page 139. For the error estimate in Theorem 30.1 page 188, the bound $C/\sqrt{h}$ in (25.8) is crucial. Note that a "twice weak BV" inequality in the sense (ii) and (iii), but of order 1 (that is C instead of $C/\sqrt{h}$ in the right hand side of (25.8)), would yield a sharp error estimate, i.e. $C_e h^{1/2}$ instead of $C_e h^{1/4}$ in (30.1) page 188.
Note that, in order to obtain (25.8), $\xi > 0$ is crucial in the CFL condition (25.3).
Recall also that (25.8) together with the $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ bound does not yield any (strong) compactness property in $L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ on a family of approximated solutions.

In the linear case (that is $f(s) = cs$ for all $s \in \mathbb{R}$, for some c in $\mathbb{R}$), the inequality (25.8) is used in the same manner as in the previous chapter; one proves that the approximate solution satisfies the weak formulation to (24.1)-(24.2) (which is equivalent to (24.3)) with an error which goes to 0 as $h \to 0$, under

condition (25.3). We deduce from this the convergence of $u_{\mathcal{T},k}$ (as $h \to 0$ and under condition (25.3)) towards the unique weak solution of (24.1)-(24.2) in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology. In fact, the convergence holds in $L^p_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ (strongly) for any $1 \leq p < \infty$, thanks to the argument developped for the study of the nonlinear case.

The nonlinear case adds an extra difficulty, as in the 1D case; it will be handled in detail in the present chapter. This difficulty arises from the fact that, if $u_{\mathcal{T},k}$ converges to u (as $h \to 0$, under condition (25.3)) and $f(u_{\mathcal{T},k})$ to μ_f, in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology, there remains to show that $\mu_f = f(u)$ and that u is the entropy weak solution to problem (24.1)-(24.2). The weak BV inequality (25.8) is used to show that, for any "entropy" function η, i.e. convex function of class C^1 from $\mathbb{R}$ to $\mathbb{R}$, with associated entropy flux ϕ, i.e. ϕ such that $\phi' = f'\eta'$, the following entropy inequality is satisfied:

$$\left\{ \begin{array}{l} \displaystyle\int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big(\mu_\eta(x,t)\varphi_t(x,t) + \mu_\phi(x,t)\mathbf{v}(x,t)\cdot\nabla\varphi(x,t) \Big) dxdt + \int_{\mathbb{R}^d} \eta(u_0(x))\varphi(x,0)dx \geq 0, \\ \forall\, \varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+), \end{array} \right. \tag{25.9}$$

where μ_η (resp. μ_ϕ) is the limit of $\eta(u_{\mathcal{T},k})$ (resp. $\phi(u_{\mathcal{T},k})$) in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ for the weak-$\star$ topology (the existence of these limits can indeed be assumed). From (25.9), it is shown that $u_{\mathcal{T},k}$ converges to u in $L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ (as $h \to 0$, k satisfying (25.3)), and that u is the entropy weak solution to problem (24.1)-(24.2). This last result uses a generalization of a result on measure valued solutions of DiPerna (see DiPerna [46], Gallouët and Herbin [71]), and is developped in section 29 page 181.

26 Stability results for the explicit scheme

26.1 L^∞ stability

Lemma 26.1 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$, let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1 and assume that (25.3) holds; let $u_{\mathcal{T},k}$ be given by (25.5), (25.4), (25.2); then,*

$$U_m \leq u_K^n \leq U_M,\ \forall n \in \mathbb{N},\ \forall K \in \mathcal{T}, \tag{26.1}$$

and

$$\|u_{\mathcal{T},k}\|_{L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)} \leq \|u_0\|_{L^\infty(\mathbb{R}^d)}. \tag{26.2}$$

Proof of Lemma 26.1

Note that (26.2) is a straightforward consequence of (26.1), which will be proved by induction. For $n = 0$, since $U_m \leq u_0 \leq U_M$ a.e., (26.1) follows from (25.2).

Let $n \in \mathbb{N}$, assume that $U_m \leq u_K^n \leq U_M$ for all $K \in \mathcal{T}$. Using the fact that div$\mathbf{v} = 0$, which yields $\displaystyle\sum_{L \in \mathcal{N}(K)} (v_{K,L}^n - v_{L,K}^n) = 0$, we can rewrite (25.4) as

$$\mathrm{m}(K)\frac{u_K^{n+1} - u_K^n}{k} + \sum_{L \in \mathcal{N}(K)} \Big(v_{K,L}^n (g(u_K^n, u_L^n) - f(u_K^n)) - v_{L,K}^n (g(u_L^n, u_K^n) - f(u_K^n)) \Big) = 0. \tag{26.3}$$

Set, for $u_K^n \neq u_L^n$,

$$\tau_{K,L}^n = v_{K,L}^n \frac{g(u_K^n, u_L^n) - f(u_K^n)}{u_K^n - u_L^n} - v_{L,K}^n \frac{g(u_L^n, u_K^n) - f(u_K^n)}{u_K^n - u_L^n},$$

and $\tau_{K,L}^n = 0$ if $u_K^n = u_L^n$.

Assumption 25.1 on g and Assumption 24.1 yields $0 \leq \tau_{K,L}^n \leq Vm(K|L)(g_1 + g_2)$. Using (26.3), we can write

$$u_K^{n+1} = \Big(1 - \frac{k}{\mathrm{m}(K)} \sum_{L\in\mathcal{N}(K)} \tau_{K,L}^n\Big) u_K^n + \frac{k}{\mathrm{m}(K)} \sum_{L\in\mathcal{N}(K)} \tau_{K,L}^n u_L^n,$$

which gives, under condition (25.3), $\inf_{L\in\mathcal{T}} u_L^n \le u_K^{n+1} \le \sup_{L\in\mathcal{T}} u_L^n$, for all $K \in \mathcal{T}$. This concludes the proof of (26.1), which, in turn, yields (26.2).

Remark 26.1 Note that the stability result (26.2) holds even if $\xi = 0$ in (25.3). However, we shall need $\xi > 0$ for the following "weak BV" inequality.

26.2 A "weak BV" estimate

In the following lemma, $B(0,R)$ denotes the ball of $\mathbb{R}^d$ of center 0 and radius R ($\mathbb{R}^d$ is always endowed with its usual scalar product).

Lemma 26.2 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1 and assume that (25.3) holds. Let $u_{\mathcal{T},k}$ be given by (25.5), (25.4), (25.2).*
Let $T > 0$, $R > 0$, $N_{T,k} = \max\{n \in \mathbb{N}, n < T/k\}$, $\mathcal{T}_R = \{K \in \mathcal{T}, K \subset B(0,R)\}$ and $\mathcal{E}_R^n = \{(K,L) \in \mathcal{T}^2, L \in \mathcal{N}(K), K|L \subset B(0,R)$ and $u_K^n > u_L^n\}$.
Then there exists $C \in \mathbb{R}$, only depending on $\mathbf{v}$, g, u_0, α, ξ, R, T such that, for $h < R$ and $k < T$,

$$\begin{aligned}\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n} \Big[v_{K,L}^n \Big(\max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(q)) + \max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(p)) \Big) + \\ v_{L,K}^n \Big(\max_{u_L^n \le p \le q \le u_K^n} (f(q) - g(p,q)) + \max_{u_L^n \le p \le q \le u_K^n} (f(p) - g(p,q)) \Big) \Big] \\ \le \frac{C}{\sqrt{h}},\end{aligned} \tag{26.4}$$

and

$$\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}_R} \mathrm{m}(K) |u_K^{n+1} - u_K^n| \le \frac{C}{\sqrt{h}}, \tag{26.5}$$

Proof of Lemma 26.2

In this proof, we shall denote by C_i ($i \in \mathbb{N}$) various quantities only depending on $\mathbf{v}$, g, u_0, α, ξ, R, T. Multiplying (26.3) by $k u_K^n$ and summing the result over $K \in \mathcal{T}_R$, $n \in \{0, \dots, N_{T,k}\}$ yields

$$B_1 + B_2 = 0, \tag{26.6}$$

with

$$B_1 = \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}_R} \mathrm{m}(K) u_K^n (u_K^{n+1} - u_K^n),$$

and

$$B_2 = \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}_R} \sum_{L\in\mathcal{N}(K)} \Big(v_{K,L}^n (g(u_K^n, u_L^n) - f(u_K^n)) u_K^n - v_{L,K}^n (g(u_L^n, u_K^n) - f(u_K^n)) u_K^n \Big).$$

Gathering the last two summations by edges in B_2 leads to the definition of B_3:

$$B_3 = \sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n} \Big[\; v_{K,L}^n \Big(u_K^n (g(u_K^n, u_L^n) - f(u_K^n)) - u_L^n (g(u_K^n, u_L^n) - f(u_L^n)) \Big) - v_{L,K}^n \Big(u_K^n (g(u_L^n, u_K^n) - f(u_K^n)) - u_L^n (g(u_L^n, u_K^n) - f(u_L^n)) \Big) \Big].$$

The expression $|B_3 - B_2|$ can be reduced to a sum of terms each of which corresponds to the boundary of a control volume which is included in $B(0, R+h) \setminus B(0, R-h)$; since the measure of $B(0, R+h) \setminus B(0, R-h)$ is less than $C_2 h$, the number of such terms is, for n fixed, lower than $(C_2 h)/(\alpha h^d) = C_3 h^{1-d}$. Thanks to (26.2), using the fact that $\mathrm{m}(\partial K) \leq (1/\alpha) h^{d-1}$, that $|\mathbf{v}(x,t)| \leq V$, that g is bounded on $[U_m, U_M]^2$, and that $g(s,s) = f(s)$, one may show that each of the non zero term in $|B_3 - B_2|$ is bounded by $C_1 h^{d-1}$. Furthermore, since $(N_{T,k} + 1)k \leq 2k$, we deduce that

$$|B_3 - B_2| \leq C_4. \tag{26.7}$$

Denoting by Φ a primitive of the function $(\cdot) f'(\cdot)$, an integration by parts yields, for all $(a,b) \in \mathbb{R}^2$,

$$\Phi(b) - \Phi(a) = \int_a^b s f'(s) ds = b(f(b) - g(a,b)) - a(f(a) - g(a,b)) - \int_a^b (f(s) - g(a,b)) ds. \tag{26.8}$$

Using (26.8), the term B_3 may be decomposed as

$$B_3 = B_4 - B_5,$$

where

$$B_4 = \sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n} \left(v_{K,L}^n \int_{u_K^n}^{u_L^n} (f(s) - g(u_K^n, u_L^n)) ds + v_{L,K}^n \int_{u_L^n}^{u_K^n} (f(s) - g(u_L^n, u_K^n)) ds \right)$$

and

$$B_5 = \sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n} (v_{K,L}^n - v_{L,K}^n) \Big(\Phi(u_K^n) - \Phi(u_L^n) \Big).$$

The term B_5 is again reduced to a sum of terms corresponding to control volumes included in $B(0, R + h) \setminus B(0, R - h)$, thanks to $\mathrm{div}\mathbf{v} = 0$; therefore, as for (26.7), there exists $C_5 \in \mathbb{R}$ such that

$$B_5 \leq C_5.$$

Let us now turn to an estimate of B_4. To this purpose, let $a, b \in \mathbb{R}$, define $\mathcal{C}(a,b) = \{(p,q) \in [a \perp b, a \top b]^2;$ $(q-p)(b-a) \geq 0\}$. Thanks to the monotonicity properties of g (and using the fact that $g(s,s) = f(s)$), the following inequality holds, for any $(p,q) \in \mathcal{C}(a,b)$:

$$\int_a^b (f(s) - g(a,b)) ds \geq \int_c^d (f(s) - g(a,b)) ds \geq \int_p^q (f(s) - g(p,q)) ds \geq 0. \tag{26.9}$$

The technical lemma 18.5 page 110 can then be applied. It states that

$$\Big| \int_p^q (\theta(s) - \theta(p)) ds \Big| \geq \frac{1}{2G} (\theta(q) - \theta(p))^2, \quad \forall p, q \in \mathbb{R},$$

for all monotone, Lipschitz continuous function $\theta : \mathbb{R} \to \mathbb{R}$, with a Lipschitz constant $G > 0$.
From Lemma 18.5, we can notice that

$$\int_p^q (f(s) - g(p,q)) ds \geq \int_p^q (g(p,s) - g(p,q)) ds \geq \frac{1}{2g_2} (f(p) - g(p,q))^2, \tag{26.10}$$

and

$$\int_p^q (f(s)-g(p,q))ds \geq \int_p^q (g(s,q)-g(p,q))ds \geq \frac{1}{2g_1}(f(q)-g(p,q))^2. \tag{26.11}$$

Multiplying (26.10) (resp. (26.11)) by $g_2/(g_1+g_2)$ (resp. $g_1/(g_1+g_2)$), taking the maximum for $(p,q) \in \mathcal{C}(a,b)$, and adding the two equations yields, with (26.9),

$$\int_a^b (f(s)-g(a,b))ds \geq \frac{1}{2(g_1+g_2)}\Big(\max_{(p,q)\in\mathcal{C}(a,b)}(f(p)-g(p,q))^2 + \max_{(p,q)\in\mathcal{C}(a,b)}(f(q)-g(p,q))^2\Big). \tag{26.12}$$

We can then deduce, from (26.12):

$$\begin{aligned} B_4 \geq \; & \frac{1}{2(g_1+g_2)}\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n} \Big[\\ & v_{K,L}^n\Big(\max_{u_L^n\leq p\leq q\leq u_K^n}(g(q,p)-f(q))^2 + \max_{u_L^n\leq p\leq q\leq u_K^n}(g(q,p)-f(p))^2\Big) + \\ & v_{L,K}^n\Big(\max_{u_L^n\leq p\leq q\leq u_K^n}(f(q)-g(p,q))^2 + \max_{u_L^n\leq p\leq q\leq u_K^n}(f(p)-g(p,q))^2\Big)\Big]. \end{aligned} \tag{26.13}$$

This gives a bound on B_2, since (with $C_6 = C_4 + C_5$):

$$B_2 \geq B_4 - C_6. \tag{26.14}$$

Let us now turn to B_1. We have

$$B_1 = -\frac{1}{2}\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}_R} \mathrm{m}(K)(u_K^{n+1}-u_K^n)^2 + \frac{1}{2}\sum_{K\in\mathcal{T}_R}\mathrm{m}(K)\Big(u_K^{N_{T,k}+1}\Big)^2 - \frac{1}{2}\sum_{K\in\mathcal{T}_R}\mathrm{m}(K)\Big(u_K^0\Big)^2. \tag{26.15}$$

Using (26.3) and the Cauchy-Schwarz inequality yields the following inequality:

$$\begin{aligned} & (u_K^{n+1}-u_K^n)^2 \leq \\ & \frac{k^2}{\mathrm{m}(K)^2}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n+v_{L,K}^n)\sum_{L\in\mathcal{N}(K)}\Big[v_{K,L}^n\Big(g(u_K^n,u_L^n)-f(u_K^n)\Big)^2 + v_{L,K}^n\Big(g(u_L^n,u_K^n)-f(u_K^n)\Big)^2\Big]. \end{aligned}$$

Then, using the CFL condition (25.3), Definition 25.1 and part (iv) of Assumption 24.1 gives

$$\begin{aligned} & \mathrm{m}(K)(u_K^{n+1}-u_K^n)^2 \leq \\ & k\frac{1-\xi}{g_1+g_2}\sum_{L\in\mathcal{N}(K)}\Big[v_{K,L}^n\Big(g(u_K^n,u_L^n)-f(u_K^n)\Big)^2 + v_{L,K}^n\Big(g(u_L^n,u_K^n)-f(u_K^n)\Big)^2\Big]. \end{aligned} \tag{26.16}$$

Summing equation (26.16) over $K\in\mathcal{T}_R$ and over $n=0,\ldots,N_{T,k}$, and reordering the summation leads to

$$\begin{aligned} \frac{1}{2}\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}_R}\mathrm{m}(K)(u_K^{n+1}-u_K^n)^2 \leq \frac{1-\xi}{2(g_1+g_2)}\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n}\Big[\\ v_{K,L}^n\Big((g(u_K^n,u_L^n)-f(u_K^n))^2 + (g(u_K^n,u_L^n)-f(u_L^n))^2\Big) + \\ v_{L,K}^n\Big((f(u_K^n)-g(u_L^n,u_K^n))^2 + (f(u_L^n)-g(u_L^n,u_K^n))^2\Big)\Big] + C_7, \end{aligned} \tag{26.17}$$

where C_7 accounts for the interfaces $K|L \subset B(0,R)$ such that $K\notin\mathcal{T}_R$ and/or $L\notin\mathcal{T}_R$ (these control volumes are included in $B(0,R+h)\setminus B(0,R-h)$).

Note that the right hand side of (26.17) is bounded by $(1-\xi)B_4 + C_7$ (from (26.13)). Using (26.6), (26.14) and (26.15) gives

$$\begin{aligned}\frac{\xi}{2(g_1+g_2)}\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n}\Big[\quad & v_{K,L}^n\Big(\max_{u_L^n\le p\le q\le u_K^n}(g(q,p)-f(q))^2 + \max_{u_L^n\le p\le q\le u_K^n}(g(q,p)-f(p))^2\Big)+\\ & v_{L,K}^n\Big(\max_{u_L^n\le p\le q\le u_K^n}(f(q)-g(p,q))^2 + \max_{u_L^n\le p\le q\le u_K^n}(f(p)-g(p,q))^2\Big)\Big]\\ \le\quad & \frac{1}{2}\sum_{K\in\mathcal{T}_R}\mathrm{m}(K)\Big(u_K^0\Big)^2 + C_6 + C_7 = C_8.\end{aligned} \tag{26.18}$$

Applying the Cauchy-Schwarz inequality to the left hand side of (26.4) and using (26.18) yields

$$\begin{aligned}\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n}\Big[\quad & v_{K,L}^n\Big(\max_{u_L^n\le p\le q\le u_K^n}(g(q,p)-f(q)) + \max_{u_L^n\le p\le q\le u_K^n}(g(q,p)-f(p))\Big)+\\ & v_{L,K}^n\Big(\max_{u_L^n\le p\le q\le u_K^n}(f(q)-g(p,q)) + \max_{u_L^n\le p\le q\le u_K^n}(f(p)-g(p,q))\Big)\Big]\\ \le\quad & C_9\Big(\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^n}(v_{K,L}^n + v_{L,K}^n)\Big)^{\frac{1}{2}}.\end{aligned} \tag{26.19}$$

Noting that

$$\sum_{(K,L)\in\mathcal{E}_R^n}(v_{K,L}^n + v_{L,K}^n) \le \sum_{K\in\mathcal{T}_{R+h}} V\mathrm{m}(\partial K) \le V\frac{1}{\alpha}h^{d-1}\frac{\mathrm{m}(B(0,R+h))}{\alpha h^d} = \frac{C_{10}}{h}$$

and $(N_{T,k}+1)k \le 2T$, one obtains (26.4) from (26.19).

Finally, since (26.3) yields

$$\mathrm{m}(K)|u_K^{n+1}-u_K^n| \le k\sum_{L\in\mathcal{N}(K)}\Big(v_{K,L}^n|g(u_K^n,u_L^n)-f(u_K^n)| + v_{L,K}^n|g(u_L^n,u_K^n)-f(u_K^n)|\Big),$$

Inequality (26.5) immediately follows from (26.4). This completes the proof of Lemma 26.2. ∎

27 Existence of the solution and stability results for the implicit scheme

This section is devoted to the time implicit scheme (given by (25.6) and (25.2)). We first prove the existence and uniqueness of the solution $\{u_K^n, n\in\mathbb{N}, K\in\mathcal{T}\}$ of (25.2), (25.6) and such that $u_K^n \in [U_m, U_M]$ for all $K\in\mathcal{T}$ and all $n\in\mathbb{N}$. Then, one gives a "weak space BV" inequality (this is equivalent to the inequality (26.4) for the explicit scheme) and a "(strong) time BV" estimate (Estimate (27.14) below). This last estimate requires that $\mathbf{v}$ does not depend on t (and it leads to the term "k" in the right hand side of (30.2) in Theorem 30.2). The error estimate, in the case where $\mathbf{v}$ depends on t, is given in Remark 30.2.

27.1 Existence, uniqueness and L^∞ stability

The following proposition gives an existence and uniqueness result of the solution to (25.2), (25.6). In this proposition, $\mathbf{v}$ may depend on t and one does not need to assume $u_0 \in BV(\mathbb{R}^d)$.

Proposition 27.1 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1.*
Then there exists a unique solution $\{u_K^n,\ n \in \mathbb{N},\ K \in \mathcal{T}\} \subset [U_m, U_M]$ to (25.2),(25.6).

Proof of Proposition 27.1

One proves Proposition 27.1 by induction. Indeed, $\{u_K^0,\ K \in \mathcal{T}\}$ is uniquely defined by (25.2) and one has $u_K^0 \in [U_m, U_M]$, for all $K \in \mathcal{T}$, since $U_m \le u_0 \le U_M$ a.e.. Assuming that, for some $n \in \mathbb{N}$, the set $\{u_K^n,\ K \in \mathcal{T}\}$ is given and that $u_K^n \in [U_m, U_M]$, for all $K \in \mathcal{T}$, the existence and uniqueness of $\{u_K^{n+1},\ K \in \mathcal{T}\}$, such that $u_K^{n+1} \in [U_m, U_M]$ for all $K \in \mathcal{T}$, solution of (25.6), must be shown.

Step 1 (uniqueness of $\{u_K^{n+1},\ K \in \mathcal{T}\}$, such that $u_K^{n+1} \in [U_m, U_M]$ for all $K \in \mathcal{T}$, solution of (25.6))
Recall that $n \in \mathbb{N}$ and $\{u_K^n,\ K \in \mathcal{T}\}$ are given. Let us consider two solutions of (25.6), respectively denoted by $\{u_K,\ K \in \mathcal{T}\}$ and $\{w_K,\ K \in \mathcal{T}\}$; therefore, $\{u_K,\ K \in \mathcal{T}\}$ and $\{w_K,\ K \in \mathcal{T}\}$ satisfy $\{u_K,\ K \in \mathcal{T}\} \subset [U_m, U_M]$, $\{w_K,\ K \in \mathcal{T}\} \subset [U_m, U_M]$,

$$\mathrm{m}(K)\frac{u_K - u_K^n}{k} + \sum_{L \in \mathcal{N}(K)} (v_{K,L}^n\ g(u_K, u_L) - v_{L,K}^n\ g(u_L, u_K)) = 0,\ \forall K \in \mathcal{T}, \tag{27.1}$$

and

$$\mathrm{m}(K)\frac{w_K - u_K^n}{k} + \sum_{L \in \mathcal{N}(K)} (v_{K,L}^n\ g(w_K, w_L) - v_{L,K}^n\ g(w_L, w_K)) = 0,\ \forall K \in \mathcal{T}. \tag{27.2}$$

Then, substracting (27.2) to (27.1), for all $K \in \mathcal{T}$,

$$\begin{aligned}
&\frac{\mathrm{m}(K)}{k}(u_K - w_K) + \sum_{L \in \mathcal{N}(K)} v_{K,L}^n (g(u_K, u_L) - g(w_K, u_L)) \\
&+ \sum_{L \in \mathcal{N}(K)} v_{K,L}^n (g(w_K, u_L) - g(w_K, w_L)) - \sum_{L \in \mathcal{N}(K)} v_{L,K}^n (g(u_L, u_K) - g(w_L, u_K)) \\
&- \sum_{L \in \mathcal{N}(K)} v_{L,K}^n (g(w_L, u_K) - g(w_L, w_K)) = 0
\end{aligned} \tag{27.3}$$

thanks to the monotonicity properties of g, (27.3) leads to

$$\begin{aligned}
&\frac{\mathrm{m}(K)}{k}|u_K - w_K| + \sum_{L \in \mathcal{N}(K)} v_{K,L}^n |g(u_K, u_L) - g(w_K, u_L)| \\
&+ \sum_{L \in \mathcal{N}(K)} v_{L,K}^n |g(w_L, u_K) - g(w_L, w_K)| \le \sum_{L \in \mathcal{N}(K)} v_{K,L}^n |g(w_K, u_L) - g(w_K, w_L)| \\
&+ \sum_{L \in \mathcal{N}(K)} v_{L,K}^n |g(u_L, u_K) - g(w_L, u_K)|.
\end{aligned} \tag{27.4}$$

Let $\varphi : \mathbb{R}^d \mapsto \mathbb{R}_+^\star$ be defined by $\varphi(x) = \exp(-\gamma|x|)$, for some positive γ which will be specified later. For $K \in \mathcal{T}$, let φ_K be the mean value of φ on K. Since φ is integrable over $\mathbb{R}^d$ (and thanks to (25.1)), one has $\sum_{K \in \mathcal{T}} \varphi_K \le (1/(\alpha h^d))\|\varphi\|_{L^1(\mathbb{R}^d)} < \infty$. Therefore the series

$$\sum_{K \in \mathcal{T}} \varphi_K \Big(\sum_{L \in \mathcal{N}(K)} v_{K,L}^n |g(w_K, u_L) - g(w_K, w_L)| \Big) \text{ and } \sum_{K \in \mathcal{T}} \varphi_K \Big(\sum_{L \in \mathcal{N}(K)} v_{L,K}^n |g(u_L, u_K) - g(w_L, u_K)| \Big)$$

are convergent (thanks to (25.1) and the boundedness of $\mathbf{v}$ on $\mathbb{R}^d$ and g on $[U_m, U_M]^2$).
Multiplying (27.4) by φ_K and summing for $K \in \mathcal{T}$ yields five convergent series which can be reordered in order to give

$$\sum_{K\in\mathcal{T}} \frac{\mathrm{m}(K)}{k}|u_K - w_K|\varphi_K \le \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} v^n_{K,L}|g(u_K,u_L) - g(w_K,u_L)||\varphi_K - \varphi_L|$$
$$+ \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} v^n_{L,K}|g(w_L,u_K) - g(w_L,w_K)||\varphi_K - \varphi_L|,$$

from which one deduces

$$\sum_{K\in\mathcal{T}} a_K|u_K - w_K| \le \sum_{K\in\mathcal{T}} b_K|u_K - w_K|, \tag{27.5}$$

with, for all $K \in \mathcal{T}$, $a_K = \frac{\mathrm{m}(K)}{k}\varphi_K$ and $b_K = \sum_{L\in\mathcal{N}(K)} (v^n_{K,L}g_1 + v^n_{L,K}g_2)|\varphi_K - \varphi_L|$.
For $K \in \mathcal{T}$, let x_K be an arbitrary point of K. Then,

$$a_K \ge \frac{1}{k}\alpha h^d \inf\{\varphi(x),\, x \in B(x_K,h)\}$$

and

$$b_K \le \frac{2V(g_1+g_2)}{\alpha} h^d \sup\{|\nabla\varphi(x)|,\, x \in B(x_K,2h)\}.$$

Therefore, taking $\gamma > 0$ small enough in order to have

$$\inf\{\varphi(y),\, y \in B(x,h)\} > C \sup\{|\nabla\varphi(y)|,\, y \in B(x,2h)\},\ \forall x \in \mathbb{R}^d \tag{27.6}$$

with $C = (2kV(g_1+g_2))/\alpha^2$, yields $a_K > b_K$ for all $K \in \mathcal{T}$. Hence (27.5) gives $u_K = w_K$, for all $K \in \mathcal{T}$. A choice of $\gamma > 0$ verifying (27.6) is always possible. Indeed, since $|\nabla\varphi(z)| = \gamma\exp(-\gamma|z|)$, taking $\gamma > 0$ such that $\gamma\exp(3\gamma h) < 1/C$ is convenient.
This concludes Step 1.

Step 2 (existence of $\{u^{n+1}_K,\, K \in \mathcal{T}\}$, such that $u^{n+1}_K \in [U_m, U_M]$ for all $K \in \mathcal{T}$, solution of (25.6)).
Recall that $n \in \mathbb{N}$ and $\{u^n_K,\, K \in \mathcal{T}\}$ are given. For $r \in \mathbb{N}^\star$, let $B_r = B(0,r) = \{x \in \mathbb{R}^d,\, |x| < r\}$ and $\mathcal{T}_r = \{K \in \mathcal{T},\, K \subset B_r\}$ (as in Lemma 26.2). Let us assume that r is large enough, say $r \ge r_0$, in order to have $\mathcal{T}_r \neq \emptyset$.
If $K \in \mathcal{T} \setminus \mathcal{T}_r$, set $u^{(r)}_K = u^n_K$. Let us first prove that there exists $\{u^{(r)}_K,\, K \in \mathcal{T}_r\} \subset [U_m, U_M]$, solution to

$$\mathrm{m}(K)\frac{u^{(r)}_K - u^n_K}{k} + \sum_{L\in\mathcal{N}(K)} (v^n_{K,L}\ g(u^{(r)}_K, u^{(r)}_L) - v^n_{L,K}\ g(u^{(r)}_L, u^{(r)}_K)) = 0,\ \forall K \in \mathcal{T}_r. \tag{27.7}$$

Then, we will prove that passing to the limit as $r \to \infty$ (up to a subsequence) leads to a solution $\{u^{n+1}_K,\, K \in \mathcal{T}\}$ to (25.6) such that $u^{n+1}_K \in [U_m, U_M]$ for all $K \in \mathcal{T}$.

For a fixed $r \ge r_0$, in order to prove the existence of $\{u^{(r)}_K,\, K \in \mathcal{T}_r\} \subset [U_m, U_M]$ solution to (27.7), a "topological degree" argument is used (see, for instance, Deimling [45] for a presentation of the degree). Let $U^n_r = \{u^n_K,\, K \in \mathcal{T}_r\}$ and assume that $U_r = \{u^{(r)}_K,\, K \in \mathcal{T}_r\}$ is a solution of (27.7). The families U_r and U^n_r may be viewed as vectors of $\mathbb{R}^N$, with $N = \mathrm{card}(\mathcal{T}_r)$. Equation (27.7) gives

$$u^{(r)}_K + \frac{k}{\mathrm{m}(K)} \sum_{L\in\mathcal{N}(K)} (v^n_{K,L}\ g(u^{(r)}_K, u^{(r)}_L) - v^n_{L,K}\ g(u^{(r)}_L, u^{(r)}_K)) = u^n_K,\ \forall K \in \mathcal{T}_r,$$

which can be written on the form

$$U_r - G_r(U_r) = U^n_r, \tag{27.8}$$

where G_r is a continuous map from $\mathbb{R}^N$ into $\mathbb{R}^N$.

One may assume that g is nondecreasing with respect to its first argument and nonincreasing with respect to its second argument on $\mathbb{R}^2$ (indeed, thanks to the monotonicity properties of g given by Assumption 25.1, it is sufficient to change, if necessary, g on $\mathbb{R}^2 \setminus [U_m, U_M]^2$, setting, for instance, $g(a,b) = g(U_m \top (U_M \perp a), U_m \top (U_M \perp b)))$. Then, since $u_K^n \in [U_m, U_M]$, for all $K \in \mathcal{T}$, and $u_K^{(r)} = u_K^n \in [U_m, U_M]$, for all $K \in \mathcal{T} \setminus \mathcal{T}_r$, it is easy to show (using $\mathrm{div}(\mathbf{v}) = 0$) that if U_r satisfies (27.8), then one has $u_K^{(r)} \in [U_m, U_M]$, for all $K \in \mathcal{T}_r$. Therefore, if $\mathcal{C}_r$ is a ball of $\mathbb{R}^N$ of center 0 and of radius great enough, Equation (27.8) has no solution on the boundary of $\mathcal{C}_r$, and one can define the topological degree of the application $Id - G_r$ associated to the set $\mathcal{C}_r$ and to the point U_r^n, that is $deg(Id - G_r, \mathcal{C}_r, U_r^n)$. Furthermore, if $\lambda \in [0,1]$, the same argument allows us to define $deg(Id - \lambda G_r, \mathcal{C}_r, U_r^n)$. Then, the property of invariance of the degree by continuous transformation asserts that $deg(Id - \lambda G_r, \mathcal{C}_r, U_r^n)$ does not depend on $\lambda \in [0,1]$. This gives

$$deg(Id - G_r, \mathcal{C}_r, U_r^n) = deg(Id, \mathcal{C}_r, U_r^n).$$

But, since $U_r^n \in \mathcal{C}_r$,

$$deg(Id, \mathcal{C}_r, U_r^n) = 1.$$

Hence

$$deg(Id - G_r, \mathcal{C}_r, U_r^n) \neq 0.$$

This proves that there exists a solution $U_r \in \mathcal{C}_r$ to (27.8). Recall also that we already proved that the components of U_r are necessarily in $[U_m, U_M]$.

In order to prove the existence of $\{u_K^{n+1},\ K \in \mathcal{T}\} \subset [U_m, U_M]$ solution to (25.6), let us pass to the limit as $r \to \infty$. For $r \geq r_0$, let $\{u_K^{(r)},\ K \in \mathcal{T}\}$ be a solution of (27.7) (given by the previous proof). Since $\{u_K^{(r)},\ r \in \mathbb{N}\}$ is included in $[U_m, U_M]$, for all $K \in \mathcal{T}$, one can find (using a "diagonal process") a sequence $(r_l)_{l \in \mathbb{N}}$, with $r_l \to \infty$, as $l \to \infty$, such that $(u_K^{r_l})_{l \in \mathbb{N}}$ converges (in $[U_m, U_M]$) for all $K \in \mathcal{T}$. One sets $u_K^{n+1} = \lim_{l \to \infty} u_K^{r_l}$. Passing to the limit in (27.7) (this is possible since for all $K \in \mathcal{T}$, this equation is satisfied for all $l \in \mathbb{N}$ large enough) shows that $\{u_K^{n+1},\ K \in \mathcal{T}\}$ is solution to (25.6).
Indeed, using the uniqueness of the solution of (25.6), one can show that $u_K^{(r)} \to u_K^{n+1}$, as $r \to \infty$, for all $K \in \mathcal{T}$.
This completes the proof of Proposition 27.1. ∎

27.2 "Weak space BV" inequality

One gives here an inequality similar to Inequality (26.4) (proved for the explicit scheme). This inequality does not make use of $u_0 \in BV(\mathbb{R}^d)$ and $\mathbf{v}$ can depend on t. Inequality (26.5) also holds but is improved in Lemma 27.3 when $u_0 \in BV(\mathbb{R}^d)$ and $\mathbf{v}$ does not depend on t.

Lemma 27.1 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1 and let $\{u_K^n, n \in \mathbb{N}, K \in \mathcal{T}\}$ be the solution of (25.6), (25.2) such that $u_K^{n+1} \in [U_m, U_M]$ for all $K \in \mathcal{T}$ and all $n \in \mathbb{N}$ (existence and uniqueness of such a solution is given by Proposition 27.1).*
Let $T > 0$, $R > 0$, $N_{T,k} = \max\{n \in \mathbb{N}, n < T/k\}$, $\mathcal{T}_R = \{K \in \mathcal{T}, K \subset B(0,R)\}$ and $\mathcal{E}_R^n = \{(K,L) \in \mathcal{T}^2, L \in \mathcal{N}(K), K|L \subset B(0,R)$ and $u_K^n > u_L^n\}$.
Then there exists $C_v \in \mathbb{R}$, only depending on $\mathbf{v}$, g, u_0, α, R, T such that, for $h < R$ and $k < T$,

$$\sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_R^{n+1}} \Big[v_{K,L}^n \Big(\max_{u_L^{n+1}\le p\le q\le u_K^{n+1}} (g(q,p)-f(q)) + \max_{u_L^{n+1}\le p\le q\le u_K^{n+1}} (g(q,p)-f(p)) \Big) + v_{L,K}^n \Big(\max_{u_L^{n+1}\le p\le q\le u_K^{n+1}} (f(q)-g(p,q)) + \max_{u_L^{n+1}\le p\le q\le u_K^{n+1}} (f(p)-g(p,q)) \Big) \Big] \le \frac{C_v}{\sqrt{h}}. \tag{27.9}$$

Furthermore, Inequality 26.5 page 161 holds.

Proof of Lemma 27.1

We multiply (25.6) by ku_K^{n+1}, and sum the result over $K \in \mathcal{T}_R$ and $n \in \{0,\dots,N_{T,k}\}$. We can then follow, step by step, the proof of Lemma 26.2 page 161 until Equation (26.15) in which the first term of the right hand side appears with the opposite sign. We can then directly conclude an inequality similar to (26.18), which is sufficient to conclude the proof of Inequality (27.9). Inequality 26.5 page 161 follows easily from (27.9). ∎

27.3 "Time BV" estimate

This section gives a so called "strong time BV estimate" (estimate (27.14)). For this estimate, the fact that $u_0 \in BV(\mathbb{R}^d)$ and that $\mathbf{v}$ does not depend on t is required. Let us begin this section with a preliminary lemma on the space $BV(\mathbb{R}^d)$.

Lemma 27.2 *Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and let $u \in BV(\mathbb{R}^d)$ (see Definition 21.19 page 141). For $K \in \mathcal{T}$, let u_K be the mean value of u over K. Then,*

$$\sum_{K|L\in\mathcal{E}} \mathrm{m}(K|L)|u_K - u_L| \le \frac{C}{\alpha^4}|u|_{BV(\mathbb{R}^d)}, \tag{27.10}$$

where C only depends on the space dimension ($d = 1$, 2 or 3).

Proof of Lemma 27.2

Lemma 27.2 is proven in two steps. In the first step, it is proved that if (27.10) holds for all $u \in BV(\mathbb{R}^d) \cap C^1(\mathbb{R}^d,\mathbb{R})$ then (27.10) holds for all $u \in BV(\mathbb{R}^d)$. In Step 2, (27.10) is proved to hold for $u \in BV(\mathbb{R}^d) \cap C^1(\mathbb{R}^d,\mathbb{R})$.

Step 1 (passing from $BV(\mathbb{R}^d) \cap C^1(\mathbb{R}^d,\mathbb{R})$ to $BV(\mathbb{R}^d)$)
Recall that $BV(\mathbb{R}^d) \subset L^1_{loc}(\mathbb{R}^d)$. Let $u \in BV(\mathbb{R}^d)$, let us regularize u by a sequence of mollifiers.
Let $\rho \in C_c^\infty(\mathbb{R}^d,\mathbb{R}_+)$ such that $\int_{\mathbb{R}^d}\rho(x)dx = 1$. Define, for all $n \in \mathbb{N}^\star$, ρ_n by $\rho_n(x) = n^d\rho(nx)$ for all $x \in \mathbb{R}^d$ and $u_n = u \star \rho_n$, that is

$$u_n(x) = \int_{\mathbb{R}^d} u(y)\rho_n(x-y)dy,\ \forall x \in \mathbb{R}^d.$$

It is well known that $(u_n)_{n\in\mathbb{N}^\star}$ is included in $C^\infty(\mathbb{R}^d,\mathbb{R})$ and converges to u in $L^1_{loc}(\mathbb{R}^d)$ as $n \to \infty$. Then, the mean value of u_n over K converges, as $n \to \infty$, to u_K, for all $K \in \mathcal{T}$. Hence, if (27.10) holds with u_n instead of u (this will be proven in Step 2) and if $|u_n|_{BV(\mathbb{R}^d)} \le |u|_{BV(\mathbb{R}^d)}$ for all $n \in \mathbb{N}^\star$, Inequality (27.10) is proved by passing to the limit as $n \to \infty$.

In order to prove $|u_n|_{BV(\mathbb{R}^d)} \le |u|_{BV(\mathbb{R}^d)}$ for all $n \in \mathbb{N}^\star$ (this will conclude step 1), let $n \in \mathbb{N}^\star$ and $\varphi \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$ such that $|\varphi(x)| \le 1$ for all $x \in \mathbb{R}^d$. A simple computation gives, using Fubini's theorem,

$$\int_{\mathbb{R}^d} u_n(x)\mathrm{div}\varphi(x)dx = \int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} u(x-y)\mathrm{div}\varphi(x)dx\Big)\rho_n(y)dy \le |u|_{BV(\mathbb{R}^d)}, \tag{27.11}$$

since, setting $\psi_y = \varphi(y+\cdot) \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$ (for all $y \in \mathbb{R}^d$),

$$\int_{\mathbb{R}^d} u(x-y)\mathrm{div}\varphi(x)dx = \int_{\mathbb{R}^d} u(z)\mathrm{div}\psi_y(z)dz \le |u|_{BV(\mathbb{R}^d)},\ \forall y \in \mathbb{R}^d,$$

and

$$\int_{\mathbb{R}^d} \rho_n(y)dy = 1.$$

Then, taking in (27.11) the supremum over $\varphi \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$ such that $|\varphi(x)| \le 1$ for all $x \in \mathbb{R}^d$ leads to $|u_n|_{BV(\mathbb{R}^d)} \le |u|_{BV(\mathbb{R}^d)}$.

Step 2 (proving (27.10) if $u \in BV(\mathbb{R}^d) \cap C^1(\mathbb{R}^d,\mathbb{R})$)
Recall that $B(x,R)$ denotes the ball of $\mathbb{R}^d$ of center x and radius R. Since $u \in C^1(\mathbb{R}^d,\mathbb{R})$,

$$\int_{\mathbb{R}^d} u(x)\mathrm{div}\varphi(x)dx = -\int_{\mathbb{R}^d} \nabla u(x)\cdot\varphi(x)dx.$$

Then $|u|_{BV(\mathbb{R}^d)} = \|(|\nabla u|)\|_{L^1(\mathbb{R}^d)}$ and we will prove (27.10) with $\|(|\nabla u|)\|_{L^1(\mathbb{R}^d)}$ instead of $|u|_{BV(\mathbb{R}^d)}$.
Let $K|L \in \mathcal{E}$, then $K \in \mathcal{T}$, $L \in \mathcal{N}(K)$ and

$$u_K - u_L = \frac{1}{\mathrm{m}(K)\mathrm{m}(L)}\int_L\int_K (u(x)-u(y))dxdy.$$

For all $x \in K$ and all $y \in L$,

$$u(x)-u(y) = \int_0^1 \nabla u(y+t(x-y))\cdot(x-y)dt.$$

Then,

$$\begin{aligned}\mathrm{m}(K)\mathrm{m}(L)|u_K-u_L| &\le \int_L\Big(\int_K\int_0^1 |\nabla u(y+t(x-y))||x-y|dtdx\Big)dy\\ &\le \int_L\Big(\int_0^1\int_K |\nabla u(y+t(x-y))||x-y|dxdt\Big)dy.\end{aligned}$$

Using $|x-y| \le 2h$ and changing the variable x in $z = x-y$ (for all fixed $y \in L$ and $t \in (0,1)$) yields

$$\mathrm{m}(K)\mathrm{m}(L)|u_K-u_L| \le 2h\int_L\Big(\int_0^1\int_{B(0,2h)} |\nabla u(y+tz)|dzdt\Big)dy,$$

which may also be written (using Fubini's theorem)

$$\mathrm{m}(K)\mathrm{m}(L)|u_K-u_L| \le 2h\int_{B(0,2h)}\Big(\int_0^1\int_L |\nabla u(y+tz)|dydt\Big)dz. \tag{27.12}$$

For all $K \in \mathcal{T}$, let x_K be an arbitrary point of K.
Then, changing the variable y in $\xi = y+tz$ (for all fixed $z \in L$ and $t \in (0,1)$) in (27.12),

$$\mathrm{m}(K)\mathrm{m}(L)|u_K-u_L| \le 2h\int_{B(0,2h)}\Big(\int_0^1\int_{B(x_L,3h)} |\nabla u(\xi)|d\xi dt\Big)dz,$$

which yields, since $\mathcal{T}$ is an adimissible mesh in the sense of Definition 25.1 page 156,

$$\mathrm{m}(K|L)|u_K - u_L| \leq \frac{2h^d}{\alpha^3 h^{2d}} \mathrm{m}(B(0,2h)) \int_{B(x_L,3h)} |\nabla u(\xi)| d\xi.$$

Therefore there exists C_1, only depending on the space dimension, such that

$$\mathrm{m}(K|L)|u_K - u_L| \leq \frac{C_1}{\alpha^3} \int_{B(x_L,3h)} |\nabla u(\xi)| d\xi, \; \forall K|L \in \mathcal{E}. \tag{27.13}$$

Let us now remark that, if $M \in \mathcal{T}$ and $L \in \mathcal{T}$, $M \cap B(x_L, 3h) \neq \emptyset$ implies $L \subset B(x_M, 5h)$. Then, for a fixed $M \in \mathcal{T}$, the number of $L \in \mathcal{T}$ such that $M \cap B(x_L, 3h) \neq \emptyset$ is less or equal to $\mathrm{m}(B(0,5h))/(\alpha h^d)$ that is less or equal C_2/α where C_2 only depends on the space dimension.
Then, summing (27.13) over $K|L \in \mathcal{E}$ leads to

$$\sum_{K|L \in \mathcal{E}} \mathrm{m}(K|L)|u_K - u_L| \leq \frac{C_1 C_2}{\alpha^4} \sum_{M \in \mathcal{T}} \int_M |\nabla u(\xi)| d\xi = \frac{C_1 C_2}{\alpha^4} \| (|\nabla u|) \|_{L^1(\mathbb{R}^d)},$$

that is (27.10) with $C = C_1 C_2$. ∎

Note that, in Lemma 27.2 the estimate (27.10) depends on α. This dependency on α is not necessary in the one dimensinal case (see (19.6) in Remark 19.4) and for particular meshes in the two and three dimensional cases. Recall also that, except if $d = 1$, the space $BV(\mathbb{R}^d)$ is not included in $L^\infty(\mathbb{R}^d)$. In particular, it is then quite easy to prove that, contrary to the 1D case given in Remark 19.4, it is not possible, for $d = 2$ or 3, to replace, in (27.10), u_K by the mean value of u over an arbitrary ball (for instance) included in K.

Let us now give the "strong time BV estimate".

Lemma 27.3 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1. Assume that $u_0 \in BV(\mathbb{R}^d)$ and that $\mathbf{v}$ does not depend on t.*
Let $\{u_K^n, n \in \mathbb{N}, K \in \mathcal{T}\}$ be the solution of (25.6), (25.2) such that $u_K^n \in [U_m, U_M]$ for all $K \in \mathcal{T}$ and all $n \in \mathbb{N}$ (existence and uniqueness of such a solution is given by Proposition 27.1 page 165).
Then, there exists C_b, only depending on $\mathbf{v}$, g, u_0 and α such that

$$\sum_{K \in \mathcal{T}} \frac{\mathrm{m}(K)}{k} |u_K^{n+1} - u_K^n| \leq C_b, \; \forall n \in \mathbb{N}. \tag{27.14}$$

Proof of lemma 27.3

Since $\mathbf{v}$ does not depend on t, one denotes $v_{K,L} = v_{K,L}^n$, for all $K \in \mathcal{T}$ and all $L \in \mathcal{N}(K)$.
For $n \in \mathbb{N}$, let

$$A_n = \sum_{K \in \mathcal{T}} \mathrm{m}(K) \frac{|u_K^{n+1} - u_K^n|}{k}$$

and

$$B_n = \sum_{K \in \mathcal{T}} | \sum_{L \in \mathcal{N}(K)} [v_{K,L} \; g(u_K^n, u_L^n) - v_{L,K} \; g(u_L^n, u_K^n)]|.$$

Since $u_0 \in BV(\mathbb{R}^d)$ and $\mathrm{div}\mathbf{v} = 0$, there exists $C_b > 0$, only depending on $\mathbf{v}$, g, u_0 and α, such that $B_0 \leq C_b$. Indeed,

$$B_0 \leq \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} V(g_1 + g_2) \mathrm{m}(K|L)|u_K^0 - u_L^0|.$$

Thanks to lemma 27.2, $B_0 \leq C_b$ with $C_b = 2V(g_1+g_2)C(1/\alpha^4)|u_0|_{BV(\mathbb{R}^d)}$, where C only depends on the space dimension ($d = 1$, 2 or 3).

From (25.6), one deduces that $B_{n+1} \leq A_n$, for all $n \in \mathbb{N}$. In order to prove Lemma 27.3, there only remains to prove that $A_n \leq B_n$ for all $n \in \mathbb{N}$ (and to conclude by induction).

Let $n \in \mathbb{N}$, in order to prove that $A_n \leq B_n$, recall that the implicit scheme (25.6) reads

$$\mathrm{m}(K)\frac{u_K^{n+1}-u_K^n}{k} + \sum_{L\in\mathcal{N}(K)} \Big(v_{K,L}\ g(u_K^{n+1},u_L^{n+1}) - v_{L,K}\ g(u_L^{n+1},u_K^{n+1}) \Big) = 0. \tag{27.15}$$

From (27.15), one deduces, for all $K \in \mathcal{T}$,

$$\begin{aligned}
&\mathrm{m}(K)\frac{u_K^{n+1}-u_K^n}{k} + \sum_{L\in\mathcal{N}(K)} v_{K,L}\ (g(u_K^{n+1},u_L^{n+1}) - g(u_K^n,u_L^{n+1})) \\
&+ \sum_{L\in\mathcal{N}(K)} v_{K,L}\ \Big(g(u_K^n,u_L^{n+1}) - g(u_K^n,u_L^n) \Big) - \sum_{L\in\mathcal{N}(K)} v_{L,K}\ \Big(g(u_L^{n+1},u_K^{n+1}) - g(u_L^n,u_K^{n+1}) \Big) \\
&- \sum_{L\in\mathcal{N}(K)} v_{L,K}\ \Big(g(u_L^n,u_K^{n+1}) - g(u_L^n,u_K^n) \Big) \\
&= - \sum_{L\in\mathcal{N}(K)} v_{K,L}\, g(u_K^n,u_L^n) + \sum_{L\in\mathcal{N}(K)} v_{L,K}\ g(u_L^n,u_K^n).
\end{aligned}$$

Using the monotonicity properties of g, one obtains for all $K \in \mathcal{T}$,

$$\begin{aligned}
&\mathrm{m}(K)\frac{|u_K^{n+1}-u_K^n|}{k} + \sum_{L\in\mathcal{N}(K)} v_{K,L}\ |g(u_K^{n+1},u_L^{n+1}) - g(u_K^n,u_L^{n+1})| \\
&+ \sum_{L\in\mathcal{N}(K)} v_{L,K}\ |g(u_L^n,u_K^{n+1}) - g(u_L^n,u_K^n)| \\
&\leq |- \sum_{L\in\mathcal{N}(K)} v_{K,L} g(u_K^n,u_L^n) + \sum_{L\in\mathcal{N}(K)} v_{L,K}\ g(u_L^n,u_K^n)| \\
&+ \sum_{L\in\mathcal{N}(K)} v_{K,L}\ |g(u_K^n,u_L^{n+1}) - g(u_K^n,u_L^n)| + \sum_{L\in\mathcal{N}(K)} v_{L,K}\ |g(u_L^{n+1},u_K^{n+1}) - g(u_L^n,u_K^{n+1})|.
\end{aligned} \tag{27.16}$$

In order to deal with convergent series, let us proceed as in the proof of proposition 27.1. For $0 < \gamma < 1$, let $\varphi_\gamma : \mathbb{R}^d \mapsto \mathbb{R}_+^\star$ be defined by $\varphi_\gamma(x) = \exp(-\gamma|x|)$.

For $K \in \mathcal{T}$, let $\varphi_{\gamma,K}$ be the mean value of φ_γ on K. As in Proposition 27.1, since φ_γ is integrable over $\mathbb{R}^d$, $\sum_{K\in\mathcal{T}} \varphi_{\gamma,K} < \infty$. Therefore, multiplying (27.16) by $\varphi_{\gamma,K}$ (for a fixed γ) and summing over $K \in \mathcal{T}$ yields six convergent series which can be reordered to give

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \mathrm{m}(K)\frac{|u_K^{n+1}-u_K^n|}{k}\varphi_{\gamma,K} \\
&\leq \sum_{K\in\mathcal{T}} |- \sum_{L\in\mathcal{N}(K)} v_{K,L} g(u_K^n,u_L^n) + \sum_{L\in\mathcal{N}(K)} v_{L,K}\ g(u_L^n,u_K^n)|\varphi_{\gamma,K} \\
&+ \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} v_{K,L}\ |g(u_K^{n+1},u_L^{n+1}) - g(u_K^n,u_L^{n+1})||\varphi_{\gamma,K} - \varphi_{\gamma,L}| \\
&+ \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} v_{L,K}\ |g(u_L^n,u_K^{n+1}) - g(u_L^n,u_K^n)||\varphi_{\gamma,K} - \varphi_{\gamma,L}|.
\end{aligned}$$

For $K \in \mathcal{T}$, let $x_K \in \overline{K}$ be such that $\varphi_{\gamma,K} = \varphi_\gamma(x_K)$. Let $K \in \mathcal{T}$ and $L \in \mathcal{N}(K)$. Then there exists $s \in (0,1)$ such that $\varphi_{\gamma,L} - \varphi_{\gamma,K} = \nabla\varphi_\gamma(x_K + s(x_L - x_K)) \cdot (x_L - x_K)$. Using $|\nabla\varphi_\gamma(x)| = \gamma\exp(-\gamma|x|)$, this yields $|\varphi_{\gamma,L} - \varphi_{\gamma,K}| \leq 2h\gamma\exp(2h\gamma)\varphi_{\gamma,K} \leq 2h\gamma\exp(2h)\varphi_{\gamma,K}$.

Then, using the assumptions 24.1 and 25.1, there exists some a only depending on k, V, h, α, g_1 and g_2 such that

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\frac{|u_K^{n+1}-u_K^n|}{k}\varphi_{\gamma,K}(1-\gamma a)$$
$$\leq \sum_{K\in\mathcal{T}} |-\sum_{L\in\mathcal{N}(K)} v_{K,L}g(u_K^n,u_L^n)+\sum_{L\in\mathcal{N}(K)} v_{L,K}\ g(u_L^n,u_K^n)|\varphi_{\gamma,K}\leq B_n.$$

Passing to the limit in the latter inequality as $\gamma\to 0$ yields $A_n\leq B_n$. This completes the proof of Lemma 27.3. ∎

28 Entropy inequalities for the approximate solution

In this section, an entropy estimate on the approximate solution is proved (Theorem 28.1), which will be used in the proofs of convergence and error estimate of the numerical scheme. In order to obtain this entropy estimate, some discrete entropy inequalities satisfied by the approximate solution are first derived.

28.1 Discrete entropy inequalities

In the case of the explicit scheme, the following lemma asserts that the scheme (25.4) satisfies a discrete entropy condition (this is classical in the study of 1D schemes, see e.g. Godlewski and Raviart [75], Godlewski and Raviart [76]).

Lemma 28.1 *Under assumption 24.1 page 153, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k>0$. Let $g\in C(\mathbb{R}^2,\mathbb{R})$ satisfying assumption 25.1 and assume that (25.3) holds.*
Let $u_{\mathcal{T},k}$ be given by (25.5), (25.4), (25.2); then, for all $\kappa\in\mathbb{R}$, $K\in\mathcal{T}$ and $n\in\mathbb{N}$, the following inequality holds:

$$\mathrm{m}(K)\frac{|u_K^{n+1}-\kappa|-|u_K^n-\kappa|}{k}+\sum_{L\in\mathcal{N}(K)}\Big[\ v_{K,L}^n\Big(g(u_K^n\top\kappa,u_L^n\top\kappa)-g(u_K^n\bot\kappa,u_L^n\bot\kappa)\Big)- \quad v_{L,K}^n\Big(g(u_L^n\top\kappa,u_K^n\top\kappa)-g(u_L^n\bot\kappa,u_K^n\bot\kappa)\Big)\Big]\ \leq 0. \tag{28.1}$$

Proof of lemma 28.1
From relation (25.4), we express u_K^{n+1} as a function of u_K^n and u_L^n, $L\in\mathcal{N}(K)$,

$$u_K^{n+1}=u_K^n+\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{L,K}^n\ g(u_L^n,u_K^n)-v_{K,L}^n\ g(u_K^n,u_L^n)).$$

The right hand side is nondecreasing with respect to u_L^n, $L\in\mathcal{N}(K)$. It is also nondecreasing with respect to u_K^n, thanks to the Courant-Friedrichs-Levy condition (25.3), and the Lipschitz continuity of g.
Therefore, for all $\kappa\in\mathbb{R}$, using $\mathrm{div}\mathbf{v}=0$, we have:

$$u_K^{n+1}\top\kappa\leq u_K^n\top\kappa+\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}\Big[v_{L,K}^n\ g(u_L^n\top\kappa,u_K^n\top\kappa)-v_{K,L}^n\ g(u_K^n\top\kappa,u_L^n\top\kappa)\Big] \tag{28.2}$$

and

$$u_K^{n+1}\bot\kappa\geq u_K^n\bot\kappa+\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{L,K}^n\ g(u_L^n\bot\kappa,u_K^n\bot\kappa)-v_{K,L}^n\ g(u_K^n\bot\kappa,u_L^n\bot\kappa)). \tag{28.3}$$

The difference between (28.2) and (28.3) leads directly to (28.1). Note that using $\mathrm{div}\mathbf{v} = 0$ leads to

$$\sum_{L\in\mathcal{N}(K)} \Big[\begin{array}{l} \mathrm{m}(K)\dfrac{|u_K^{n+1}-\kappa|-|u_K^n-\kappa|}{k}+ \\ v_{K,L}^n\Big(g(u_K^n\top\kappa,u_L^n\top\kappa)-f(u_K^n\top\kappa)-g(u_K^n\bot\kappa,u_L^n\bot\kappa)+f(u_K^n\bot\kappa)\Big)- \\ v_{L,K}^n\Big(g(u_L^n\top\kappa,u_K^n\top\kappa)-f(u_K^n\top\kappa)-g(u_L^n\bot\kappa,u_K^n\bot\kappa)+f(u_K^n\bot\kappa)\Big)\Big] \le 0. \end{array} \tag{28.4}$$

■

For the implicit scheme, one obtains the same kind of discrete entropy inequalities.

Lemma 28.2 *Under assumption 24.1 page 153, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and $k>0$. Let $g\in C(\mathbb{R}^2,\mathbb{R})$ satisfying assumption 25.1.*
Let $\{u_K^n,\ n\in\mathbb{N},\ K\in\mathcal{T}\}\subset[U_m,U_M]$ be the solution of (25.6),(25.2) (the existence and uniqueness of such a solution is given by Proposition 27.1). Then, for all $\kappa\in\mathbb{R}$, $K\in\mathcal{T}$ and $n\in\mathbb{N}$, the following inequality holds:

$$\begin{array}{l} \mathrm{m}(K)\dfrac{|u_K^{n+1}-\kappa|-|u_K^n-\kappa|}{k}+\displaystyle\sum_{L\in\mathcal{N}(K)}\Big[v_{K,L}^n\Big(g(u_K^{n+1}\top\kappa,u_L^{n+1}\top\kappa)-g(u_K^{n+1}\bot\kappa,u_L^{n+1}\bot\kappa)\Big) \\ \qquad -v_{L,K}^n\Big(g(u_L^{n+1}\top\kappa,u_K^{n+1}\top\kappa)-g(u_L^{n+1}\bot\kappa,u_K^{n+1}\bot\kappa)\Big)\Big]\le 0. \end{array} \tag{28.5}$$

Proof of lemma 28.2
Let $\kappa\in\mathbb{R}$, $K\in\mathcal{T}$ and $n\in\mathbb{N}$. Equation (25.6) may be written as

$$u_K^{n+1}=u_K^n-\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(u_K^{n+1},u_L^{n+1})-v_{L,K}^n\ g(u_L^{n+1},u_K^{n+1})).$$

The right hand side of this last equation is nondecreasing with respect to u_K^n and with respect to u_L^{n+1} for all $L\in\mathcal{N}(K)$. Thus,

$$u_K^{n+1}\le u_K^n\top\kappa-\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(u_K^{n+1},u_L^{n+1}\top\kappa)-v_{L,K}^n\ g(u_L^{n+1}\top\kappa,u_K^{n+1})).$$

Writing $\kappa=\kappa-\dfrac{k}{\mathrm{m}(K)}\displaystyle\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(\kappa,\kappa)-v_{L,K}^n\ g(\kappa,\kappa))$, one may remark that

$$\kappa\le u_K^n\top\kappa-\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(\kappa,u_L^{n+1}\top\kappa)-v_{L,K}^n\ g(u_L^{n+1}\top\kappa,\kappa)).$$

Therefore, since $u_K^{n+1}\top\kappa=u_K^{n+1}$ or κ,

$$u_K^{n+1}\top\kappa\le u_K^n\top\kappa-\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(u_K^{n+1}\top\kappa,u_L^{n+1}\top\kappa)-v_{L,K}^n\ g(u_L^{n+1}\top\kappa,u_K^{n+1}\top\kappa)). \tag{28.6}$$

A similar argument yields

$$u_K^{n+1}\bot\kappa\ge u_K^n\bot\kappa-\frac{k}{\mathrm{m}(K)}\sum_{L\in\mathcal{N}(K)}(v_{K,L}^n\ g(u_K^{n+1}\bot\kappa,u_L^{n+1}\bot\kappa)-v_{L,K}^n\ g(u_L^{n+1}\bot\kappa,u_K^{n+1}\bot\kappa)). \tag{28.7}$$

Hence, substracting (28.7) to (28.6) gives (28.5). ■

28.2 Continuous entropy estimates for the approximate solution

For $\Omega = \mathbb{R}^d$ or $\mathbb{R}^d \times \mathbb{R}_+$, we denote by $\mathcal{M}(\Omega)$ the set of positive measures on Ω, that is of σ-additive applications from the Borel σ-algebra of Ω in $\overline{\mathbb{R}}_+$. If $\mu \in \mathcal{M}(\Omega)$ and $\psi \in C_c(\Omega)$, one sets $\langle \mu, \psi \rangle = \int \psi d\mu$. The following theorems investigate the entropy inequalities which are satisfied by the approximate solutions $u_{\mathcal{T},k}$ in the case of the time explicit scheme (Theorem 28.1) and in the case of the time implicit scheme (Theorem 28.2).

Theorem 28.1 *Under assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy assumption 25.1 and assume that (25.3) holds.*
Let $u_{\mathcal{T},k}$ be given by (25.5), (25.4), (25.2); then there exist $\mu_{\mathcal{T},k} \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}_+)$ and $\mu_{\mathcal{T}} \in \mathcal{M}(\mathbb{R}^d)$ such that

$$\left\{ \begin{array}{lll} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big(& |u_{\mathcal{T},k}(x,t) - \kappa|\varphi_t(x,t) + & \\ & (f(u_{\mathcal{T},k}(x,t)\top\kappa) - f(u_{\mathcal{T},k}(x,t)\bot\kappa))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)dxdt & + \\ \displaystyle\int_{\mathbb{R}^d} & |u_0(x) - \kappa|\varphi(x,0)dx & \geq \\ \displaystyle -\int_{\mathbb{R}^d\times\mathbb{R}_+} & \Big(|\varphi_t(x,t)| + |\nabla\varphi(x,t)|\Big)d\mu_{\mathcal{T},k}(x,t) - \displaystyle\int_{\mathbb{R}^d}\varphi(x,0)d\mu_{\mathcal{T}}(x), & \\ & \forall\kappa\in\mathbb{R},\quad \forall\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+). & \end{array} \right. \tag{28.8}$$

The measures $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{T}}$ verify the following properties:

1. *For all $R > 0$ and $T > 0$, there exists C depending only on $\mathbf{v}$, g, u_0, α, ξ, R and T such that, for $h < R$ and $k < T$,*

$$\mu_{\mathcal{T},k}(B(0,R)\times[0,T]) \leq C\sqrt{h}. \tag{28.9}$$

2. *The measure $\mu_{\mathcal{T}}$ is the measure of density $|u_0(\cdot) - u_{\mathcal{T},0}(\cdot)|$ with respect to the Lebesgue measure, where $u_{\mathcal{T},0}$ is defined by $u_{\mathcal{T},0}(x) = u_K^0$ for a.e. $x \in K$, for all $K \in \mathcal{T}$.*

   If $u_0 \in BV(\mathbb{R}^d)$, then there exists D, only depending on u_0 and α, such that

$$\mu_{\mathcal{T}}(\mathbb{R}^d) \leq Dh. \tag{28.10}$$

Remark 28.1

1. Let u be the weak entropy solution to (24.1)-(24.2). Then (28.8) is satisfied with u instead of $u_{\mathcal{T},k}$ and $\mu_{\mathcal{T},k} = 0$ and $\mu_{\mathcal{T}} = 0$.

2. Let $BV_{loc}(\mathbb{R}^d)$ be the set of $v \in L^1_{loc}(\mathbb{R}^d)$ such that the restriction of v to Ω belongs to $BV(\Omega)$ for all open bounded subset Ω of $\mathbb{R}^d$.

   An easy adaptation of the following proof gives that if $u_0 \in BV_{loc}(\mathbb{R}^d)$ instead of $BV(\mathbb{R}^d)$ (in the second item of Theorem 28.1) then, for all $R > 0$, there exists D, only depending on u_0, α and R, such that $\mu_{\mathcal{T}}(B(0,R)) \leq Dh$.

Proof of Theorem 28.1
Let $\varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$ and $\kappa \in \mathbb{R}$.
Multiplying (28.4) by $k\varphi_K^n = (1/\mathrm{m}(K))\int_{nk}^{(n+1)k}\int_K \varphi(x,t)dxdt$ and summing the result for all $K \in \mathcal{T}$ and $n \in \mathbb{N}$ yields

$$T_1 + T_2 \leq 0,$$

with

$$T_1 = \sum_{n\in\mathbb{N}}\sum_{K\in\mathcal{T}} \frac{|u_K^{n+1}-\kappa| - |u_K^n-\kappa|}{k} \int_{nk}^{(n+1)k}\int_K \varphi(x,t)dxdt, \tag{28.11}$$

and

$$\begin{aligned} T_2 = k\sum_{n\in\mathbb{N}}\sum_{(K,L)\in\mathcal{E}_n} \Big[& \\ & v_{K,L}^n\varphi_K^n\Big(g(u_K^n\top\kappa, u_L^n\top\kappa) - f(u_K^n\top\kappa) - g(u_K^n\bot\kappa, u_L^n\bot\kappa) + f(u_K^n\bot\kappa)\Big) \\ & -v_{K,L}^n\varphi_L^n\Big(g(u_K^n\top\kappa, u_L^n\top\kappa) - f(u_L^n\top\kappa) - g(u_K^n\bot\kappa, u_L^n\bot\kappa) + f(u_L^n\bot\kappa)\Big) \\ & -v_{L,K}^n\varphi_K^n\Big(g(u_L^n\top\kappa, u_K^n\top\kappa) - f(u_K^n\top\kappa) - g(u_L^n\bot\kappa, u_K^n\bot\kappa) + f(u_K^n\bot\kappa)\Big) \\ & +v_{L,K}^n\varphi_L^n\Big(g(u_L^n\top\kappa, u_K^n\top\kappa) - f(u_L^n\top\kappa) - g(u_L^n\bot\kappa, u_K^n\bot\kappa) + f(u_L^n\bot\kappa)\Big)\Big], \end{aligned} \tag{28.12}$$

where $\mathcal{E}_n = \{(K,L)\in\mathcal{T}^2,\ u_K^n > u_L^n\}$.
One has to prove

$$T_{10} + T_{20} \le \int_{\mathbb{R}^d\times\mathbb{R}_+} \Big(|\varphi_t(x,t)| + |\nabla\varphi(x,t)|\Big) d\mu_{\mathcal{T},k}(x,t) + \int_{\mathbb{R}^d}\varphi(x,0)d\mu_{\mathcal{T}}(x), \tag{28.13}$$

for some convenient measures $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{T}}$, and T_{10}, T_{20} defined as follows

$$\begin{aligned} T_{10} &= -\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} |u_{\mathcal{T},k}(x,t)-\kappa|\varphi_t(x,t)dxdt - \int_{\mathbb{R}^d}|u_0(x)-\kappa|\varphi(x,0)dx, \\ T_{20} &= -\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big((f(u_{\mathcal{T},k}(x,t)\top\kappa) - f(u_{\mathcal{T},k}(x,t)\bot\kappa))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)dxdt. \end{aligned} \tag{28.14}$$

In order to prove (28.13), one compares T_1 and T_{10} (this will give $\mu_{\mathcal{T}}$, and a part of $\mu_{\mathcal{T},k}$) and one compares T_2 and T_{20} (this will give another part of $\mu_{\mathcal{T},k}$).
Inequality (26.5) (in the comparison of T_1 and T_{10}) and Inequality (26.4) (in the comparison of T_2 and T_{20}) will be used in order to obtain (28.9).

Comparison of T_1 and T_{10}

Using the definition of $u_{\mathcal{T},k}$ and introducing the function $u_{\mathcal{T},0}$ (defined by $u_{\mathcal{T},0}(x) = u_K^0$, for a.e. $x\in K$, for all $K\in\mathcal{T}$) yields

$$\begin{aligned} T_{10} = \sum_{n\in\mathbb{N}}\sum_{K\in\mathcal{T}} \frac{|u_K^{n+1}-\kappa| - |u_K^n-\kappa|}{k} \int_{nk}^{(n+1)k}\int_K \varphi(x,(n+1)k)dxdt \ + \\ \int_{\mathbb{R}^d}(|u_{\mathcal{T},0}(x)-\kappa| - |u_0(x)-\kappa|)\varphi(x,0)dx. \end{aligned}$$

The function $|\cdot - \kappa|$ is Lipschitz continuous with a Lipschitz constant equal to 1, we then obtain

$$\begin{aligned} |T_1 - T_{10}| \le \sum_{n\in\mathbb{N}}\sum_{K\in\mathcal{T}} \frac{|u_K^{n+1}-u_K^n|}{k} \int_{nk}^{(n+1)k}\int_K |\varphi(x,(n+1)k) - \varphi(x,t)|dxdt \ + \\ \int_{\mathbb{R}^d}|u_0(x) - u_{\mathcal{T},0}(x)|\varphi(x,0)dx, \end{aligned}$$

which leads to

$$\begin{aligned} |T_1 - T_{10}| \le \sum_{n\in\mathbb{N}}\sum_{K\in\mathcal{T}} |u_K^{n+1}-u_K^n| \int_{nk}^{(n+1)k}\int_K |\varphi_t(x,t)|dxdt \ + \\ \int_{\mathbb{R}^d}|u_0(x) - u_{\mathcal{T},0}(x)|\varphi(x,0)dx. \end{aligned} \tag{28.15}$$

Inequality (28.15) gives

$$|T_1 - T_{10}| \leq \int_{\mathbb{R}^d \times \mathbb{R}_+} |\varphi_t(x,t)| d\nu_{\mathcal{T},k}(x,t) + \int_{\mathbb{R}^d} \varphi(x,0) d\mu_{\mathcal{T}}(x), \tag{28.16}$$

where the measures $\mu_{\mathcal{T}} \in \mathcal{M}(\mathbb{R}^d)$ and $\nu_{\mathcal{T},k} \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}_+)$ are defined, by their action on $C_c(\mathbb{R}^d)$ and $C_c(\mathbb{R}^d \times \mathbb{R}_+)$, as follows

$$\langle \mu_{\mathcal{T}}, \psi \rangle = \int_{\mathbb{R}^d} |u_0(x) - u_{\mathcal{T},0}(x)| \psi(x) dx, \ \forall \psi \in C_c(\mathbb{R}^d),$$

$$\langle \nu_{\mathcal{T},k}, \psi \rangle = \sum_{n \in \mathbb{N}} \sum_{K \in \mathcal{T}} |u_K^{n+1} - u_K^n| \int_{nk}^{(n+1)k} \int_K \psi(x,t) dx dt,$$
$$\forall \psi \in C_c(\mathbb{R}^d \times \mathbb{R}_+).$$

The measures $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T},k}$ are absolutely continuous with respect to the Lebesgue measure. Indeed, one has $d\mu_{\mathcal{T}}(x) = |u_0(x) - u_{\mathcal{T},0}(x)| dx$ and $d\nu_{\mathcal{T},k}(x,t) = (\sum_{n \in \mathbb{N}} \sum_{K \in \mathcal{T}} |u_K^{n+1} - u_K^n| 1_{K \times [nk,(n+1)k)}) dx dt$ (where 1_Ω denotes the characteristic function of Ω for any Borel subset Ω of $\mathbb{R}^{d+1}$).

If $u_0 \in BV(\mathbb{R}^d)$, the measure $\mu_{\mathcal{T}}$ verifies (28.10) with some D only depending on $|u_0|_{BV(\mathbb{R}^d)}$ and α (this is classical result which is given in Lemma 28.3 below for the sake of completeness).

The measure $\nu_{\mathcal{T},k}$ satisfies (28.9), with $\nu_{\mathcal{T},k}$ instead of $\mu_{\mathcal{T},k}$, thanks to (26.5) and condition (25.3). Indeed, for $R > 0$ and $T > 0$,

$$\nu_{\mathcal{T},k}(B(0,R) \times [0,T]) = \int_0^T \int_{B(0,R)} \sum_{n \in \mathbb{N}} \sum_{K \in \mathcal{T}} |u_K^{n+1} - u_K^n| 1_{K \times [nk,(n+1)k)} dx dt,$$

which yields, with $\mathcal{T}_{2R} = \{K \in \mathcal{T},\ K \subset B(0,2R)\}$ and $N_{T,k} k < T \leq (N_{T,k}+1)k$, $h < R$ and $k < T$,

$$\nu_{\mathcal{T},k}(B(0,R) \times [0,T]) \leq k \sum_{n=0}^{N_{T,k}} \sum_{K \in \mathcal{T}_{2R}} \mathrm{m}(K) |u_K^{n+1} - u_K^n| \leq \frac{k C_1}{\sqrt{h}},$$

where C_1 is given by lemma 26.2 and only depends on $\mathbf{v}$, g, u_0, α, ξ, R, T. Finally, since the condition (25.3) gives $k \leq C_2 h$, where C_2 only depends on $\mathbf{v}$, g, u_0, α, ξ, the last inequality yields, for $h < R$ and $k < T$,

$$\nu_{\mathcal{T},k}(B(0,R) \times [0,T]) \leq C_3 \sqrt{h}, \tag{28.17}$$

with $C_3 = C_1 C_2$.

Comparison of T_2 and T_{20}

Using $\mathrm{div}\mathbf{v} = 0$, and gathering (28.14) by interfaces, we get

$$T_{20} = -\sum_{n \in \mathbb{N}} \sum_{(K,L) \in \mathcal{E}_n} \Big[\Big((f(u_K^n \top \kappa) - f(u_K^n \bot \kappa)) - (f(u_L^n \top \kappa) - f(u_L^n \bot \kappa)) \Big) \int_{K|L} \int_{nk}^{(n+1)k} \Big(\mathbf{v}(x,t) \cdot \boldsymbol{n}_{K,L}\, \varphi(x,t) \Big) d\gamma(x) dt \Big]. \tag{28.18}$$

Define, for all $K \in \mathcal{T}$, all $L \in \mathcal{N}(K)$ and all $n \in \mathbb{N}$,

$$(v\varphi)_{K,L}^{n,+} = \frac{1}{k} \int_{nk}^{(n+1)k} \int_{K|L} (\mathbf{v}(x,t) \cdot \boldsymbol{n}_{K,L})^+ \varphi(x,t) d\gamma(x) dt$$

and

$$(v\varphi)_{K,L}^{n,-} = \frac{1}{k} \int_{nk}^{(n+1)k} \int_{K|L} (\mathbf{v}(x,t) \cdot \boldsymbol{n}_{K,L})^- \varphi(x,t) d\gamma(x) dt.$$

Note that $(v\varphi)^{n,+}_{K,L} = (v\varphi)^{n,-}_{L,K}$. Then, (28.18) gives

$$\begin{aligned}
T_{20} = k\sum_{n\in\mathbb{N}}\sum_{(K,L)\in\mathcal{E}_n}\Big[& \\
&(v\varphi)^{n,+}_{K,L}\Big(g(u^n_K\top\kappa,u^n_L\top\kappa)-f(u^n_K\top\kappa)-g(u^n_K\perp\kappa,u^n_L\perp\kappa)+f(u^n_K\perp\kappa)\Big)\\
&-(v\varphi)^{n,-}_{L,K}\Big(g(u^n_K\top\kappa,u^n_L\top\kappa)-f(u^n_L\top\kappa)-g(u^n_K\perp\kappa,u^n_L\perp\kappa)+f(u^n_L\perp\kappa)\Big)\\
&-(v\varphi)^{n,-}_{K,L}\Big(g(u^n_L\top\kappa,u^n_K\top\kappa)-f(u^n_K\top\kappa)-g(u^n_L\perp\kappa,u^n_K\perp\kappa)+f(u^n_K\perp\kappa)\Big)\\
&+(v\varphi)^{n,+}_{L,K}\Big(g(u^n_L\top\kappa,u^n_K\top\kappa)-f(u^n_L\top\kappa)-g(u^n_L\perp\kappa,u^n_K\perp\kappa)+f(u^n_L\perp\kappa)\Big)\Big].
\end{aligned}\tag{28.19}$$

Let us introduce some terms related to the difference between φ on $K\in\mathcal{T}$ and $K|L\in\mathcal{E}$,

$$r^{n,+}_{K,L} = |v^n_{K,L}\varphi^n_K - (v\varphi)^{n,+}_{K,L}|$$

and

$$r^{n,-}_{K,L} = |v^n_{L,K}\varphi^n_K - (v\varphi)^{n,-}_{K,L}|.$$

Then, from (28.12) and (28.19),

$$\begin{aligned}
|T_2 - T_{20}| \le \sum_{n\in\mathbb{N}} k\sum_{(K,L)\in\mathcal{E}_n}\Big[& \\
&r^{n,+}_{K,L}\Big(g(u^n_K\top\kappa,u^n_L\top\kappa)-f(u^n_K\top\kappa)+g(u^n_K\perp\kappa,u^n_L\perp\kappa)-f(u^n_K\perp\kappa)\Big) \quad +\\
&r^{n,-}_{L,K}\Big(g(u^n_K\top\kappa,u^n_L\top\kappa)-f(u^n_L\top\kappa)+g(u^n_K\perp\kappa,u^n_L\perp\kappa)-f(u^n_L\perp\kappa)\Big) \quad +\\
&r^{n,-}_{K,L}\Big(f(u^n_K\top\kappa)-g(u^n_L\top\kappa,u^n_K\top\kappa)+f(u^n_K\perp\kappa)-g(u^n_L\perp\kappa,u^n_K\perp\kappa)\Big) \quad +\\
&r^{n,+}_{L,K}\Big(f(u^n_L\top\kappa)-g(u^n_L\top\kappa,u^n_K\top\kappa)+f(u^n_L\perp\kappa)-g(u^n_L\perp\kappa,u^n_K\perp\kappa)\Big)\Big].
\end{aligned}\tag{28.20}$$

For all $(K,L)\in\mathcal{E}_n$, the following inequality holds:

$$0 \le g(u^n_K\top\kappa,u^n_L\top\kappa) - f(u^n_K\top\kappa) \le \max_{u^n_L\le p\le q\le u^n_K}(g(q,p)-f(q)),$$

more precisely, one has $g(u^n_K\top\kappa,u^n_L\top\kappa) - f(u^n_K\top\kappa) = 0$, if $\kappa\ge u^n_K$, and one has $g(u^n_K\top\kappa,u^n_L\top\kappa) - f(u^n_K\top\kappa) = g(q,p)-f(q)$ with $p=\kappa$ and $q=u^n_K$ if $\kappa\in[u^n_L,u^n_K]$, and with $p=u^n_L$ and $q=u^n_K$ if $\kappa\le u^n_L$. In the same way, we can assert that

$$0 \le g(u^n_K\perp\kappa,u^n_L\perp\kappa) - f(u^n_K\perp\kappa) \le \max_{u^n_L\le p\le q\le u^n_K}(g(q,p)-f(q)).$$

The same analysis can be applied to the six other terms of (28.20).
To conclude the estimate on $|T_2-T_{20}|$, there remains to estimate the two quantities $r^{n,\pm}_{K,L}$. This will be done with convenient measures applied to $|\nabla\varphi|$ and $|\varphi_t|$. To estimate $r^{n,+}_{K,L}$, for instance, one remarks that

$$r^{n,+}_{K,L} \le \frac{1}{k^2\mathrm{m}(K)}\int_{nk}^{(n+1)k}\int_{nk}^{(n+1)k}\int_K\int_{K|L}|\varphi(x,t)-\varphi(y,s)|(\mathbf{v}(y,s)\cdot\boldsymbol{n}_{K,L})^+ d\gamma(y)dxdtds.$$

Hence

$$\begin{aligned}
r^{n,+}_{K,L} \le \frac{1}{k^2\mathrm{m}(K)}\int_{nk}^{(n+1)k}\int_{nk}^{(n+1)k}\int_K\int_{K|L}\int_0^1 & |\nabla\varphi(x+\theta(y-x),t+\theta(s-t))\cdot(y-x)+\\
&\varphi_t(x+\theta(y-x),t+\theta(s-t))(s-t)|(\mathbf{v}(y,s)\cdot\boldsymbol{n}_{K,L})^+ d\theta d\gamma(y)dxdtds
\end{aligned}$$

which yields

$$r_{K,L}^{n,+} \le \frac{1}{k^2 \mathrm{m}(K)} \int_{nk}^{(n+1)k} \int_{nk}^{(n+1)k} \int_K \int_{K|L} \int_0^1 \Big(h|\nabla\varphi(x+\theta(y-x), t+\theta(s-t))| + k|\varphi_t(x+\theta(y-x), t+\theta(s-t))| \Big) (\mathbf{v}(y,s)\cdot \boldsymbol{n}_{K,L})^+ d\theta d\gamma(y) dx dt ds.$$

This leads to the definition of a measure $\mu_{K,L}^{n,+}$, given by its action on $C_c(\mathbb{R}^d \times \mathbb{R}_+)$:

$$\langle \mu_{K,L}^{n,+}, \psi \rangle = \frac{2}{k^2 \mathrm{m}(K)} \int_{nk}^{(n+1)k} \int_{nk}^{(n+1)k} \int_K \int_{K|L} \int_0^1 \Big((h+k)\psi(x+\theta(y-x), t+\theta(s-t)) \Big) (\mathbf{v}(y,s)\cdot \boldsymbol{n}_{K,L})^+ d\theta d\gamma(y) dx dt ds, \ \forall \psi \in C_c(\mathbb{R}^d \times \mathbb{R}_+),$$

in order to have $2r_{K,L}^{n,+} \le \langle \mu_{K,L}^{n,+}, |\nabla\varphi| + |\varphi_t| \rangle$.
We define in the same way $\mu_{K,L}^{n,-}$, changing $(\mathbf{v}(y,s)\cdot \boldsymbol{n}_{K,L})^+$ in $(\mathbf{v}(y,s)\cdot \boldsymbol{n}_{K,L})^-$. We finally define the measure $\tilde{\nu}_{\mathcal{T},k}$ by

$$\begin{aligned} \langle \tilde{\nu}_{\mathcal{T},k}, \psi \rangle = \sum_{n\in\mathbb{N}} k \sum_{(K,L)\in\mathcal{E}_n} \Big[& \Big(\max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(q)) \Big) \langle \mu_{K,L}^{n,+}, \psi \rangle \\ & + \Big(\max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(p)) \Big) \langle \mu_{L,K}^{n,-}, \psi \rangle \\ & + \Big(\max_{u_L^n \le p \le q \le u_K^n} (f(q) - g(p,q)) \Big) \langle \mu_{K,L}^{n,-}, \psi \rangle \\ & + \Big(\max_{u_L^n \le p \le q \le u_K^n} (f(p) - g(p,q)) \Big) \langle \mu_{L,K}^{n,+}, \psi \rangle \Big]. \end{aligned} \tag{28.21}$$

Since $2r_{K,L}^{n,\pm} \le \langle \mu_{K,L}^{n,\pm}, |\nabla\varphi| + |\varphi_t| \rangle$, (28.20) and (28.21) leads to $|T_2 - T_{20}| \le \langle \tilde{\nu}_{\mathcal{T},k}, |\nabla\varphi| + |\varphi_t| \rangle$. Therefore, setting $\mu_{\mathcal{T},k} = \nu_{\mathcal{T},k} + \tilde{\nu}_{\mathcal{T},k}$, using (28.16) and $T_1 + T_2 \le 0$,

$$T_{10} + T_{20} \le \int_{\mathbb{R}^d \times \mathbb{R}_+} \Big(|\varphi_t(x,t)| + |\nabla\varphi(x,t)| \Big) d\mu_{\mathcal{T},k}(x,t) + \int_{\mathbb{R}^d} \varphi(x,0) d\mu_{\mathcal{T}}(x),$$

which is (28.13) and yields (28.8).

There remains to prove (28.9).
For all $K \in \mathcal{T}$, let x_K be an arbitrary point of K. For all $K \in \mathcal{T}$, all $K \in \mathcal{N}(K)$ and all $n \in \mathbb{N}$, the supports of the measures $\mu_{K,L}^{n,\pm}$ are included in the closed set $\bar{B}(x_K, h) \cap [nk, (n+1)k]$. Furthermore,

$$\mu_{K,L}^{n,+}(\mathbb{R}^d \times \mathbb{R}_+) \le 2v_{K,L}^n(h+k) \text{ and } \mu_{K,L}^{n,-}(\mathbb{R}^d \times \mathbb{R}_+) \le 2v_{L,K}^n(h+k).$$

Then, for all $R > 0$ and $T > 0$, the definition of $\mu_{\mathcal{T},k}$ (i.e. $\mu_{\mathcal{T},k} = \nu_{\mathcal{T},k} + \tilde{\nu}_{\mathcal{T},k}$)) leads to

$$\begin{aligned} & \mu_{\mathcal{T},k}(B(0,R)\times[0,T]) \le C_3\sqrt{h} \\ & + 2(h+k) \sum_{n=0}^{N_{T,k}} k \sum_{(K,L)\in\mathcal{E}_{2R}^n} \Big[v_{K,L}^n \Big(\max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(q)) + \max_{u_L^n \le p \le q \le u_K^n} (g(q,p) - f(p)) \Big) \\ & + v_{L,K}^n \Big(\max_{u_L^n \le p \le q \le u_K^n} (f(q) - g(p,q)) + \max_{u_L^n \le p \le q \le u_K^n} (f(p) - g(p,q)) \Big) \Big], \end{aligned}$$

for $h < R$ and $k < T$, where $C_3\sqrt{h}$ is the bound of $\nu_{\mathcal{T},k}(B(0,R)\times[0,T])$ given in (28.17). Therefore, thanks to Lemma 26.2,

$$\mu_{\mathcal{T},k}(B(0,R)\times[0,T]) \le C_3\sqrt{h} + (1+C_2)h\frac{C_4}{\sqrt{h}} = C\sqrt{h},$$

where C only depends on $\mathbf{v}$, g, u_0, α, ξ, R and T. The proof of Theorem 28.1 is complete. ∎

The following theorem investigates the case of the implicit scheme.

Theorem 28.2 *Under Assumption 24.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 and $k>0$. Let $g\in C(\mathbb{R}^2,\mathbb{R})$ satisfy Assumption 25.1.*
Let $\{u_K^n, n\in\mathbb{N}, K\in\mathcal{T}\}$, such that $u_K^n\in[U_m,U_M]$ for all $K\in\mathcal{T}$ and $n\in\mathbb{N}$, be the solution of (25.6),(25.2) (existence and uniqueness of such a solution are given by Proposition 27.1). Let $u_{\mathcal{T},k}$ be given by (25.5). Assume that $\mathbf{v}$ does not depend on t and that $u_0\in BV(\mathbb{R}^d)$.
Then, there exist $\mu_{\mathcal{T},k}\in\mathcal{M}(\mathbb{R}^d\times\mathbb{R}_+)$ and $\mu_{\mathcal{T}}\in\mathcal{M}(\mathbb{R}^d)$ such that

$$\left\{\begin{array}{ll} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big(& |u_{\mathcal{T},k}(x,t)-\kappa|\varphi_t(x,t)+ \\ & \displaystyle(f(u_{\mathcal{T},k}(x,t)\top\kappa)-f(u_{\mathcal{T},k}(x,t)\bot\kappa))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)dxdt \quad + \\ \displaystyle\int_{\mathbb{R}^d} & |u_0(x)-\kappa|\varphi(x,0)dx \qquad\qquad \geq \\ \displaystyle-\int_{\mathbb{R}^d\times\mathbb{R}_+} & \Big(|\varphi_t(x,t)|+|\nabla\varphi(x,t)|\Big)d\mu_{\mathcal{T},k}(x,t)-\displaystyle\int_{\mathbb{R}^d}\varphi(x,0)d\mu_{\mathcal{T}}(x), \\ & \\ & \forall\kappa\in\mathbb{R},\quad \forall\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+). \end{array}\right. \tag{28.22}$$

The measures $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{T}}$ verify the following properties:

1. *For all $R>0$ and $T>0$, there exists C, only depending on $\mathbf{v}$, g, u_0, α, R, T such that, for $h<R$ and $k<T$,*
$$\mu_{\mathcal{T},k}(B(0,R)\times[0,T])\leq C(k+\sqrt{h}). \tag{28.23}$$
2. *The measure $\mu_{\mathcal{T}}$ is the measure of density $|u_0(\cdot)-u_{\mathcal{T},0}(\cdot)|$ with respect to the Lebesgue measure and there exists D, only depending on u_0 and α, such that*
$$\mu_{\mathcal{T}}(\mathbb{R}^d)\leq Dh. \tag{28.24}$$

Proof of Theorem 28.2
Similarly to the proof of Theorem 28.1, we introduce a test function $\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+)$ and a real number $\kappa\in\mathbb{R}$. We multiply (28.5) by $(1/\mathrm{m}(K))\int_{nk}^{(n+1)k}\int_K\varphi(x,t)dxdt$, and sum the result for all $K\in\mathcal{T}$ and $n\in\mathbb{N}$. We then define T_1 and T_2 such that $T_1+T_2\leq 0$ using equations (28.11) and (28.12) in which we replace u_K^n by u_K^{n+1} and u_L^n by u_L^{n+1}. Therefore we get (28.16), where the measure $\nu_{\mathcal{T},k}$ is such that for all $T>0$, there exists C_1 only depending on $\mathbf{v}$, g, u_0 α and T, such that, for $k<T$,

$$\nu_{\mathcal{T},k}(\mathbb{R}^d\times[0,T])\leq C_1k,$$

using Lemma 27.3 page 170, which is available if $\mathbf{v}$ does not depend on t (and for which one needs that $u_0\in BV(\mathbb{R}^d)$).
The treatment of T_2 is very similar to that of Theorem 28.1, replacing u_K^n by u_K^{n+1} and u_L^n by u_L^{n+1}. But, since $\mathbf{v}$ does not depend on t, the bounds on $r_{K,L}^{n,\pm}$ are simpler. Indeed,

$$r_{K,L}^{n,\pm}\leq\frac{1}{k\mathrm{m}(K)}\int_{nk}^{(n+1)k}\int_K\int_{K|L}|\varphi(x,t)-\varphi(y,t)|(\mathbf{v}(y)\cdot\boldsymbol{n}_{K,L})^\pm d\gamma(y)dxdt.$$

Now $2r_{K,L}^{n,\pm}\leq\langle\mu_{K,L}^{n,\pm},|\nabla\varphi|\rangle$ where $\mu_{K,L}^{n,\pm}$ is defined by

$$\begin{aligned}\langle\mu_{K,L}^{n,\pm},\psi\rangle=\frac{2}{k\mathrm{m}(K)}\int_{nk}^{(n+1)k}\int_K\int_{K|L}\int_0^1\Big(h\,\psi(x+\theta(y-x),t)\Big)\\ (\mathbf{v}(y)\cdot\boldsymbol{n}_{K,L})^\pm d\theta d\gamma(y)dxdt,\ \forall\psi\in C_c(\mathbb{R}^d\times\mathbb{R}_+).\end{aligned}$$

With this definition of $\mu_{K,L}^{n,\pm}$, the bound on $\tilde{\nu}_{\mathcal{T},k}$ (defined by (28.21), replacing u_K^n by u_K^{n+1} and u_L^n by u_L^{n+1}) becomes, thanks to Lemma 27.1 page 167,

$$\tilde{\nu}_{\mathcal{T},k}(B(0,R)\times[0,T]) \leq C_2\sqrt{h},$$

for $h < R$ and $k < T$, where C_2 only depends on $\mathbf{v}$, g, u_0, α, R and T.
Hence, defining (as in Theorem 28.1) $\mu_{\mathcal{T},k} = \nu_{\mathcal{T},k} + \tilde{\nu}_{\mathcal{T},k}$, for all $R > 0$ and all $T > 0$ there exists C, only depending on $\mathbf{v}$, g, u_0, α, R, T such that, for $h < R$ and $k < T$,

$$\mu_{\mathcal{T},k}(B(0,R)\times[0,T]) \leq C(k+\sqrt{h}),$$

which is (28.23) and concludes the proof of Theorem 28.2. ∎

Remark 28.2 In the case where $\mathbf{v}$ depends on t, Lemma 27.3 cannot be used. However, it is easy to show (the proof follows that of Theorem 28.1) that Theorem 28.2 is true if (28.23) is replaced by

$$\mu_{\mathcal{T},k}(B(0,R)\times[0,T]) \leq C(\frac{k}{\sqrt{h}}+\sqrt{h}), \tag{28.25}$$

which leads to the result given in Remark 30.2. The estimate (28.25) may be obtained without assuming that $u_0 \in BV(\mathbb{R}^d)$ (it is sufficient that $u_0 \in L^\infty(\mathbb{R}^d)$).

For the sake of completeness we now prove a lemma which gives the bound on the measure $\mu_{\mathcal{T}}$ in the two last theorems.

Lemma 28.3 *Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and let $u \in BV(\mathbb{R}^d)$ (see Definition 21.19 page 141). For $K \in \mathcal{T}$, let u_K be the mean value of u over K. Define $u_{\mathcal{T}}$ by $u_{\mathcal{T}}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Then,*

$$\|u - u_{\mathcal{T}}\|_{L^1(\mathbb{R}^d)} \leq \frac{C}{\alpha^2} h |u|_{BV(\mathbb{R}^d)}, \tag{28.26}$$

where C only depends on the space dimension ($d = 1$, 2 or 3).

Proof of Lemma 28.3

The proof is very similar to that of Lemma 27.2 and we will mainly refer to the proof of Lemma 27.2.

First, remark that if (28.26) holds for all $u \in BV(\mathbb{R}^d)\cap C^1(\mathbb{R}^d,\mathbb{R})$ then (28.26) holds for all $u \in BV(\mathbb{R}^d)$. Indeed, let $u \in BV(\mathbb{R}^d)$, it is proven in Step 1 of the proof of Lemma 27.2 that there exists a sequence $(u_n)_{n\in\mathbb{N}} \subset C^\infty(\mathbb{R}^d,\mathbb{R})$ such that $u_n \to u$ in $L^1_{loc}(\mathbb{R}^d)$, as $n \to \infty$, and $\|u_n\|_{BV(\mathbb{R}^d)} \leq \|u\|_{BV(\mathbb{R}^d)}$ for all $n \in \mathbb{N}$. One may also assume, up to a subsequence, that $u_n \to u$ a.e. on $\mathbb{R}^d$. Then, if (28.26) is true with u_n instead of u, passing to the limit in (28.26) (for u_n) as $n \to \infty$ leads to (28.26) (for u) thanks to Fatou's lemma.

Let us now prove (28.26) if $u \in BV(\mathbb{R}^d) \cap C^1(\mathbb{R}^d,\mathbb{R})$ (this concludes the proof of Lemma 28.3). Since $u \in C^1(\mathbb{R}^d,\mathbb{R})$,

$$|u|_{BV(\mathbb{R}^d)} = \|(|\nabla u|)\|_{L^1(\mathbb{R}^d)};$$

hence we shall prove (28.26) with $\|(|\nabla u|)\|_{L^1(\mathbb{R}^d)}$ instead of $|u|_{BV(\mathbb{R}^d)}$.
For $K \in \mathcal{T}$,

$$\int_K |u(x) - u_K|dx \leq \frac{1}{\mathrm{m}(K)} \int_K (\int_K |u(x) - u(y)|dx)dy.$$

Then, following the lines of Step 2 of Lemma 27.2,

$$\int_K |u(x) - u_K|dx \leq \frac{1}{\mathrm{m}(K)} h \int_{B(0,h)} (\int_0^1 \int_K |\nabla u(y+tz)|dydt)dz. \tag{28.27}$$

For all $K \in \mathcal{T}$, let x_K be an arbitrary point of K.

Then, changing the variable y in $\xi = y + tz$ (for all fixed $z \in K$ and $t \in (0,1)$) in (28.27),

$$\int_K |u(x) - u_K| dx \leq \frac{1}{\mathrm{m}(K)} h \int_{B(0,h)} \Big(\int_0^1 \int_{B(x_K, 2h)} |\nabla u(\xi)| d\xi dt \Big) dz,$$

which yields, since $\mathcal{T}$ is an admissible mesh in the sense of Definition 25.1 page 156,

$$\int_K |u(x) - u_K| dx \leq \frac{1}{\alpha h^d} \mathrm{m}(B(0,h)) h \int_{B(x_K, 2h)} |\nabla u(\xi)| d\xi.$$

Therefore there exists C_1, only depending on the space dimension, such that

$$\int_K |u(x) - u_K| dx \leq \frac{C_1}{\alpha} h \int_{B(x_K, 2h)} |\nabla u(\xi)| d\xi, \ \forall K \in \mathcal{T}. \tag{28.28}$$

As in Lemma 27.2, for a fixed $M \in \mathcal{T}$, the number of $K \in \mathcal{T}$ such that $M \cap B(x_K, 2h) \neq \emptyset$ is less or equal to $\mathrm{m}(B(0,4h))/(\alpha h^d)$ that is less or equal to C_2/α where C_2 only depends on the space dimension. Then, summing (28.28) over $K \in \mathcal{T}$ leads to

$$\sum_{K \in \mathcal{T}} \int_K |u(x) - u_K| dx \leq \frac{C_1 C_2}{\alpha^2} h \sum_{M \in \mathcal{T}} \int_M |\nabla u(\xi)| d\xi = \frac{C_1 C_2}{\alpha^2} h \| (|\nabla u|) \|_{L^1(\mathbb{R}^d)},$$

that is (28.26) with $C = C_1 C_2$. ∎

29 Convergence of the scheme

This section is devoted to the proof of the existence and uniqueness of the entropy weak solution and of the convergence of the approximate solution towards the entropy weak solution as the mesh size and time step tend to 0. This proof will be performed in two steps. We first prove in section 29.1 the convergence of the approximate solution towards an entropy process solution which is defined in Definition 29.1 below (note that the convergence also yields the existence of an entropy process solution).

Definition 29.1 A function μ is an entropy process solution to problem (24.1)-(24.2) if μ satisfies

$$\left\{ \begin{array}{l} \mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)), \\ \displaystyle\int_{\mathbb{R}^d} \int_0^{+\infty} \int_0^1 \Big(\eta(\mu(x,t,\alpha)) \varphi_t(x,t) + \Phi(\mu(x,t,\alpha)) \mathbf{v}(x,t) \cdot \nabla \varphi(x,t) \Big) d\alpha dt dx \\ \displaystyle + \int_{\mathbb{R}^d} \eta(u_0(x)) \varphi(x,0) dx \geq 0, \\ \text{for any } \varphi \in C_c^1(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+), \\ \text{for any convex function } \eta \in C^1(\mathbb{R}, \mathbb{R}), \text{ and } \Phi \in C^1(\mathbb{R}, \mathbb{R}) \text{ such that } \Phi' = f' \eta'. \end{array} \right. \tag{29.1}$$

Remark 29.1 From an entropy weak solution u to problem (24.1)-(24.2), one may easily construct an entropy process solution to problem (24.1)-(24.2) by setting $\mu(x,t,\alpha) = u(x,t)$ for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$. Reciprocally, if μ is an entropy process solution to problem (24.1)-(24.2) such that there exists $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $\mu(x,t,\alpha) = u(x,t)$, for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$, then u is an entropy weak solution to problem (24.1)-(24.2).

In section 29.2, we show the uniqueness of the entropy process solution, which, thanks to remark 29.1, also yields the existence and uniqueness of the entropy weak solution. This allows us to state and prove, in section 29.3, the convergence of the approximate solution towards the entropy weak solution.

We now give a useful characterization of an entropy process solution in terms of Krushkov's entropies (as for the entropy weak solution).

Proposition 29.1 *A function μ is an entropy process solution of problem (24.1)-(24.2) if and only if,*

$$\left\{\begin{array}{l} \mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)), \\ \displaystyle\int_{\mathbb{R}^d}\int_0^{+\infty}\int_0^1 \Big(|\mu(x,t,\alpha)-\kappa|\varphi_t(x,t) + \Phi(\mu(x,t,\alpha),\kappa)\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)d\alpha dt dx \\ \qquad\qquad + \displaystyle\int_{\mathbb{R}^d}|u_0(x)-\kappa|\varphi(x,0)dx \geq 0, \\ \qquad\qquad \forall \kappa \in \mathbb{R},\ \forall \varphi \in C_c^1(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+), \end{array}\right. \tag{29.2}$$

where we set $\Phi(a,b) = f(a\top b) - f(a\perp b)$, for all a, $b \in \mathbb{R}$.

Proof of Proposition 29.1

The proof of this result is similar to the case of classical entropy weak solutions. The characterization (29.2) can be obtained from (29.1), by using regularizations of the function $|\cdot - \kappa|$. Conversely, (29.1) may be obtained from (29.2) by approximating any convex function $\eta \in C^1(\mathbb{R},\mathbb{R})$ by functions of the form: $\eta_n(\cdot) = \displaystyle\sum_{i=1}^n \alpha_i^{(n)}|\cdot - \kappa_i^{(n)}|$, with $\alpha_i^{(n)} \geq 0$. ∎

29.1 Convergence towards an entropy process solution

Let $\alpha > 0$ and $0 < \xi < 1$. Let $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$ be a sequence of admissible meshes in the sense of Definition 25.1 page 156 and time steps. Note that $\mathcal{T}_m$ is admissible with α independent of m. Assume that k_m satisfies (25.3), for $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, and that size$(\mathcal{T}_m) \to 0$ as $m \to \infty$.
By Lemma 26.1 page 160, the sequence $(u_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$ of approximate solutions defined by the finite volume scheme (25.2) and (25.4) page 157, with $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, is bounded in $L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$; therefore, there exists $\mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))$ such that $u_{\mathcal{T}_m,k_m}$ converges, as m tends to ∞, towards μ in the nonlinear weak-$\star$ sense (see Definition 32.1 page 200 and Proposition 32.1 page 201), that is:

$$\begin{array}{l} \displaystyle\lim_{m\to\infty}\int_{\mathbb{R}^d}\int_{\mathbb{R}_+}\theta(u_{\mathcal{T}_m,k_m}(x,t))\varphi(x,t)dtdx = \int_{\mathbb{R}^d}\int_{\mathbb{R}_+}\int_0^1\theta(\mu(x,t,\alpha))\varphi(x,t)d\alpha dtdx, \\ \forall\varphi \in L^1(\mathbb{R}\times\mathbb{R}_+^\star),\ \forall\theta\in C(\mathbb{R},\mathbb{R}). \end{array} \tag{29.3}$$

Taking for θ, in (29.3), the Krushkov entropies (namely $\theta = |\cdot - \kappa|$, for all $\kappa \in \mathbb{R}$) and the associated functions defining the entropy fluxes (namely $\theta = f(\cdot,\kappa) = f(\cdot\top\kappa) - f(\cdot\perp\kappa)$) and using Theorem 28.1 (that is passing to the limit, as $m \to \infty$, in (28.8) written with $u_{\mathcal{T},k} = u_{\mathcal{T}_m,k_m}$) yields that μ is an entropy process solution. Hence the following result holds:

Proposition 29.2 *Under assumptions 24.1, let $\alpha > 0$ and $0 < \xi < 1$. Let $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$ be a sequence of admissible meshes in the sense of Definition 25.1 page 156 and time steps. Note that $\mathcal{T}_m$ is admissible with α independent of m. Assume that k_m satisfy (25.3), for $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, and that* size$(\mathcal{T}_m) \to 0$ *as $m \to \infty$.*
Then there exists a subsequence, still denoted by $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$, and a function $\mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))$ such that

1. *the approximate solution defined by (25.4), (25.2) and (25.5) with $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, that is $u_{\mathcal{T}_m,k_m}$, converges towards μ in the nonlinear weak-$\star$ sense, i.e. (29.3) holds,*
2. *μ is an entropy process solution of (24.1)-(24.2).*

Remark 29.2 The same theorem can be proved for the implicit scheme without condition (25.3) (and thus without ξ).

Remark 29.3 Note that a consequence of Proposition 29.2 is the existence of an entropy process solution to Problem (24.1)-(24.2).

29.2 Uniqueness of the entropy process solution

In order to show the uniqueness of an entropy process solution, we shall use the characterization of an entropy process solution given in proposition 29.1.

Theorem 29.1 *Under Assumption 24.1, the entropy process solution μ of problem (24.1),(24.2), as defined in Definition 29.1 page 181, is unique. Moreover, there exists a function $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $u(x,t) = \mu(x,t,\alpha)$, for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$. (Hence, with Proposition 29.2 and Remark 29.1, there exists a unique entropy weak solution to Problem (24.1)-(24.2).)*

Proof of Theorem 29.1

Let μ and ν be two entropy process solutions to Problem (24.1)-(24.2). Then, one has $\mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))$, $\nu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))$ and

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_0^{+\infty}\int_0^1 \Big(|\mu(x,t,\alpha)-\kappa|\varphi_t(x,t) \\
&+(f(\mu(x,t,\alpha)\top\kappa)-f(\mu(x,t,\alpha)\bot\kappa))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)d\alpha dt dx \\
&+\int_{\mathbb{R}^d}|u_0(x)-\kappa|\varphi(x,0)dx \geq 0,\ \forall\kappa\in\mathbb{R},\ \forall\varphi\in C_c^1(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+),
\end{aligned} \tag{29.4}$$

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_0^{+\infty}\int_0^1 \Big(|\nu(y,s,\beta)-\kappa|\varphi_s(y,s) \\
&+(f(\nu(y,s,\beta)\top\kappa)-f(\nu(y,s,\beta)\bot\kappa))\mathbf{v}(y,s)\cdot\nabla\varphi(y,s)\Big)d\beta ds dy \\
&+\int_{\mathbb{R}^d}|u_0(y)-\kappa|\varphi(y,0)dy \geq 0,\ \forall\kappa\in\mathbb{R},\ \forall\varphi\in C_c^1(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+).
\end{aligned} \tag{29.5}$$

The proof of Theorem 29.1 contains 2 steps. In Step 1, it is proven that

$$\begin{aligned}
&\int_0^1\int_0^1\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big[|\mu(x,t,\alpha)-\nu(x,t,\beta)|\psi_t(x,t) \\
&+\Big(f(\mu(x,t,\alpha)\top\nu(x,t,\beta))-f(\mu(x,t,\alpha)\bot\nu(x,t,\beta))\Big)\mathbf{v}(x,t)\cdot\nabla\psi(x,t)\Big]dxdtd\alpha d\beta \geq 0, \\
&\forall\psi\in C_c^1(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+).
\end{aligned} \tag{29.6}$$

In Step 2, it is proven that $\mu(x,t,\alpha) = \nu(x,t,\beta)$ for a.e. $(x,t,\alpha,\beta) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1) \times (0,1)$. We then deduce that there exists $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $\mu(x,t,\alpha) = u(x,t)$ for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$ (therefore u is necessarily the unique entropy weak solution to (24.1)-(24.2)).

Step 1 (proof of relation (29.6))

In order to prove relation (29.6), a sequence of mollifiers in $\mathbb{R}$ and $\mathbb{R}^d$ is introduced .

Let $\rho \in C_c^\infty(\mathbb{R}^d, \mathbb{R}_+)$ and $\bar\rho \in C_c^\infty(\mathbb{R}, \mathbb{R}_+)$ be such that

$$\{x \in \mathbb{R}^d;\ \rho(x) \neq 0\} \subset \{x \in \mathbb{R}^d;\ |x| \leq 1\},$$

$$\{x \in \mathbb{R};\ \bar\rho(x) \neq 0\} \subset [-1, 0] \tag{29.7}$$

and

$$\int_{\mathbb{R}^d}\rho(x)dx = 1,\ \int_{\mathbb{R}}\bar\rho(x)dx = 1.$$

For $n \in \mathbb{N}^\star$, define $\rho_n = n^d\rho(nx)$ for all $x \in \mathbb{R}^d$ and $\bar\rho_n = n\bar\rho(nx)$ for all $x \in \mathbb{R}$.

Let $\psi \in C_c^1(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$. For $(y,s,\beta) \in \mathbb{R}^d \times \mathbb{R}_+ \times (0,1)$, let us take, in (29.4), $\varphi(x,t) = \psi(x,t)\rho_n(x-y)\bar\rho_n(t-s)$ and $\kappa = \nu(y,s,\beta)$. Then, integrating the result over $\mathbb{R}^d \times \mathbb{R}_+ \times (0,1)$ leads to

$$A_1 + A_2 + A_3 + A_4 + A_5 \geq 0, \tag{29.8}$$

where

$$A_1 = \int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[|\mu(x,t,\alpha) - \nu(y,s,\beta)| \\ \psi_t(x,t)\rho_n(x-y)\bar{\rho}_n(t-s) \Big] dxdtdyds d\alpha d\beta,$$

$$A_2 = \int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[|\mu(x,t,\alpha) - \nu(y,s,\beta)| \\ \psi(x,t)\rho_n(x-y)\bar{\rho}_n'(t-s) \Big] dxdtdyds d\alpha d\beta,$$

$$A_3 = \int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[\Big(f(\mu(x,t,\alpha) \top \nu(y,s,\beta)) - f(\mu(x,t,\alpha) \bot \nu(y,s,\beta)) \Big) \\ \mathbf{v}(x,t) \cdot \nabla \psi(x,t) \rho_n(x-y) \bar{\rho}_n(t-s) \Big] dxdtdyds d\alpha d\beta,$$

$$A_4 = \int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[\Big(f(\mu(x,t,\alpha) \top \nu(y,s,\beta)) - f(\mu(x,t,\alpha) \bot \nu(y,s,\beta)) \Big) \\ \mathbf{v}(x,t) \cdot \nabla \rho_n(x-y) \psi(x,t) \bar{\rho}_n(t-s) \Big] dxdtdyds d\alpha d\beta$$

and

$$A_5 = \int_0^1 \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} |u_0(x) - \nu(y,s,\beta)| \psi(x,0) \rho_n(x-y) \bar{\rho}_n(-s) dyds dx d\beta.$$

Passing to the limit in (29.8) as $n \to \infty$ (using (29.5) for the study of $A_2 + A_4$ and A_5) will give (29.6). Let us first consider A_1 and A_3. Note that, using (29.7),

$$\int_{\mathbb{R}^d} \int_0^\infty \rho_n(x-y)\bar{\rho}_n(t-s) ds dy = 1, \, \forall x \in \mathbb{R}^d, \, \forall t \in \mathbb{R}_+.$$

Then,

$$\begin{aligned} &|A_1 - \int_0^1 \int_0^1 \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big[|\mu(x,t,\alpha) - \nu(x,t,\beta)| \psi_t(x,t) \Big] \, dxdtd\alpha d\beta| \\ &\leq \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[|\nu(x,t,\beta) - \nu(y,s,\beta)| |\psi_t(x,t)| \rho_n(x-y)\bar{\rho}_n(t-s) \Big] dxdtdyds d\beta \\ &\leq \|\psi_t\|_{L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)} \varepsilon(n,S), \end{aligned}$$

with $S = \{(x,t) \in \mathbb{R}^d \times \mathbb{R}_+; \, \psi(x,t) \neq 0\}$ and

$$\varepsilon(n,S) = \sup\{ \|\nu - \nu(\cdot + \eta, \cdot + \tau, \cdot)\|_{L^1(S \times (0,1))}; \; |\eta| \leq \frac{1}{n}, \, 0 \leq \tau \leq \frac{1}{n} \}.$$

Since $\nu \in L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_+ \times [0,1])$ and S is bounded, one has $\varepsilon(n,S) \to 0$ as $n \to \infty$. Hence,

$$A_1 \to \int_0^1 \int_0^1 \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big[|\mu(x,t,\alpha) - \nu(x,t,\beta)| \psi_t(x,t) \Big] dxdtd\alpha d\beta, \text{ as } n \to \infty. \tag{29.9}$$

Similarly, let M be the Lipschitz constant of f on $[-D,D]$ where $D = \max\{\|\mu\|_\infty, \|\nu\|_\infty\}$, with $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))}$,

$$|A_3 - \int_0^1 \int_0^1 \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big(f(\mu(x,t,\alpha) \top \nu(x,t,\beta)) - f(\mu(x,t,\alpha) \bot \nu(x,t,\beta)) \Big) \\ \mathbf{v}(x,t) \cdot \nabla \psi(x,t) dxdtd\alpha d\beta| \leq 2MV \| (|\nabla \psi|) \|_{L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)} \, \varepsilon(n,S),$$

which yields

$$A_3 \to \int_0^1 \int_0^1 \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big(f(\mu(x,t,\alpha)\top\nu(x,t,\beta)) - f(\mu(x,t,\alpha)\bot\nu(x,t,\beta)) \Big) \mathbf{v}(x,t)\cdot\nabla\psi(x,t) dx dt d\alpha d\beta, \text{ as } n\to\infty. \tag{29.10}$$

Let us now consider $A_2 + A_4$.
For $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+ \times (0,1)$, let us take $\varphi(y,s) = \psi(x,t)\rho_n(x-y)\bar\rho_n(t-s)$ and $\kappa = \mu(x,t,\alpha)$ in (29.5). Integrating the result over $\mathbb{R}^d \times \mathbb{R}_+ \times (0,1)$ leads to

$$-A_2 - B_4 \geq 0, \tag{29.11}$$

with

$$A_4 - B_4 = \int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[\Big(f(\mu(x,t,\alpha)\top\nu(y,s,\beta)) - f(\mu(x,t,\alpha)\bot\nu(y,s,\beta)) \Big) (\mathbf{v}(x,t) - \mathbf{v}(y,s))\cdot\nabla\rho_n(x-y)\psi(x,t)\bar\rho_n(t-s) \Big] dx dt dy ds d\alpha d\beta.$$

Note that $B_4 = A_4$ if $\mathbf{v}$ is constant (and one directly obtains (29.13) below). In the general case, in order to prove that $A_4 - B_4 \to 0$ as $n\to\infty$ (which then gives (29.13)), let us remark that, using $\mathrm{div}\mathbf{v} = 0$,

$$\int_0^1 \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[\Big(f(\mu(x,t,\alpha)\top\nu(x,t,\beta)) - f(\mu(x,t,\alpha)\bot\nu(x,t,\beta)) \Big) (\mathbf{v}(x,t) - \mathbf{v}(y,s))\cdot\nabla\rho_n(x-y)\psi(x,t)\bar\rho_n(t-s) \Big] dx dt dy ds d\alpha d\beta = 0. \tag{29.12}$$

Indeed, the latter equality follows from an integration by parts for the variable $y \in \mathbb{R}^d$. Then, substracting the left hand side of (29.12) to $A_4 - B_4$ and using the regularity of $\mathbf{v}$, there exists C_1, only depending on M, $\mathbf{v}$ and ψ, such that $|A_4 - B_4| \leq C_1\varepsilon(n,S)$. This gives $A_4 - B_4 \to 0$ as $n\to\infty$ and, thanks to (29.11),

$$\limsup_{n\to\infty} (A_2 + A_4) \leq 0. \tag{29.13}$$

Finally, let us consider A_5.
For $x \in \mathbb{R}^d$, let us take $\varphi(y,s) = \psi(x,0)\rho_n(x-y)\int_s^\infty \bar\rho_n(-\tau)d\tau$ and $\kappa = u_0(x)$ in (29.5). Integrating the resulting inequality with respect to $x \in \mathbb{R}^d$ gives

$$-A_5 + B_{5a} + B_{5b} \geq 0, \tag{29.14}$$

with

$$B_{5a} = -\int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(\nu(y,s,\beta)\top u_0(x)) - f(\nu(y,s,\beta)\bot u_0(x))) \mathbf{v}(y,s)\cdot\nabla\rho_n(x-y)\psi(x,0)\bar\rho_n(-\tau) d\tau dy dx ds d\beta,$$

$$B_{5b} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \psi(x,0)\rho_n(x-y)|u_0(x) - u_0(y)| dy dx.$$

Let $S_0 = \{x \in \mathbb{R}^d;\ \psi(x,0) \neq 0\}$ and

$$\varepsilon_0(n,S_0) = \sup\{ \int_{S_0} |u_0(x) - u_0(x+\eta)| dx;\ |\eta| \leq \frac{1}{n} \},$$

so that $B_{5b} \leq \|\psi(\cdot,0)\|_{L^\infty(\mathbb{R}^d)}\varepsilon_0(n,S_0)$.
Since $u_0 \in L^1_{loc}(\mathbb{R}^d)$ and since S_0 is bounded, one has $\varepsilon_0(n,S_0) \to 0$ as $n\to\infty$. Then, $B_{5b} \to 0$ as $n\to\infty$.

Let us now prove that $B_{5a} \to 0$ as $n \to \infty$ (then, (29.14) will give (29.15) below). Note that $B_{5a} = -B_{5c} + (B_{5a} + B_{5c})$ with

$$B_{5c} = \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(\nu(y,s,\beta) \top u_0(y)) - f(\nu(y,s,\beta) \bot u_0(y))) \\ \mathbf{v}(y,s) \cdot \nabla \rho_n(x-y) \psi(x,0) \bar{\rho}_n(-\tau) d\tau dy dx ds d\beta.$$

Integrating by parts for the x variable yields

$$B_{5c} = \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(\nu(y,s,\beta) \top u_0(y)) - f(\nu(y,s,\beta) \bot u_0(y))) \\ \mathbf{v}(y,s) \cdot \nabla \psi(x,0) \rho_n(x-y) \bar{\rho}_n(-\tau) d\tau dy dx ds d\beta.$$

Noting that the integration with respect to s is reduced to $[0, 1/n]$, $B_{5c} \to 0$ as $n \to \infty$.
There remains to study $B_{5a} + B_{5c}$. Noting that $|f(a \top b) - f(a \top c)| \leq \bar{M}|b-c|$ and $|f(a \bot b) - f(a \bot c)| \leq \bar{M}|b-c|$ if b, $c \in [-\bar{D}, \bar{D}]$, where $\bar{D} = \|u_0\|_{L^\infty(\mathbb{R}^d)}$ and $\bar{M}$ is the Lipschitz constant to f on $[-\bar{D}, \bar{D}]$,

$$|B_{5a} + B_{5c}| \leq 2\bar{M} V \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty |u_0(x) - u_0(y)| |\nabla \rho_n(x-y)| \psi(x,0) \bar{\rho}_n(-\tau) d\tau dy dx ds,$$

which yields the existence of C_2, only depending on $\bar{M}$, V and ψ, such that

$$|B_{5a} + B_{5c}| \leq C_2 \int_0^{\frac{1}{n}} \int_{S_0} \int_{B(0,\frac{1}{n})} |u_0(x) - u_0(x-z)| n^{d+1} dz dx ds.$$

Therefore, $|B_{5a} + B_{5c}| \leq C_3 \varepsilon_0(n, S_0)$, with some C_3 only depending on $\bar{M}$, V and ψ. Since $\varepsilon_0(n, S_0) \to 0$ as $n \to \infty$, one deduces $|B_{5a} + B_{5c}| \to 0$ as $n \to \infty$. Hence, $B_{5a} \to 0$ as $n \to \infty$ and (29.14) yields

$$\limsup_{n \to \infty} A_5 \leq 0. \tag{29.15}$$

It is now possible to conclude Step 1. Passing to the limit as $n \to \infty$ in (29.8) and using (29.9), (29.10), (29.13) and (29.15) yields (29.6).

Step 2 (proof of $\mu = \nu$ and conclusion)
Let $R > 0$ and $T > 0$. One sets $\omega = VM$ (recall that V is given in Assumption 24.1 and that M is given in Step 1).
Let $\varphi \in C_c^1(\mathbb{R}_+, [0,1])$ be a function such that $\varphi(r) = 1$ if $r \in [0, R + \omega T]$, $\varphi(r) = 0$ if $r \in [R + \omega T + 1, \infty)$ and $\varphi'(r) \leq 0$, for all $r \in \mathbb{R}_+$.
One takes, in (29.6), ψ defined by

$$\begin{cases} \psi(x,t) = \varphi(|x| + \omega t) \frac{T-t}{T}, \text{ for } x \in \mathbb{R}^d \text{ and } t \in [0,T], \\ \psi(x,t) = 0, \text{ for } x \in \mathbb{R}^d \text{ and } t \geq T. \end{cases}$$

The function ψ is not in $C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$, but, using a usual regularization technique, it may be proved that such a function can be considered in (29.6), in which case Inequality (29.6) reads

$$\int_0^1 \int_0^1 \int_0^T \int_{\mathbb{R}^d} \Big[|\mu(x,t,\alpha) - \nu(x,t,\beta)| \Big(\frac{T-t}{T} \omega \varphi'(|x| + \omega t) - \frac{1}{T} \varphi(|x| + \omega t) \Big) + \\ \Big(f(\mu(x,t,\alpha) \top \nu(x,t,\beta)) - f(\mu(x,t,\alpha) \bot \nu(x,t,\beta)) \Big) \frac{T-t}{T} \varphi'(|x| + \omega t) \mathbf{v}(x,t) \cdot \frac{x}{|x|} \Big] dx dt d\alpha d\beta \geq 0.$$

Since $\omega = VM$ and $\varphi' \leq 0$, one has $(f(a \top b) - f(a \bot b)) \varphi'(|x| + \omega t) \mathbf{v}(x,t) \cdot (x/|x|) \leq |a-b| \omega(-\varphi'(|x| + \omega t))$, for a.e. $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+^\star$ and all a, $b \in [-D, D]$ (D is defined in Step 1). Therefore, since $\varphi(|x| + \omega t) = 1$ if $(x,t) \in B(0,R) \times [0,T]$, the preceding inequality gives

$$\int_0^1 \int_0^1 \int_0^T \int_{B(0,R)} |\mu(x,t,\alpha) - \nu(x,t,\beta)| dx dt d\alpha d\beta \leq 0,$$

which yields, since R and T are arbitrary, $\mu(x,t,\alpha) = \nu(x,t,\beta)$ for a.e. $(x,t,\alpha,\beta) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1) \times (0,1)$.

Let us now deduce also from this uniqueness result that there exists $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $\mu(x,t,\alpha) = u(x,t)$, for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$ (then it is easy to see, with Definition 29.1, that u is the entropy weak solution to Problem (24.1)-(24.2)).

Indeed, it is possible to take, in the preceeding proof, $\mu = \nu$ (recall that the proposition 29.2 gives the existence of an entropy process solution to Problem (24.1)-(24.2), see Remark 29.3). This yields $\mu(x,t,\alpha) = \mu(x,t,\beta)$ for a.e. $(x,t,\alpha,\beta) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1) \times (0,1)$. Then, for a.e. $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+^\star$, one has

$$\mu(x,t,\alpha) = \mu(x,t,\beta) \text{ for a.e. } (\alpha,\beta) \in (0,1) \times (0,1)$$

and, for a.e. $\alpha \in (0,1)$,

$$\mu(x,t,\alpha) = \mu(x,t,\beta) \text{ for a.e. } \beta \in (0,1).$$

Thus, defining u from $\mathbb{R}^d \times \mathbb{R}_+^\star$ to $\mathbb{R}$ by

$$u(x,t) = \int_0^1 \mu(x,t,\beta) d\beta,$$

one obtains $\mu(x,t,\alpha) = u(x,t)$, for a.e. $(x,t,\alpha) \in \mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1)$, and u is the entropy weak solution to Problem (24.1)-(24.2). This completes the proof of Theorem 29.1. ∎

29.3 Convergence towards the entropy weak solution

We now know that there exists a unique entropy process solution to problem (24.1)-(24.2) page 153, which is identical to the entropy weak solution of problem (24.1)-(24.2); we may now prove the convergence of the approximate solution given by the finite volume scheme (25.4), (25.2) and (25.5) towards the entropy weak solution as the mesh size tends to 0.

Theorem 29.2 *Under Assumptions 24.1 page 153, let $\alpha \in \mathbb{R}_+^\star$ and $\xi \in (0,1)$ be given. For an admissible mesh $\mathcal{T}$ in the sense of Definition 25.1 page 156 and for $k > 0$ satisfying (25.3) (note that α and ξ are fixed), let $u_{\mathcal{T},k}$ be the solution to (25.4), (25.2) and (25.5).*
Then, $u_{\mathcal{T},k} \to u$ in $L^p_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ for all $p \in [1,\infty)$, as $h = \text{size}(\mathcal{T}) \to 0$, where u is the entropy weak solution to (24.1)-(24.2) page 153.

Proof of Theorem 29.2

In order to prove that $u_{\mathcal{T},k} \to u$ (in $L^p_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ for all $p \in [1,\infty)$, as $h = \text{size}(\mathcal{T}) \to 0$), let us proceed by a classical way of contradiction which uses the uniqueness of the entropy process solution to Problem (24.1)-(24.2) page 153. Assume that there exists $1 \leq p_0 < \infty$, $\varepsilon > 0$, $\bar{\omega}$ a compact subset of $\mathbb{R}^d$, $T > 0$ and a sequence $((\mathcal{T}_m, k_m))_{m \in \mathbb{N}}$ such that, for any $m \in \mathbb{N}$, $\mathcal{T}_m$ is an admissible mesh, k_m satisfies (25.3) (with $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$, note that α and ξ are independent of m), $\text{size}(\mathcal{T}_m) \to 0$ as $m \to \infty$ and

$$\int_0^T \int_{\bar{\omega}} |u_{\mathcal{T}_m,k_m} - u|^{p_0} dx dt \geq \varepsilon, \ \forall m \in \mathbb{N}, \tag{29.16}$$

where $u_{\mathcal{T}_m,k_m}$ is the solution to (25.4), (25.2) and (25.5) with $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$ and u is the entropy weak solution to (24.1)-(24.2).

Using Proposition 29.2, there exists a subsequence of the sequence $((\mathcal{T}_m, k_m))_{m \in \mathbb{N}}$, still denoted by $((\mathcal{T}_m, k_m))_{m \in \mathbb{N}}$, and a function $\mu \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star \times (0,1))$ such that

1. $u_{\mathcal{T}_m,k_m} \to \mu$, as $m \to \infty$, in the nonlinear weak-$\star$ sense, that is:

$$\lim_{m\to\infty} \int_0^\infty \int_{\mathbb{R}^d} \theta(u_{\mathcal{T}_m,k_m}(x,t))\varphi(x,t)dxdt = \int_0^1 \int_0^\infty \int_{\mathbb{R}^d} \theta(\mu(x,t,\alpha))\varphi(x,t)dxdtd\alpha, \quad \forall \varphi \in L^1(\mathbb{R}^d \times \mathbb{R}_+^\star),\ \forall \theta \in C(\mathbb{R},\mathbb{R}), \tag{29.17}$$

2. μ is an entropy process solution to (24.1)-(24.2).

By Theorem 29.1 page 183, one has $\mu(\cdot,\cdot,\alpha) = u$, for a.e. $\alpha \in [0,1]$ (and u is the entropy weak solution to (24.1)-(24.2)). Taking first $\theta(s) = s^2$ in (29.17) and then $\theta(s) = s$ and φu instead of φ in (29.17) one obtains:

$$\int_0^\infty \int_{\mathbb{R}^d} (u_{\mathcal{T}_m,k_m}(x,t)) - u(x,t))^2 \varphi(x,t)dxdt \to 0, \text{ as } m \to \infty, \tag{29.18}$$

for any function $\varphi \in L^1(\mathbb{R}^d \times (0,T))$. From (29.18), and thanks to the L^∞-bound on $(u_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$, one deduces the convergence of $(u_{\mathcal{T}_m,k_m})_{m\in\mathbb{N}}$ towards u in $L^p_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$ for all $p \in [1,\infty)$, which is in contradiction with (29.16).
This completes the proof of our convergence theorem. ∎

Remark 29.4

1. Theorem 29.2 is also true with the implicit scheme instead of the explicit scheme (that is (25.6) and (25.7) instead of (25.4) and (25.5)) without the condition (25.3) (and thus without ξ).
2. The following section improves this convergence result and gives an error estimate.

30 Error estimate

30.1 Statement of the results

This section is devoted to the proof of an error estimate of time explicit and time implicit finite volume approximations to the solution $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ of Problem (24.1)-(24.2) page 153. Assuming that $u_0 \in BV(\mathbb{R}^d)$, a "$h^{1/4}$" error estimate is shown for a large variety of finite volume monotone flux schemes such as those which were presented in Section 25 page 156.

Under Assumption 24.1 page 153, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and $k > 0$. Let $g \in C(\mathbb{R}^2,\mathbb{R})$ satisfying Assumption 25.1.
Let u be the entropy weak solution of (24.1)-(24.2) and let $u_{\mathcal{T},k}$ be the solution of the time explicit scheme (25.4), (25.2), (25.5), assuming that (25.3) holds, or $u_{\mathcal{T},k}$ be the solution of the time implicit scheme (25.6), (25.2), (25.7). Our aim is to give an error estimate between u and $u_{\mathcal{T},k}$.

In the case of the explicit scheme, one proves, in this section, the following theorem.

Theorem 30.1 *Under Assumption 24.1 page 153, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and $k > 0$. Let $g \in C(\mathbb{R}^2,\mathbb{R})$ satisfy Assumption 25.1 and assume that condition (25.3) holds. Let u be the unique entropy weak solution of (24.1)-(24.2) and $u_{\mathcal{T},k}$ be given by (25.5), (25.4), (25.2). Assume $u_0 \in BV(\mathbb{R}^d)$. Then, for all $R > 0$ and all $T > 0$ there exists $C_e \in \mathbb{R}_+$, only depending on R, T, $\mathbf{v}$, g, u_0, α and ξ, such that the following inequality holds:*

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)|dxdt \le C_e h^{\frac{1}{4}}. \tag{30.1}$$

(Recall that $B(0,R) = \{x \in \mathbb{R}^d,\ |x| < R\}$.)

In Theorem 30.1, u_0 is assumed to belong to $BV(\mathbb{R}^d)$ (recall that $u_0 \in BV(\mathbb{R}^d)$ if $\sup\{\int u_0(x)\mathrm{div}\varphi(x)dx$, $\varphi \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)$; $|\varphi(x)| \leq 1$, $\forall x \in \mathbb{R}^d\} < \infty$). This assumption allows us to obtain an $h^{1/4}$ estimate in (30.1). If $u_0 \notin BV(\mathbb{R}^d)$ (but u_0 still belongs to $L^\infty(\mathbb{R}^d)$), one can also give an error estimate which depends on the functions $\varepsilon(r, S)$ and $\varepsilon_0(r, S)$ defined in (30.16) and (30.23).

A slight improvement of Theorem 30.1 (and also Theorem 30.2 below) is possible. Using the fact that $u \in C(\mathbb{R}_+, L^1_{loc}(\mathbb{R}^d))$ and thus $u(\cdot, t)$ is defined for all $t \in \mathbb{R}_+$, Theorem 30.1 remains true with

$$\int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)|dx \leq C_e h^{1/4}, \ \forall t \in [0,T],$$

instead of (30.1). The proof of such a result may be handled with an adaptation of the proof of a uniqueness of the entropy process solution given for instance in Eymard, Gallouët and Herbin [54], see Vila [155] and Cockburn, Coquel and LeFloch [32] for some similar results.

In some cases, it is possible to obtain $h^{1/2}$, instead of $h^{1/4}$, in Theorem 30.1. This is the case, for instance, when the mesh $\mathcal{T}$ is composed of rectangles ($d = 2$) and when $\mathbf{v}$ does not depend on (x, t), since, in this case, one obtains a "BV estimate" on $u_{\mathcal{T},k}$. In this case, the right hand sides of inequalities (26.4) and (26.5), proven above, are changed from $C/\sqrt{h}$ to C, so that the right hand side of (28.9) becomes Ch instead of $C\sqrt{h}$, which in turn yields $C_e h^{1/2}$ in (30.1) instead of $C_e h^{1/4}$. It is, however, still an open problem to know whether it is possible to obtain an error estimate with $h^{1/2}$, instead of $h^{1/4}$, in Theorem 30.1 (under the hypotheses of Theorem 30.1), even in the case where $\mathbf{v}$ does not depend on (x, t) (see Cockburn and Gremaud [34] for an attempt in this direction).

Remark 30.1 Theorem 30.1 (and also Theorem 30.2) remains true with some slightly more general assumption on g, instead of 25.1, in order to allow g to depend on $\mathcal{T}$ and k. Indeed, in (25.4), one can replace $g(u_K^n, u_L^n)$ (and $g(u_L^n, u_K^n)$) by $g_{K,L}(u_K^n, u_L^n, \mathcal{T}, k)$ (and $g_{L,K}(u_L^n, u_K^n, \mathcal{T}, k)$). Assume that, for all $K \in \mathcal{T}$ and all $L \in \mathcal{N}(K)$, the function $(a, b) \mapsto g_{K,L}(a, b, \mathcal{T}, k)$, from $[U_m, U_M]^2$ to $\mathbb{R}$, is nondecreasing with respect to a, nonincreasing with respect to b, Lipschitz continuous uniformly with respect to K and L and that $g_{K,L}(a, a, \mathcal{T}, k) = f(a)$ for all $a \in [U_m, U_M]$ (recall that $U_m \leq u_0 \leq U_M$ a.e. on $\mathbb{R}^d$). Then Theorem 30.1 remains true.

However, note that condition (25.3) and C_e in the estimate (30.1) of Theorem 30.1 depend on the Lipschitz constants of $g_{K,L}(\cdot, \cdot, \mathcal{T}, k)$ on $[U_m, U_M]^2$. An interesting form for $g_{K,L}$ is $g_{K,L}(a, b, \mathcal{T}, k) = c_{K,L}(\mathcal{T}, k)f(a)$ $+ (1-c_{K,L}(\mathcal{T}, k))\ f(b) + D_{K,L}(\mathcal{T}, k)\ (a-b)$, with some $c_{K,L}(\mathcal{T}, k) \in [0, 1]$ and $D_{K,L}(\mathcal{T}, k) \geq 0$. In order to obtain the desired properties on $g_{K,L}$, it is sufficient to take $\max\{|f'(s)|,\ s \in [U_m, U_M]\} \leq D_{K,L}(\mathcal{T}, k) \leq D$ (for all K, L), with some $D \in \mathbb{R}$. The Lipschitz constants of $g_{K,L}$ on $[U_m, U_M]^2$ only depend on D, f, U_m and U_M.

For instance, a "Lax-Friedrichs type" scheme consists, roughly speaking, in taking $D_{K,L}(\mathcal{T}, k)$ of order "h/k". The desired properties on $g_{K,L}$ are satisfied, provided that $k/h \leq C$, with some C depending on $\max\{|f'(s)|,\ s \in [U_m, U_M]\}$. Note, however, that the condition $k/h \leq C$ is not sufficient to give a real "$h^{1/4}$" estimate, since the coefficient C_e in (30.1) depends on D. Taking, for example, k of order "h^2" leads to an estimate "$C_e h^{1/4}$" which do not goes to 0 as h goes to 0 (indeed, it is known, in this case, that the approximate solution does not converge towards the entropy weak solution to (24.1)-(24.2)). One obtains a real "$h^{1/4}$" estimate, in the case of that "Lax-Friedrichs type" scheme, by taking $C_1 \leq (k/h) \leq C_2$. In order to avoid the condition $C_1 \leq (k/h)$ (note that $(k/h) \leq C_2$ is imposed by the Courant-Friedrichs-Levy condition 25.3), a possibility is to take $D_{K,L}(\mathcal{T}, k) = D = \max\{|f'(s)|,\ s \in [U_m, U_M]\}$ (this is related to the "modified Lax-Friedrichs " of Example 21.1 page 135 in the 1D case). Then D only depends on f and u_0 and, in the estimate "$C_e h^{1/4}$" of Theorem 30.1, C_e only depends on R, T, $\mathbf{v}$, f, u_0, α and ξ, which leads to a convergence result at rate "$h^{1/4}$" as $h \to 0$ (with fixed α and ξ).

In the case of the implicit scheme, one proves the following theorem.

Theorem 30.2 *Under Assumption 24.1 page 153, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and $k > 0$. Let $g \in C(\mathbb{R}^2, \mathbb{R})$ satisfy Assumption 25.1. Let u be the unique entropy weak solution of (24.1)-(24.2). Assume that $u_0 \in BV(\mathbb{R}^d)$ and that $\mathbf{v}$ does not depend on t.*

Let $\{u_K^n,\ n \in \mathbb{N},\ K \in \mathcal{T}\}$ be the unique solution to (25.6) and (25.2) such that $u_K^n \in [U_m, U_M]$ for all $K \in \mathcal{T}$ and $n \in \mathbb{N}$ (existence and uniqueness of such a solution is given by Proposition 27.1). Let $u_{\mathcal{T},k}$ be defined by (25.7).
Then, for all $R > 0$ and $T > 0$, there exists C_e, only depending on R, T, $\mathbf{v}$, g, u_0 and α, such that the following inequality holds:

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dx dt \leq C_e (k + h^{\frac{1}{2}})^{\frac{1}{2}}. \tag{30.2}$$

Remark 30.2 Note that, in Theorem 30.2, there is no restriction on k (this is usual for an implicit scheme), and one obtains an "$h^{1/4}$" error estimate for some "large" k, namely if $k \leq h^{1/2}$. In Theorem 30.2, if $\mathbf{v}$ depends on t and $u_0 \in L^\infty(\mathbb{R}^d)$ (but u_0 not necessarily in $BV(\mathbb{R}^d)$), one can also give an error estimate. Indeed one obtains

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dx dt \leq C_e (\frac{k}{h^{\frac{1}{2}}} + h^{\frac{1}{2}})^{\frac{1}{2}},$$

which yields an "$h^{1/4}$" error estimate if k is of order "h".

Theorem 30.1 (resp. Theorem 30.2) is an easy consequence of Theorem 28.1 (resp. 28.2) and of a quite general theorem of comparison between the entropy weak solution to (24.1)-(24.2) and an approximate solution. This theorem of comparison (Theorem 30.3) may be used in other frameworks (for instance, to compare the entropy weak solution to (24.1)-(24.2) and the approximate solution obtained with a parabolic regularization of (24.1)). It is stated and proved in Section 30.3 where the proofs of theorems 30.1 and 30.2 are also given. First, in Section 30.2, two preliminary lemmata are given. Indeed, Lemma 30.2 is the crucial part of the two following sections.

30.2 Preliminary lemmata

Let us first give a classical lemma on the space BV.

Lemma 30.1 *Let $u \in BV_{loc}(\mathbb{R}^p)$, $p \in \mathbb{N}^\star$, that is $u \in L^1_{loc}(\mathbb{R}^p)$ and the restriction of u to Ω belongs to $BV(\Omega)$ for all open bounded subset Ω of $\mathbb{R}^p$ (see Definition 21.19 page 141 for the definition of $BV(\Omega)$). Then, for all bounded subset Ω of $\mathbb{R}^p$ and for all $a > 0$,*

$$\|u(\cdot + \eta) - u\|_{L^1(\Omega)} \leq |\eta| |u|_{BV(\Omega_a)},\ \forall \eta \in \mathbb{R}^p,\ |\eta| \leq a, \tag{30.3}$$

where $\Omega_a = \{x \in \mathbb{R}^p;\ d(x,\Omega) < a\}$ and $d(x,\Omega) = \inf\{|x - y|,\ y \in \Omega\}$ is the distance from x to Ω.

Proof of Lemma 30.1
Let Ω be a bounded subset of $\mathbb{R}^p$ and $\eta \in \mathbb{R}^p$. The following equality classically holds:

$$\|u(\cdot + \eta) - u\|_{L^1(\Omega)} = \sup\{\int_\Omega (u(x+\eta) - u(x))\varphi(x) dx,\ \varphi \in C_c^\infty(\Omega, \mathbb{R}),\ \|\varphi\|_{L^\infty(\Omega)} \leq 1\}. \tag{30.4}$$

Let $\varphi \in C_c^\infty(\Omega, \mathbb{R})$ such that $\|\varphi\|_{L^\infty(\Omega)} \leq 1$.
Since $\varphi(x) = 0$ if $x \in \Omega_{|\eta|} \setminus \Omega$ (recall that $\Omega_{|\eta|} = \{x \in \mathbb{R}^p;\ d(x,\Omega) < \eta\}$),

$$\int_\Omega u(x)\varphi(x) dx = \int_{\Omega_{|\eta|}} u(x)\varphi(x) dx.$$

Similarly, using an obvious change of variables,

$$\int_\Omega u(x+\eta)\varphi(x) dx = \int_{\Omega_{|\eta|}} u(x)\varphi(x-\eta) dx.$$

Therefore,

$$\int_\Omega (u(x+\eta)-u(x))\varphi(x)dx = \int_{\Omega_{|\eta|}} u(x)(\varphi(x-\eta)-\varphi(x))dx = -\int_{\Omega_{|\eta|}} u(x)(\int_0^1 \nabla\varphi(x-s\eta)\cdot\eta ds)dx$$

and, with Fubini's theorem,

$$\int_\Omega (u(x+\eta)-u(x))\varphi(x)dx = \int_0^1 (\int_{\Omega_{|\eta|}} u(x)\nabla\varphi(x-s\eta)\cdot\eta dx)ds. \tag{30.5}$$

For all $s\in(0,1)$, Define $\psi_s\in C_c^\infty(\Omega_{|\eta|},\mathbb{R}^p)$ by $\psi_s(x)=\varphi(x-s\eta)\eta$; since $\psi_s\in C_c^\infty(\Omega_{|\eta|},\mathbb{R}^p)$ and $|\psi_s(x)|\le|\eta|$ for all $x\in\mathbb{R}^p$, the definition of $|u|_{BV(\Omega_{|\eta|})}$ yields

$$\int_{\Omega_{|\eta|}} u(x)\nabla\varphi(x-s\eta)\cdot\eta dx = \int_{\Omega_{|\eta|}} u(x)\mathrm{div}\psi_s(x)dx \le |\eta||u|_{BV(\Omega_{|\eta|})}.$$

Then, (30.5) gives

$$\int_\Omega (u(x+\eta)-u(x))\varphi(x)dx \le |\eta||u|_{BV(\Omega_{|\eta|})}. \tag{30.6}$$

Taking in (30.6) the supremum over $\varphi\in C_c^\infty(\Omega,\mathbb{R})$ such that $\|\varphi\|_{L^\infty(\Omega)}\le 1$ yields, thanks to (30.4),

$$\|u(\cdot+\eta)-u\|_{L^1(\Omega)} \le |\eta||u|_{BV(\Omega_{|\eta|})},\ \forall\eta\in\mathbb{R}^p,$$

and (30.3) follows, since $\Omega_{|\eta|}\subset\Omega_a$ if $|\eta|\le a$. ∎

Remark 30.3 Let us give an application of the lemma 30.1 which will be quite useful further on. Let $u\in BV_{loc}(\mathbb{R}^p)$, $p\in\mathbb{N}^\star$. Let ψ, $\varphi\in C_c(\mathbb{R}^p,\mathbb{R}_+)$, $a>0$ and $0<\varepsilon<a$ such that $\int_{\mathbb{R}^p}\varphi(x)dx=1$ and $\varphi(x)=0$ for all $x\in\mathbb{R}^p$, $|x|>\varepsilon$. Let $S=\{x\in\mathbb{R}^p,\ \psi(x)\neq 0\}$.
Then,

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} |u(x)-u(y)|\psi(x)\varphi(x-y)dydx \le \varepsilon\|\psi\|_{L^\infty(\mathbb{R}^p)}|u|_{BV(S_a)}, \tag{30.7}$$

where $S_a=\{x\in\mathbb{R}_p,\ d(x,S)<a\}$.
Indeed, Lemma 30.1 gives

$$\|u(\cdot+\eta)-u\|_{L^1(S)} \le |\eta||u|_{BV(S_a)},\ \forall\eta\in\mathbb{R}^p, |\eta|\le a. \tag{30.8}$$

Using a change of variables in the left hand side of (30.7),

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} |u(x)-u(y)|\psi(x)\varphi(x-y)dydx \le \|\psi\|_{L^\infty(\mathbb{R}^p)}\int_{B(0,\varepsilon)}(\int_S |u(x)-u(x-z)|dx)\varphi(z)dz.$$

Then, (30.8) yields

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} |u(x)-u(y)|\psi(x)\varphi(x-y)dydx \le \varepsilon\|\psi\|_{L^\infty(\mathbb{R}^p)}|u|_{BV(S_a)}\int_{\mathbb{R}^p}\varphi(z)dz,$$

which gives (30.7).

Lemma 30.2 *Under assumption 24.1, let $u_0 \in BV(\mathbb{R}^d)$ and $\tilde{u} \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $U_m \le \tilde{u} \le U_M$ a.e. on $\mathbb{R}^d \times \mathbb{R}_+^\star$. Assume that there exist $\mu \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}_+)$ and $\mu_0 \in \mathcal{M}(\mathbb{R}^d)$ such that*

$$\left\{\begin{array}{ll} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big(& |\tilde{u}(x,t)-\kappa|\varphi_t(x,t)+ \\ & (f(\tilde{u}(x,t)\top\kappa)-f(\tilde{u}(x,t)\bot\kappa))\mathbf{v}(x,t)\cdot\nabla\varphi(x,t)\Big)dxdt \quad + \\ \displaystyle\int_{\mathbb{R}^d} & |u_0(x)-\kappa|\varphi(x,0)dx \quad \ge \\ \displaystyle -\int_{\mathbb{R}^d\times\mathbb{R}_+} & \Big(|\varphi_t(x,t)|+|\nabla\varphi(x,t)|\Big)d\mu(x,t)-\displaystyle\int_{\mathbb{R}^d}|\varphi(x,0)|d\mu_0(x), \\ & \forall\kappa\in\mathbb{R},\quad \forall\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+). \end{array}\right. \tag{30.9}$$

Let u be the unique entropy weak solution of (24.1)-(24.2) (i.e. $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ is the unique solution to (30.9) with u instead of $\tilde{u}$ and $\mu = 0$, $\mu_0 = 0$).
Then for all $\psi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$ there exists C only depending on ψ (more precisely on $\|\psi\|_\infty$, $\|\psi_t\|_\infty$, $\|\nabla\psi\|_\infty$, and on the support of ψ), $\mathbf{v}$, f, and u_0, such that

$$\left\{\begin{array}{ll} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} & \Big[|\tilde{u}(x,t)-u(x,t)|\psi_t(x,t)+ \\ & \Big(f(\tilde{u}(x,t)\top u(x,t))-f(\tilde{u}(x,t)\bot u(x,t))\Big)\ (\mathbf{v}(x,t)\cdot\nabla\psi(x,t))\Big]\ dxdt \ \ge \\ & -C(\mu_0(\{\psi(\cdot,0)\neq 0\})+(\mu(\{\psi\neq 0\}))^{\frac{1}{2}}+\mu(\{\psi\neq 0\})), \end{array}\right. \tag{30.10}$$

where $\{\psi \neq 0\} = \{(x,t) \in \mathbb{R}^d \times \mathbb{R}_+,\ \psi(x,t) \neq 0\}$ and $\{\psi(\cdot,0) \neq 0\} = \{x \in \mathbb{R}^d,\ \psi(x,0) \neq 0\}$. (Note that $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(\mathbb{R}^d\times\mathbb{R}_+^\star)}$.)

Proof of Lemma 30.2

The proof of Lemma 30.2 is close to that of step 1 in the proof of Theorem 29.1. Let us first define mollifiers in $\mathbb{R}$ and $\mathbb{R}^d$. For $p = 1$ and $p = d$, one defines $\rho_p \in C_c^\infty(\mathbb{R}^p, \mathbb{R})$ satisfying the following properties:

$$\mathrm{supp}(\rho_p) = \{x \in \mathbb{R}^p;\ \rho_p(x) \neq 0\} \subset \{x \in \mathbb{R}^p; |x| \le 1\},$$

$$\rho_p(x) \ge 0,\ \forall x \in \mathbb{R}^p,$$

$$\int_{\mathbb{R}^p} \rho_p(x)dx = 1$$

and furthermore, for $p = 1$,

$$\rho_1(x) = 0,\ \forall x \in \mathbb{R}_+. \tag{30.11}$$

For $r \in \mathbb{R}$, $r \ge 1$, one defines $\rho_{p,r}(x) = r^p\rho_p(rx)$, for all $x \in \mathbb{R}^p$.
Using the mollifiers $\rho_{p,r}$ will allow to choose convenient test functions in (30.9) (which are the inequalities satisfied by $\tilde{u}$) and in the analogous inequalities satisfied by u which are

$$\left\{\begin{array}{l} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big[|u(y,s)-\kappa|\varphi_s(y,s)+\Big(f(u(y,s)\top\kappa)-f(u(y,s)\bot\kappa)\Big)\mathbf{v}(y,s)\cdot\nabla\varphi(y,s)\Big]dyds+ \\ \displaystyle\int_{\mathbb{R}^d}|u_0(y)-\kappa|\varphi(y,0)dy \ge 0,\quad \forall\kappa\in\mathbb{R},\quad \forall\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+). \end{array}\right. \tag{30.12}$$

Indeed, the main tool is to take $\kappa = u(y,s)$ in (30.9), $\kappa = \tilde{u}(x,t)$ in (30.12) and to introduce mollifiers in order to have y close to x and s close to t.
Let $\psi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$, and let $\varphi : (\mathbb{R}^d \times \mathbb{R}_+)^2 \to \mathbb{R}_+$ be defined by:

$$\varphi(x,t,y,s) = \psi(x,t)\rho_{d,r}(x-y)\rho_{1,r}(t-s).$$

Note that, for any $(y,s) \in \mathbb{R}^d \times \mathbb{R}_+$, one has $\varphi(\cdot,\cdot,y,s) \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$ and, for any $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$, one has $\varphi(x,t,\cdot,\cdot) \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$. Let us take $\varphi(\cdot,\cdot,y,s)$ as test function φ in (30.9) and $\varphi(x,t,\cdot,\cdot)$ as test function φ in (30.12). We take, in (30.9), $\kappa = u(y,s)$ and we take, in (30.12), $\kappa = \tilde{u}(x,t)$. We then integrate (30.9) for $(y,s) \in \mathbb{R}^d \times \mathbb{R}_+$, and (30.12) for $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$. Adding the two inequalities yields

$$E_{11} + E_{12} + E_{13} + E_{14} \geq -E_2, \tag{30.13}$$

where

$$E_{11} = \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[|\tilde{u}(x,t) - u(y,s)| \psi_t(x,t)\rho_{d,r}(x-y)\rho_{1,r}(t-s) \Big] dxdtdyds,$$

$$E_{12} = \int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big[\Big(f(\tilde{u}(x,t)\top u(y,s)) - f(\tilde{u}(x,t)\bot u(y,s)) \Big) \mathbf{v}(x,t) \cdot \nabla\psi(x,t)\rho_{d,r}(x-y)\rho_{1,r}(t-s) \Big] dxdtdyds,$$

$$E_{13} = -\int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} \Big(f(\tilde{u}(x,t)\top u(y,s)) - f(\tilde{u}(x,t)\bot u(y,s)) \Big) \psi(x,t) (\mathbf{v}(y,s) - \mathbf{v}(x,t)) \cdot \nabla\rho_{d,r}(x-y)\rho_{1,r}(t-s) dxdtdyds,$$

$$E_{14} = \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathbb{R}^d} |u_0(x) - u(y,s)| \psi(x,0)\rho_{d,r}(x-y)\rho_{1,r}(-s) dydsdx$$

and

$$E_2 = \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d \times \mathbb{R}_+} \Big(|\rho_{d,r}(x-y)(\psi_t(x,t)\rho_{1,r}(t-s) + \psi(x,t)\rho'_{1,r}(t-s))| + |\rho_{1,r}(t-s)(\nabla\psi(x,t)\rho_{d,r}(x-y) + \psi(x,t)\nabla\rho_{d,r}(x-y))| \Big) d\mu(x,t)dyds + \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\psi(x,0)\rho_{d,r}(x-y)\rho_{1,r}(-s)| d\mu_0(x)dyds. \tag{30.14}$$

One may be surprised by the fact that the inequation (30.13) is obtained without using the initial condition which is satisfied by the entropy weak solution u of (24.1)-(24.2). Indeed, this initial condition appears only in the third term of the left hand side of (30.12); since $\varphi(x,t,\cdot,0) = 0$ for all $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$, the third term of the left hand side of (30.12) is zero when $\varphi(x,t,\cdot,\cdot)$ is chosen as a test function in (30.12). However, the fact that u satisfies the initial condition of (24.1)-(24.2) will be used later in order to get a bound on E_{14}.

Let us now study the five terms of (30.13). One sets $S = \{\psi \neq 0\} = \{(x,t) \in \mathbb{R}^d \times \mathbb{R}_+;\ \psi(x,t) \neq 0\}$ and $S_0 = \{\psi(\cdot,0) \neq 0\} = \{x \in \mathbb{R}^d;\ \psi(x,0) \neq 0\}$. In the following, the notation C_i ($i \in \mathbb{N}$) will refer to various real quantities only depending on $\|\psi\|_\infty$, $\|\psi_t\|_\infty$, $\|\nabla\psi\|_\infty$, S, S_0, $\mathbf{v}$, f, and u_0.
Equality (30.14) leads to

$$E_2 \leq (r+1)C_1\mu(S) + C_2\mu_0(S_0). \tag{30.15}$$

Let us handle the term E_{11}. For all $x \in \mathbb{R}^d$ and for all $t \in \mathbb{R}_+$, one has, using (30.11),

$$\int_{\mathbb{R}^d} \int_0^\infty \rho_{d,r}(x-y)\rho_{1,r}(t-s) dsdy = 1.$$

Then,

$$|E_{11} - \int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big[|\tilde{u}(x,t)-u(x,t)|\psi_t(x,t)\Big]\,dxdt| \le$$
$$\int_0^\infty\int_{\mathbb{R}^d}\int_0^\infty\int_{\mathbb{R}^d}\Big[|u(x,t)-u(y,s)||\psi_t(x,t)|\rho_{d,r}(x-y)\rho_{1,r}(t-s)\Big]dxdtdyds \le \|\psi_t\|_\infty\varepsilon(r,S),$$

with

$$\varepsilon(r,S) = \sup\{\|u-u(\cdot+\eta,\cdot+\tau)\|_{L^1(S)},\ |\eta|\le\frac{1}{r},\ 0\le\tau\le\frac{1}{r}\}. \tag{30.16}$$

Since $u_0 \in BV(\mathbb{R}^d)$, the function u (entropy weak solution to (24.1)-(24.2)) belongs to $BV(\mathbb{R}^d\times(-T,T))$, for all $T>0$, setting, for instance, $u(.,t)=u_0$ for $t<0$ (see Krushkov [94] or Chainais-Hillairet [23] where this result is proven passing to the limit on numerical schemes).
Then, Lemma 30.1 gives, since $r\ge 1$, (taking $p=d+1$, $\Omega=S$ and $a=\sqrt{2}$ in Lemma 30.1,)

$$\varepsilon(r,S) \le \frac{C_3}{r}. \tag{30.17}$$

Hence,

$$|E_{11} - \int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big[|\tilde{u}(x,t)-u(x,t)|\psi_t(x,t)\Big]\,dxdt| \le \frac{C_4}{r}. \tag{30.18}$$

In the same way, using $|f(a\top b)-f(a\top c)|\le M|b-c|$ and $|f(a\bot b)-f(a\bot c)|\le M|b-c|$ for all a, b, $c\in[U_m,U_M]$ where M is the Lipschitz constant of f in $[U_m,U_M]$,

$$\begin{aligned}|E_{12} - \int_{\mathbb{R}_+}\int_{\mathbb{R}^d}\Big(f(\tilde{u}(x,t)\top u(x,t)) - f(\tilde{u}(x,t)\bot u(x,t))\Big)\\ (\mathbf{v}(x,t)\cdot\nabla\psi(x,t))dxdt| \le C_5\varepsilon(r,S) \le \tfrac{C_6}{r}.\end{aligned} \tag{30.19}$$

Let us now turn to E_{13}. We compare this term with

$$\begin{aligned}E_{13b} = -\int_0^\infty\int_{\mathbb{R}^d}\int_0^\infty\int_{\mathbb{R}^d}\Big(f(\tilde{u}(x,t)\top u(x,t)) - f(\tilde{u}(x,t)\bot u(x,t))\Big)\psi(x,t)\\ (\mathbf{v}(y,s)-\mathbf{v}(x,t))\cdot\nabla\rho_{d,r}(x-y)\rho_{1,r}(t-s)\,dxdtdyds.\end{aligned}$$

Since $\mathrm{div}(\mathbf{v}(\cdot,s)-\mathbf{v}(x,t))=0$ (on $\mathbb{R}^d$) for all $x\in\mathbb{R}^d$, $t\in\mathbb{R}_+$ and $s\in\mathbb{R}_+$, one has $E_{13b}=0$. Therefore, substracting E_{13b} from E_{13} yields

$$\begin{aligned}E_{13} \le C_7\int_0^\infty\int_{\mathbb{R}^d}\int_0^\infty\int_{\mathbb{R}^d}|u(x,t)-u(y,s)|\psi(x,t)\\ |(\mathbf{v}(y,s)-\mathbf{v}(x,t))\cdot\nabla\rho_{d,r}(x-y)|\rho_{1,r}(t-s)\,dxdtdyds.\end{aligned} \tag{30.20}$$

The right hand side of (30.20) is then smaller than $C_8\varepsilon(r,S)$, since $|(\mathbf{v}(y,s)-\mathbf{v}(x,t))\cdot\nabla\rho_{d,r}(x-y)|$ is bounded by C_9r^d (noting that $|x-y|\le 1/r$). Then, with (30.17), one has

$$E_{13} \le \frac{C_{10}}{r}. \tag{30.21}$$

In order to estimate E_{14}, let us take in (30.12), for $x\in\mathbb{R}^d$ fixed, $\varphi=\varphi(x,\cdot,\cdot)$, with

$$\varphi(x,y,s) = \psi(x,0)\rho_{d,r}(x-y)\int_s^\infty\rho_{1,r}(-\tau)d\tau,$$

and $\kappa=u_0(x)$. Note that $\varphi(x,\cdot,\cdot)\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}_+,\mathbb{R}_+)$. We then integrate the resulting inequality with respect to $x\in\mathbb{R}^d$. We get

$$-E_{14} + E_{15} + E_{16} \geq 0,$$

with

$$E_{15} = -\int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(u(y,s)\top u_0(x)) - f(u(y,s)\perp u_0(x))) \\ \mathbf{v}(y,s) \cdot (\psi(x,0)\nabla\rho_{d,r}(x-y))\rho_{1,r}(-\tau)d\tau dydxds,$$

$$E_{16} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_0^\infty \psi(x,0)\rho_{d,r}(x-y)\rho_{1,r}(-\tau)|u_0(x) - u_0(y)|d\tau dydx.$$

To bound E_{15}, one introduces E_{15b} defined as

$$E_{15b} = \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(u(y,s)\top u_0(y)) - f(u(y,s)\perp u_0(y))) \\ (\mathbf{v}(y,s) \cdot \nabla\rho_{d,r}(x-y))\psi(x,0)\rho_{1,r}(-\tau)d\tau dydxds.$$

Integrating by parts for the x variable yields

$$E_{15b} = -\int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty (f(u(y,s)\top u_0(y)) - f(u(y,s)\perp u_0(y))) \\ (\mathbf{v}(y,s) \cdot \nabla\psi(x,0))\rho_{d,r}(x-y)\rho_{1,r}(-\tau)d\tau dydxds.$$

Then, noting that the time support of this integration is reduced to $s \in [0, 1/r]$, one has

$$E_{15b} \leq \frac{C_{11}}{r}. \tag{30.22}$$

Furthermore, one has

$$|E_{15} + E_{15b}| \leq C_{12} \int_0^\infty \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_s^\infty |u_0(x) - u_0(y)||\mathbf{v}(y,s) \cdot \nabla\rho_{d,r}(x-y)|\psi(x,0)\rho_{1,r}(-\tau)d\tau dydxds,$$

which is bounded by $C_{13}\varepsilon_0(r, S_0)$, since the time support of the integration is reduced to $s \in [0, 1/r]$, where $\varepsilon_0(r, S_0)$ is defined by

$$\varepsilon_0(r, S_0) = \sup\{\int_{S_0} |u_0(x) - u_0(x+\eta)|dx;\ |\eta| \leq \frac{1}{r}\}. \tag{30.23}$$

Since $u_0 \in BV(\mathbb{R}^d)$, one has (thanks to Lemma 30.1) $\varepsilon_0(r, S_0) \leq C_{14}/r$ and therefore, with (30.22), $E_{15} \leq C_{15}/r$.

Since $u_0 \in BV(\mathbb{R}^d)$, again thanks to Lemma 30.1, see remark 30.3, the term E_{16} is also bounded by C_{16}/r.

Hence, since $E_{14} \leq E_{15} + E_{16}$,

$$E_{14} \leq \frac{C_{17}}{r}. \tag{30.24}$$

Using (30.13), (30.15), (30.18), (30.19),(30.21), (30.24), one obtains

$$\int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \Big[|\tilde{u}(x,t) - u(x,t)|\psi_t(x,t) + \\ \Big(f(\tilde{u}(x,t)\top u(x,t)) - f(\tilde{u}(x,t)\perp u(x,t))\Big)\ (\mathbf{v}(x,t) \cdot \nabla\psi(x,t))\Big]\ dxdt\ \geq \\ -C_1(r+1)\mu(S) - C_2\mu_0(S_0) - \tfrac{C_{18}}{r},$$

which, taking $r = 1/\sqrt{\mu(S)}$ if $0 < \mu(S) \leq 1$ ($r \to \infty$ if $\mu(S) = 0$ and $r = 1$ if $\mu(S) > 1$), gives (30.10). This concludes the proof of the lemma 30.2. ∎

30.3 Proof of the error estimates

Let us now prove a quite general theorem of comparison between the entropy weak solution to (24.1)-(24.2) and an approximate solution, from which theorems 30.1 and 30.2 will be deduced.

Theorem 30.3 *Under assumption 24.1, let $u_0 \in BV(\mathbb{R}^d)$ and $\tilde{u} \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ such that $U_m \leq \tilde{u} \leq U_M$ a.e. on $\mathbb{R}^d \times \mathbb{R}_+^\star$. Assume that there exist $\mu \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}_+)$ and $\mu_0 \in \mathcal{M}(\mathbb{R}^d)$ such that (30.9) holds. Let u be the unique entropy weak solution of (24.1)-(24.2) (note that $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ is solution to (30.9) with u instead of $\tilde{u}$ and $\mu = 0$, $\mu_0 = 0$).*
Then, for all $R > 0$ and all $T > 0$ there exists C_e and $\bar{R}$, only depending on R, T, $\mathbf{v}$, f and u_0, such that the following inequality holds:

$$\int_0^T \int_{B(0,R)} |\tilde{u}(x,t) - u(x,t)| dx dt \leq C_e(\mu_0(B(0,\bar{R})) + [\mu(B(0,\bar{R}) \times [0,T])]^{\frac{1}{2}} + \mu(B(0,\bar{R}) \times [0,T])).$$

Recall that $B(0,R) = \{x \in \mathbb{R}^d;\ |x| < R\}$.

Proof of Theorem 30.3

The proof of Theorem 30.3 is close to that of Step 2 in the proof of Theorem 29.1. It uses Lemma 30.2 page 192, the proof of which is given in section 30.2 above.
Let $R > 0$ and $T > 0$. One sets $\omega = VM$, where V is given in Assumption 24.1 and M is the Lipschitz constant of f in $[U_m, U_M]$ (indeed, since $f \in C^1(\mathbb{R}, \mathbb{R})$, one has $M = \sup\{|f'(s)|;\ s \in [U_m, U_M]\}$).
Let $\rho \in C_c^1(\mathbb{R}_+, [0,1])$ be a function such that $\rho(r) = 1$ if $r \in [0, R + \omega T]$, $\rho(r) = 0$ if $r \in [R + \omega T + 1, \infty)$ and $\rho'(r) \leq 0$, for all $r \in \mathbb{R}_+$ (ρ only depends on R, T, $\mathbf{v}$, f and u_0).
One takes, in (30.10), ψ defined by

$$\begin{cases} \psi(x,t) = \rho(|x| + \omega t)\frac{T-t}{T}, \text{ for } x \in \mathbb{R}^d \text{ and } t \in [0,T], \\ \psi(x,t) = 0, \text{ for } x \in \mathbb{R}^d \text{ and } t \geq T. \end{cases}$$

Note that $\rho(|x| + \omega t) = 1$, if $(x,t) \in B(0,R) \times [0,T]$.
The function ψ is not in $C_c^\infty(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}_+)$, but, using a usual regularization technique, it may be proved that such a function can be considered in (30.10), in which case Inequality (30.10) reads, with $\bar{R} = R + \omega T + 1$,

$$\int_0^T \int_{\mathbb{R}^d} \Big[|\tilde{u}(x,t) - u(x,t)| \Big(\frac{T-t}{T} \omega \rho'(|x| + \omega t) - \frac{1}{T} \rho(|x| + \omega t) \Big) + \Big(f(\tilde{u}(x,t) \top u(x,t)) - f(\tilde{u}(x,t) \perp u(x,t)) \Big) \frac{T-t}{T} \rho'(|x| + \omega t)(\mathbf{v}(x,t) \cdot \frac{x}{|x|}) \Big] dx dt \geq -C(\mu_0(B(0,\bar{R})) + (\mu(B(0,\bar{R}) \times [0,T]))^{\frac{1}{2}} + \mu(B(0,\bar{R}) \times [0,T])),$$

where C only depends on R, T, $\mathbf{v}$, f and u_0.
Since $\omega = VM$ and $\rho' \leq 0$, one has

$$\Big(f(\tilde{u}(x,t) \top u(x,t)) - f(\tilde{u}(x,t) \perp u(x,t)) \Big) \frac{T-t}{T} \rho'(|x| + \omega t)(\mathbf{v}(x,t) \cdot \frac{x}{|x|}) \Big) \leq |\tilde{u}(x,t) - u(x,t)| \frac{T-t}{T} \omega(-\rho'(|x| + \omega t)),$$

and therefore, since $\rho(|x| + \omega t) = 1$, if $(x,t) \in B(0,R) \times [0,T]$,

$$\int_0^T \int_{B(0,R)} |\tilde{u}(x,t) - u(x,t)| dx dt \leq CT(\mu_0(B(0,\bar{R})) + (\mu(B(0,\bar{R}) \times [0,T]))^{\frac{1}{2}} + \mu(B(0,\bar{R}) \times [0,T])).$$

This completes the proof of Theorem 30.3. ∎

Let us now conclude with the proofs of theorems 30.1 page 188 (which gives an error estimate for the time explicit numerical scheme (25.4), (25.2) page 157) and 30.2 page 189 (which gives an error estimate for the time implicit numerical scheme (25.6), (25.2) page 157). There are easy consequences of theorems 28.1 and 28.2 and of Theorem 30.3.

PROOF of Theorem 30.1

Under the assumptions of Theorem 30.1, let $\tilde{u} = u_{\mathcal{T},k}$. Thanks to the L^∞ estimate on $u_{\mathcal{T},k}$ (Lemma 26.1) and to Theorem 28.1, $\tilde{u} = u_{\mathcal{T},k}$ satisfies the hypotheses of Theorem 30.3 with $\mu = \mu_{\mathcal{T},k}$ and $\mu_0 = \mu_{\mathcal{T}}$ (the measures $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{T}}$ are given in Theorem 28.1).
Let $R > 0$ and $T > 0$. Then, Theorem 30.3 gives the existence of C_1 and $\bar{R}$, only depending on R, T, $\mathbf{v}$, f and u_0, such that

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dxdt \leq C_1(\mu_{\mathcal{T}}(B(0,\bar{R})) + [\mu_{\mathcal{T},k}(B(0,\bar{R}) \times [0,T])]^{\frac{1}{2}} + \mu_{\mathcal{T},k}(B(0,\bar{R}) \times [0,T])). \tag{30.25}$$

For h small enough, say $h \leq R_0$, one has $h < \bar{R}$ and $k < T$ (thanks to condition 25.3, note that R_0 only depends on R, T, $\mathbf{v}$, g, u_0, α and ξ).
Then, for $h < R_0$, Theorem 28.1 gives, with (30.25),

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dxdt \leq C_1(Dh + \sqrt{C}h^{\frac{1}{4}} + C\sqrt{h}) \leq C_2 h^{\frac{1}{4}},$$

where C_2 only depends on R, T, $\mathbf{v}$, g, u_0, α and ξ.
This gives the desired estimate (30.1) of Theorem 30.1 for $h < R_0$.
There remains the case $h \geq R_0$. This case is trivial since, for $h \geq R_0$,

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dxdt \leq 2\max\{-U_m, U_M\} \mathrm{m}(B(0,R) \times (0,T)) \leq C_3 (R_0)^{\frac{1}{4}} \leq C_3 h^{\frac{1}{4}},$$

for some C_3 only depending on R, T, $\mathbf{v}$, g, u_0, α and ξ.
This completes the proof of Theorem 30.1. ∎

PROOF of Theorem 30.2

The proof of Theorem 30.2 is very similar to that of Theorem 30.1 and we follow the proof of Theorem 30.1.
Under the assumptions of Theorem 30.2, using Theorem 28.2 instead of Theorem 28.1 gives that $\tilde{u} = u_{\mathcal{T},k}$ satisfies the hypotheses of Theorem 30.3 with $\mu = \mu_{\mathcal{T},k}$ and $\mu_0 = \mu_{\mathcal{T}}$ (the measures $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{T}}$ are given in Theorem 28.2).
Let $R > 0$ and $T > 0$. Theorem 30.3 gives the existence of C_1 and $\bar{R}$, only depending on R, T, $\mathbf{v}$, f and u_0, such that (30.25) holds.
For $h < \bar{R}$ and $k < T$ Theorem 28.1 gives with (30.25),

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dxdt \leq C_1(Dh + \sqrt{C}(k + h^{\frac{1}{2}})^{\frac{1}{2}} + C(k + h^{\frac{1}{2}})) \leq C_2(k + h^{\frac{1}{2}})^{\frac{1}{2}},$$

where C_2 only depends on R, T, $\mathbf{v}$, g, u_0, α.
This gives the desired estimate (30.2) of Theorem 30.2 for $h < \bar{R}$ and $k < T$.
There remains the cases $h \geq \bar{R}$ and $k \geq T$. These cases are trivial since

$$\int_0^T \int_{B(0,R)} |u_{\mathcal{T},k}(x,t) - u(x,t)| dxdt \leq 2\max\{-U_m, U_M\} \mathrm{m}(B(0,R) \times (0,T)) \leq C_3 \inf\{\bar{R}^{\frac{1}{4}}, T^{\frac{1}{2}}\}$$

for some C_3 only depending on R, T, $\mathbf{v}$, g, u_0.
This completes the proof of Theorem 30.2. ∎

30.4 Remarks and open problems

Theorem 30.1 page 188 gives an error estimate of order $h^{1/4}$ for the approximate solution of a nonlinear hyperbolic equation of the form $u_t + \mathrm{div}(\mathbf{v}f(u)) = 0$, with initial data in $L^\infty \cap BV$ by the explicit finite volume scheme (25.4) and (25.2) page 157, under a usual CFL condition $k \leq Ch$ (see (25.3) page 157).

Note that, in fact, the same estimate holds if u_0 is only locally BV. More generally, if the initial data u_0 is only in L^∞, then one still obtains an error estimate in terms of the quantities

$$\varepsilon(r, S) = \sup\{\int_S |u(x,t) - u(x+\eta, t+\tau)|dxdt;\ |\eta| \leq \frac{1}{r},\ 0 \leq \tau \leq \frac{1}{r}\}$$

and

$$\varepsilon_0(r, S_0) = \sup\{\int_{S_0} |u_0(x) - u_0(x+\eta)|dx;\ |\eta| \leq \frac{1}{r}\},$$

see (30.16) page 194 and (30.23) page 195. This is again an obvious consequence of Theorem 28.1 page 174 and Theorem 30.3 page 196.

We also considered the implicit schemes, which seem to be much more widely used in industrial codes in order to ensure their robustness. The implicit case required additional work in order
(i) to prove the existence of the solution to the finite volume scheme,
(ii) to obtain the "strong time BV" estimate (27.14) if $\mathbf{v}$ does not depend on t.
For $\mathbf{v}$ depending on t, Remark 30.2 yields an estimate of order $h^{1/4}$ if k behaves as h; however, in the case where $\mathbf{v}$ does not depend on t, then an estimate of order $h^{1/4}$ is obtained (in Theorem 30.2) for a behaviour of k as $\sqrt{h}$; Indeed, recent numerical experiments suggest that taking k of the order of $\sqrt{h}$ yields results of the same precision than taking k of the order of h, with an obvious reduction of the computational cost.

Note that the method described here may also be extended to higher order schemes for the same equation, see Chainais-Hillairet [22]; other methods have been used for error estimates for higher order schemes with a nonlinearity of the form $F(u)$, as in Noëlle [117]. However, it is still an open problem, to our knowledge, to improve the order of the error estimate in the case of higher order schemes.

31 Boundary conditions

In this section, a generalization of Theorem 23.1 is presented for the multidimensional scalar case together with a rough sketch of proof. For the sake of simplicity, one considers $d = 2$ (the extension to $d = 3$ is straightforward) and a flux function under the form $v(x,t)f(u)$, with $\mathrm{div}(v(\cdot,t)) = 0$ (see [157] for the general case of a flux function $f(x,t,u)$). This leads to the following equation:

$$u_t + \mathrm{div}(vf(u)) = 0,\ \text{in } \Omega \times (0,T), \tag{31.1}$$

where Ω is a bounded polygonal open set of $\mathbf{R}^2$, $T > 0$, $f \in C^1(\mathbf{R}, \mathbf{R})$ (or $f : \mathbf{R} \to \mathbf{R}$ Lipschitz continuous) and $v \in C^1(\mathbf{R}^2 \times [0,T]) \to \mathbf{R}^2$ with $\mathrm{div}(v(\cdot,t)) = 0$ in $\mathbf{R}^2$ for all $t \in [0,T]$. The unknown is $u : \Omega \times (0,T) \to \mathbf{R}$.

Let $u_0 \in L^\infty(\Omega)$ and $\overline{u} \in L^\infty(\partial\Omega \times (0,T))$. Let $A, B \in \mathbf{R}$ be such that $A \leq u_0 \leq B$ a.e. on Ω and $A \leq \overline{u} \leq B$ a.e. on $\partial\Omega \times (0,T)$. Following the work of [122], an entropy weak solution of (31.1) with the initial condition u_0 and the (weak) boundary condition $\overline{u}$ is a solution of (31.2):

$$
\begin{aligned}
&u \in L^\infty(\Omega \times (0,T)),\\
&\int_0^T \int_\Omega [(u-\kappa)^\pm \varphi_t + \mathrm{sign}_\pm(u-\kappa)(f(u)-f(\kappa))v \cdot \mathrm{grad}\varphi] dxdt\\
&\qquad + M \int_0^T \int_{\partial\Omega} (\overline{u}(t)-\kappa)^\pm \varphi(x,t) d\gamma(x) dt\\
&\qquad + \int_\Omega (u_0-\kappa)^\pm \varphi(x,0) dx \geq 0,\\
&\forall \kappa \in [A,B],\ \forall \varphi \in C_c^1(\overline{\Omega} \times [0,T), \mathbf{R}_+),
\end{aligned}
\tag{31.2}
$$

where $d\gamma(x)$ stands for the integration with respect to the one dimensional Lebesgue measure on the boundary of Ω and M is such that $\|v\|_\infty |f(s_1)-f(s_2)| \leq M|s_1-s_2|$ for all $s_1, s_2 \in [A,B]$, where$\|v\|_\infty = \sup_{(x,t)\in\Omega\times[0,T]} |v(x,t)|$ (and $|\cdot|$ denotes here the Euclidean norm in $\mathbf{R}^2$).

Remark 31.1

1. *If u satisfies the family of inequalities (31.2), it is possible to prove that u is a solution of (31.1) (on a weak form), u satisfies some entropy inequalities in $\Omega\times(0,T)$, namely $|u-\kappa|_t+\mathrm{div}(v(f(\max(u,\kappa))-f(\min(u,\kappa)))) \leq 0$ for all $\kappa \in \mathbf{R}$, but also on the boundary of $\partial\Omega$ and on $t=0$. u satisfies the initial condition ($u(\cdot,0)=u_0$) and u satisfies partially the boundary condition. For instance, if $f'>0$ and u is regular enough, then $u(x,t)=\overline{u}(x,t)$ if $x \in \partial\Omega$, $t \in (0,T)$ and $v(x,t)\cdot n(x,t) < 0$, where n is the outward normal vector to $\partial\Omega$.*
2. *Let $\overline{M} \geq 1$. It is interesting to remark that u is solution of (31.2) if and only if u is solution of (31.2) where the term $\int_\Omega (u_0-\kappa)^\pm\varphi(x,0)dx$ is replaced by $\overline{M}\int_\Omega (u_0-\kappa)^\pm\varphi(x,0)dx$.*

A sketch of proof of existence and uniqueness of the solution of (31.2) together with the convergence of numerical approximations is now given, following [157].

Step 1: Approximate solution. With a quite general mesh of Ω (with triangles, for instance), denoted by $\mathcal{T}$, and a time step k, it is possible to define an approximate solution, denoted by $u_{\mathcal{T},k}$, using some numerical fluxes (on the edges of the mesh) satisfying conditions similar to (C1)-(C3) in Sect. 23.1. Under a so called CFL condition (like $k \leq (1-\zeta)\frac{h}{L}$ in Sect. 23.1), it is easy to prove that $A \leq u_{\mathcal{T},k} \leq B$ a.e. on $\Omega\times(0,T)$. Unfortunately, it does not seem easy to obtain directly a strong compactness result on the familly of approximate solutions (alhough this strong compactness result is true, as we shall see below).

Step 2: Weak compactness. Using only this L^∞ bound on $u_{\mathcal{T},k}$, one can assume (for convenient subsequences of sequences of approximate solutions) that $u_{\mathcal{T},k} \to u$, as the mesh size goes to zero (with the CFL condition), in a "non linear weak-$\star$ sense" (similar to the convergence towards young measures, see [53] for instance), that is $u \in L^\infty(\Omega\times(0,T)\times(0,1))$ and

$$
\int_0^T \int_\Omega g(u_{\mathcal{T},k}(x,t))\varphi(x,t)dxdt \to \int_0^1 \int_0^T \int_\Omega g(u(x,t,\alpha))\varphi(x,t)dxdtd\alpha,
$$
$$
\text{for all } \varphi \in L^1(\Omega\times(0,T)).
$$

Step 3: Passing to the limit. Using the monotonicity of the numerical fluxes, the approximate solutions satisfy some discrete entropy inequalities. Passing to the limit in these inequalities gives that u (defined in Step 2) is solution of some inequalities which are very similar to (31.2), namely:

$$
\begin{aligned}
&u \in L^\infty(\Omega \times (0,T) \times (0,1)),\\
&\int_0^1 \int_0^T \int_\Omega [(u-\kappa)^\pm \varphi_t + \mathrm{sign}_\pm(u-\kappa)(f(u)-f(\kappa))v \cdot \mathrm{grad}\varphi] dxdtd\alpha\\
&\qquad + M \int_0^T \int_{\partial\Omega} (\overline{u}(t)-\kappa)^\pm \varphi(x,t) d\gamma(x) dt\\
&\qquad + \int_\Omega (u_0-\kappa)^\pm \varphi(x,0) dx \geq 0,\\
&\forall \kappa \in [A,B],\ \forall \varphi \in C_c^1(\overline{\Omega} \times [0,T), \mathbf{R}_+),
\end{aligned}
\tag{31.3}
$$

For this step, one chooses M not only greater than the Lipschitz constant of $\|v\|_\infty f$ on $[A,B]$, but also greater than the Lipschitz constant (on $[A,B]^2$) of the numerical fluxes associated to the edges of the meshes (the equivalent of L in Theorem 23.1). This choice of M is possible since the unique solution of (31.2) does not depend on M provided that M is greater than the Lipschitz constant of $\|v\|_\infty f$ on $[A,B]$ and since it is possible to choose numerical fluxes (namely, Godunov flux, for instance) such as the Lipschitz constant of these numerical fluxes is bounded by the Lipschitz constant of $\|v\|_\infty f$ (then, the present method leads to an existence result with M only greater than the Lipschitz constant of $\|v\|_\infty f$ on $s \in [A,B]$, passing to the limit on approximate solutions given with these numerical fluxes).

Step 4: Uniqueness of the solution of (31.3). In this step, the "doubling variables" method of Krushkov is used to prove the uniqueness of the solution of (31.3). Indeed, if u and w are two solutions of (31.3), the doubling variables method leads to:

$$\begin{aligned}&\int_0^1\int_0^1\int_0^T\int_\Omega |u(x,t,\alpha)-w(x,t,\beta)|\varphi_t\,dxdtd\alpha d\beta\\ &+\int_0^1\int_0^1\int_0^T\int_\Omega (f(\max(u,w))-f(\min(u,w)))v\cdot \mathrm{grad}\varphi\,dxdtd\alpha d\beta \geq 0\\ &\qquad\qquad \forall \varphi \in C_c^1(\overline{\Omega}\times[0,T),\mathbf{R}_+),\end{aligned} \tag{31.4}$$

Taking $\varphi(x,t)=(T-t)^+$ in (31.4) (which is, indeed, possible) gives that u does not depend on α, v does not depend on β and $u=v$ a.e. on $\Omega\times(0,T)$. As a result, u is also the unique solution of (31.2).

Step 5: Conclusion. Step 4 gives, in particular, the uniqueness of the solution of (31.2). It gives also that the non linear weak-$\star$ limit of sequences of approximate solutions is solution of (31.2) and, therefore, the existence of the solution of (31.2). Furthermore, since the non linear weak-$\star$ limit of sequences of approximate solution does not depend on α, it is quite easy to deduce that this limit is "strong" in $L^p(\Omega\times(0,T))$ for any $p\in[1,\infty)$ (see [53], for instance) and, thanks to the uniqueness of the limit, the convergence holds without extraction of subsequences.

32 Nonlinear weak-$\star$ convergence

The notion of nonlinear weak-$\star$ convergence was used in Section 29.3. We give here the definition of this type of convergence and we prove that a bounded sequence of L^∞ converges, up to a subsequence, in the nonlinear weak-$\star$ sense.

Definition 32.1 (Nonlinear weak-$\star$ convergence)
Let Ω be an open subset of $\mathbb{R}^N$ $(N\geq 1)$, $(u_n)_{n\in\mathbb{N}}\subset L^\infty(\Omega)$ and $u\in L^\infty(\Omega\times(0,1))$. The sequence $(u_n)_{n\in\mathbb{N}}$ converges towards u in the "nonlinear weak-$\star$ sense" if

$$\begin{aligned}&\int_\Omega g(u_n(x))\varphi(x)dx \to \int_0^1\int_\Omega g(u(x,\alpha))\varphi(x)dxd\alpha,\ \text{as } n\to+\infty,\\ &\qquad\qquad \forall\varphi\in L^1(\Omega),\ \forall g\in C(\mathbb{R},\mathbb{R}).\end{aligned} \tag{32.1}$$

Remark 32.1 Let Ω be an open subset of $\mathbb{R}^N$ $(N\geq 1)$, $(u_n)_{n\in\mathbb{N}}\subset L^\infty(\Omega)$ and $u\in L^\infty(\Omega\times(0,1))$ such that $(u_n)_{n\in\mathbb{N}}$ converges towards u in the nonlinear weak-$\star$ sense. Then, in particular, the sequence $(u_n)_{n\in\mathbb{N}}$ converges towards v in $L^\infty(\Omega)$, for the weak-$\star$ topology, where v is defined by

$$v(x)=\int_0^1 u(x,\alpha)d\alpha,\ \text{for a.e. } x\in\Omega.$$

Therefore, the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$ (thanks to the Banach-Steinhaus theorem). The following proposition gives that, up to a subsequence, a bounded sequence of $L^\infty(\Omega)$ converges in the nonlinear weak-$\star$ sense.

Proposition 32.1 *Let Ω be an open subset of $\mathbb{R}^N$ ($N \geq 1$) and $(u_n)_{n\in\mathbb{N}}$ be a bounded sequence of $L^\infty(\Omega)$. Then there exists a subsequence of $(u_n)_{n\in\mathbb{N}}$, which will still be denoted by $(u_n)_{n\in\mathbb{N}}$, and a function $u \in L^\infty(\Omega \times (0,1))$ such that the subsequence $(u_n)_{n\in\mathbb{N}}$ converges towards u in the nonlinear weak-$\star$ sense.*

Proof

This proposition is classical in the framework of "Young measures" and we only sketch the proof for the sake of completeness.

Let $(u_n)_{n\in\mathbb{N}}$ be a bounded sequence of $L^\infty(\Omega)$ and $r \geq 0$ such that $\|u_n\|_{L^\infty(\Omega)} \leq r, \forall n \in \mathbb{N}$.

Step 1 (diagonal process)

Thanks to the separability of the set of continuous functions defined from $[-r, r]$ into $\mathbb{R}$ (this set is endowed with the uniform norm) and the sequential weak-$\star$ relative compactness of the bounded sets of $L^\infty(\Omega)$, there exists (using a diagonal process) a subsequence, which will still be denoted by $(u_n)_{n\in\mathbb{N}}$, such that, for any function $g \in C(\mathbb{R}, \mathbb{R})$, the sequence $(g(u_n))_{n\in\mathbb{N}}$ converges in $L^\infty(\Omega)$ for the weak-$\star$ topology towards a function $\mu_g \in L^\infty(\Omega)$.

Step 2 (Young measure)

In this step, we prove the existence of a family $(m_x)_{x\in\Omega}$ such that

1. for all $x \in \Omega$, m_x is a probability on $\mathbb{R}$ whose support is included in $[-r, +r]$ (i.e. m_x is a σ-additive application from the Borel σ-algebra of $\mathbb{R}$ in $\mathbb{R}_+$ such that $m_x(\mathbb{R}) = 1$ and $m_x(\mathbb{R} \setminus [-r, r]) = 0$),
2. $\mu_g(x) = \int_{\mathbb{R}} g(s)dm_x(s)$ for a.e. $x \in \Omega$ and for all $g \in C(\mathbb{R}, \mathbb{R})$.

The family $m = (m_x)_{x\in\Omega}$ is called a "Young measure".

Let us first claim that it is possible to define $\mu_g \in L^\infty(\Omega)$ for $g \in C([-r, r], \mathbb{R})$ by setting $\mu_g = \mu_f$ where $f \in C(\mathbb{R}, \mathbb{R})$ is such that $f = g$ on $[-r, r]$. Indeed, this definition is meaningful since if f and h are two elements of $C(\mathbb{R}, \mathbb{R})$ such that $f = g$ on $[-r, r]$ then μ_f and μ_h are the same element of $L^\infty(\Omega)$ (i.e. $\mu_f = \mu_h$ a.e. on Ω) thanks to the fact that $-r \leq u_n \leq r$ a.e. on Ω and for all $n \in \mathbb{N}$.

For $x \in \Omega$, let

$$E_x = \{g \in C([-r, r], \mathbb{R});\ \lim_{h\to 0} \frac{1}{\mathrm{m}(B(0,h))} \int_{B(x,h)} \mu_g(z)dz \text{ exists in } \mathbb{R}\},$$

where $B(x, h)$ is the ball of center x and radius h (note that $B(x, h) \subset \Omega$ for h small enough).

If $g \in E_x$, we set

$$\bar{\mu}_g(x) = \lim_{h\to 0} \frac{1}{\mathrm{m}(B(0,h))} \int_{B(x,h)} \mu_g(z)dz.$$

Then, we define T_x from E_x in $\mathbb{R}$ by $T_x(g) = \bar{\mu}_g(x)$. It is easily seen that E_x is a vector space which contains the constant functions, that T_x is a linear application from E_x to $\mathbb{R}$ and that T_x is nonnegative (i.e. $g(s) \geq 0$ for all $s \in \mathbb{R}$ implies $T_x(g) \geq 0$). Hence, using a modified version of the Hahn-Banach theorem, one can prolonge T_x into a linear nonnegative application $\overline{T}_x$ defined on the whole set $C([-r, r], \mathbb{R})$. By a classical Riesz theorem, there exists a (nonnegative) measure m_x on the Borel sets of $[-r, r]$ such that

$$\overline{T}_x(g) = \int_{-r}^{r} g(s)dm_x(s), \forall g \in C([-r, r], \mathbb{R}). \tag{32.2}$$

If $g(s) = 1$ for all $s \in [-r, r]$, the function g belongs to E_x and $\bar{\mu}_g(x) = 1$ (note that $\mu_g = 1$ a.e. on Ω). Hence, from (32.2), m_x is a probability over $[-r, r]$, and therefore a probability over $\mathbb{R}$ by prolonging it by 0 outside of $[-r, r]$. This gives the first item on the family $(m_x)_{x\in\Omega}$.

Let us prove now the second item on the family $(m_x)_{x\in\Omega}$. If $g \in C([-r, r], \mathbb{R})$ then $g \in E_x$ for a.e. $x \in \Omega$ and $\mu_g(x) = \bar{\mu}_g(x)$ for a.e. $x \in \Omega$ (this is a classical result, since $\mu_g \in L^1_{loc}(\Omega)$, see Rudin [129]). Therefore, $\mu_g(x) = T_x(g) = \overline{T}_x(g)$ for a.e. $x \in \Omega$. Hence,

$$\mu_g(x) = \int_{-r}^{r} g(s) dm_x(s) \text{ for a.e. } x \in \Omega,$$

for all $g \in C([-r, r], \mathbb{R})$ and therefore for all $g \in C(\mathbb{R}, \mathbb{R})$. Finally, since the support of m_x is included in $[-r, r]$,

$$\mu_g(x) = \int_{\mathbb{R}} g(s) dm_x(s) \text{ for a.e. } x \in \Omega, \forall g \in C(\mathbb{R}, \mathbb{R}).$$

This completes Step 2.

Step 3 (construction of u)

It is well known that, if $\bar{m}$ is a probability on $\mathbb{R}$, one has

$$\int_{\mathbb{R}} g(s) d\bar{m}(s) = \int_0^1 g(u(\alpha)) d\alpha, \forall g \in \mathcal{M}_b, \tag{32.3}$$

where $\mathcal{M}_b$ is the set of bounded measurable functions from $\mathbb{R}$ to $\mathbb{R}$ and with

$$u(\alpha) = \sup\{c \in \mathbb{R};\ \bar{m}((-\infty, c)) < \alpha\}, \forall \alpha \in (0, 1).$$

Note that the function u is measurable, nondecreasing and left continuous. Furthermore, if the support of $\bar{m}$ is included in $[a, b]$ (for some a, $b \in \mathbb{R}$, $a < b$) then $u(\alpha) \in [a, b]$ for all $\alpha \in (0, 1)$ and (32.3) holds for all $g \in C(\mathbb{R}, \mathbb{R})$.

Applying this result to the measures m_x leads to the definition of u as

$$u(x, \alpha) = \sup\{c \in \mathbb{R};\ m_x((-\infty, c)) < \alpha\}, \forall \alpha \in (0, 1), \forall x \in \Omega.$$

For all $x \in \Omega$, the function $u(x, \cdot)$ is measurable (from $(0, 1)$ to $\mathbb{R}$), nondecreasing, left continuous and takes its values in $[-r, r]$. Furthermore,

$$\mu_g(x) = \int_0^1 g(u(x, \alpha)) d\alpha \text{ for a.e. } x \in \Omega, \forall g \in C(\mathbb{R}, \mathbb{R}).$$

Therefore,

$$\int_\Omega g(u_n(x))\varphi(x)dx \to \int_\Omega \Big(\int_0^1 g(u(x, \alpha))d\alpha\Big)\varphi(x)dx, \text{ as } n \to \infty,$$
$$\forall \varphi \in L^1(\Omega), \forall g \in C(\mathbb{R}, \mathbb{R}).$$

In order to conclude the proof of Proposition 32.1, there remains to show that modifying u on a negligible set leads to a function (still denoted by u) measurable with respect to $(x, \alpha) \in \Omega \times (0, 1)$. Indeed, this mesurability is needed in order to assert for instance, applying Fubini's Theorem (see RUDIN [129]), that

$$\int_\Omega \Big(\int_0^1 g(u(x, \alpha))d\alpha\Big)\varphi(x)dx = \int_0^1 \Big(\int_\Omega g(u(x, \alpha))\varphi(x)dx\Big)d\alpha,$$

for all $\varphi \in L^1(\Omega)$ and for all $g \in C(\mathbb{R}, \mathbb{R})$.

For all $g \in C(\mathbb{R}, \mathbb{R})$, one chooses for μ_g (which belongs to $L^\infty(\Omega)$) a bounded measurable function from Ω to $\mathbb{R}$.

Let us define $\mathcal{E} = \{g_{a,b};\ a, b \in \mathbb{Q},\ a < b\}$ where $g_{a,b} \in C(\mathbb{R}, \mathbb{R})$ is defined by

$$g_{a,b}(x) = 1 \text{ if } x \le a,$$
$$g_{a,b}(x) = \tfrac{x-b}{a-b} \text{ if } a < x < b,$$
$$g_{a,b}(x) = 0 \text{ if } x \ge b.$$

Since $\mathcal{E}$ is a countable subset of $C(\mathbb{R}, \mathbb{R})$, there exists a Borel subset A of Ω such that $\mathrm{m}(A) = 0$ and

$$\mu_g(x) = \int_{\mathbb{R}} g(s) dm_x(s),\ \forall x \in \Omega \setminus A,\ \forall g \in \mathcal{E}. \tag{32.4}$$

Define for all $\alpha \in (0,1)$ $v(.,\alpha)$ by

$$v(x,\alpha) = 0 \text{ if } x \in A,$$
$$v(x,\alpha) = \sup\{c \in \mathbb{R},\ m_x((-\infty,c)) < \alpha\} \text{ if } x \in \Omega \setminus A,$$

so that $u = v$ on $(\Omega \setminus A) \times (0,1)$ (and then $u = v$ a.e. on $\Omega \times (0,1)$).
Let us now prove that v is measurable from $\Omega \times (0,1)$ to $\mathbb{R}$ (this will conclude the proof of Proposition 32.1).
Since $v(x,.)$ is left continuous on $(0,1)$ for all $x \in \Omega$, proving that $v(.,\alpha)$ is measurable (from Ω to $\mathbb{R}$) for all $\alpha \in (0,1)$ leads to the mesurability of v on $\Omega \times (0,1)$ (this is also classical, see RUDIN [129]).
There remains to show the mesurability of $v(.,\alpha)$ for all $\alpha \in (0,1)$.
Let $\alpha \in (0,1)$ (in the following, α is fixed). Let us set $w = v(.,\alpha)$ and define, for $c \in \mathbb{R}$,

$$f_c(x) = m_x((-\infty,c)) - \alpha,\ x \in \Omega \setminus A,$$

so that $v(x,\alpha) = w(x) = \sup\{c \in \mathbb{R},\ f_c(x) < 0\}$ for all $x \in \Omega \setminus A$.
Using (32.4) leads to

$$m_x((-\infty,c)) = \sup\{\mu_g(x),\ g \leq 1_{(-\infty,c)} \text{ and } g \in \mathcal{E}\},\ \forall x \in \Omega \setminus A.$$

Then, the function $f_c : \Omega \setminus A \to \mathbb{R}$ is measurable as the supremum of a countable set of measurable functions (recall that μ_g is measurable for all $g \in \mathcal{E}$).
In order to prove the measurability of w (from Ω to $\mathbb{R}$), it is sufficient to prove that $\{x \in \Omega \setminus A;\ w(x) \geq a\}$ is a Borel set, for all $a \in \mathbb{R}$ (recall that $w = 0$ on A).
Let $a \in \mathbb{R}$, since $f_c(x)$ is nondecreasing with respect to c, one has

$$\{x \in \Omega \setminus A;\ w(x) \geq a\} = \cap_{n>0}\{x \in \Omega \setminus A;\ f_{a-\frac{1}{n}}(x) < 0\}.$$

Then $\{x \in \Omega \setminus A;\ w(x) \geq a\}$ is measurable, thanks to the measurability of f_c for all $c \in \mathbb{R}$.
This concludes the proof of Proposition 32.1. ∎

Remark 32.2 Let Ω be an open subset of $\mathbb{R}^N$ $(N \geq 1)$, $(u_n)_{n\in\mathbb{N}} \subset L^\infty(\Omega)$ and $u \in L^\infty(\Omega \times (0,1))$ such that $(u_n)_{n\in\mathbb{N}}$ converges towards u in the nonlinear weak-$\star$ sense. Assume that u does not depend on α, i.e. there exists $v \in L^\infty(\Omega)$ such that $u(x,\alpha) = v(x)$ for a.e. $(x,\alpha) \in \Omega \times (0,1)$. Then, it is easy to prove that $(u_n)_{n\in\mathbb{N}}$ converges towards u in $L^p(B)$ for all $1 \leq p < \infty$ and all bounded subset B of Ω. Indeed, let B be a bounded subset of Ω. Taking, in (32.1), $g(s) = s^2$ (for all $s \in \mathbb{R}$) and $\varphi = 1_B$ and also $g(s) = s$ (for all $s \in \mathbb{R}$) and $\varphi = 1_B v$ leads to

$$\int_B (u_n(x) - v(x))^2 dx \to 0,\ \text{as } n \to \infty.$$

This proves that $(u_n)_{n\in\mathbb{N}}$ converges towards u in $L^2(B)$. The convergence of $(u_n)_{n\in\mathbb{N}}$ towards u in $L^p(B)$ for all $1 \leq p < \infty$ is then an easy consequence of the $L^\infty(\Omega)$ bound on $(u_n)_{n\in\mathbb{N}}$ (see Remark 32.1).

33 A stabilized finite element method

In this section, we shall try to compare the finite element method to the finite volume method for the discretization of a nonlinear hyperbolic equation. It is well known that the use of the finite element is not straightforward in the case of hyperbolic equations, since the lack of coerciveness of the operator yields a lack of stability of the finite element scheme. There are several techniques to stabilize these schemes,

which are beyond the scope of this work. Here, as in SELMIN [134], we are interested in viewing the finite element as a finite volume method, by writing it in a conservative form, and using a stabilization as in the third item of Example 21.1 page 135.

Let $F \in C^1(\mathbb{R}, \mathbb{R}^2)$, consider the following scalar conservation law:

$$u_t(x,t) + \operatorname{div}(F(u))(x,t) = 0,\ x \in \mathbb{R}^2,\ t \in \mathbb{R}_+, \tag{33.1}$$

with an initial condition. Let $\mathcal{T}$ be a triangular mesh of $\mathbb{R}^2$, well suited for the finite element method. Let $\mathcal{S}$ denote the set of nodes of this mesh, and let $(\phi_j)_{j\in\mathcal{S}}$ be the classical piecewise bilinear shape functions. Following the finite element principles, let us look for an approximation of u in the space spanned by the shape functions ϕ_j; hence, at time $t_n = nk$ (where k is the time step), we look for an approximate solution of the form

$$u(.,t_n) = \sum_{j\in\mathcal{S}} u_j^n \phi_j;$$

multiplying (33.1) by ϕ_i, integrating over $\mathbb{R}^2$, approximating $F(\sum_{j\in\mathcal{S}} u_j^n\phi_j)$ by $\sum_{j\in\mathcal{S}} F(u_j^n)\phi_j$, using the explicit Euler scheme for the time discretization) and the mass lumping technique on the mass matrix as described in Remark 16.3 yields the following scheme:

$$\frac{u_i^{n+1}-u_i^n}{k}\int_{\mathbb{R}^2}\phi_i(x)dx - \sum_{j\in\mathcal{S}} F(u_j^n)\cdot\int_{\mathbb{R}^2}\phi_j(x)\nabla\phi_i(x)dx = 0,$$

which reads, noting that $\displaystyle\int \phi_j(x)\nabla\phi_i(x)dx = -\int \phi_i(x)\nabla\phi_j(x)dx$ and that $\displaystyle\sum_{j\in\mathcal{S}}\nabla\phi_j(x) = 0$,

$$\frac{u_i^{n+1}-u_i^n}{k}\int_{\mathbb{R}^2}\phi_i(x)dx + \sum_{j\in\mathcal{S}}(F(u_i^n)+F(u_j^n))\cdot\int_{\mathbb{R}^2}\phi_i(x)\nabla\phi_j(x)dx = 0.$$

This last equality may also be written

$$\frac{u_i^{n+1}-u_i^n}{k}\int_{\mathbb{R}^2}\phi_i(x)dx + \sum_{j\in\mathcal{S}} E_{i,j} = 0,$$

where

$$E_{i,j} = \frac{1}{2}(F(u_i^n)+F(u_j^n))\cdot\int_{\mathbb{R}^2}(\phi_i(x)\nabla\phi_j(x) - \phi_j(x)\nabla\phi_i(x))dx.$$

Note that $E_{j,i} = -E_{i,j}$.

This is a centered and therefore unstable scheme. One way to stabilize it is to replace $E_{i,j}^n$ by

$$\tilde{E}_{i,j}^n = E_{i,j}^n + D_{i,j}(u_i^n - u_j^n),$$

where $D_{i,j} = D_{j,i}$ (in order for the scheme to remain "conservative") and $D_{i,j} \geq 0$ is chosen large enough so that $E_{i,j}^n$ is a nondecreasing function of u_i^n and a nonincreasing function of u_j^n, which ensure the stability of the scheme, under a so called CFL condition, and does not change the "consistency" (see (21.7) page 135 and Remark 30.1 page 189).

34 Moving meshes

For some evolution problems the use of time variable control volumes is advisable, e.g. when the domain of study changes with time. This is the case, for instance, for the simulation of a flow in a porous medium, when the porous medium is heterogeneous and its geometry changes with time. In this case, the mesh is

required to move with the medium. The influence of the moving mesh on the finite volume formulation can be explained by considering the following simple transport equation:

$$u_t(x,t) + \mathrm{div}(u\mathbf{v})(x,t) = 0, \quad x \in \mathbb{R}^2, \quad t \in \mathbb{R}_+, \tag{34.1}$$

where $\mathbf{v}$ depends on the unknown u (and possibly on other unknowns). Let k be the time step, and set $t_n = nk$, $n \in \mathbb{N}$. Let $\mathcal{T}(t)$ be the mesh at time t. Since the mesh moves, the elements of the mesh vary in time. For a fixed $n \in \mathbb{N}$, let $R(K,t)$ be the domain of $\mathbb{R}^2$ occupied by the element K ($K \in \mathcal{T}(t_n)$) at time t, $t \in [t_n, t_{n+1}]$, that is $R(K,t_n) = K$. Let $\mathbf{v}_s(x,t)$ be the velocity of the displacement of the mesh at point $x \in \mathbb{R}^2$ and for all $t \in [t_n, t_{n+1}]$ (note that $\mathbf{v}_s(x,t) \in \mathbb{R}^2$). Let u_K^n and u_K^{n+1} be the discrete unknowns associated to element K at times t_n and t_{n+1} (they can be considered as the approximations of the mean values of $u(\cdot,t_n)$ and $u(\cdot,t_{n+1})$ over $R(K,t_n)$ and $R(K,t_{n+1})$ respectively). The discretization of (34.1) must take into account the evolution of the mesh in time. In order to do so, let us first consider the following differential equation with initial condition:

$$\begin{array}{rcl} \dfrac{\partial y}{\partial t}(x,t) & = & -\mathbf{v}_s(y(x,t),t), \quad t \in [t_n, t_{n+1}], \\ y(x,t_n) & = & x. \end{array} \tag{34.2}$$

Under suitable assumptions on $\mathbf{v}_s$ (assume for instance that $\mathbf{v}_s$ is continuous, Lipschitz continuous with respect to its first variable and that the Lipschitz constant is integrable with respect to its second variable), the problem (34.2) has, for all $x \in \mathbb{R}^2$, a unique (global) solution. For $x \in \mathbb{R}^2$, define the function $y(x,\cdot)$ from $[t_n, t_{n+1}]$ to $\mathbb{R}^2$ as the solution of problem (34.2). Let $(\varphi_p)_{p\in\mathbb{N}} \subset C_c^1(\mathbb{R}^2, \mathbb{R}_+)$ such that $0 \le \varphi_p(x) \le 1$ for $x \in \mathbb{R}^2$ and for all $p \in \mathbb{N}$, and such that $\varphi_p \to 1_K$ a.e. as $p \to +\infty$. Multiplying (34.1) by $\psi_p(x,t) = \varphi_p(y(x,t))$ and integrating over $\mathbb{R}^2$ yields

$$\int_{\mathbb{R}^2} \Big(\frac{\partial(u\psi_p)}{\partial t}(x,t) + u(x,t)\nabla\varphi_p(y(x,t)) \cdot \mathbf{v}_s(y(x,t),t) - (u\mathbf{v})(x,t) \cdot \nabla\psi_p(x,t) \Big) dx = 0. \tag{34.3}$$

Using the explicit Euler discretization in time on Equation (34.3) and denoting by $u^n(x)$ a (regular) approximate value of $u(x,t_n)$ yields

$$\begin{array}{c} \displaystyle\int_{\mathbb{R}^2} \frac{1}{k}\Big(u^{n+1}(x)\psi_p(x,t_{n+1}) - u^n(x)\psi_p(x,t_n) \Big) dx + \\ \displaystyle\int_{\mathbb{R}^2} u^n(x)(\mathbf{v}_s(x,t_n) - \mathbf{v}(x,t_n)) \cdot \nabla\varphi_p(x) dx = 0, \end{array}$$

which also gives (noting that $\psi_p(x,t) = \varphi_p(y(x,t))$)

$$\begin{array}{c} \displaystyle\int_{\mathbb{R}^2} \frac{1}{k}\Big(u^{n+1}(x)\varphi_p(y(x,t_{n+1})) - u^n(x)\varphi_p(y(x,t_n)) \Big) dx - \\ \displaystyle\int_{\mathbb{R}^2} \mathrm{div}(u^n(\mathbf{v}_s - \mathbf{v}))(x,t_n) \cdot \varphi_p(x) dx = 0. \end{array} \tag{34.4}$$

Letting p tend to infinity and noting that $1_K(y(x,t_n)) = 1_{R(K,t_n)}(x)$ and $1_K(y(x,t_{n+1})) = 1_{R(K,t_{n+1})}(x)$, (34.4) becomes

$$\frac{1}{k}\Big(\int_{R(K,t_{n+1})} u^{n+1}(x)dx - \int_{R(K,t_n)} u^n(x)dx \Big) + \int_{R(K,t_n)} \mathrm{div}((\mathbf{v} - \mathbf{v}_s)u^n)(x,t_n)dx = 0,$$

which can also be written

$$\begin{array}{c} \dfrac{1}{k}(u_K^{n+1}\mathrm{m}(R(K,t_{n+1})) - u_K^n\mathrm{m}(R(K,t_n))) + \\ \displaystyle\int_{\partial R(K,t_n)} (\mathbf{v} - \mathbf{v}_s)(x,t_n) \cdot \boldsymbol{n}_K(x,t_n) u^n(x) d\gamma(x) = 0, \end{array}$$

where $u_K^n = [1/\mathrm{m}(R(K,t_n))] \int_{R(K,t_n)} u^n(x)dx$ and $u_K^{n+1} = [1/\mathrm{m}(R(K,t_{n+1}))] \int_{R(K,t_{n+1})} u^{n+1}(x)dx$. Recall that $\boldsymbol{n}_K$ denotes the normal to ∂K, outward to K. The complete discretization of the problem uses some additional equations (on $\mathbf{v}$, $\mathbf{v}_s \ldots$).

Remark 34.1 The above considerations concern a pure convection equation. In the case of a convection-diffusion equation, such a moving mesh may become non-admissible in the sense of definitions 9.1 page 37 or 10.1 page 63. It is an interesting open problem to understand what should be done in that case.

Chapter 7

Systems

In chapters 2 to 6, the finite volume was successively investigated for the discretization of elliptic, parabolic, and hyperbolic equations. In most scientific models, however, systems of equations have to be discretized. These may be partial differential equations of the same type or of different types, and they may also be coupled to ordinary differential equations or algebraic equations.

The discretization of systems of elliptic equations by the finite volume method is straightforward, following the principles which were introduced in chapters 2 and 3. Examples of the performance of the finite volume method for systems of elliptic equations on rectangular meshes, with "unusual" source terms (in particular, with source terms located on the edges or interfaces of the mesh) may be found in e.g. Angot [3] (see also references therein), Fiard, Herbin [66] (where a comparison to a mixed finite element formulation is also performed). Parabolic systems are treated similarly as elliptic systems, with the addition of a convenient time discretization.

A huge literature is devoted to the discretization of hyperbolic systems of equations, in particular to systems related to the compressible Euler equations, using structured or unstructured meshes. We shall give only a short insight on this subject in Section 35, without any convergence result. Indeed, very few theoretical results of convergence of numerical schemes are known on this subject. We refer to Godlewski and Raviart [76] and references therein for a more complete description of the numerical schemes for hyperbolic systems.

Finite volume methods are also well adapted to the discretization of systems of equations of different types (for instance, an elliptic or parabolic equation coupled with hyperbolic equations). Some examples are considered in sections 36 page 218 and 37 page 222. The classical case of incompressible Navier-Stokes (for which, generally, staggered grids are used) and examples which arise in the simulation of a multiphase flow in a porous medium are described. The latter example also serves as an illustration of how to deal with algebraic equations and inequalities.

35 Hyperbolic systems of equations

Let us consider a hyperbolic system consisting of m equations (with $m \geq 1$). The unknown of the system is a function $u = (u_1, \ldots, u_m)^t$, from $\overline{\Omega} \times [0,T]$ to $\mathbb{R}^m$, where Ω is an open set of $\mathbb{R}^d$ (i.e. $d \geq 1$ is the space dimension), and u is a solution of the following system:

$$\begin{aligned} &\frac{\partial u_i}{\partial t}(x,t) + \sum_{j=1}^{d} \frac{\partial G_{i,j}}{\partial x_j}(x,t) = g_i(x,t,u(x,t)), \\ &x = (x_1, \ldots, x_d) \in \Omega,\ t \in (0,T),\ i = 1, \ldots, m, \end{aligned} \tag{35.1}$$

where

$$G_{i,j}(x,t) = F_{i,j}(x,t,u(x,t)),$$

and the functions $F_j = (F_{1,j}, \ldots, F_{m,j})^t$ $(j = 1, \ldots, d)$ and $g = (g_1, \ldots, g_m)^t$ are given functions from $\overline{\Omega} \times [0, T] \times \mathbb{R}^m$ (indeed, generally, a part of $\mathbb{R}^m$, instead of $\mathbb{R}^m$) to $\mathbb{R}^m$. The function $F = (F_1, \ldots, F_d)$ is assumed to satisfy the usual hyperbolicity condition, that is, for any (unit) vector of $\mathbb{R}^d$, $\mathbf{n}$, the derivative of $F \cdot \mathbf{n}$ with respect to its third argument (which can be considered as an $m \times m$ matrix) has only real eigenvalues and is diagonalizable.
Note that in real applications, diffusion terms may also be present in the equations, we shall omit them here. In order to complete System (35.1), an initial condition for $t = 0$ and adequate boundary conditions for $x \in \partial\Omega$ must be specified.

In the first section (Section 35.1), we shall only briefly describe the general method of discretization by finite volume and some classical schemes. In the subsequent sections, some possible treatments of difficulties appearing in real simulations will be given.

35.1 Classical schemes

Let us first describe some classical finite volume schemes for the discretization of (35.1) with initial and boundary conditions, using the concepts and notations which were introduced in chapter 6. Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 25.1 page 156 and k be the time step, which is assumed to be constant (the generalization to a variable time step is easy). We recall that the interface, $K|L$, between any two elements K and L of $\mathcal{T}$ is assumed to be included in a hyperplane of $\mathbb{R}^d$. The discrete unknowns are the u_K^n, $K \in \mathcal{T}$, $n \in \{0, \ldots, N_k + 1\}$, with $N_k \in \mathbb{N}$, $(N_k + 1)k = T$. For $K \in \mathcal{T}$, let $\mathcal{N}(K)$ be the set of its neighbours, that is the set of elements L of $\mathcal{T}$ such that the $(d-1)$ Lebesgue measure of $K|L$ is positive. For $L \in \mathcal{N}(K)$, let $\mathbf{n}_{K,L}$ be the unit normal vector to $K|L$ oriented from K to L. Let $t_n = nk$, for $n \in \{0, \ldots, N_k + 1\}$.
A finite volume scheme reads

$$\mathrm{m}(K)\frac{u_K^{n+1} - u_K^n}{k} + \sum_{L \in \mathcal{N}(K)} \mathrm{m}(K|L) F_{K,L}^n = \mathrm{m}(K) g_K^n, \qquad K \in \mathcal{T},\ n \in \{0, \ldots, N_k\}, \tag{35.2}$$

where

1. $\mathrm{m}(K)$ (resp. $\mathrm{m}(K|L)$) denotes the d (resp. $d-1$) Lebesgue measure of K (resp. $K|L$),
2. the quantity g_K^n, which depends on u_K^n (or u_K^{n+1} or u_K^n and u_K^{n+1}), for $K \in \mathcal{T}$, is some "consistent" approximation of g on element K, between times t_n and t_{n+1} (we do not discuss this approximation here).
3. the quantity $F_{K,L}^n$, which depends on the set of discrete unknowns u_M^n (or u_M^{n+1} or u_M^n and u_M^{n+1}) for $M \in \mathcal{T}$, is an approximation of $F \cdot \mathbf{n}_{K,L}$ on $K|L$ between times t_n and t_{n+1}.

In order to obtain a "good" scheme, this approximation of $F \cdot \mathbf{n}_{K,L}$ has to be consistent, conservative (that is $F_{K,L}^n$=-$F_{L,K}^n$) and must ensure some stability properties on the approximate solution given by the scheme (indeed, one also needs some consistency with respect to entropies, when entropies exist...). Except in the scalar case, it is not so easy to see what kind of stability properties is needed.... Indeed, in the scalar case, that is $m = 1$, taking $g = 0$ and $\Omega = \mathbb{R}^d$ (for simplicity), it is essentially sufficient to have an L^∞ estimate (that is a bound on u_K^n independent of K, n, and of the time and space discretizations) and a "touch" of "BV estimate" (see, for instance, chapters 5 and 6 and Chainais-Hillairet [22] for more precise assumptions). In the case $m > 1$, it is not generally possible to give stability properties from which a mathematical proof of convergence could be deduced. However, it is advisable to require some stability properties such as the positivity of some quantities depending on the unknowns; in the case of flows, the required stability may be the positivity of the density, energy, pressure...; the positivity of these quantities may be essential for the computation of $F(u)$ or for its hyperbolicity.

The computation of $F^n_{K,L}$ is often performed, at each "interface", by solving the following 1D (for the space variable) system (where, for simplicity, the possible dependency of F with respect to x and t is omitted):

$$\frac{\partial u}{\partial t}(z,t) + \frac{\partial f_{K,L}(u)}{\partial z}(z,t) = 0, \tag{35.3}$$

where $f_{K,L}(u)(z,t) = F\cdot\mathbf{n}_{K,L}(u(z,t))$, for all $z \in \mathbb{R}$ and $t \in (0,T)$, which gives consistency, conservativity (and, hopefully, stability) of the final scheme (that is (35.2)). To be more precise, in the case of lower order schemes, $F^n_{K,L}$ may be taken as: $F^n_{K,L} = F.\mathbf{n}_{K,L}(w)$ where w is the solution for $z = 0$ of (35.3) with initial conditions $u(x,0) = u^n_K$ if $x < 0$ and $u(x,0) = u^n_L$ if $x > 0$. Note that the variable z lies in $\mathbb{R}$, so that the multidimensional problem has therefore been transformed (as in chapter 6) into a succession of one-dimensional problems. Hence, in the following, we shall mainly keep to the case $d = 1$.

Let us describe two classical schemes, namely the Godunov scheme and the Roe scheme, in the case $d = 1$, $\Omega = \mathbb{R}$, $F(x,t,u) = F(u)$ and $g = 0$ (but $m \geq 1$), in which case System (35.1) becomes

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial F(u)}{\partial x}(x,t) = 0, \quad x \in \mathbb{R},\ t \in (0,T). \tag{35.4}$$

in order to complete this system, an initial condition must be specified, the discretization of which is standard.
Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 20.1 page 128, that is $\mathcal{T} = (K_i)_{i\in\mathbb{Z}}$, with K_i=($x_{i-1/2}$,$x_{i+1/2}$), with $x_{i-1/2} < x_{i+1/2}$, $i \in \mathbb{Z}$. One sets $h_i = x_{i+1/2} - x_{i-1/2}$, $i \in \mathbb{Z}$. The discrete unknowns are u^n_i, $i \in \mathbb{Z}$, $n \in \{0,\ldots,N_k+1\}$ and the scheme (35.2) then reads

$$h_i\frac{u^{n+1}_i - u^n_i}{k} + F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} = 0, \quad i \in \mathbb{Z},\ n \in \{0,\ldots,N_k\}, \tag{35.5}$$

where $F^n_{i+1/2}$ is a consistent approximation of $F(u(x_{i+1/2},t_n)$. This scheme is clearly conservative (in the sense defined above). Let us consider explicit schemes, so that $F^n_{i+1/2}$ is a function of u^n_j, $j \in \mathbb{Z}$. The principle of the Godunov scheme Godunov [77] is to take $F^n_{i+1/2} = F(w)$ where w is the solution, for $x = 0$ (and any $t > 0$), of the following (Riemann) problem

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial F(u)}{\partial x}(x,t) = 0, \quad x \in \mathbb{R},\ t \in \mathbb{R}_+, \tag{35.6}$$

$$\begin{array}{c} u(x,0) = u^n_i, \text{ if } x < 0, \\ u(x,0) = u^n_{i+1}, \text{ if } x > 0. \end{array} \tag{35.7}$$

Then, w depends on u^n_i, u^n_{i+1} and F.
The time step is limited by the so called "CFL condition", which reads $k \leq Lh_i$, for all $i \in \mathbb{Z}$, where L is given by F and the initial condition. The quantity u^{n+1}_i, given by the Godunov scheme, see Godunov [77], is, for all $i \in \mathbb{Z}$, the mean value on K_i of the exact solution at time k of (35.4) with the initial condition (at time $t = 0$) u_0 defined, a.e. on $\mathbb{R}$, by $u_0(x) = u^n_i$ if $x_{i-1/2} < x < x_{i+1/2}$.

The Godunov scheme is an efficient scheme (consistent, conservative, stable), sometimes too diffusive (especially if k is far from Lh_i defined above), but easy improvements are possible, such as the MUSCL technique, see below and Section 22. Its principal drawback is its difficult implementation for many problems, indeed the computation of $F(w)$ can be impossible or too expensive. For instance, this computation may need a non trivial parametrization of the non linear waves. Note also that F is generally not given directly as a function of u (the components of u are called "conservative unknowns") but as a function of some "physical" unknowns (for instance, pressure, velocity, energy...), and the passage from u to these physical unknowns (or the converse) is often not so easy...it may be the consequence of expensive and implicit calculations, using, for instance, Newton's algorithm.

Due to this difficulty of implementation, some "Godunov type" schemes were developed (see HARTEN, LAX and VAN LEER [81]). The idea is to take, for u_i^{n+1}, the mean value on K_i of an *approximate* solution at time k of (35.4) with the initial condition (at time $t = 0$), u_0, defined by $u_0(x) = u_i^n$, if $x_{i-1/2} < x < x_{i+1/2}$. In order for the scheme to be written under the conservative form (35.5), with a consistent approximation of the fluxes, this approximate solution must satisfy some consistency relation (another relation is needed for the consistency with entropies). One of the best known of this family of schemes is the Roe scheme (see ROE [127] and ROE [128]), where this approximate solution is computed by the solution of the following linearized Riemann problems:

$$\frac{\partial u(x,t)}{\partial t} + A(u_i^n, u_{i+1}^n)\frac{\partial u(x,t)}{\partial x} = 0,\ x \in \mathbb{R},\ t \in \mathbb{R}_+, \tag{35.8}$$

$$\begin{array}{c} u(x,0) = u_i^n,\ \text{if } x < 0, \\ u(x,0) = u_{i+1}^n,\ \text{if } x > 0, \end{array} \tag{35.9}$$

where $A(\cdot,\cdot)$ is an $m \times m$ matrix, continuously depending on its two arguments, with only real eigenvalues, diagonalizable and satisfying the so called "Roe condition":

$$A(u,v)(u-v) = F(u) - F(v), \forall u, v \in \mathbb{R}^m. \tag{35.10}$$

Thanks to (35.10), the Roe scheme can be written as (35.5) with

$$\begin{array}{c} F_{i+\frac{1}{2}}^n = F(u_i^n) + A^-(u_i^n, u_{i+1}^n)(u_i^n - u_{i+1}^n) \\ (= F(u_{i+1}^n) + A^+(u_i^n, u_{i+1}^n)(u_i^n - u_{i+1}^n)), \end{array} \tag{35.11}$$

where $A^\pm$ are the classical nonnegative and nonpositive parts of the matrix A: let A be a matrix with only real eigenvalues, $(\lambda_p)_{p=1,\ldots,m}$, and diagonalizable, let $(\varphi_p)_{p=1,\ldots,m}$ be a basis of $\mathbb{R}^m$ associated to these eigenvalues. Then, the matrix A^+ is the matrix which has the same eigenvectors as A and has $(\max\{\lambda_p, 0\})_{p=1,\ldots,m}$ as corresponding eigenvalues. The matrix A^- is $(-A)^+$.
Roe's scheme was proved to be an efficient scheme, often less expensive than Godunov's scheme, with, more or less the same limitation on the time step, the same diffusion effect and some lack of entropy consistency, which can be corrected. It has some properties of consistency and stability. Its main drawback is the difficulty of the computation of a matrix $A(u,v)$ satisfying (35.10). For instance, when it is possible to compute and diagonalize the derivative of F, $DF(u)$, one can take $A(u,v) = DF(u^\star)$, but the difficulty is to find $u^\star$ such that (35.10) holds (note that this condition is crucial in order to ensure conservativity of Roe's scheme). In some difficult cases, the Roe matrix is computed approximately by using a "limited expansion" with respect to some small parameter.

35.2 Rough schemes for complex hyperbolic systems

The aim of this section is to present some discretization techniques for "complex" hyperbolic systems. In many applications, the expressions of g and F which appear in (35.1) are rather "complex", and it is difficult or impossible to use classical schemes such as the 1D Godunov or Roe schemes or their standard extensions, for multidimensional problems, using 1D solvers on the interfaces of the mesh. This is the case of gas dynamics (Euler equations) with real gas, for which the state law (pressure as a function of density and internal energy) is tabulated or given by some complex analytical expressions. This is also the case when modelling multiphase flows in pipe-lines: the function F is difficult to handle and highly depends on x and u, because, for instance, of changes of the geometry and slope of the pipe, of changes of the friction law or, more generally, of the varying nature of the flow. Most of the attempts given below were developed for this last situation. Other interesting cases of "complexity" are the treatment of boundary conditions (mathematical literature is rather scarce on this subject, see Section 35.4 for a first insight), and the way to handle the case where the eigenvalues (of the derivative of $F \cdot \mathbf{n}$ with respect to its third argument) are of very different magnitude, see Section 35.3. Another case of complexity is the

treatment of nonconservative terms in the equations. One refers, for instance, to Brun, Hérard, Leal De Sousa and Uhlmann [17] and references therein, for this important case.

Possible modifications of Godunov and Roe schemes (including "classical" improvements to avoid excessive artificial diffusion) are described now to handle "complex" systems. Because of the complexity of the models, the justification of the schemes presented here is rather numerical than mathematical. Many variations have also been developed, which are not presented here. Note that other approaches are also possible, see e.g. Ghidaglia, Kumbaro and Le Coq [74]. For simplicity, one considers the case $d = 1$, $\Omega = \mathbb{R}$, $F(x,t,u) = F(u)$ and $g = 0$ (but $m \geq 1$) described in Section 35.1, with the same notations. The Godunov and Roe schemes can both be written under the form (35.5) with $F^n_{i+1/2}$ computed as a function of u^n_i and u^n_{i+1}; both schemes are consistent (in the sense of Section 35.1, i.e. consistency of the "fluxes") since $F^n_{i+1/2} = F(u)$ if $u^n_i = u^n_{i+1} = u$.

Going further along this line of thought yields (among other possibilities, see below) the "VFRoe" scheme which is (35.5), that is:

$$h_i \frac{u_i^{n+1} - u_i^n}{k} + F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} = 0, \quad i \in \mathbb{Z}\,,\ n \in \{0, \ldots, N_k\}, \tag{35.12}$$

with $F^n_{i+1/2} = F(w)$, where w is the solution of the linearized Riemann problem (35.8), (35.9), with $A(u^n_i, u^n_{i+1}) = DF(w^\star)$, that is:

$$\frac{\partial u(x,t)}{\partial t} + DF(w^\star)\frac{\partial u(x,t)}{\partial x} = 0,\ x \in \mathbb{R},\ t \in \mathbb{R}_+, \tag{35.13}$$

$$\begin{array}{l} u(x,0) = u^n_i,\ \text{if } x < 0, \\ u(x,0) = u^n_{i+1},\ \text{if } x > 0, \end{array} \tag{35.14}$$

where $w^\star$ is some value between u^n_i and u^n_{i+1} (for instance, $w^\star = (1/2)(u^n_i + u^n_{i+1})$). In this scheme, the Roe condition (35.10) is not required (note that it is naturally conservative, thanks to its finite volume origin). Hence, the VFRoe scheme appears to be a simplified version of the Godunov and Roe schemes. The study of the scalar case ($m = 1$) shows that, in order to have some stability, at least as much as in Roe's scheme, the choice of $w^\star$ is essential. In practice, the choice $w^\star = (1/2)(u^n_i + u^n_{i+1})$ is often adequate, at least for regular meshes.

Remark 35.1 In Roe's scheme, the Roe condition (35.10) ensures conservativity. The VFRoe scheme is "naturally" conservative, and therefore no such condition is needed. Also note that the VFRoe scheme yields precise approximations of the shock velocities, without Roe's condition.

Numerical tests show the good behaviour of the VFRoe scheme. Its two main flaws are a lack of entropy consistency (as in Roe's scheme) and a large diffusion effect (as in the Godunov and Roe schemes). The first drawback can be corrected, as for Roe's scheme, with a nonparametric entropy correction inspired from Harten, Hyman and Lax [82] (see Masella, Faille, and Gallouët [106]). The two drawbacks can be corrected with a classical MUSCL technique, which consists in replacing, in (35.9) page 210, u^n_i and u^n_{i+1} by $u^n_{i+1/2,-}$ and $u^n_{i+1/2,+}$, which depend on $\{u^n_j,\ j = i-1, i, i+1, i+2\}$ (see, for instance, Section 22 page 146 and Godlewski and Raviart [76] or LeVeque [100]). For stability reasons, the computation of the gradient of the unknown (cell by cell) and of the "limiters" is performed on some "physical" quantities (such as density, pressure, velocity for Euler equations) instead of u. The extension of the MUSCL technique to the case $d > 1$ is more or less straightforward.

This MUSCL technique improves the space accuracy (in the truncation error) and the numerical results are significantly better. However, stability is sometimes lost. Indeed, considering the linear scalar equation, one remarks that the scheme is antidiffusive when the limiters are not active, this might lead to a loss of stability. The time step must then be reduced (it is reduced by a factor 10 in severe situations...).

In order to allow larger time steps, the time accuracy should be improved by using, for instance, an order 2 Runge-Kutta scheme (in the severe situations suggested above, the time step is then multiplied by a factor 4). Surprisingly, this improvement of time accuracy is used to gain stability rather than precision...

Several numerical experiments (see MASELLA, FAILLE, and GALLOUËT [106]) were performed which prove the efficiency of the VFRoe scheme, such as the classical Sod tests (SOD [137]). The shock velocities are exact, there are no oscillations.... For these tests, the treatment of the boundary conditions is straightforward. Throughout these experiments, the use of a MUSCL technique yields a significant improvement, while the use of a higher order time scheme is not necessary. In one of the Sod tests, the entropy correction is needed.
A comparison between the VFRoe scheme and the Godunov scheme was performed by J. M. Hérard (personal communication) for the Euler equations on a Van Der Wals gas, for which a matrix satisfying (35.10) seems difficult to find. The numerical results are better with the VFRoe schem, which is also much cheaper computationally. An improvment of the VFRoe scheme is possible, using, instead of (35.13)-(35.14), linearized Riemann problems associated to a nonconservative form of the initial system, namely System (35.4) or more generally System (35.1), for the computation of w (which gives the flux $F^n_{i+1/2}$ in (35.12) by the formula $F^n_{i+1/2} = F(w)$), see for instance BUFFARD, GALLOUËT and HÉRARD [18] for a simple example.
In some more complex cases, the flux F may also highly, and not continuously, depend on the space variable x. In the space discretization, it is "natural" to set the discontinuities of F with respect to x on the boundaries of the mesh. The function F may change drastically from K_i to K_{i+1}. In this case, the implementation of the VFRoe scheme yields two additional difficulties:

(i) The matrix $A(u^n_i, u^n_{i+1})$ in the linearized Riemann problem (35.8), (35.9) now depends on x: $A(u^n_i, u^n_{i+1}) = D_u F(x, w^\star)$, where $w^\star$ is some value between u^n_i and u^n_{i+1} and $D_u F$ denotes the derivative of F with respect to its "u" argument.

(ii) once the solution, w, of the linearized problem (35.8) (35.9), for $x = 0$ and any $t > 0$, is calculated, the choice $F^n_{i+1/2} = F(x, w)$ again depends on x.

The choice of $F^n_{i+1/2}$ (point (ii)) may be solved by remarking that, in Roe's scheme, $F^n_{i+1/2}$ may be written (thanks to (35.10)) as

$$F^n_{i+\frac{1}{2}} = \frac{1}{2}(F(u^n_i) + F(u^n_{i+1})) + \frac{1}{2}A^n_{i+\frac{1}{2}}(u^n_i - u^n_{i+1}), \tag{35.15}$$

where $A^n_{i+1/2} = |A(u^n_i, u^n_{i+1})|$, and $|A| = A^+ + A^-$.
Under this form, the second term of the right hand side of (35.15) appears to be a stabilization term, which does not affect the consistency. Indeed, in the scalar case ($m = 1$), one has $A^n_{i+1/2} = |F(u^n_i) - F(u^n_{i+1})|/|u^n_i - u^n_{i+1}|$, which easily yields the L^∞ stability of the scheme (but not the consistency with respect to the entropies). Moreover, the scheme is stable and consistent with respect to the entropies, under a Courant-Friedrichs-Levy (CFL) condition, if $F^n_{i+1/2}$ is nondecreasing with respect to u^n_i and nonincreasing with respect to u^n_{i+1}, which holds if $A^n_{i+1/2} \geq \sup\{|F'(s)|,\ s \in [u^n_i, u^n_{i+1}] \text{ or } [u^n_{i+1}, u^n_i]\}$.
This remark suggests a slightly different version of the VFRoescheme (closer to Roe's scheme), which is the scheme (35.12)-(35.14), taking

$$F^n_{i+1/2} = \frac{1}{2}(F(u^n_i) + F(u^n_{i+1})) + \frac{1}{2}|DF(w^\star)|(u^n_i - u^n_{i+1}),$$

in (35.12), instead of $F^n_{i+1/2} = F(w)$. Note that it is also possible to take other convex combinations of $F(u^n_i)$ and $F(u^n_{i+1})$ in the latter expression of $F^n_{i+1/2}$, without modifying the consistency of the scheme.

When F depends on x, the discontinuities of F being on the boundaries of the control volumes, the generalization of (35.15) is obvious, except for the choice of $A^n_{i+1/2}$. The quantity $F(u^n_i)$ is replaced by

$F(x_i, u_i^n)$, where x_i is the center of K_i. Let us now turn to the choice of a convenient matrix $A_{i+1/2}^n$ for this modified VFRoe scheme, when F highly depends on x. A first possible choice is

$$A_{i+1/2}^n = (1/2)(|D_u F(x_i, u_i^n)| + |D_u F(x_{i+1}, u_{i+1}^n)|).$$

The following slightly different choice for $A_{i+1/2}^n$ seems, however, to give better numerical results (see Faille and Heintzé [60]). Let us define

$$A_i = D_u F(x_i, u_i^n), \forall\, i \in \mathbb{Z}$$

(for the determination of $A_{i+1/2}^n$ the fixed index n is omitted). Let $(\lambda_p^{(i)})_{p=1,\dots,m}$ be the eigenvalues of A_i (with $\lambda_{p-1}^{(i)} \leq \lambda_p^{(i)}$, for all p) and $(\varphi_p^{(i)})_{p=1,\dots,m}$ a basis of $\mathbb{R}^m$ associated to these eigenvalues. Then, the matrix $A_{i+1/2}^{(-)}$ [resp. $A_{i+1/2}^{(+)}$] is the matrix which has the same eigenvectors as A_i [resp. A_{i+1}] and has $(\max\{|\lambda_p^{(i)}|, |\lambda_p^{(i+1)}|\})_{p=1,\dots,m}$ as corresponding eigenvalues. The choice of $A_{i+1/2}^n$ is

$$A_{i+\frac{1}{2}}^n = \frac{\lambda}{2}(A_{i+\frac{1}{2}}^{(-)} + A_{i+\frac{1}{2}}^{(+)}), \tag{35.16}$$

where λ is a parameter, the "normal" value of which is 1. Numerically, larger values of λ, say $\lambda = 2$ or $\lambda = 3$, are sometimes needed, in severe situations, to obtain enough stability. Too large values of λ yield too much artificial diffusion.

The new scheme is then (35.12)-(35.14), taking

$$F_{i+1/2}^n = \frac{1}{2}\Big(F(x_i, u_i^n) + F(x_i, u_{i+1}^n)\Big) + \frac{1}{2}A_{i+\frac{1}{2}}^n(u_i^n - u_{i+1}^n). \tag{35.17}$$

where $A_{i+1/2}^n$ is defined by (35.16). It has, more or less, the same properties as the Roe and VFRoe schemes but allows the simulation of more complex systems. It needs a MUSCL technique to reduce diffusion effects and order 2 Runge-Kutta for stability. It was implemented for the simulation of multiphase flows in pipe lines (see Faille and Heintzé [60]). The other difficulties encountered in this case are the treatment of the boundary conditions and the different magnitude of the eigenvalues, which are discussed in the next sections.

35.3 Partial implicitation of explicit scheme

In the modelling of flows, where "propagation" phenomena and "convection" phenomena coexist, the Jacobian matrix of F often has eigenvalues of different magnitude, the "large" eigenvalues (large meaning "far from 0", positive or negative) corresponding to the propagation phenomena and "small" eigenvalues corresponding to the "convection" phenomena . Large and small eigenvalues may differ by a factor 10 or 100.

With the explicit schemes described in the previous sections, the time step is limited by the CFL condition corresponding to the large eigenvalues. Roughly speaking, with the notations of Section 35.1, this condition is (for all $i \in \mathbb{Z}$) $k \leq |\lambda|^{-1} h_i$, where λ is the largest eigenvalue. In some cases, this limitation can be unsatisfactory for two reasons. Firstly, the time step is too small and implies a prohibitive computational cost. Secondly, the discontinuities in the solutions, associated to the small eigenvalues, are not sharp because the time step is far from the CFL condition of the small eigenvalues (however, this can be somewhat corrected with a MUSCL method). This is in fact a major problem when the discontinuities associated to the small eigenvalues need to be computed precisely. It is the case of interest here.

A first method to avoid the time step limitation is to take a "fully implicit" version of the schemes developed in the previous sections, that is $F_{i+1/2}^n$ function of u_j^{n+1}, $j \in \mathbb{Z}$, instead of u_j^n, $j \in \mathbb{Z}$ (the terminology "fully implicit" is by opposition to "linearly implicit", see below and Fernandez [63]).

However, in order to be competitive with explicit schemes, the fully implicit scheme is used with large time steps. In practice, this prohibits the use of a MUSCL technique in the computation of the solution at time t_{n+1} by, for instance, a Newton algorithm. This implicit scheme is therefore very diffusive and will smear discontinuities.
A second method consists in splitting the system into two systems, the first one is associated with the "small" eigenvalues, and the second one with the "large" eigenvalues (in the case of the Euler equations, this splitting may correspond to a "convection" system and a "propagation" system). At each time step, the first system is solved with an explicit scheme and the second one with an implicit scheme. Both use the same time step, which is limited by the CFL condition of the small eigenvalues. Using a MUSCL technique and an order 2 Runge-Kutta method for the first system yields sharp discontinuities associated to the small eigenvalues. This method is often satisfactory, but is difficult to handle in the case of severe boundary conditions, since the convenient boundary conditions for each system may be difficult to determine.
Another method, developed by E. Turkel (see TURKEL [145]), in connexion with Roe's scheme, uses a change of variables in order to reduce the ratio between large and small eigenvalues.

Let us now describe a partially linearly implicit method ("turbo" scheme) which was successfully tested for multiphase flows in pipe lines (see FAILLE and HEINTZÉ [60]) and other cases (see FERNANDEZ [63]). For the sake of simplicity, the method is described for the last scheme of Section 35.2, i.e. the scheme defined by (35.12)- (35.14), where $F^n_{i+\frac{1}{2}}$ is defined by (35.17) and (35.16) (recall that F may depend on x).

Assume that $I \subset \{1, \dots, m\}$ is the set of index of large eigenvalues (and does not depend on i). The aim here is to "implicit" the unknowns coresponding to the large eigenvalues only: let $\tilde{A}_i$, $\tilde{A}^{(-)}_{i+1/2}$ and $\tilde{A}^{(+)}_{i+1/2}$ be the matrix having the same eigenvectors as A_i, $A^{(-)}_{i+1/2}$ and $A^{(+)}_{i+1/2}$, with the same large eigenvalues (i.e. corresponding to $p \in I$) and 0 as small eigenvalues. Let

$$\tilde{A}^n_{i+1/2} = (\lambda/2)(\tilde{A}^{(-)}_{i+1/2} + \tilde{A}^{(+)}_{i+1/2}).$$

Then, the partially linearly implicit scheme is obtained by replacing $F^n_{i+1/2}$ in (35.5) by $\tilde{F}^n_{i+1/2}$ defined by

$$\begin{aligned}\tilde{F}^n_{i+\frac{1}{2}} = F^n_{i+\frac{1}{2}} + \tfrac{1}{2}(\tilde{A}_i(u_i^{n+1} - u_i^n) + \tilde{A}_{i+1}(u_{i+1}^{n+1} - u_{i+1}^n)) \\ + \tfrac{1}{2}\tilde{A}^n_{i+\frac{1}{2}}(u_i^{n+1} - u_i^n + u_{i+1}^n - u_{i+1}^{n+1}).\end{aligned}$$

In order to obtain sharp discontinuities corresponding to the small eigenvalues, a MUSCL technique is used for the computation of $F^n_{i+1/2}$. Then, again for stability reasons, it is preferable to add an order 2 Runge-Kutta method for the time discretization. Although it is not so easy to implement, the order 2 Runge-Kutta method is needed to enable the use of "large" time steps. The time step is, in severe situations, very close to that given by the usual CFL condition corresponding to the small eigenvalues, and can be considerably larger than that given by the large eigenvalues (see FAILLE and HEINTZÉ [60] for several tests).

35.4 Boundary conditions

In many simulations of real situations, the treatment of the boundary conditions is not easy (in particular in the case of sign change of eigenvalues). We give here a classical possible mean (see e.g. KUMBARO [95] and DUBOIS and LEFLOCH [47]) of handling boundary conditions (a more detailed description may be found in MASELLA [105] for the case of multiphase flows in pipe lines).

Let us consider now the system (35.4) where "$x \in \mathbb{R}$" is replaced by "$x \in \Omega$" with $\Omega = (0, 1)$. In order for the system to be well-posed, an initial condition (for $t = 0$) and some convenient boundary conditions

for $x = 0$ and $x = 1$ are needed; these boundary conditions will appear later in the discretization (we do not detail here the mathematical analysis of the problem of the adequacy of the boundary conditions, see e.g. SERRE [135] and references therein). Let us now explain the numerical treatment of the boundary condition at $x = 0$.

With the notations of Section 35.1, the space mesh is given by $\{K_i,\, i \in \{0, \ldots, N_{\mathcal{T}}\}\}$, with $\sum_{i=1}^{N_{\mathcal{T}}} h_i = 1$. Using the finite volume scheme (35.5) with $i \in \{1, \ldots, N_{\mathcal{T}}\}$ instead of $i \in \mathbb{Z}$ needs, for the computation of u_1^{n+1}, with $\{u_i^n,\, i \in \{1, \ldots, N_{\mathcal{T}}\}\}$ given, a value for $F_{1/2}^n$ (which corresponds to the flux at point $x = 0$ and time $t = t_n$).

For the sake of simplicity, consider only the case of the Roe and VFRoe schemes. Then, the "interior fluxes", that is $F_{i+1/2}^n$ for $i \in \{1, \ldots, N_{\mathcal{T}} - 1\}$, are determined by using matrices $A(u_i^n, u_{i+1}^n)$ ($i \in \{1, \ldots, N_{\mathcal{T}} - 1\}$). In the case of the Roe scheme, $F_{i+1/2}^n$ is given by (35.11) or (35.15) and $A(\cdot, \cdot)$ satisfies the Roe condition (35.10). In the case of the VFRoe scheme, $F_{i+1/2}^n$ is given through the resolution of the linearized Riemann problem (35.8), (35.9) with e.g. $A(u_i^n, u_{i+1}^n) = DF((1/2)(u_i^n + u_{i+1}^n))$. In order to compute $F_{1/2}^n$, a possibility is to take the same method as for the interior fluxes; this requires the determination of some u_0^n. In some cases (e.g. when all the eigenvalues of $D_u F(u)$ are nonnegative), the given boundary conditions at $x = 0$ are sufficient to determine the value u_0^n, or directly $F_{1/2}^n$, but this is not true in the general case.... In the general case, there are not enough given boundary conditions to determine u_0^n and missing equations need to be introduced. The idea is to use an iterative process. Since $A(u_0^n, u_1^n)$ is diagonalizable and has only real eigenvalues, let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $A(u_0^n, u_1^n)$ and $\varphi_1, \ldots, \varphi_m$ a basis of $\mathbb{R}^m$ associated to these eigenvalues. Then the vectors u_0^n and u_1^n may be decomposed on this basis, this yields

$$u_0^n = \sum_{i=1}^{m} \alpha_{0,i}\ \varphi_i,\ u_1^n = \sum_{i=1}^{m} \alpha_{1,i}\ \varphi_i.$$

Assume that the number of negative eigenvalues of $A(u_0^n, u_1^n)$ does not depend on u_0^n (this is a simplifying assumption); let p be the number of negative eigenvalues and $m - p$ the number of positive eigenvalues of $A(u_0^n, u_1^n)$.
Then, the number of (scalar) given boundary conditions is (hopefully ...) $m - p$. Therefore, one takes, for u_0^n, the solution of the (nonlinear) system of m (scalar) unknowns, and m (scalar) equations. The m unknowns are the components of u_0^n and the m equations are obtained with the $m - p$ boundary conditions and the p following equations:

$$\alpha_{0,i} = \alpha_{1,i},\ \text{if } \lambda_i < 0. \tag{35.18}$$

Note that the quantities $\alpha_{0,i}$ depend on $A(u_0^n, u_1^n)$; the resulting system is therefore nonlinear and may be solved with, for instance, a Newton algorithm.

Other possibilities around this method are possible. For instance, another possibility, perhaps more natural, consists in writing the $m-p$ boundary conditions on $u_{1/2}^n$ instead of u_0^n and to take (35.18) with the components of $u_{1/2}^n$ instead of those of u_0^n, where $u_{1/2}^n$ is the solution at $x = 0$ of (35.8), (35.9) with $i = 0$. With the VFRoe scheme, the flux at the boundary $x = 0$ is then $F_{1/2}^n = F(u_{1/2}^n)$. In the case of a linear system with linear boundary conditions and with the VFRoe scheme, this method gives the same flux $F_{1/2}^n$ as the preceding method, the value $u_{1/2}^n$ is completely determined although u_0^n is not completely determined.

In the case of the scheme described in the second part of Section 35.2, the following "simpler" possibility was implemented. For this scheme, $F_{i+1/2}^n$ is given, for $i \in \{1, \ldots, N_{\mathcal{T}} - 1\}$, by (35.15) with (35.16). Then, the idea is to take the same equation for the computation of $F_{1/2}^n$ but to compute u_0^n as above (that is with $m - p$ boundary conditions and (35.18)) with the choice $A(u_0^n, u_1^n) = D_u F(x_1, u_1^n)$.

This method of computation of the boundary fluxes gives good results but is not adapted to all cases (for instance, if p changes during the Newton iterations or if the number of boundary conditions is not equal to $m - p \ldots$). Some particular methods, depending on the problems under consideration, have to be developped.

We now give an attempt for the justification of this treatment of the boundary conditions, at least for a linear system with linear boundary conditions.
Consider the system

$$\begin{array}{l} u_t(x,t) + u_x(x,t) = 0,\ x \in (0,1),\ t \in \mathbb{R}_+, \\ v_t(x,t) - v_x(x,t) = 0,\ x \in (0,1),\ t \in \mathbb{R}_+, \end{array} \tag{35.19}$$

with the boundary conditions

$$\begin{array}{l} u(0,t) + \alpha v(0,t) = 0,\ t \in \mathbb{R}_+, \\ v(1,t) + \beta u(1,t) = 0,\ t \in \mathbb{R}_+, \end{array} \tag{35.20}$$

and the initial conditions

$$\begin{array}{l} u(x,0) = u_0(x),\ x \in (0,1), \\ v(x,0) = v_0(x),\ x \in (0,1), \end{array} \tag{35.21}$$

where $\alpha \in \mathbb{R}^\star$, $\beta \in \mathbb{R}^\star$, $u_0 \in L^\infty(\Omega)$ and $v_0 \in L^\infty(\Omega)$ are given. It is well known that the problem (35.19)-(35.21) admits a unique weak solution (entropy conditions are not necessary to obtain uniqueness of the solution of this linear system).

A stable numerical scheme for the discretization of the problem (35.19)-(35.21) will add some numerical diffusion terms. It seems quite natural to assume that this diffusion does not lead a coupling between the two equations of (35.19). Then, roughly speaking, the numerical scheme will consist in an approximation of the following parabolic system:

$$\begin{array}{l} u_t(x,t) + u_x(x,t) - \varepsilon u_{xx}(x,t) = 0,\ x \in (0,1),\ t \in \mathbb{R}_+, \\ v_t(x,t) - v_x(x,t) - \eta v_{xx}(x,t) = 0,\ x \in (0,1),\ t \in \mathbb{R}_+, \end{array} \tag{35.22}$$

for some $\varepsilon > 0$ and $\eta > 0$ depending on the mesh (and time step) and $\varepsilon \to 0$, $\eta \to 0$ as the space and time steps tend to 0.
In order to be well posed, this parabolic system has to be completed with the initial conditions (35.21) and (for all $t > 0$) four boundary conditions, i.e. two conditions at $x = 0$ and two conditions at $x = 1$. This is also the case for the numerical scheme which may be viewed as a discretization of (35.22). There are two boundary conditions given by (35.20). Hence two other boundary conditions must be found, one at $x = 0$ and the other at $x = 1$.

If these two additional conditions are, for instance, $v(0,t) = u(1,t) = 0$, then the (unique) solution to (35.20)-(35.22) with these two additional conditions does not converge, as $\varepsilon \to 0$ and $\eta \to 0$, to the weak solution of (35.19)-(35.21). This negative result is also true for a large choice of other additional boundary conditions. However, if the additional boundary conditions are (wisely) chosen to be $v_x(0,t) = u_x(1,t) = 0$, the solution to (35.20)-(35.22) with these two additional conditions converges to the weak solution of (35.19)-(35.21).

The numerical treatment of the boundary conditions described above may be viewed as a discretization of (35.20) and $v_x(0,t) = u_x(1,t) = 0$; this remark gives a formal justification to such a choice.

35.5 Staggered grids

For some systems of equations it may be "natural" (in the sense that the discretization seems simpler) to associate different grids to different unknowns of the problem. To each unknown is associated an equation and this equation is integrated over the elements (which are the control volumes) of the corresponding mesh, and then discretized by using one discrete unknown per control volume (and time step, for evolution problems). This is the case, for instance, of the well known discretization of the incompressible Navier-Stokes equations with staggered grids, see PATANKAR [123] and Section 36.2.
Let us now give an example in order to show that staggered grids should be avoided in the case of nonlinear hyperbolic systems since they may yield some kind of "instability". As an illustration, let us consider the following "academic" problem:

$$\begin{array}{l} u_t(x,t) + (vu)_x(x,t) = 0,\ x \in \mathbb{R},\ t \in \mathbb{R}_+, \\ v_t(x,t) + (v^2)_x(x,t) = 0,\ x \in \mathbb{R},\ t \in \mathbb{R}_+, \\ u(x,0) = u_0(x),\ x \in \mathbb{R}, \\ v(x,0) = u_0(x),\ x \in \mathbb{R}, \end{array} \tag{35.23}$$

where u_0 is a bounded function from $\mathbb{R}$ to $[0,1]$. Taking $u = v$ equal to the weak entropy solution of the Bürgers equation (namely $u_t + (u^2)_x = 0$), with initial condition u_0, leads to a solution of the problem (35.23). One would expect a numerical scheme to give an approximation of this solution. Note that the solution of the Bürgers equation, with initial condition u_0, also takes its values in $[0,1]$, and hence, a "good" numerical scheme can be expected to give approximate solutions taking values in $[0,1]$. Let us show that this property is not satisfied when using staggered grids.

Let k be the time step and h be the (uniform) space step. Let $x_i = ih$ and $x_{i+1/2} = (i+1/2)h$, for $i \in \mathbb{Z}$. Define, for $i \in \mathbb{Z}$, $K_i = (x_{i-1/2}, x_{i+1/2})$ and $K_{i+1/2} = (x_i, x_{i+1})$.
The mesh associated to u is $\{K_i, i \in \mathbb{Z}\}$ and the mesh associated to v is $\{K_{i+1/2}, i \in \mathbb{Z}\}$. Using the principle of staggered grids, the discrete unknowns are u_i^n, $i \in \mathbb{Z}$, $n \in \mathbb{N}^\star$, and $v_{i+1/2}^n$, $i \in \mathbb{Z}$, $n \in \mathbb{N}^\star$. The discretization of the initial conditions is, for instance,

$$\begin{array}{l} u_i^0 = \dfrac{1}{h}\displaystyle\int_{K_i} u_0(x)dx,\ i \in \mathbb{Z}, \\ v_{i+\frac{1}{2}}^0 = \dfrac{1}{h}\displaystyle\int_{K_{i+\frac{1}{2}}} u_0(x)dx,\ i \in \mathbb{Z}. \end{array} \tag{35.24}$$

The second equation of (35.23) does not depend on u. It seems reasonable to discretize this equation with the Godunov scheme, which is here the upstream scheme, since u_0 is nonnegative. The discretization of the first equation of (35.23) with the principle of staggered grids is easy. Since $v_{i+1/2}^n$ is always nonnegative, we also take an upstream value for u at the extremities of the cell K_i. Then, with the explicit Euler scheme in time, the scheme becomes

$$\begin{array}{l} \dfrac{1}{k}(u_i^{n+1} - u_i^n) + \dfrac{1}{h}(v_{i+\frac{1}{2}}^n u_i^n - v_{i-\frac{1}{2}}^n u_{i-1}^n) = 0,\ i \in \mathbb{Z},\ n \in \mathbb{N}, \\ \dfrac{1}{k}(v_{i+\frac{1}{2}}^{n+1} - v_{i+\frac{1}{2}}^n) + \dfrac{1}{h}((v_{i+\frac{1}{2}}^n)^2 - (v_{i-\frac{1}{2}}^n)^2) = 0,\ i \in \mathbb{Z},\ n \in \mathbb{N}. \end{array} \tag{35.25}$$

It is easy to show that, whatever k and h, there exists u_0 (function from $\mathbb{R}$ to $[0,1]$) such that $\sup\{u_i^1, i \in \mathbb{Z}\}$ is strictly larger than 1. In fact, it is possible to have, for instance, $\sup\{u_i^1, i \in \mathbb{Z}\} = 1 + k/(2h)$. In this sense the scheme (35.25) appears to be unstable. Note that the same phenomenon exists with the implicit Euler scheme instead of the explicit Euler scheme . Hence staggered grids do not seem to be the best choice for nonlinear hyperbolic systems.

36 Incompressible Navier-Stokes Equations

The discretization of the stationary Navier-Stokes equations by the finite volume method is presented in this section. We first recall the classical discretization on cartesian staggered grids. We then study, in the linear case of the Stokes equations, a finite volume method on a staggered triangular grid, for which we show, in a particular case, the convergence of the method.

36.1 The continuous equation

Let us consider here the stationary Navier-Stokes equations:

$$\begin{aligned} &-\nu\Delta u^{(i)}(x) + \sum_{j=1}^{d} u^{(j)}(x)\frac{\partial u^{(i)}}{\partial x_j}(x) + \frac{\partial p}{\partial x_i}(x) = f^{(i)}(x),\ x\in\Omega,\ \forall i = 1,\ldots,d,\\ &\sum_{i=1}^{d}\frac{\partial u^{(i)}}{\partial x_i}(x) = 0,\ x\in\Omega. \end{aligned} \tag{36.1}$$

with Dirichlet boundary condition

$$u^{(i)}(x) = 0,\ x\in\partial\Omega,\ \forall i = 1,\ldots,d, \tag{36.2}$$

under the following assumption:

Assumption 36.1

(i) Ω *is an open bounded connected polygonal subset of* $\mathbb{R}^d, d = 2, 3$,
(iii) $\nu > 0$,
(iii) $f^{(i)} \in L^2(\Omega),\ \forall i = 1,\ldots,d$.

In the above equations, $u^{(i)}$ represents the ith component of the velocity of a fluid, ν the kinematic viscosity and p the pressure. The unknowns of the problem are $u^{(i)}$, $i \in \{1,\ldots,d\}$ and p. The number of unknown functions from Ω to $\mathbb{R}$ which are to be computed is therefore $d+1$. Note that (36.1) yields $d+1$ (scalar) equations.

We shall also consider the Stokes equations, which are obtained by neglecting the nonlinear convection term.

$$\begin{aligned} &-\nu\Delta u^{(i)}(x) + \frac{\partial p}{\partial x_i}(x) = f^{(i)}(x),\ x\in\Omega,\ \forall i = 1,\ldots,d,\\ &\sum_{i=1}^{d}\frac{\partial u^{(i)}}{\partial x_i} = 0,\ x\in\Omega. \end{aligned} \tag{36.3}$$

There exist several convenient mathematical formulations of (36.1)-(36.2) and (36.3)-(36.2), see e.g. Temam [141]. Let us give one of them for the Stokes problem. Let

$$V = \{u = (u^{(1)},\ldots,u^{(d)})^t \in (H_0^1(\Omega))^d,\ \sum_{i=1}^{d}\frac{\partial u^{(i)}}{\partial x_i} = 0\}.$$

Under assumption 36.1, there exists a unique function u such that

$$\begin{aligned} &u\in V,\\ &\nu\sum_{i=1}^{d}\int_\Omega \nabla u^{(i)}(x)\cdot\nabla v^{(i)}(x)dx = \sum_{i=1}^{d}\int_\Omega f^{(i)}(x)v^{(i)}(x)dx,\ \forall v = (v^{(1)},\ldots,v^{(d)})^t \in V. \end{aligned} \tag{36.4}$$

Equation (36.4) yields the existence of $p \in L^2$ (unique if $\int_\Omega p(x)dx = 0$) such that

$$-\nu\Delta u^{(i)} + \frac{\partial p}{\partial x_i} = f^{(i)} \text{ in } \mathcal{D}'(\Omega),\ \forall i \in \{1,\dots,d\}. \tag{36.5}$$

In the following, we shall study finite volume schemes for the discretization of Problem (36.1)-(36.2) and (36.3)-(36.2). Note that the Stokes equations may also be successfully discretized by the finite element method, see e.g. Girault and Raviart [73] and references therein.

36.2 Structured staggered grids

The discretization of the incompressible Navier-Stokes equations with staggered grids is classical (see Patankar [123]): the idea is to associate different control volume grids to the different unknowns. In the two-dimensional case, the meshes consist in rectangles. Consider, for instance, the mesh, say $\mathcal{T}$, for the pressure p. Then, considering that the discrete unknowns are located at the centers of the elements of their associated mesh, the discrete unknowns for p are, of course, located at the centers of the element of $\mathcal{T}$. The meshes are staggered such that the discrete unknowns for the x-velocity are located at the centers of the edges of $\mathcal{T}$ parallel to the y-axis, and the discrete unknowns for the y-velocity are located at the centers of the edges of $\mathcal{T}$ parallel to the x-axis. The two equations of "momentum" are associated to the x and y-velocity (and integrated over the control volumes of the considered mesh) and the "divergence free" equation is associated to the pressure (and integrated over the control volume of $\mathcal{T}$). Then the discretization of all the terms of the equations is straightforward, except for the convection terms (in the momentum equations) which, eventually, have to be discretized according to the Reynolds number (upstream or centered discretization...). The convergence analysis of this so-called "MAC" (Marker and Cell) is performed in Nicolaides [114] in the linear case and Nicolaides and Wu [116] in the case of the Navier-Stokes equations.

36.3 A finite volume scheme on unstructured staggered grids

Let us now turn to the case of unstructured grids; the scheme we shall study uses the same control volumes for all the components of the velocity. The pressure unknowns are located at the vertices, and a Galerkin expansion is used for the approximation of the pressure. Note that other finite volume schemes have been proposed for the discretization of the Stokes and incompressible Navier-Stokes equations on unstructured grids (Botta and Hempel [14]), but, to our knowledge, no proof of convergence has been given yet.

We again use the notion of admissible mesh, introduced in Definition 9.1 page 37, in the particular case of triangles, if $d = 2$, or tetrahedra, if $d = 3$. We limit the description below to the case $d = 2$ and to the Stokes equations. Let Ω be an open bounded polygonal connected subset Ω of $\mathbb{R}^2$. Let $\mathcal{T}$ be a mesh of Ω consisting of triangles, satisfying the properties required for the finite element method (see e.g. Ciarlet [29]), with acute angles only. Defining, for all $K \in \mathcal{T}$, the point x_K as the intersection of the orthogonal bisectors of the sides of the triangle K yields that $\mathcal{T}$ is an admissible mesh in the sense of Definition 9.1 page 37. Let $\mathcal{S}_\mathcal{T}$ be the set of vertices of $\mathcal{T}$. For $S \in \mathcal{S}_\mathcal{T}$, let ϕ_S be the shape function associated to S in the piecewise linear finite element method for the mesh $\mathcal{T}$. For all $K \in \mathcal{T}$, let $\mathcal{S}_K \subset \mathcal{S}_\mathcal{T}$ be the set of the vertices of K.
A possible finite volume scheme using a Galerkin expansion for the pressure is defined by the following equations, with the notations of Definition 9.1 page 37:

$$\begin{aligned}\nu \sum_{\sigma\in\mathcal{E}_K} F^{(i)}_{K,\sigma} + \sum_{S\in\mathcal{S}_K} p_S \int_K \frac{\partial \phi_S}{\partial x_i}(x)dx =& \mathrm{m}(K) f^{(i)}_K,\\ &\forall K \in \mathcal{T},\ \forall i = 1,\dots,d,\end{aligned} \tag{36.6}$$

$$\begin{aligned}&F^{(i)}_{K,\sigma} = \tau_\sigma(u^{(i)}_K - u^{(i)}_L),\ \text{if } \sigma \in \mathcal{E}_{\text{int}},\ \sigma = K|L,\ i = 1,\dots,d,\\ &F^{(i)}_{K,\sigma} = \tau_\sigma u^{(i)}_K,\ \text{if } \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K,\ i = 1,\dots,d,\end{aligned} \tag{36.7}$$

$$\sum_{K\in\mathcal{T}}\sum_{i=1}^{d} u_K^{(i)} \int_K \frac{\partial \phi_S}{\partial x_i}(x)dx = 0,\ \forall S \in \mathcal{S}_\mathcal{T}, \tag{36.8}$$

$$\int_\Omega \sum_{S\in\mathcal{S}_\mathcal{T}} p_S \phi_S(x)dx = 0, \tag{36.9}$$

$$f_K^{(i)} = \frac{1}{\mathrm{m}(K)} \int_K f(x)dx,\ \forall K \in \mathcal{T}. \tag{36.10}$$

The discrete unknowns of (36.6)-(36.10) are $u_K^{(i)}$, $K \in \mathcal{T}$, $i = 1, \ldots, d$ and p_S, $S \in \mathcal{S}_\mathcal{T}$.
The approximate solution is defined by

$$p_\mathcal{T} = \sum_{S\in\mathcal{S}_\mathcal{T}} p_S \phi_S, \tag{36.11}$$

$$u_\mathcal{T}^{(i)}(x) = u_K^{(i)},\ \text{ a.e. } x \in K,\ \forall K \in \mathcal{T},\ \forall i = 1, \ldots, d. \tag{36.12}$$

The proof of the convergence of the scheme is not straightforward in the general case. We shall prove in the following proposition the convergence of the discrete velocities given by the finite volume scheme (36.6)-(36.10) in the simple case of a mesh consisting of equilateral triangles.

Proposition 36.1 *Under Assumption 36.1, let $\mathcal{T}$ be a triangular finite element mesh of Ω, with acute angles only, and let, for all $K \in \mathcal{T}$, x_K be the intersection of the orthogonal bisectors of the sides of the triangle K (hence $\mathcal{T}$ is an admissible mesh in the sense of Definition 9.1 page 37). Then, there exists a unique solution to (36.6)-(36.10), denoted by $\{u_K^{(i)},\ K \in \mathcal{T},\ i = 1, \ldots, d\}$ and $\{p_S,\ S \in \mathcal{S}_\mathcal{T}\}$. Furthermore, if the elements of $\mathcal{T}$ are equilateral triangles, then $u_\mathcal{T} \to u$ in $(L^2(\Omega))^d$, as $\mathrm{size}(\mathcal{T}) \to 0$, where u is the (unique) solution to (36.4) and $u_\mathcal{T} = (u_\mathcal{T}^{(1)}, \ldots, u_\mathcal{T}^{(d)})^d$ is defined by (36.12).*

Proof of Proposition 36.1.

Step 1 (estimate on $u_\mathcal{T}$)

Let $\mathcal{T}$ be an admissible mesh, in the sense of Proposition 36.1, and $\{u_K^{(i)},\ K \in \mathcal{T},\ i = 1, \ldots, d\}$, $\{p_S,\ S \in \mathcal{S}_\mathcal{T}\}$ be a solution of (36.6)-(36.8) with (36.10).
Multiplying the equations (36.6) by $u_K^{(i)}$, summing over $i = 1, \ldots, d$ and $K \in \mathcal{T}$ and using (36.8) yields

$$\nu \sum_{i=1}^{d} \sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma u^{(i)})^2 = \sum_{i=1}^{d} \sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^{(i)} f_K^{(i)}, \tag{36.13}$$

with $D_\sigma u^{(i)} = |u_L^{(i)} - u_K^{(i)}|$ if $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$, $i \in \{1, \ldots, d\}$ and $D_\sigma u^{(i)} = |u_K^{(i)}|$ if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$, $i \in \{1, \ldots, d\}$.
In step 2, the existence and the uniqueness of the solution of (36.6)-(36.10) will be essentially deduced from (36.13).
Using the discrete Poincaré inequality (9.13) in (36.13) gives an L^2 estimate and an estimate on the "discrete H_0^1 norm" on the component of the approximate velocities, as in Lemma 9.2 page 42, that is:

$$\|u_\mathcal{T}^{(i)}\|_{1,\mathcal{T}} \le C,\ \|u_\mathcal{T}^{(i)}\|_{L^2(\Omega)} \le C, \forall i \in \{1, \ldots, d\},$$

where C only depends on Ω, vu and $f^{(i)}$, $i = 1, \ldots, d$.
As in Theorem 9.1 page 45 (thanks to Lemma 9.3 page 44 and Theorem 14.2 page 94), this estimate gives the relative compactness in $(L^2(\Omega))^d$ of the set of approximate solutions $u_\mathcal{T}$, for $\mathcal{T}$ in the set of admissible meshes in the sense of Proposition 36.1. It also gives that if $u_{\mathcal{T}_n} \to u$ in $(L^2(\Omega))^d$, as $n \to \infty$, where $u_{\mathcal{T}_n}$ is the solution associated to the mesh $\mathcal{T}_n$, and $\mathrm{size}(\mathcal{T}_n) \to 0$ as $n \to \infty$, then $u \in (H_0^1(\Omega))^d$. This will be used in Step 3 in order to prove the convergence of $u_\mathcal{T}$ to the solution of (36.4).

Step 2 (existence and uniqueness of $u_\mathcal{T}$ and $p_\mathcal{T}$)
Let $\mathcal{T}$ be an admissible mesh, in the sense of Proposition 36.1. Replace, in the right hand side of (36.8), "0" by "g_S" with some $\{g_S,\, S \in \mathcal{S}_\mathcal{T}\} \subset \mathbb{R}$. Eliminating $F_{K,\sigma}^{(i)}$, the system (36.6)-(36.8) becomes a linear system with as many equations as unknowns. The sets of unknowns are $\{u_K^{(i)},\, K \in \mathcal{T},\, i = 1, \ldots, d\}$ and $\{p_S,\, S \in \mathcal{S}_\mathcal{T}\}$. Ordering the equations and the unknowns yields a matrix, say A, defining this system.

Let us determine the kernel of A; let $f_K^{(i)} = 0$ and $g_S = 0$ for all $K \in \mathcal{T}$, all $S \in \mathcal{S}_\mathcal{T}$ and all $i \in \{1, \ldots, d\}$. Then, (36.13) leads to $u_K^{(i)} = 0$ for all $K \in \mathcal{T}$ and all $i \in \{1, \ldots, d\}$. Turning back to (36.6) yields that $p_\mathcal{T}$ (defined by (36.11)) is constant on K for all $K \in \mathcal{T}$. Therefore, since Ω is connected, $p_\mathcal{T}$ is constant on Ω. Hence, the dimension of the kernel of A is 1 and so is the codimension of the range of A. In order to determine the range of A, note that

$$\sum_{S \in \mathcal{S}_\mathcal{T}} \varphi_S(x) = 1,\ \forall x \in \Omega.$$

Then, a necessary condition in order that the linear system (36.6)-(36.8) has a solution is

$$\sum_{S \in \mathcal{S}_\mathcal{T}} g_S = 0 \tag{36.14}$$

and, since the codimension of the range of A is 1, this condition is also sufficient. Therefore, under the condition (36.14), the linear system (36.6)-(36.8) has a solution, this solution is unique up to an additive constant for $p_\mathcal{T}$. In the particular case $g_S = 0$ for all $S \in \mathcal{S}_\mathcal{T}$, this yields that (36.6)-(36.10) has a unique solution.

Step 3 (convergence of $u_\mathcal{T}$ to u)
In this step the convergence of $u_\mathcal{T}$ towards u in $(L^2(\Omega))^d$ as $\mathrm{size}(\mathcal{T}) \to 0$ is shown for meshes consisting of equilateral triangles. Let $(\mathcal{T}_n)_{n \in \mathbb{N}}$ be a sequence of meshes (such as defined in Proposition 36.1) consisting of equilateral triangles and let $(u_{\mathcal{T}_n})_{n \in \mathbb{N}}$ be the associated solutions. Assume that $\mathrm{size}(\mathcal{T}_n) \to 0$ and $u_{\mathcal{T}_n} \to u$ in $(L^2(\Omega))^d$ as $n \to \infty$. Thanks to the compactness result of Step 1, proving that u is the solution of (36.4) is sufficient to conclude this step and to conclude Proposition 36.1.
By Step 1, $u \in (H_0^1(\Omega))^d$. It remains to show that $u \in V$ (which is the first part of (36.4)) and that u satisfies the second part of (36.4).

For the sake of simplicity of the notations, let us omit, from now on, the index n in $\mathcal{T}_n$ and let $h = \mathrm{size}(\mathcal{T})$. Note that x_K (which is the intersection of the orthogonal bisectors of the sides of the triangle K) is the center of gravity of K, for all $K \in \mathcal{T}$. Let $\varphi = (\varphi^{(1)}, \ldots, \varphi^{(d)})^t \in V$ and assume that the functions $\varphi^{(i)}$ are regular functions with compact support in Ω, say $\varphi^{(i)} \in C_c^\infty(\Omega)$ for all $i \in \{1, \ldots, d\}$. There exists $C > 0$ only depending on φ such that

$$|\varphi^{(i)}(x_K) - \frac{1}{\mathrm{m}(K)} \int_K \varphi^{(i)}(x) dx| \leq C h^2, \tag{36.15}$$

for all $K \in \mathcal{T}$ and $i = 1, \ldots, d$. Let us proceed as in the proof of convergence of the finite volume scheme for the Dirichlet problem (Theorem 9.1 page 45).
Assume that h is small enough so that $\varphi(x) = 0$ for all x such that $x \in K$, $K \in \mathcal{T}$ and $\mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}} \neq \emptyset$. Note that $(\partial \phi_S)/(\partial x_i)$ is constant in each $K \in \mathcal{T}$ and that

$$\sum_{i=1}^d \int_\Omega \frac{\partial \phi_S}{\partial x_i}(x) \varphi^{(i)}(x) dx = -\int_\Omega \phi_S(x) \sum_{i=1}^d \frac{\partial \varphi^{(i)}}{\partial x_i}(x) dx = 0.$$

Then,

$$\sum_{i=1}^d \sum_{K \in \mathcal{T}} \sum_{S \in \mathcal{S}_K} p_S \int_K \frac{\partial \phi_S}{\partial x_i}(x) dx \frac{1}{\mathrm{m}(K)} \int_K \varphi^{(i)}(x) dx = 0.$$

Therefore, multiplying the equations (36.6) by $(1/\mathrm{m}(K))\int_K \varphi^{(i)}(x)dx$, for each $i = 1, \ldots, d$, summing the results over $K \in \mathcal{T}$ and $i \in \{i \ldots, d\}$ yields

$$\nu \sum_{i=1}^{d} \sum_{K|L \in \mathcal{E}_{\rm int}} \tau_{K|L}(u_L^{(i)} - u_K^{(i)})\left(\frac{1}{\mathrm{m}(L)}\int_L \varphi^{(i)}(x)dx - \frac{1}{\mathrm{m}(K)}\int_K \varphi^{(i)}(x)dx\right) = \sum_{i=1}^{d}\sum_{K\in\mathcal{T}} f_K^{(i)} \int_K \varphi^{(i)}(x)dx. \tag{36.16}$$

Passing to the limit in (36.16) as $n \to \infty$ and using (36.15) gives, in the same way as for the Dirichlet problem (see Theorem 9.1 page 45), that u satisfies the equation given in (36.4), at least for $v \in V \cap (C_c^\infty(\Omega))^d$. Then, since $V \cap (C_c^\infty(\Omega))^d$ is dense (for the $(H_0^1(\Omega))^d$-norm) in V (see, for instance, Lions [102] for a proof of this result), u satisfies the equation given in (36.4).

Since $u \in (H_0^1(\Omega))^d$, it remains to show that u is divergence free. Let $\varphi \in C_c^\infty(\Omega)$. Multiplying (36.8) by $\varphi(S)$, summing over $S \in \mathcal{S}_\mathcal{T}$ and noting that the function $\sum_{S\in\mathcal{S}_\mathcal{T}} \varphi(S)\phi_S$ converges to φ in $H^1(\Omega)$, one obtains that u is divergence free and then belongs to V. This completes the proof that u is the (unique) solution of (36.4) and concludes the proof of Proposition 36.1. ∎

37 Flows in porous media

37.1 Two phase flow

This section is devoted to the discretization of a system which may be viewed as an elliptic equation coupled to a hyperbolic equation. This system appears in the modelling of a two phase flow in a porous medium. Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3, and let a and b be functions of class C^1 from $\mathbb{R}$ to $\mathbb{R}_+$. Assume that a is nondecreasing and b is nonincreasing. Let g and $\overline{u}$ be bounded functions from $\partial\Omega \times \mathbb{R}_+$ to $\mathbb{R}$, and u_0 be a bounded function from Ω to $\mathbb{R}$. Consider the following problem:

$$\begin{aligned}
u_t(x,t) - \mathrm{div}(a(u)\nabla p)(x,t) &= 0, \quad (x,t) \in \Omega \times \mathbb{R}_+, \\
(1-u)_t(x,t) - \mathrm{div}(b(u)\nabla p)(x,t) &= 0, \quad (x,t) \in \Omega \times \mathbb{R}_+, \\
\nabla p(x,t)\cdot \boldsymbol{n}(x) &= g(x,t), \quad (x,t) \in \partial\Omega \times \mathbb{R}_+, \\
u(x,t) &= \overline{u}(x,t), \quad (x,t) \in \partial\Omega \times \mathbb{R}_+ \ ; \ g(x,t) \geq 0, \\
u(x,0) &= u_0(x), \quad x \in \Omega,
\end{aligned} \tag{37.1}$$

where $\boldsymbol{n}$ is the normal to $\partial\Omega$, outward to Ω. The unknowns of this system are the functions p and u (from $\Omega \times \mathbb{R}_+$ to $\mathbb{R}$). Adding the two first equations of (37.1), this system may be viewed as an elliptic equation with respect to the unknown p, for a given u (note that there is no time derivative in this equation), with a Neumann condition, coupled to a hyperbolic equation with respect to the unknown u (for a given p). Note that, for the elliptic problem with the Neumann condition, the compatibility condition on g reads

$$\int_{\partial\Omega} M(u(x,t))g(x,t)d\gamma(x) = 0, \quad t \in \mathbb{R}_+,$$

where $M = a + b$. It is not known whether the system (37.1) has a solution, except in the simple case where the function M is a positive constant (which is, however, already an interesting case for real applications).

In order to discretize (37.1), let $\mathcal{T}$ be an admissible mesh of Ω in the sense of Definition 10.1 page 63 and $k > 0$ be the time step. The discrete unknowns are p_K^n and u_K^n for $K \in \mathcal{T}$ and $n \in \mathbb{N}^\star$. The discretization of the initial condition is

$$u_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x)dx, \ K \in \mathcal{T}.$$

In order to take into account the boundary condition on u, define, with $t_n = nk$,

$$\overline{u}_K^n = \frac{1}{k\,\mathrm{m}(\partial K \cap \partial\Omega)} \int_{\partial K \cap \partial\Omega} \int_{t_n}^{t_{n+1}} \overline{u}(x,t)d\gamma(x)dt, \ K \in \mathcal{T}, \quad n \in \mathbb{N}.$$

The scheme will use an "upstream choice" of $a(u)$ and $b(u)$ on each "interface" of the mesh, that is, for all $K \in \mathcal{T}$, $L \in \mathcal{N}(K)$,

$$\begin{array}{l} (a(u))_{K,L}^n = a(u_K^n) \ \text{ if } \ p_K^{n+1} \geq p_L^{n+1} \\ (a(u))_{K,L}^n = a(u_L^n) \ \text{ if } \ p_K^{n+1} < p_L^{n+1}, \\ (b(u))_{K,L}^n = b(u_K^n) \ \text{ if } \ p_K^{n+1} \geq p_L^{n+1} \\ (b(u))_{K,L}^n = b(u_L^n) \ \text{ if } \ p_K^{n+1} < p_L^{n+1}, \end{array}$$

The discrete equations are, for all $K \in \mathcal{T}$, $n \in \mathbb{N}$,

$$\begin{array}{l} \mathrm{m}(K)\dfrac{u_K^{n+1} - u_K^n}{k} - \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K|L}(p_L^{n+1} - p_K^{n+1})(a(u))_{K,L}^n \\ -\dfrac{a(\overline{u}_K^n)}{k} \displaystyle\int_{\partial K \cap \partial\Omega} \int_{t_n}^{t_{n+1}} g^+(x,t)d\gamma(x)dt + \dfrac{a(u_K^n)}{k} \int_{\partial K \cap \partial\Omega} \int_{t_n}^{t_{n+1}} g^-(x,t)d\gamma(x)dt = 0, \\ -\mathrm{m}(K)\dfrac{u_K^{n+1} - u_K^n}{k} - \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K|L}(p_L^{n+1} - p_K^{n+1})(b(u))_{K,L}^n \\ -\dfrac{b(\overline{u}_K^n)}{k} \displaystyle\int_{\partial K \cap \partial\Omega} \int_{t_n}^{t_{n+1}} g^+(x,t)d\gamma(x)dt + \dfrac{b(u_K^n)}{k} \int_{\partial K \cap \partial\Omega} \int_{t_n}^{t_{n+1}} g^-(x,t)d\gamma(x)dt = 0. \end{array}$$

Recall that $g^+(x,t) = \max\{g(x,t), 0\}$, $g^- = (-g)^+$ and $\tau_{K|L} = \mathrm{m}(K|L)/d_{K|L}$ (see Definition 9.1 page 37). This finite volume scheme gives very good numerical results under a usual stability condition on the time step with respect to the space mesh. It can be generalized to more complicated systems (in particular, for the simulation of multiphase flows in porous medium such as the "black oil" case of reservoir engineering, see Eymard [48]). It is possible to prove the convergence of this scheme in the case where the function M is constant and the function g does not depend on t. In this case, the scheme may be written as a finite volume scheme for a stationary diffusion equation with respect to the unknown p (which does not depend on t) and an upstream finite volume scheme for a hyperbolic equation with respect to the unknown u. The proof of this convergence is given below (Theorem 37.1) under the assumptions that $a(u) = u$ and $b(u) = 1-u$ (see also Vignal [151]). Note that the elliptic equation with respect to the pressure may also be discretized with a finite element method, and coupled to the finite volume scheme for the hyperbolic equation. This coupling of finite elements and finite volumes was introduced in Forsyth [68], where it is called "CVFE" (Control Volume Finite Element), in Sonier and Eymard [138] and in Eymard and Gallouët [49], where the convergence of the finite element-finite volume scheme is shown under the same assumptions.

37.2 Compositional multiphase flow

Let us now turn to the study of a system of partial differential equations which arises in the simulation of a multiphase flow in a porous medium (the so called "Black Oil" case in petroleum engineering, see e.g. Eymard [48]). This system consists in a parabolic equation coupled with hyperbolic equations and algebraic equations and inequalities (these algebraic equations and inequalities are given by an assumption of thermodynamical equilibrium). It may be written, for $x \in \Omega$ and $t \in \mathbb{R}_+$, as:

$$\frac{\partial}{\partial t}(\rho_1(p)u)(x,t) - \mathrm{div}(f_1(u,v,c)\nabla p)(x,t) = 0, \tag{37.2}$$

$$\frac{\partial}{\partial t}(\rho_2(p,c)(1-u-v)(1-c))(x,t) - \mathrm{div}(f_2(u,v,c)\nabla p)(x,t) = 0, \tag{37.3}$$

$$\frac{\partial}{\partial t}(\rho_2(p,c)(1-u-v)c + \rho_3(p)v)(x,t) - \mathrm{div}(f_3(u,v,c)\nabla p)(x,t) = 0, \tag{37.4}$$

$$(v(x,t) = 0 \text{ and } c(x,t) \le f(p(x,t)) \text{ or } (c(x,t) = f(p(x,t)) \text{ and } v(x,t) \ge 0), \tag{37.5}$$

where Ω is a given open bounded polygonal subset of $\mathbb{R}^d$ ($d = 2$ or 3), f_1, f_2, f_3 are given functions from $\mathbb{R}^3$ to $\mathbb{R}_+$, f, ρ_1, ρ_3 are given functions from $\mathbb{R}$ to $\mathbb{R}_+$ and ρ_2 is a given function from $\mathbb{R}^2$ to $\mathbb{R}_+$. The problem is completed by initial and boundary conditions which are omitted here. The unknowns of this problem are the functions u, v, c, p from $\Omega \times \mathbb{R}_+$ to $\mathbb{R}$.

In order to discretize this problem, let k be the time step (as usual, k may in fact be variable) and $\mathcal{T}$ be a cartesian mesh of Ω. Following the ideas (and notations) of the previous chapters, the discrete unknowns are u_K^n, v_K^n, c_K^n and p_K^n, for $K \in \mathcal{T}$ and $n \in \mathbb{N}^\star$ and it is quite easy to discretize (37.2)-(37.4) with a classical finite volume method. Note that the time discretization of the unknown p must generally be implicit while the time discretization of the unknowns u, v, c may be explicit or implicit. The explicit choice requires a usual restriction on the time step (linearly with respect to the space step). The only new problem is the discretization of (37.5), which is now described.

Let $n \in \mathbb{N}$. The discrete unknowns at time t_{n+1}, namely u_K^{n+1}, v_K^{n+1}, c_K^{n+1} and p_K^{n+1}, $K \in \mathcal{T}$, have to be computed from the discrete unknowns at time t_n, namely u_K^n, v_K^n, c_K^n and p_K^n, $K \in \mathcal{T}$. Even if the time discretization of (37.2)-(37.4) is explicit with respect to the unknowns u, v and c, the system of discrete equations (with unknowns u_K^{n+1}, v_K^{n+1}, c_K^{n+1} and p_K^{n+1}, $K \in \mathcal{T}$) is nonlinear, whatever the discretization of (37.5). It can be solved by, say, a Newton process. Let $l \in \mathbb{N}$ be the index of the "Newton iteration", and $u_K^{n+1,l}$, $v_K^{n+1,l}$, $c_K^{n+1,l}$ and $p_K^{n+1,l}$ ($K \in \mathcal{T}$) be the computed unknowns at iteration l. As usual, these unknowns are, for $l = 0$, taken equal to u_K^n, v_K^n, c_K^n and p_K^n. In order to discretize (37.5), a "phase index" is introduced; it is denoted by i_K^n, for all $K \in \mathcal{T}$ and $n \in \mathbb{N}$ and it is defined by:

$$\begin{array}{l} \text{if } i_K^n = 0 \text{ then } v_K^n = 0 \ \ (\text{ and } c_K^n \le f(p_K^n)), \\ \text{if } i_K^n = 1 \text{ then } c_K^n = f(p_K^n) \ \ (\text{ and } v_K^n \ge 0). \end{array}$$

In the Newton process for the computation of the unknowns at time t_{n+1}, a "phase index", denoted by $i_K^{n+1,l}$ is also introduced, with $i_K^{n+1,0} = i_K^n$. This phase index is used in the computation of $u_K^{n+1,l+1}$, $v_K^{n+1,l+1}$, $c_K^{n+1,l+1}$, $p_K^{n+1,l+1}$ and $i_K^{n+1,l+1}$ ($K \in \mathcal{T}$), starting from $u_K^{n+1,l}$, $v_K^{n+1,l}$, $c_K^{n+1,l}$, $p_K^{n+1,l}$ and $i_K^{n+1,l}$. Setting $v_K^{n+1,l+1} = 0$ if $i_K^{n+1,l} = 0$, and $c_K^{n+1,l+1} = f(p_K^{n+1,l+1})$ if $i_K^{n+1,l} = 1$, the computation of (intermediate) values of $u_K^{n+1,l+1}$, $v_K^{n+1,l+1}$, $c_K^{n+1,l+1}$, $p_K^{n+1,l+1}$ is possible with a "Newton iteration" on (37.2), (37.3), (37.4) (note that the number of unknowns is equal to the number of equations). Then, for each $K \in \mathcal{T}$, three cases are possible:

1. if $c_K^{n+1,l+1} \le f(p_K^{n+1,l+1})$ and $v_K^{n+1,l+1} \ge 0$, then set $i_K^{n+1,l+1} = i_K^{n+1,l}$,
2. if $c_K^{n+1,l+1} > f(p_K^{n+1,l+1})$ (and necessarily $i_K^{n+1,l} = 0$), then set $c_K^{n+1,l+1} = f(p_K^{n+1,l+1})$ and $i_K^{n+1,l+1} = 1$,
3. if $v_K^{n+1,l+1} < 0$ (and necessarily $i_K^{n+1,l} = 1$), then set $v_K^{n+1,l+1} = 0$ and $i_K^{n+1,l+1} = 0$.

This yields the final values of $u_K^{n+1,l+1}$, $v_K^{n+1,l+1}$, $c_K^{n+1,l+1}$, $p_K^{n+1,l+1}$ and $i_K^{n+1,l+1}$ ($K \in \mathcal{T}$).

When the "convergence" of the Newton process is achieved, say at iteration $l^\star$, the values of the unknowns at time t_{n+1} are found. They are taken equal to those indexed by $(n+1, l^\star)$ (for u, v, c, p, i). It can be proved, under convenient hypotheses on the function f (which are realistic in the applications), that there is no "oscillation" of the "phase index" during the Newton iterations performed from time t_n to time t_{n+1} (see Eymard and Gallouët [50]). This method, using the phase index, was also successfully adapted for the treatment of the obstacle problem and the Signorini problem, see Herbin and Marchand [87].

37.3 A simplified case

The aim of this section and of the following sections is the study of the convergence of two coupled finite volume schemes, for the system of equations $u_t - \mathrm{div}(u\nabla p) = 0$ and $\Delta p = 0$, defined on an open set Ω. A finite volume mesh $\mathcal{T}$ is used for the discretization in space, together with an explicit Euler time discretization. Similar results are in VIGNAL [151] and VIGNAL and VERDIÈRE [153] where the case of different space meshes for the two equations is also studied.

We assume that the following assumption is satisfied.

Assumption 37.1 *Let Ω be an open polygonal bounded connected subset of $\mathbb{R}^d$, $d = 2$ or 3, and $\partial\Omega$ its boundary. We denote by $\boldsymbol{n}$ the normal vector to $\partial\Omega$ outward to Ω.*
Let $g \in L^2(\partial\Omega)$ be a function such that

$$\int_{\partial\Omega} g(x)d\gamma(x) = 0,$$

and let $\partial\Omega^+ = \{x \in \partial\Omega,\ g(x) \geq 0\}$, $\Omega^+ = \Omega \cup \partial\Omega^+$ and $\partial\Omega^- = \{x \in \partial\Omega,\ g(x) \leq 0\}$. Let $u_0 \in L^\infty(\Omega)$ and $\bar{u} \in L^\infty(\partial\Omega^+ \times \mathbb{R}_+^\star)$ represent respectively the initial condition and the boundary condition for the unknown u.

The set

$$\mathcal{D}(\Omega^+ \times \mathbb{R}_+) = \{\varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}, \mathbb{R}),\ \varphi = 0 \text{ on } \partial\Omega^- \times \mathbb{R}_+\}$$

will be the set of test functions for Equation (37.10) in the weak formulation of the problem, which is given below.

Definition 37.1 A pair $(u, p) \in L^\infty(\Omega \times \mathbb{R}_+^\star) \times H^1(\Omega)$ (u is the saturation, p is the pressure) is a weak solution of

$$\left\{\begin{array}{ll} \Delta p(x) = 0, & \forall x \in \Omega, \\ \nabla p(x) \cdot \boldsymbol{n}(x) = g(x), & \forall x \in \partial\Omega, \\ u_t(x,t) - \mathrm{div}(u\nabla p)(x,t) = 0, & \forall x \in \Omega, \forall t \in \mathbb{R}_+, \\ u(x,0) = u_0(x), & \forall x \in \Omega, \\ u(x,t) = \bar{u}(x,t), & \forall x \in \partial\Omega^+, \forall t \in \mathbb{R}_+. \end{array}\right. \tag{37.6}$$

if it verifies

$$p \in H^1(\Omega), \tag{37.7}$$

$$u \in L^\infty(\Omega \times \mathbb{R}_+^\star), \tag{37.8}$$

$$\int_\Omega \nabla p(x) \cdot \nabla X(x)dx - \int_{\partial\Omega} X(x)g(x)d\gamma(x) = 0, \forall X \in H^1(\Omega). \tag{37.9}$$

and

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}_+}\int_\Omega u(x,t)(\varphi_t(x,t) - \nabla p(x) \cdot \nabla\varphi(x,t)dxdt + \int_\Omega u_0(x)\varphi(x,0)dx + \\ \displaystyle\int_{\mathbb{R}_+}\int_{\partial\Omega^+} \bar{u}(x,t)\varphi(x,t)g(x)d\gamma(x)dt = 0, \forall\varphi \in \mathcal{D}(\Omega^+ \times \mathbb{R}_+). \end{array} \tag{37.10}$$

Under Assumption 37.1, a classical result gives the existence of $p \in H^1(\Omega)$ and the uniqueness of ∇p where p is the solution of (37.7),(37.9), which is a variational formulation of the classical Neumann problem. Additional hypotheses on the function g are necessary to get the uniqueness of $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+^\star)$ solution of (37.10). The existence of u results from the convergence of the scheme, but not its uniqueness,

which could be obtained thanks to regularity properties of ∇p. We shall assume such regularity, which ensures the uniqueness of the function u and allows an error estimate between the finite volume scheme approximation of the pressure and the exact pressure. In fact, for the sake of simplicity, we assume (in Assumption 37.2 below) that $p \in C^2(\overline{\Omega})$. This is a rather "strong" assumption which can be weakened. However, a convergence result (such as in Theorem 37.1) with the only assumption $p \in H^1(\Omega)$ seems not easy to obtain. Note also that similar results of convergence (for the "pressure scheme" and for the "saturation scheme") are possible with an open bounded connected subset of $\mathbb{R}^d$ with a C^2 boundary (instead of an open bounded connected polygonal subset of $\mathbb{R}^d$) using Definition 18.4 page 116 of admissible meshes.

Assumption 37.2 *The pressure p, weak solution in $H^1(\Omega)$ to (37.9), belongs to $C^2(\overline{\Omega})$.*

Remark 37.1 The solution (u, p) of (37.7)-(37.10) is also a weak solution of

$$(1-u)_t(x,t) - \operatorname{div}((1-u)\nabla p)(x,t) = 0.$$

Remark 37.2 The finite volume scheme will ensure the conservation of each of the quantities u and $1-u$. It can be extended to more complex phenomena such as compressibility, thermodynamic equilibrium... (see Section 37.2)

Remark 37.3 The proof which is given here can easily be extended to the case of the existence of a source term which reads

$$\begin{array}{ll}
-\Delta p(x) = v(x), & x \in \Omega, \\
\nabla p(x) \cdot \boldsymbol{n}(x) = g(x), & x \in \partial\Omega, \\
u_t(x,t) - \operatorname{div}(u\nabla p)(x,t) + u(x,t)v^-(x) = s(x,t)v^+(x), & x \in \Omega,\ t \in \mathbb{R}_+, \\
u(x,0) = u_0(x), & x \in \Omega, \\
u(x,t) = \bar{u}(x,t), & x \in \partial\Omega^+,\ t \in \mathbb{R}_+,
\end{array}$$

where $v \in L^2(\Omega)$ with $\displaystyle\int_{\partial\Omega} g(x)d\gamma(x) + \int_\Omega v(x)dx = 0$ and $s \in L^\infty(\Omega \times \mathbb{R}_+^\star)$. All modifications which are connected to such terms will be stated in remarks.

37.4 The scheme for the simplified case

Let Ω be an open polygonal bounded connected subset of $\mathbb{R}^d$. Let $\mathcal{T}$ be an admissible mesh, in the sense of Definition 10.1 page 63, and let $h = \operatorname{size}(\mathcal{T})$. Assume furthermore that there exists $\alpha > 0$ such that $d_\sigma \geq \alpha h$ for all $\sigma \in \mathcal{E}_{\rm int}$.

The pressure finite volume scheme

We first define the approximate pressure, using the finite volume scheme defined in section 10 page 63 (that is (10.6)-(10.8)).

(i) The values G_K, for $K \in \mathcal{T}$, are defined by

$$\begin{array}{l}
G_K = \displaystyle\int_{\partial K \cap \partial\Omega} g(x)d\gamma(x) \text{ if } \mathrm{m}(\partial K \cap \partial\Omega) \neq 0, \\
G_K = 0, \text{ if } \mathrm{m}(\partial K \cap \partial\Omega) = 0.
\end{array} \tag{37.11}$$

(ii) The scheme is defined by

$$-\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(p_L - p_K\Big) = G_K,\ \forall K \in \mathcal{T}, \tag{37.12}$$

and

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)p_K = 0. \tag{37.13}$$

We recall that, from lemma 10.1 page 64, there exists a unique function $p_{\mathcal{T}} \in X(\mathcal{T})$ defined by $p_{\mathcal{T}}(x) = p_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$, where $(p_K)_{K\in\mathcal{T}}$ satisfy equations (37.11)-(37.13). Then, using Theorem 10.1 page 69, there exist C_1 and C_2, only depending on p and Ω, such that

$$\|p_{\mathcal{T}} - p\|_{L^2(\Omega)} \leq C_1 h \tag{37.14}$$

and

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L)d_{K|L}\Big(\frac{p_L - p_K}{d_{K|L}} - \frac{1}{\mathrm{m}(K|L)}\int_{K|L} \nabla p(x)\cdot \mathbf{n}_{K,L} d\gamma(x)\Big)^2 \leq (C_2 h)^2. \tag{37.15}$$

Last but not least, using lemma 10.6 page 74, there exists C_3, only depending on g and Ω, such that

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(p_L - p_K)^2 \leq (C_3)^2. \tag{37.16}$$

The saturation finite volume scheme

Let us now turn to the finite volume discretization of the hyperbolic equation (37.10). In order to write the scheme, let us introduce the following notations: let

$$G_K^{(+)} = \int_{\partial K\cap\partial\Omega} g^+(x)d\gamma(x) \text{ and } G_K^{(-)} = \int_{\partial K\cap\partial\Omega} g^-(x)d\gamma(x),$$

so that $G_K^{(+)} - G_K^{(-)} = G_K$. Let

$$G^{(+)} = \int_{\partial\Omega} g^+(x)d\gamma(x) = \sum_{K\in\mathcal{T}} G_K^{(+)}$$

(note that $G^{(+)}$ does not depend on $\mathcal{T}$). The scheme (37.12) may also be written

$$\sum_{L\in\mathcal{N}(K)} \tau_{K|L}\Big(p_L - p_K\Big) + G_K^{(+)} - G_K^{(-)} = 0,\ \forall K \in \mathcal{T}. \tag{37.17}$$

Remark 37.4 In the case of the problem with source terms, the right hand side of the equation (37.12) is replaced by $G_K + V_K^{(+)} - V_K^{(-)}$ with

$$V_K^{(\pm)} = \int_K v^{\pm}(x)dx.$$

Then, in the equation (37.17) the quantities $G_K^{(\pm)}$ are replaced by $G_K^{(\pm)} + V_K^{(\pm)}$.

Let $\xi \in (0,1)$. Given an admissible mesh $\mathcal{T}$, the time step is defined by a real value $k > 0$ such that

$$k \leq \inf_{K\in\mathcal{T}} \{\frac{\mathrm{m}(K)\ (1-\xi)}{\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ + G_K^{(+)}}. \tag{37.18}$$

Remark 37.5 Since the right hand side of (37.18) has a strictly positive lower bound, it is always possible to find values $k > 0$ which satisfy (37.18). Roughly speaking, the condition (37.18) is a linear condition between the time step and the size of the mesh. Let us explain this point in more detail: in most practical cases, function g is regular enough so that $|p_L - p_K|/d_{K|L}$ is bounded by some C only depending on g and Ω. Assume furthermore that the mesh $\mathcal{T}$ is admissible in the sense of Definition 10.1 page 63 and that, for some $\alpha > 0$, $d_{K,\sigma} \geq \alpha h$, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$. Then the condition $k \leq Dh$, with $D = ((1-\xi)\alpha)/(d(C + \|g\|_{L^\infty(\partial\Omega)}))$, implies the condition (37.18). Note also that for all $g \in L^2(\partial\Omega)$ we already have a bound for $|p_\mathcal{T}|_{1,\mathcal{T}}$ (but this does not yield a bound on $|p_L - p_K|/d_{K|L}$). Finally, note that condition (37.18) is easy to implement in practise, since the values $\tau_{K|L}$ and p_K are available by the pressure scheme.

Remark 37.6 In the problem with source terms, the condition (37.18) will be modified as follows:

$$k \leq \inf_{K\in\mathcal{T}} \quad \frac{\mathrm{m}(K)\ (1-\xi)}{\displaystyle\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ + G_K^{(+)} + V_K^{(+)}}.$$

The initial condition is discretized by:

$$u_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x)dx,\ \forall K \in \mathcal{T}. \tag{37.19}$$

We extend the definition of $\bar{u}$ by 0 on $\partial\Omega^- \times \mathbb{R}_+$, and we define $\bar{u}_K^n$, for $K \in \mathcal{T}$ and $n \in \mathbb{N}$, by

$$\begin{array}{l} \bar{u}_K^n = \dfrac{1}{k\,\mathrm{m}(\partial K \cap \partial\Omega)} \displaystyle\int_{nk}^{(n+1)k} \int_{\partial K\cap\partial\Omega} \bar{u}(x,t)d\gamma(x)dt,\ \text{if } \mathrm{m}(\partial K \cap \partial\Omega) \neq 0, \\ \bar{u}_K^n = 0,\ \text{if } \mathrm{m}(\partial K \cap \partial\Omega) = 0. \end{array} \tag{37.20}$$

Hence the following function may be defined on $\partial\Omega \times \mathbb{R}_+$:

$$\bar{u}_{\mathcal{T},k}(x,t) = \bar{u}_K^n,\ \forall x \in \partial K \cap \partial\Omega, \forall K \in \mathcal{T}, \forall t \in [nk, (n+1)k),\ n \in \mathbb{N}.$$

The finite volume discretization of the hyperbolic equation (37.10) is then written as the following relation between u_K^{n+1} and all u_L^n, $L \in \mathcal{T}$.

$$\mathrm{m}(K)(u_K^{n+1} - u_K^n) - k\ \Big[\sum_{L\in\mathcal{N}(K)} \tau_{K|L}u_{K,L}^n(p_L - p_K) + \bar{u}_K^n G_K^{(+)} - u_K^n G_K^{(-)}\Big] = 0,\ \forall K \in \mathcal{T},\ \forall n \in \mathbb{N}, \tag{37.21}$$

in which the upstream value $u_{K,L}^n$ is defined by

$$\begin{array}{l} u_{K,L}^n = u_K^n,\ \text{if } p_K \geq p_L, \\ u_{K,L}^n = u_L^n,\ \text{if } p_L > p_K. \end{array} \tag{37.22}$$

The approximate solution, denoted by $u_{\mathcal{T},k}$, is defined a.e. from $\Omega \times \mathbb{R}_+ \to$ to $\mathbb{R}$ by

$$u_{\mathcal{T},k}(x,t) = u_K^n,\ \forall x \in K,\ \forall K \in \mathcal{T},\ \forall t \in [nk, (n+1)k),\ \forall n \in \mathbb{N}. \tag{37.23}$$

Remark 37.7 In the case of source terms, the following term is defined:

$$s_K^n = \frac{1}{\mathrm{m}(K)k} \int_{nk}^{(n+1)k} \int_K s(x,t)dxdt$$

and the term $k(s_K^n V_K^{(+)} - u_K^n V_K^{(-)})$ is added to the right hand side of (37.21) .

37.5 Estimates on the approximate solution

Estimate in $L^\infty(\Omega \times \mathbb{R}_+^\star)$

Lemma 37.1 *Under the assumptions 37.1 and 37.2, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63 and $k > 0$ satisfying (37.18). Then, the function $u_{\mathcal{T},k}$ defined by (37.11)-(37.13) and (37.19)-(37.23) satisfies*

$$\|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times\mathbb{R}_+^\star)} \le \max\{\|u_0\|_{L^\infty(\Omega)}, \|\bar{u}\|_{L^\infty(\partial\Omega^+\times\mathbb{R}_+^\star)}\}. \tag{37.24}$$

Proof of Lemma 37.1

Relation (37.21) can be written as

$$\begin{aligned} u_K^{n+1} = \quad & u_K^n\big[1 - \frac{k}{\mathrm{m}(K)}\big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_K - p_L)^- + G_K^{(-)}\big)\big] + \\ & \frac{k}{\mathrm{m}(K)}\big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L} u_L^n (p_L - p_K)^+ + G_K^{(+)}\bar{u}_K^n\big). \end{aligned}$$

Using

$$\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ + G_K^{(+)} = \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_K - p_L)^- + G_K^{(-)},$$

and Inequality (37.18), the term u_K^{n+1} may be expressed as a linear combination of terms u_L^n, $L \in \mathcal{T}$, and $\bar{u}_K^n$, with positive coefficients. Thanks to relation (37.17), the sum of these coefficients is equal to 1. The estimate (37.24) follows by an easy induction. ∎

Remark 37.8 In the case of source terms, Lemma 37.1 remains true with the following estimate instead of (37.24):

$$\|u_{\mathcal{T},k}\|_{L^\infty(\Omega\times\mathbb{R}_+^\star)} \le \max\{\|u_0\|_{L^\infty(\Omega)}, \|\bar{u}\|_{L^\infty(\partial\Omega^+\times\mathbb{R}_+^\star)}, \|s\|_{L^\infty(\Omega\times\mathbb{R}_+^\star)}\}.$$

Weak BV estimate

Lemma 37.2 *Under the assumptions 37.1 and 37.2, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 10.1 page 63. Let $h = \mathrm{size}(\mathcal{T})$ and $\alpha > 0$ be such that $d_\sigma \ge \alpha h$ for all $\sigma \in \mathcal{E}_{\mathrm{int}}$. Let $k > 0$ satisfying (37.18). Let $\{u_K^n,\ K \in \mathcal{T},\ n \in \mathbb{N}\}$ be the solution to (37.19)-(37.22) with $\{p_K,\ K \in \mathcal{T}\}$ given by (37.11)-(37.13). Let $T > k$ be a given real value, and let $N_{T,k}$ be the integer value such that $N_{T,k}k < T \le (N_{T,k}+1)k$. Then there exists H, which only depends on T, Ω, u_0, $\bar{u}$, g, α and ξ, such that the following inequality holds:*

$$k\sum_{n=0}^{N_{T,k}} \sum_{K|L\in\mathcal{E}_{\mathrm{int}}} \tau_{K|L}|p_K - p_L||u_K^n - u_L^n| + k\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} G_K^{(+)}|u_K^n - \bar{u}_K^n| \le \frac{H}{\sqrt{h}}. \tag{37.25}$$

Proof of Lemma 37.2

For $n \in \mathbb{N}$ and $K \in \mathcal{T}$, multiplying (37.21) by u_K^n yields

$$\mathrm{m}(K)(u_K^{n+1}u_K^n - u_K^n u_K^n) - k\big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L} u_{K,L}^n u_K^n (p_L - p_K) + \bar{u}_K^n u_K^n G_K^{(+)} - (u_K^n)^2 G_K^{(-)}\big) = 0. \tag{37.26}$$

Writing $u_K^{n+1}u_K^n - u_K^n u_K^n = -\frac{1}{2}(u_K^{n+1} - u_K^n)^2 - \frac{1}{2}(u_K^n)^2 + \frac{1}{2}(u_K^{n+1})^2$ and summing (37.26) on $K \in \mathcal{T}$ and $n \in \{0, \ldots, N_{T,k}\}$ gives

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}\mathrm{m}(K)(u_K^{n+1}-u_K^n)^2+\frac{1}{2}\sum_{K\in\mathcal{T}}\mathrm{m}(K)((u_K^{N_{T,k}+1})^2-(u_K^0)^2)\\
&-k\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}\Big(\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)+\bar{u}_K^n u_K^n G_K^{(+)}-(u_K^n)^2G_K^{(-)}\Big)\quad=0.
\end{aligned}
\tag{37.27}
$$

Using (37.22) gives, for all $K\in\mathcal{T}$,

$$
-\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)=\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n)^2(p_K-p_L)^+-\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_L^n u_K^n(p_L-p_K)^+.
$$

Then,

$$
-\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)=\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}((u_K^n)^2-u_L^n u_K^n)(p_K-p_L)^+.
$$

Therefore, since $(u_K^n)^2-u_K^n u_L^n=\frac{1}{2}(u_K^n-u_L^n)^2+\frac{1}{2}((u_K^n)^2-(u_L^n)^2)$,

$$
\begin{aligned}
-\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)=\;&\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n-u_L^n)^2(p_K-p_L)^+\\
&+\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n)^2(p_K-p_L)^+\\
&-\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_L^n)^2(p_K-p_L)^+\\
=\;&\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n-u_L^n)^2(p_K-p_L)^+\\
&+\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n)^2(p_K-p_L)
\end{aligned}
$$

and, using (37.17),

$$
\begin{aligned}
-\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)=\;&\tfrac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\tau_{K|L}(u_K^n-u_L^n)^2(p_K-p_L)^+\\
&+\tfrac{1}{2}\sum_{K\in\mathcal{T}}G_K^{(+)}(u_K^n)^2-\frac{1}{2}\sum_{K\in\mathcal{T}}G_K^{(-)}(u_K^n)^2.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&-k\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}\Big(\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n u_K^n(p_L-p_K)+\bar{u}_K^n u_K^n G_K^{(+)}-(u_K^n)^2G_K^{(-)}\Big)\quad=\\
&\tfrac{1}{2}k\sum_{n=0}^{N_{T,k}}\Big(\sum_{K|L\in\mathcal{E}_{\mathrm{int}}}\tau_{K|L}|p_K-p_L|(u_K^n-u_L^n)^2+\sum_{K\in\mathcal{T}}G_K^{(+)}(u_K^n-\bar{u}_K^n)^2\Big)\qquad-\\
&\tfrac{1}{2}k\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}(G_K^{(+)}(\bar{u}_K^n)^2-G_K^{(-)}(u_K^n)^2).
\end{aligned}
\tag{37.28}
$$

Using (37.21), we get

$$
\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}\mathrm{m}(K)(u_K^{n+1}-u_K^n)^2=\sum_{n=0}^{N_{T,k}}\sum_{K\in\mathcal{T}}\frac{k^2}{\mathrm{m}(K)}\Big(\sum_{L\in\mathcal{N}(K)}\tau_{K|L}u_{K,L}^n(p_L-p_K)+\bar{u}_K^n G_K^{(+)}-u_K^n G_K^{(-)}\Big)^2.
$$

Then, for all $K \in \mathcal{T}$, using again (37.17) and the definition (37.22),

$$\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)^2 =$$
$$\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} \frac{k^2}{\mathrm{m}(K)} \Big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(u_L^n - u_K^n)(p_L - p_K)^+ + G_K^{(+)}(\bar{u}_K^n - u_K^n)\Big)^2.$$

The Cauchy-Schwarz inequality yields

$$\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)^2 \le \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} \frac{k^2}{\mathrm{m}(K)} \Big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ + G_K^{(+)}\Big)$$
$$\Big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ (u_L^n - u_K^n)^2 + G_K^{(+)}(\bar{u}_K^n - u_K^n)^2\Big).$$

Using the stability condition (37.18) and reordering the summations gives

$$\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)^2 \le \sum_{n=0}^{N_{T,k}} k\,(1-\xi)$$
$$\Big(\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_L - p_K|(u_L^n - u_K^n)^2 + \sum_{K\in\mathcal{T}} G_K^{(+)}(\bar{u}_K^n - u_K^n)^2\Big). \tag{37.29}$$

Using (37.27), (37.28) and (37.29), we obtain

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)((u_K^{N_{T,k}+1})^2 - (u_K^0)^2)$$
$$+\xi k \sum_{n=0}^{N_{T,k}} \Big(\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L|(u_K^n - u_L^n)^2 + \sum_{K\in\mathcal{T}} G_K^{(+)}(u_K^n - \bar{u}_K^n)^2\Big) \tag{37.30}$$
$$-k \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} (G_K^{(+)}(\bar{u}_K^n)^2 - G_K^{(-)}(u_K^n)^2) \le 0.$$

Then, setting $C_4 = \mathrm{m}(\Omega)\|u_0\|^2_{L^\infty(\Omega)} + 2TG^{(+)}\|\bar{u}\|^2_{L^\infty(\partial\Omega^+\times\mathbb{R}_+^\star)}$ which only depends on Ω, u_0, T, g and $\bar{u}$,

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)(u_K^{N_{T,k}+1})^2 + k \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} G_K^{(-)}(u_K^n)^2 \le C_4$$

(this inequality will not be used in the sequel) and

$$k \sum_{n=0}^{N_{T,k}} \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L|(u_K^n - u_L^n)^2 + k \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} G_K^{(+)}(u_K^n - \bar{u}_K^n)^2 \le \frac{C_4}{\xi}. \tag{37.31}$$

The Cauchy-Schwarz inequality yields

$$k \sum_{n=0}^{N_{T,k}} \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L||u_K^n - u_L^n| + k \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} G_K^{(+)}|u_K^n - \bar{u}_K^n| \qquad \le$$
$$(k \sum_{n=0}^{N_{T,k}} \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L|(u_K^n - u_L^n)^2 + k \sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} G_K^{(+)}(u_K^n - \bar{u}_K^n)^2)^{\frac12} \tag{37.32}$$
$$\Big(k \sum_{n=0}^{N_{T,k}} (\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L| + \sum_{K\in\mathcal{T}} G_K^{(+)})\Big)^{\frac12}$$

The expression W, defined by $W = \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}|p_K - p_L|$, verifies

$$W \le (\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L})^{\frac{1}{2}} (\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(p_K - p_L)^2)^{\frac{1}{2}} \le C_3 (\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L})^{\frac{1}{2}} \tag{37.33}$$

using (37.16). Recall that C_3 only depends on g and Ω.
Since

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L} \le (\sum_{K|L\in\mathcal{E}_{\rm int}} {\rm m}(K|L) d_{K|L}) \frac{1}{\alpha^2 h^2} \le \frac{d{\rm m}(\Omega)}{\alpha^2 h^2} \tag{37.34}$$

and

$$\sum_{K\in\mathcal{T}} G_K^{(+)} = \int_{\partial\Omega} g^+(x) d\gamma(x),$$

we finally conclude that (37.25) holds.

Remark 37.9 In the case of source terms, one adds the term $k\sum_{n=0}^{N_{T,k}} \sum_{K\in\mathcal{T}} V_K^{(+)} |u_K^n - s_K^n|$ in the left hand side of (37.25) (and H also depends on v and s).

37.6 Theorem of convergence

We already know, by the results of section 10 page 63, that the pressure scheme converges. Let us now prove the convergence of the saturation scheme (37.21). Thanks to the estimate (37.24) in $L^\infty(\Omega\times\mathbb{R}_+^\star)$ (Lemma 37.1), for any sequence of meshes and time steps, such that the size of the mesh tends to 0, we can extract a subsequence such that the approximate saturation converges to a function u in $L^\infty(\Omega\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology. We have to show that u is the (unique) solution of (37.8), (37.10) (the uniqueness of the solution is given by Assumption 37.2).

Theorem 37.1 *Under assumptions 37.1 and 37.2, let $\xi \in (0,1)$ and $\alpha > 0$ be given. For an admissible mesh $\mathcal{T}$, in the sense of Definition 10.1 page 63, such that $d_\sigma \ge \alpha\,{\rm size}(\mathcal{T})$ for all $\sigma \in \mathcal{E}_{\rm int}$ and for a time step $k > 0$ satisfying (37.18), let $u_{\mathcal{T},k}$ be defined by (37.11)-(37.13) and (37.19)-(37.23). Then $u_{\mathcal{T},k}$ converges to the solution u of (37.8), (37.10) in $L^\infty(\Omega\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology, as* ${\rm size}(\mathcal{T}) \to 0$.

Proof of Theorem 37.1

In the case $g(x) = 0$ for a.e. (for the $(d-1)$-dimensional Lebesgue measure) $x \in \partial\Omega$, the proof of Theorem 37.1 is easy. Indeed, $\nabla p(x) = 0$ for a.e. $x \in \Omega$ and, for any mesh and time step, $p_K - p_L = 0$ for all K, $L \in \mathcal{T}$. Then, $u_K^n = u_K^0$ for all $K \in \mathcal{T}$ and all $n \in \mathbb{N}$. Therefore, it is easy to prove that the sequence $u_{\mathcal{T},k}$ converges, as ${\rm size}(\mathcal{T}) \to 0$ (for any $k\ldots$), to u, defined by $u(x,t) = u_0(x)$ for a.e. $(x,t) \in \Omega\times\mathbb{R}_+$; note that u is the unique solution to (37.8), (37.10).

Let us now assume that g is not the null function in $L^2(\partial\Omega)$.

Let $(\mathcal{T}_m, k_m)_{m\in\mathbb{N}}$ be a sequence of space meshes and time steps. For all $m \in \mathbb{N}$, assume that $\mathcal{T}_m$ is an admissible mesh in the sense of Definition 10.1, that $d_\sigma \ge \alpha{\rm size}(\mathcal{T}_m)$ for all $\sigma \in \mathcal{E}_{\rm int}$ and that $k_m > 0$ satisfies (37.18) (with $k = k_m$ and $\mathcal{T} = \mathcal{T}_m$). Assume also that ${\rm size}(\mathcal{T}_m) \to 0$ as $m \to \infty$.
Let u_m be the function $u_{\mathcal{T},k}$ defined by (37.11)-(37.13) and (37.19)-(37.23), for $\mathcal{T} = \mathcal{T}_m$ and $k = k_m$. By Lemma 37.1, the sequence $(u_m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(\Omega\times\mathbb{R}_+^\star)$. In order to prove that the sequence $(u_m)_{m\in\mathbb{N}}$ converges in $L^\infty(\Omega\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology to the solution of (37.8), (37.10), using a classical contradiction argument, it is sufficient to prove that if $u_m \to u$ in $L^\infty(\Omega\times\mathbb{R}_+^\star)$ for the weak-$\star$ topology then the function u is a solution of (37.8), (37.10).

Let us proceed in two steps. In the first step, it is proved that $k_m \to 0$ as $m \to \infty$. Then, in the second step, it is proved that the function u is a solution of (37.8), (37.10).
From now on, the index "m" is omitted.

Step 1 (proof of $k \to 0$ as $m \to \infty$)
The proof that $k \to 0$ (as $m \to \infty$) uses (37.18) and the fact that size($\mathcal{T}$) $\to 0$.Indeed, define

$$A_{\mathcal{T}} = \sum_{K|L \in \mathcal{E}_{\rm int}} \mathrm{m}(K|L)|p_K - p_L|,$$

and, for $\sigma \in \mathcal{E}_{\rm int}$, define χ_σ from $\Omega \times \Omega$ to $\{0,1\}$ by

$$\chi_\sigma(x,y) = 1, \text{ if } \sigma \cap [x,y] \neq \emptyset,$$

$$\chi_\sigma(x,y) = 0, \text{ if } \sigma \cap [x,y] = \emptyset.$$

Let $\eta \in \mathbb{R}^d \setminus \{0\}$ and $\bar{\omega} \subset \Omega$ be a compact set such that $d(\bar{\omega}, \Omega^c) \geq \eta$. Recall that $p_{\mathcal{T}}$ is defined by $p_{\mathcal{T}}(x) = p_K$ for a.e. $x \in K$ and all $K \in \mathcal{T}$. For a.e. $x \in \bar{\omega}$ one has

$$|p_{\mathcal{T}}(x+\eta) - p_{\mathcal{T}}(x)| \leq \sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} \chi_\sigma(x, x+\eta)|p_K - p_L|,$$

integrating this inequality over $\bar{\omega}$ yields, using $\int_{\bar{\omega}} \chi_\sigma(x, x+\eta)dx \leq |\eta|\mathrm{m}(\sigma)$,

$$\|p_{\mathcal{T}}(\cdot + \eta) - p_{\mathcal{T}}\|_{L^1(\bar{\omega})} \leq |\eta| A_{\mathcal{T}}. \tag{37.35}$$

Assume $A_{\mathcal{T}} \to 0$ as $m \to \infty$. Then, since $p_{\mathcal{T}} \to p$ in $L^1(\Omega)$, one deduces from (37.35) that $\nabla p = 0$ a.e. on Ω which is impossible (since g is not the null function in $L^2(\partial\Omega)$). By the same way, it is also impossible that $A_{\mathcal{T}} \to 0$ for a subsequence. Then there exists $a > 0$ (only depending on the sequence $(p_{\mathcal{T}})_{m \in \mathbb{N}}$, recall that $p_{\mathcal{T}} = p_{\mathcal{T}_m}$ since we omit the index m) such that $A_{\mathcal{T}} \geq a$ for all $m \in \mathbb{N}$.
Therefore, since $A_{\mathcal{T}} = \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \mathrm{m}(K|L)(p_L - p_K)^+ \geq a$, there exists $K \in \mathcal{T}$ such that

$$\sum_{L \in \mathcal{N}(K)} \mathrm{m}(K|L)(p_L - p_K)^+ \geq a \frac{\mathrm{m}(K)}{\mathrm{m}(\Omega)},$$

Then, since $\tau_{K|L} = \mathrm{m}(K|L)/d_{K|L}$ and $d_{K|L} \leq 2h$,

$$\sum_{L \in \mathcal{N}(K)} \tau_{K|L}(p_L - p_K)^+ \geq a \frac{\mathrm{m}(K)}{2h\mathrm{m}(\Omega)},$$

which yields, using (37.18),

$$k \leq (1-\xi)\mathrm{m}(\Omega)\frac{2}{a}h.$$

Hence $k \to 0$ as $m \to \infty$ (since $h \to 0$ as $m \to \infty$). This concludes Step 1.

Step 2 (proof of u solution to (37.10))
Let $\varphi \in \mathcal{D}(\Omega^+ \times \mathbb{R}_+)$. Let $T > 0$ such that, for all $t > T-1$ and all $x \in \Omega$, $\varphi(x,t) = 0$. Let $m \in \mathbb{N}$ such that $h < 1$ and $k < 1$ (thanks to Step 1, this is true for m large enough). Recall that we denote $\mathcal{T} = \mathcal{T}_m$, $h = \mathrm{size}(\mathcal{T}_m)$ and $k = k_m$. Let $N_{T,k} \in \mathbb{N}$ be such that $N_{T,k}k < T \leq (N_{T,k}+1)k$. Multiplying equation (37.21) by $\varphi(x_K, nk)$ and summing the result on $K \in \mathcal{T}$ and $n \in \mathbb{N}$ yields

$$E_{1,m} + E_{2,m} = 0,$$

with

$$E_{1,m} = \sum_{n=0}^{N_{T,k}} \sum_{K \in \mathcal{T}} \mathrm{m}(K)(u_K^{n+1} - u_K^n)\varphi(x_K, nk)$$

and

$$E_{2,m} = -\sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} \Big(\sum_{L\in\mathcal{N}(K)} \tau_{K|L} u_{K,L}^n (p_L - p_K) + G_K^{(+)} \bar{u}_K^n - G_K^{(-)} u_K^n \Big) \varphi(x_K, nk).$$

It is shown below that

$$\lim_{m\to\infty} E_{1,m} = T_1, \tag{37.36}$$

where

$$T_1 = -\int_{\mathbb{R}_+} \int_\Omega u(x,t)\varphi_t(x,t)dxdt - \int_\Omega u_0(x)\varphi(x,0)dx,$$

and that

$$\lim_{m\to\infty} E_{2,m} = T_2, \tag{37.37}$$

where

$$T_2 = \int_{\mathbb{R}_+} \int_\Omega u(x,t)\nabla p(x)\cdot\nabla\varphi(x,t)dxdt - \int_{\mathbb{R}_+}\int_{\partial\Omega} \bar{u}(x,t)\varphi(x,t)g(x)d\gamma(x)dt.$$

Then, passing to the limit in $E_{1,m} + E_{2,m} = 0$ proves that u is the (unique) solution of (37.8), (37.10) and concludes the proof of Theorem 37.1.

Let us first prove (37.36). Writing $E_{1,m}$ in the following way:

$$E_{1,m} = \sum_{n=1}^{N_{T,k}} \sum_{K\in\mathcal{T}} \mathrm{m}(K) \frac{\varphi(x_K,(n-1)k) - \varphi(x_K,nk)}{k} u_K^n - \sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^0 \varphi(x_K,0),$$

the assertion (37.36) is easily proved, in the same way as, for instance, in the proof of Theorem 18.1 page 113.

Let us prove now (37.37). To this purpose, we need auxiliary expressions, which make use of the convergence of the approximate pressure to the continuous one. Define $E_{3,m}$ and $E_{4,m}$ by

$$\begin{aligned} E_{3,m} &= \sum_{n=0}^{N_{T,k}} k \sum_{K|L\in\mathcal{E}_{\rm int}} (u_K^n - u_L^n) \frac{p_L - p_K}{d_{K|L}} \int_{K|L} \varphi(x,nk)d\gamma(x) \\ &\quad + \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} (u_K^n - \bar{u}_K^n) \int_{\partial K\cap\partial\Omega} g(x)\varphi(x,nk)d\gamma(x) \end{aligned}$$

and

$$E_{4,m} = \sum_{n\in\mathbb{N}} \int_{nk}^{(n+1)k} \Big(\int_\Omega u_{\mathcal{T},k}(x,t)\nabla p(x)\cdot\nabla\varphi(x,nk)dx - \int_{\partial\Omega} \bar{u}_{\mathcal{T},k}(x,t)\varphi(x,nk)g(x)d\gamma(x)\Big)dt.$$

We have $E_{4,m}\to T_2$ as $m\to\infty$ thanks to the convergence of $u_{\mathcal{T},k}$ to u in $L^\infty(\Omega\times\mathbb{R})$ for the weak-$\star$ topology and to the convergence of $\bar{u}_{\mathcal{T},k}$ to $\bar{u}$ in $L^\infty(\partial\Omega^+\times\mathbb{R}_+)$ for the weak-$\star$ topology (the latter convergence holds also in $L^p(\partial\Omega^+\times(0,S))$ for all $1\le p<\infty$ and all $0<S<\infty$). Let us prove that $|E_{3,m} - E_{4,m}|\to 0$ as $m\to\infty$ (which gives $E_{3,m}\to T_2$ as $m\to\infty$).

using the equation satisfied by p leads to

$$\begin{aligned} E_{4,m} &= \sum_{n=0}^{N_{T,k}} k \sum_{K|L\in\mathcal{E}_{\rm int}} (u_K^n - u_L^n) \int_{K|L} \varphi(x,nk)\nabla p(x)\cdot \mathbf{n}_{K,L} d\gamma(x) \\ &\quad + \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} (u_K^n - \bar{u}_K^n) \int_{\partial K\cap\partial\Omega} g(x)\varphi(x,nk) d\gamma(x). \end{aligned}$$

Therefore,

$$\begin{aligned} E_{3,m} - E_{4,m} &= \sum_{n=0}^{N_{T,k}} k \sum_{K|L\in\mathcal{E}_{\rm int}} (u_K^n - u_L^n) \int_{K|L} (\frac{p_L - p_K}{d_{K|L}} - \nabla p(x)\cdot \mathbf{n}_{K,L})\varphi(x,nk) d\gamma(x) \\ &= \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} u_K^n \Big(\sum_{L\in\mathcal{N}(K)} \int_{K|L} (\frac{p_L - p_K}{d_{K|L}} - \nabla p(x)\cdot \mathbf{n}_{K,L})\varphi(x,nk) d\gamma(x)\Big). \end{aligned}$$

Using the equation satisfied by the pressure in (37.6) and the pressure scheme (37.12) yields

$$E_{3,m} - E_{4,m} = \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} u_K^n \Big(\sum_{L\in\mathcal{N}(K)} \int_{K|L} (\frac{p_L - p_K}{d_{K|L}} - \nabla p(x)\cdot \mathbf{n}_{K,L})(\varphi(x,nk) - \varphi(x_K,nk)) d\gamma(x)\Big).$$

Thanks to the regularity of φ and p, there exists $C_5 > 0$, only depending on p, and C_6, only depending on φ, such that, for all $K|L \in \mathcal{E}_{\rm int}$,

$$|\frac{p_L - p_K}{d_{K|L}} - \nabla p(x)\cdot \mathbf{n}_{K,L})| \le |\frac{p_L - p_K}{d_{K|L}} - \frac{1}{{\rm m}(K|L)} \int_\sigma \nabla p(x)\cdot \mathbf{n}_{K,L} d\gamma(x)| + C_5 h,\ \forall x \in K|L$$

and, for all $K \in \mathcal{T}$,

$$|\varphi(x,nk) - \varphi(x_K,nk)| \le C_6 h,\ \forall x \in \overline{K},\ \forall n \in \mathbb{N}.$$

Thus,

$$\begin{aligned} |E_{3,m} - E_{4,m}| &\le \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} |u_K^n| \Big(\sum_{L\in\mathcal{N}(K)} |\tau_{K|L}(p_L - p_K) - \int_{K|L} \nabla p(x)\cdot \mathbf{n}_{K,L} d\gamma(x)|\Big) C_6 h \\ &\quad + \sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} |u_K^n| (\sum_{L\in\mathcal{N}(K)} {\rm m}(K|L) C_6 C_5 h^2), \end{aligned}$$

which leads to $|E_{3,m} - E_{4,m}| \to 0$ as $m \to \infty$, using (37.15), (37.34) and the Cauchy-Schwarz inequality. In order to prove that $E_{2,m} \to T_2$ as $m \to \infty$ (which concludes the proof of Theorem 37.1), let us show that $|E_{2,m} - E_{3,m}| \to 0$ as $m \to \infty$.
We get, using (37.17) and (37.22)

$$\begin{aligned} E_{2,m} = &\ -\sum_{n=0}^{N_{T,k}} k \sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(u_L^n - u_K^n)(p_L - p_K)\varphi(x_K,nk) \\ &\ -\sum_{n=0}^{N_{T,k}} k \sum_{K\in\mathcal{T}} (\bar{u}_K^n - u_K^n) G_K^{(+)} \varphi(x_K,nk). \end{aligned}$$

This yields

$$E_{3,m} - E_{2,m} = \sum_{n=0}^{N_{T,k}} k \sum_{K|L \in \mathcal{E}_{\rm int}} \tau_{K|L}(u_K^n - u_L^n)(p_L - p_K)\phi_{K,L}^n + \sum_{n=0}^{N_{T,k}} k \sum_{K \in \mathcal{T}} (u_K^n - \bar{u}_K^n) G_K^{(+)} \phi_K^n, \tag{37.38}$$

where

$$\phi_{K,L}^n = \frac{1}{\mathrm{m}(K|L)} \int_{K|L} \varphi(x, nk) d\gamma(x) - \varphi(x_K, nk), \ \forall K \in \mathcal{T}, \ \forall L \in \mathcal{N}(K)$$

and

$$G_K^{(+)} \phi_K^n = \int_{\partial K \cap \partial\Omega} \varphi(x, nk) g(x) d\gamma(x) - G_K^{(+)} \varphi(x_K, nk).$$

We recall that, for all $x \in \partial\Omega$, $\varphi(x, nk) g^+(x) = \varphi(x, nk) g(x)$, by definition of $\mathcal{D}(\Omega^+ \times \mathbb{R}_+)$. Therefore, there exists C_7, which only depends on φ, such that $|\phi_{K,L}^n| \leq C_7 h$ and $G_K^{(+)} |\phi_K^n| \leq G_K^{(+)} C_7 h$, for all $K \in \mathcal{T}$, $L \in \mathcal{N}(K)$ and all $n \in \mathbb{N}$. Therefore, using Lemma 37.2, we get $|E_{3,m} - E_{2,m}| \leq C_7 h \frac{H}{\sqrt{h}}$ which yields $|E_{2,m} - E_{3,m}| \to 0$ and then $E_{2,m} \to T_2$ as $m \to \infty$. This concludes the proof of Theorem 37.1. ∎

Remark 37.10 In the case of source terms, the convergence theorem 37.1 still holds. There are some minor modifications in the proof. The definitions of $E_{2,m}$, $E_{3,m}$ and $E_{4,m}$ change. In the definition of $E_{2,m}$, the quantity $G_K^{(+)} \bar{u}_K^n - G_K^{(-)} u_K^n$ is replaced by $G_K^{(+)} \bar{u}_K^n - G_K^{(-)} u_K^n + V_K^{(+)} s_K^n - V_K^{(-)} u_K^n$. In the definition of $E_{3,m}$ one adds

$$\sum_{n=0}^{N_{T,k}} k \sum_{K \in \mathcal{T}} (u_K^n - s_K^n) \int_K v^+(x) \varphi(x, nk) dx.$$

The quantity $E_{3,m} - E_{4,m}$ does not change and in order to prove $E_{3,m} - E_{2,m} \to 0$ it is sufficient to remark that there exists C_8, only depending on φ, such that

$$\left| \int_K \varphi(x, nk) v^+(x) dx - V_K^{(+)} \varphi(x_K, nk) \right| \leq V_K^{(+)} C_8 h.$$

38 Boundary conditions

In the industrial context, efficient numerical simulators are often developped after a long "trial and error" procedure. The efficiency of the simulators may be evaluated, for instance, by the fact that the solution satisfies some natural constraints and that it is in agreement with experimental data. In some cases, estimates on the approximate solutions allow to obtain the convergence of some sequences of approximate solutions as the discretization size tends to 0. However, it is not easy to give the answer to the following question: "Which problem is the limit of the approximate solutions the unique solution to ?".

This paper will focus on the problem of boundary conditions needed in the discretization of non linear hyperbolic equations or systems of equations; this problem is not yet clearly understood in many cases. Two different cases will be presented: a two phase flow in a pipeline and a two phase flow in a porous medium.

38.1 A two phase flow in a pipeline

Description of the system A "simple" model for a two phase flow in a pipeline (see [60], for instance) leads to a 3×3 system of conservations laws. The unknown w is a function from $(0,1) \times \mathbf{R}_+$ in $\mathbf{R}^3$, solution of the following system:

$$w_t + (F(w))_x = 0,\ x \in (0,1),\ t \in \mathbf{R}_+, \tag{38.1}$$

where $(\cdot)_t$ and $(\cdot)_x$ denote the derivatives with respect to t and x variables. The first two equations of (38.1) give the mass conservation of the 2 phases (gas and liquid) and the third one is the momentum equation for the mixture. The expression of the given function $F : \mathbf{R}^3 \to \mathbf{R}^3$ is quite complicated. It takes into account thermodynamical laws and a hydrodynamical law. System (38.1) is hyperbolic: for any $w \in \mathbf{R}^3$, the Jacobian matrix $DF(w)$ is diagonalizable in $\mathbf{R}$. The three eigenvalues can be ordered: $\lambda_1(w) < \lambda_2(w) < \lambda_3(w)$. In real situations, the first eigenvalue, $\lambda_1(w)$ is negative and the third, $\lambda_3(w)$, is positive (they correspond to some "pressure waves" which are related to a "sound velocity"). The second eigenvalue, $\lambda_2(w)$, corresponds to some mean velocity between the two phases and can change sign. One can also note that the field related to this second eigenvalue is quite complicated because it is not, in general, a genuinely non linear field or a linearly degenerate field. In petroleum engineering, the wave associated to this second eigenvalue is a "void fraction wave"; engineers require a good representation of this wave in the numerical simulations.

Remark 38.1 *In real situations, the function F in System (38.1) also depends on x, in order to take into account, for instance, the variation in the slope of the pipeline. Moreover, some source terms have to be added to the system, in order to take into account, for instance, some friction terms.*

In order to complete System (38.1), an initial condition is prescribed:

$$w(x,0) = w_0(x),\ x \in (0,1), \tag{38.2}$$

and it is also necessary to give some boundary conditions. This appears to be not so easy. Indeed, classically, a general principle is that the number of boundary conditions needs to be equal to the number of positive eigenvalues of the Jacobian matrix at $x = 0$ and to the number of negative eigenvalues of the Jacobian matrix at $x = 1$ (and these boundary conditions have to satisfy some compatibility conditions). However, this principle is not so easy to understand when an eigenvalue changes sign during the simulation (or in the case of a null eigenvalue). A very interesting case is the so called "severe slugging" case in a pipeline. For this case, there are always two positive eigenvalues at $x = 0$ and two natural boundary conditions are prescribed at $x = 0$, namely the fluxes of gas and liquid; these boundary conditions can be taken constant in time. At $x = 1$, there is one natural boundary condition, namely the pressure (which is the same for the two phases, in this model), to be prescribed. It can also be constant in time. The true physical solution, which is measured by experiments (and the aim is to modelize these experiments), is periodical in time and it appears that, at $x = 1$, the first eigenvalue is always positive and the third one is always negative but the second eigenvalue changes sign during the simulation. In the sequel, one presents different ways to take into account the boundary conditions and one gives a convergence result in a simplified case.

Discretization of the problem In order to discretize Problem (38.1), (38.2) and some boundary conditions, which will be introduced later, let $h = \frac{1}{N}$ (with $N \in \mathbf{N}^\star$) be the mesh size and $k > 0$ be the time step (assumed to be constant, for the sake of simplicity). The discrete unknown are the values $w_i^n \in \mathbf{R}^3$ for $i \in \{1, \ldots, N\}$ and $n \in \mathbf{N}$. The discretization of the initial condition leads to

$$w_i^0 = \frac{1}{h} \int_{(i-1)h}^{ih} w_0(x)dx,\ i \in \{1, \ldots, N\}. \tag{38.3}$$

For the computation of w_i^n for $n > 0$, one uses an explicit, 3-points scheme:

$$\frac{h}{k}(w_i^{n+1} - w_i^n) + F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n = 0,\ i \in \{1, \dots, N\},\ n \in \mathbf{N}. \tag{38.4}$$

For $i \in 1, \dots, N-1$, one takes $F_{i+\frac{1}{2}}^n = g(w_i^n, w_{i+1}^n)$, where g is the numerical flux. It has to satisfy, in particular, the classical consistency condition, namely $g(a,a) = F(a)$, and needs to be chosen in order to obtain some stability properties for the numerical scheme under a so called CFL condition on the time step (see Sect. 23 for the study of a scalar model). In the case of two phase flow in a pipeline, the classical numerical fluxes such as the Godunov flux (see [77]) or the Roe flux (see [128]) may not be implemented, because of computational difficulties. A convenient choice is obtained with a simplified Roe flux, namely $g(a,b) = \frac{g(a)+g(b)}{2} + \frac{1}{2}|A(a,b)|(a-b)$, where $A(a,b)$ is some appoximation of the Jacobian matrix, depending on a and b, but not satisfying the so called Roe condition, see [60].

Remark 38.2 *In fact, for the simulation of a two phase flow in a pipeline, the magnitude of the so-called fast eigenvalues, λ_1 and λ_3, is much greater than that of λ_2; the choice in [60] is to use an implicit scheme with respect to the fast eigenvalues, whereas the eingevalue λ_2, which corresponds to the void fraction wave, is handled with an explicit second order discretization, since the void fraction wave needs to be simulated precisely (see [60] for details).*

Let us now define the fluxes $F_{\frac{1}{2}}^n$ and $F_{N+\frac{1}{2}}^n$ at the boundary.

Boundary conditions for the discretized problem In order to compute $F_{\frac{1}{2}}^n$ (and similarily $F_{N+\frac{1}{2}}^n$) a good way is to know, or to determine, some artificial value $w_0^n \in \mathbf{R}^3$ (and $w_{N+1}^n \in \mathbf{R}^3$) and to take $F_{\frac{1}{2}}^n = g_0(w_0^n, w_1^n)$ (and $F_{N+\frac{1}{2}}^n = g_1(w_N^n, w_{N+1}^n)$). The numerical fluxes g_0 and g_1 can be chosen equal to g, but this is not at all necessary (see the convergence result of sections 23 and 31); in fact, there are numerous situations where one should take g_0 and g_1 different from g. Indeed, the scheme is often very sensitive to the computation of the boundary fluxes and it is often worthwhile to use a more precise, but also more expensive numerical flux (such as the Godunov flux, for instance) for the computation of the boundary fluxes than for the computation of the interior fluxes. The difficulty is now to determine these artificial values, w_0^n and w_{N+1}^n.

Remark 38.3 *In some cases, the choice of w_0^n and w_{N+1}^n is quite easy. A well known example is given by the wall-boundary condition for the Euler equations (with a perfect gas state law or a more general state law). For the sake of simplicity, let us mention the one-dimensional case; the generalization to a multi-dimensional case is quite easy. The Euler equations may be written the form (38.1), corresponding to conservation of mass, momentum and energy, with $w = (\rho, \rho u, E)^t$, where ρ is the density of the fluid, u its velocity, and E its energy. The wall-boundary condition at $x = 0$ is $u = 0$, and the only component to compute for the boundary condition is the second component of $F_{\frac{1}{2}}^n$ which is equal here to the pressure at $x = 0$ (since $u = 0$ at the wall), say $p_{\frac{1}{2}}^n$. The value w_1^n may be computed from the values ρ_1^n, u_1^n and p_1^n. A natural choice for w_0^n is to take $\rho_0^n = \rho_1^n$, $u_0^n = -u_1^n$ and $p_0^n = p_1^n$. The flux $F_{\frac{1}{2}}^n$ (that is the value $p_{\frac{1}{2}}^n$) is then obtained with $F_{\frac{1}{2}}^n = g_0(w_0^n, w_1^n)$ and a convenient choice of the numerical flux g_0. We suggest to choose g_0 as the Godunov flux (or as a linearized Godunov flux, see [19] for instance). Numerical tests which were performed in [19] show that this choice is very satisfactory, even in the difficult case of a strong depressurization at the boundary. These tests also show that the pressure obtained with the Roe flux is not so satisfactory and neither is the choice $p_{\frac{1}{2}}^n = p_1^n$ which may seem natural (in particular, in 2D simulations, using a dual mesh obtained with a finite element primal mesh).*

In most cases, however, the choice of w_0^n and w_{N+1}^n is not so easy. A possible method, which is described in [53], is now layed out, for a fixed n and g_0 given:

1. Compute $DF(w_1^n)$, its eigenvalues $\{\lambda_1, \lambda_2, \lambda_3\}$ and a basis of $\mathbf{R}^3$, $\{\varphi_1, \varphi_2, \varphi_3\}$, such that $DF(w_1^n)\varphi_i = \lambda_i \varphi_i$, $i = 1, 2, 3$.

2. Write w_1^n on the basis $\{\varphi_1, \varphi_2, \varphi_3\}$, namely $w_1^n = \alpha_1\varphi_1 + \alpha_2\varphi_2 + \alpha_3\varphi_3$,
3. Let p be the number of positive eigenvalues, compute $w_0^n = \beta_1\varphi_1 + \beta_2\varphi_2 + \beta_3\varphi_3$ and $F^n_{\frac{1}{2}} = g_0(w_0^n, w_1^n)$, where the three unknowns β_1, β_2, and β_3 are determined by the p equations stating the boundary conditions (note that these equations involve the components of $F^n_{\frac{1}{2}}$) and by the $3-p$ equalities $\beta_i = \alpha_i$ for $\lambda_i < 0$.

This method leads, at each time step, to a non linear system of 3 equations with 3 unknowns (except if $\lambda_i = 0$ for some i), namely β_1, β_2 and β_3; note that some compatibility conditions are needed in order that this non linear system has a solution. Several variants of this method are possible. For instance, a boundary condition may be imposed on w_0^n rather than $F^n_{\frac{1}{2}}$. A similar method is, of course, possible at point $x = 1$ (changing the role of positive and negative eigenvalues).

This method is not always satisfactory. In the case of severe slugging for the simulation of two phase flow in a pipeline, the method seems to perform well at $x = 0$, where the eigenvalues λ_1 and λ_2 are always positive and the two boundary conditions (gas and liquid fluxes) are convenient. However, at $x = 1$, the second eigenvalue sometimes becomes negative and one needs a second boundary condition (the first one is a condition on the pressure). A natural condition seems to be $Q_l = 0$, where Q_l is the second component of the flux F, that is the liquid flux, but this condition does not lead to good results. Other possible choices of this additional boundary condition at $x = 1$ were tested and did not give good results. A possible interpretation of this problem is the fact that the sign of λ_2 is computed with w_N^n. Roughly speaking, it is "too late" when $\lambda_2(w_N^n)$ becomes negative (see Sect. 23 for the study of a simple scalar case). Indeed, good results (in agreement with experiments) are obtained with the unilateral condition $Q_l \geq 0$ (whatever the sign of $\lambda_2(w_N^n)$). It consists in using the preceeding method (for the boundary condition at $x = 1$) and in replacing, in the numerical scheme (38.4), the second component of $F^n_{N+\frac{1}{2}}$ by its positive part. Then, if $\lambda_2(w_N^n) < 0$, two boundary conditions are given at $x = 1$ (pressure and $Q_l = 0$) and if $\lambda_2(w_N^n) \geq 0$, one boundary condition is given at $x = 1$ (pressure) but, in (38.4), the second component of $F^n_{N+\frac{1}{2}}$ is replaced by its positive part.

We studied in Section 23 page 147 the sense of this boundary condition in the simplified scalar case.

38.2 Two phase flow in a porous medium

A second example is given by the modelization of a two phase flow, oil and water (for instance), in a porous medium. Phases are immiscible. Compressibility and capillarity effects are neglected. The model is obtained using the conservation of mass for each phase and Darcy's law. This study is limited to the one dimensional case. In this case the pressure can be eliminated and the problem is reduced to a single equation, namely (23.1) with :

$$f(u) = \frac{f_1(u)(\alpha + \beta f_2(u))}{f_1(u) + f_2(u)}. \tag{38.5}$$

The unknown is the saturation of one phase, say water, and is denoted by u. The quantity α is the total flux, which is constant in space, thanks to the incompressibility of the phases. One assumes also that it is constant in time and positive. The quantity β is the difference between the densities of the phases. The functions f_1 and f_2 are the mobilities of the phases. The function f_1 is nondecreasing, regular and satisfies $f_1(0) = 0$. The function f_2 is nonincreasing, regular and satisfies $f_2(1) = 0$. The function $f_1 + f_2$ is bounded from below by a positive number.

Remark 38.4 *For the equivalent two or three dimensional model, the pressure cannot be eliminated and the resulting model is a coupled system of two partial differential equations and two unknowns (pressure and saturation). The problem to which the limit of the approximate solutions is solution is then much more complicated to determine. See [49] for a partial study of this question.*

Here again, an initial condition is prescribed, namely (23.2), with $u_0 \in L^\infty((0,1))$, $0 \le u_0 \le 1$ a.e.. The boundary condition will be given later.

The numerical scheme is as in Sect. 23.1; it is given by (23.3) and (23.4) with (23.5). The choice of the numerical flux, g, satisfying (C1)-(C3), is usually given, for this model, using an "upwinding phase by phase", that is (see [15], for instance) :

$$\begin{aligned} g(a,b) &= \frac{f_1(a)(\alpha+\beta f_2(a))}{f_1(a)+f_2(a)} \text{ if } -\alpha+\beta f_1(a) \le 0 \\ g(a,b) &= \frac{f_1(a)(\alpha+\beta f_2(b))}{f_1(a)+f_2(b)} \text{ if } -\alpha+\beta f_1(a) > 0. \end{aligned} \tag{38.6}$$

Let us then define $f^n_{\frac{1}{2}}$ and $f^n_{N+\frac{1}{2}}$. On considers here the case of an injection of pure water at $x = 0$. Then :

$$f^n_{\frac{1}{2}} = \alpha,\ n \ge 0. \tag{38.7}$$

At $x = 1$, The boundary condition is quite complicated. A simple example is (see [56] for a more complete study):

$$f^n_{N+\frac{1}{2}} = \frac{f_1(u^n_N)\alpha}{f_1(u^n_N)+f_2(u^n_N)}. \tag{38.8}$$

Then, the approximate solution is given with (23.3)-(23.5), g given by (38.6), and (38.7)-(38.8).

In order to prove that the approximate solutions converge, as h and k go to zero, and to determine the problem which the limit of the approximate solutions is the unique solution to, one proceeds as in Sect. 23.3. One has to find g_0 and g_1 satisfying (C1)-(C3) and $\overline{u}, \overline{\overline{u}} \in L^\infty(\mathbf{R}_+)$ such that $f^n_{\frac{1}{2}}$ and $f^n_{N+\frac{1}{2}}$, respectively defined by (38.7) and (38.8), satisfy (23.6). This is again performed in [56]. The most interesting case is obtained for $\beta f_1(1) > \alpha$ and when the function f is increasing on $(0, u_M)$ and decreasing on $(u_M, 1)$, as in Sect. 23.3. In fact, the main point is the existence of a unique $u_m \in (0,1)$such that $f(u_m) = f(1) = \alpha$ and that f is increasing on $[0, u_m]$ and greater or equal to α on $[u_m, 1]$. Then, it is quite easy to prove that (38.7) gives

$$f^n_{\frac{1}{2}} = \alpha = g_G(u_m, u^n_1),$$

where g_G is the Godunov flux given in Sect. 23.3.

For the boundary condition at $x = 1$, it is possible to construct (see [56]) a function $g_1 : [0,1]^2 \to \mathbf{R}$ satisfying (C1)-(C3) such that (38.8) gives :

$$f^n_{N+\frac{1}{2}} = g_1(u^n_N, 1).$$

It is now possible to use Theorem 23.1.

Let L be a common Lipschitz constant for g (given by (38.6)), g_G and g_1 (on $[0,1]^2$) and let $\zeta > 0$. If $k \le (1-\zeta)\frac{h}{L}$, the approximate solution $u_{h,k}$, that is the solution defined by (23.3)-(23.5) (with g given by 38.6), and by the boundary fluxes (38.7)-(38.8), takes its values in $[0,1]$ and converges towards the unique solution of (38.9) in $L^p_{loc}([0,1]\times\mathbf{R}_+)$ for any $1 \le p < \infty$, as $h \to 0$:

$$\begin{aligned} &u \in L^\infty((0,1)\times(0,\infty)), \\ &\int_0^\infty \int_0^1 [(u-\kappa)^\pm \varphi_t + \operatorname{sign}_\pm(u-\kappa)(f(u)-f(\kappa))\varphi_x]dxdt \\ &\quad + M\int_0^\infty (u_m-\kappa)^\pm \varphi(0,t)dt + M\int_0^\infty (1-\kappa)^\pm \varphi(1,t)dt \\ &\qquad + \int_0^1 (u_0-\kappa)^\pm \varphi(x,0)dx \ge 0, \\ &\forall \kappa \in [0,1],\ \forall \varphi \in C^1_c([0,1]\times[0,\infty), \mathbf{R}_+), \end{aligned} \tag{38.9}$$

where M is a bound for $|f'|$ on $[0,1]$ (f is given by (38.5). As in Sect. 23.3. It is possible to give the sense of the boundary condition if u is regular enough. Indeed, let u be a regular solution of (38.9). Then, u satisfies the boundary conditions in the sense given by [9], that is :

$$\text{sign}(u(0,t)-u_m)(f(u(0,t))-f(\kappa))\leq 0,\ \forall \kappa\in[u_m,u(0,t)],\ \text{for a.e. } t\in\mathbf{R}_+,$$

$$\text{sign}(u(1,t)-1)(f(u(1,t))-f(\kappa))\geq 0,\ \forall \kappa\in[1,u(1,t)],\ \text{for a.e. } t\in\mathbf{R}_+,$$

with $[a,b]=\{ta+(1-t)b,\ t\in[0,1]\}$ and $\text{sign}(s)=1$ for $s>0$, $\text{sign}(s)=-1$ for $s<0$, $\text{sign}(0)=0$. This gives $u(0,t)=u_m$ or $u(0,t)=1$ and $u(1,t)\leq u_m$ or $u(1,t)=1$. In particular, at $x=0$, one has $f(u(0,t))=\alpha$ (only water is injected) and, at $x=1$, $f(u(1,t))<\alpha$ if $u(1,t)<u_m$ (which states that there is some oil production).

Bibliography

[1] Angermann, L. (1996), Finite volume schemes as non-conforming Petrov-Galerkin approximations of primal-dual mixed formulations, Report 181, Institüt für Angewandte Mathematik, Universtät Erlangen-Nurnberg.

[2] Amiez, G. and P.A. Gremaud (1991), On a numerical approach to Stefan-like problems, *Numer. Math.* **59**, 71-89.

[3] Angot, P. (1989), Contribution à l'etude des transferts thermiques dans les systèmes complexes, application aux composants électroniques, Thesis, Université de Bordeaux 1.

[4] Agouzal, A., J. Baranger, J.-F.Maitre and F. Oudin (1995), Connection between finite volume and mixed finite element methods for a diffusion problem with non constant coefficients, with application to Convection Diffusion, *East-West Journal on Numerical Mathematics.*, **3**, 4, 237-254.

[5] Arbogast, T., M.F. Wheeler and I. Yotov(1997), Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences, *SIAM J. Numer. Anal.* **34**, 2, 828-852.

[6] Atthey, D.R. (1974), A Finite Difference Scheme for Melting Problems, *J. Inst. Math. Appl.* **13**, 353-366.

[7] R.E. Bank and D.J. Rose (1986), Error estimates for the box method, *SIAM J. Numer. Anal.*, 777-790

[8] Baranger, J., J.-F. Maitre and F. Oudin (1996), Connection between finite volume and mixed finite element methods, *Modél. Math. Anal. Numér.*, 30, 3, 4, 444-465.

[9] Bardos, C., LeRoux, A.Y., Nédélec, J.C. (1979): First order quasilinear equations with boundary conditions. Comm. Partial Differential Equations, **9**, 1017–1034

[10] Barth, T.J. (1994), Aspects of unstructured grids and finite volume solvers for the Euler and Navier-Stokes equations, *Von Karmann Institute Lecture.*

[11] Belmouhoub R. (1996) Modélisation tridimensionnelle de la genèse des bassins sédimentaires, Thesis, Ecole Nationale Supérieure des Mines de Paris, 1996.

[12] Berger, A.E., H. Brezis and J.C.W. Rogers (1979), A Numerical Method for Solving the Problem $u_t - \Delta f(u) = 0$, *RAIRO Numerical Analysis*, **13**, 4, 297-312.

[13] Bertsch, M., R. Kersner and L.A. Peletier (1995), Positivity versus localization in degenerate diffusion equations, Nonlinear Analysis TMA, **9**, 9, 987-1008.

[14] Botta, N. and D. Hempel (1996), A finite volume projection method for the numerical solution of the incompressible Navier-Stokes equations on triangular grids, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 355-363.

[15] Brenier, Y., Jaffré, J. (1991): Upstream differencing for multiphase flow in reservoir simulation. SIAM J. Num. Ana. **28**, 685–696

[16] Brezis, H. (1983), *Analyse Fonctionnelle: Théorie et Applications* (Masson, Paris).

[17] Brun, G., J. M. Hérard, L. Leal De Sousa and M. Uhlmann (1996), Numerical modelling of turbulent compressible flows using second order models, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 338-346.

[18] Buffard, T., T. Gallouët and J. M. Hérard (1998), Un schéma simple pour les équations de Saint-Venant, *C. R. Acad. Sci., Série I*,**326**, 386-390.

[19] Buffard, T., Gallouët, T., Hérard, J.M. (2000): A sequel to a rough Godunov scheme : Application to real gases. Computers and Fluids, **29**, 813–847

[20] Cai, Z. (1991), On the finite volume element method, *Numer. Math.*, **58**, 713-735.

[21] Cai, Z., J. Mandel and S. Mc Cormick (1991), The finite volume element method for diffusion equations on general triangulations, *SIAM J. Numer. Anal.*, **28**, 2, 392-402.

[22] Chainais-Hillairet, C. (1996), First and second order schemes for a hyperbolic equation: convergence and error estimate, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 137-144.

[23] Chainais-Hillairet, C. (1999), Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate, *Modél. Math. Anal. Numér.*, 33, 1, 129-156.

[24] Champier, S. and T. Gallouët (1992), Convergence d'un schéma décentré amont pour une équation hyperbolique linéaire sur un maillage triangulaire, *Modél. Math. Anal. Numér.* **26**, 7, 835-853.

[25] Champier, S., T. Gallouët and R. Herbin (1993), Convergence of an Upstream finite volume Scheme on a Triangular Mesh for a Nonlinear Hyperbolic Equation, *Numer. Math.* 66, 139-157.

[26] Chavent, G. and J. Jaffré (1990), Mathematical Models and Finite Element for Reservoir Simulation, *Studies in Mathematics and its Applications* (North Holland, Amsterdam).

[27] Chevrier, P. and Galley, H. (1993), A Van Leer finite volume scheme for the Euler equations on unstructured meshes, *Modél. Math. Anal. Numér.*, **27**, 2, 183-201.

[28] Chou, S. and Q. Li Error estimates in L^2, H^1, and L^∞ in covolume methods for elliptic and parabolic problems: a unified approach, *Math. Comp.*, **69**, 2000, 103-120.

[29] Ciarlet, P.G. (1978), *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam).

[30] Ciarlet, P.G. (1991), *Basic error estimates for elliptic problems* in: *Handbook of Numerical Analysis* **II** (North-Holland,Amsterdam) 17-352.

[31] Ciavaldini, J.F. (1975), Analyse numérique d'un problème de Stefan à deux phases par une méthode d'éléments finis, *SIAM J. Numer. Anal.*, **12**, 464-488.

[32] Cockburn, B., F. Coquel and P. LeFloch (1994), An error estimate for finite volume methods for multidimensional conservation laws. *Math. Comput.* **63**, 207, 77-103.

[33] Cockburn, B., F. Coquel and P. LeFloch (1995), Convergence of the finite volume method for multidimensional conservation laws, *SIAM J. Numer. Anal.* **32**, 687-705. s

[34] Cockburn, B. and P. A. Gremaud (1996), A priori error estimates for numerical methods for scalar conservation laws. I. The general approach, *Math. Comput.* **65**,522-554.

[35] Cockburn, B. and P. A. Gremaud (1996), Error estimates for finite element methods for scalar conservation laws, *SIAM J. Numer. Anal.* **33**, 2, 522-554.

[36] Conway E. and J. Smoller (1966), Global solutions of the Cauchy problem for quasi-linear first-order equations in several space variables, *Comm. Pure Appl. Math.*, **19**, 95-105.

[37] Coquel, F. and P. LeFloch (1996), An entropy satisfying MUSCL scheme for systems of conservation laws, *Numer. Math.* **74**, 1-33.

[38] Cordes C. and M. Putti (1998), Finite element approximation of the diffusion operator on tetrahedra, SIAM J. Sci. Comput, **19**, 4, 1154-1168.

[39] Coudière, Y., T. Gallouët and R. Herbin (1998), Discrete Sobolev Inequlities and L^p error estimates for approximate finite volume solutions of convection diffusion equations, submitted.

[40] Coudière, Y., J.P. Vila, and P. Villedieu (1996), Convergence of a finite volume scheme for a diffusion problem, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 161-168.

[41] Y. Coudière, J.P. Vila and P. Villedieu (1999), Convergence rate of a finite volume scheme for a two-dimensional convection diffusion problem, Math. Model. Numer. Anal. 33 (1999), no. 3, 493–516.

[42] Courbet, B. and J. P. Croisille (1998), Finite-volume box-schemes on triangular meshes, M2AN, vol 32, 5, 631-649, 1998.

[43] Crandall, M.G. and A. Majda (1980), Monotone Difference Approximations for Scalar Conservation Laws, *Math. Comput.* **34**, 149, 1-21.

[44] Dahlquist, G. and A. Björck (1974), *Numerical Methods*, Prentice Hall Series in Automatic Computation.

[45] Deimling, K. (1980), *Nonlinear Functional Analysis*, (Springer, New York).

[46] DiPerna, R. (1985), Measure-valued solutions to conservation laws, *Arch. Rat. Mech. Anal.*, 88, 223-270.

[47] Dubois, F. and P. LeFloch (1988), Boundary conditions for nonlinear hyperbolic systems of conservation laws, *Journal of Differential Equations*, 71, 93-122.

[48] Eymard, R. (1992), Application à la simulation de réservoir des méthodes volumes-éléments finis; problèmes de mise en oeuvre, Cours CEA-EDF-INRIA.

[49] Eymard, R. and T. Gallouët (1993), Convergence d'un schéma de type eléments finis - volumes finis pour un système couplé elliptique - hyperbolique, *Modél. Math. Anal. Numér.* **27**, 7, 843-861.

[50] Eymard, R. and T. Gallouët (1991), Traitement des changements de phase dans la modélisation de gisements pétroliers, *Journées numériques de Besançon.*

[51] Eymard, R., Gallouët, T. (2003): H-convergence and numerical schemes for elliptic equations. SIAM Journal on Numerical Analysis, **41**, Number 2, 539–562

[52] Eymard, R., T. Gallouët, M. Ghilani and R. Herbin (1997), Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes, *IMA Journal of Numerical Analysis*, 18, 563-594.

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

[54] Eymard, R., T. Gallouët and R. Herbin (1995), Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation, *Chin. Ann. of Math.*, **16B** 1, 1-14.

[55] Eymard, R., T. Gallouët and R. Herbin (1999), Convergence of finite volume approximations to the solutions of semilinear convection diffusion reaction equations, Numer. Math., 82, 91-116.

[56] Eymard, R., Gallouët, T., Vovelle, J.: Boundary conditions in the numerical approximation of some physical problems via finite volume schemes. Accepted for publication in "Journal of CAM"

[57] Eymard R., M. Ghilani (1994), Convergence d'un schéma implicite de type éléments finis-volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique, *C.R. Acad. Sci. Paris Serie I,*, **319**, 1095-1100.

[58] Faille, I. (1992), Modélisation bidimensionnelle de la genèse et la migration des hydrocarbures dans un bassin sédimentaire, Thesis, Université de Grenoble.

[59] Faille, I. (1992), A control volume method to solve an elliptic equation on a two-dimensional irregular meshing, *Comp. Meth. Appl. Mech. Engrg.*, 100, 275-290.

[60] Faille, I. and Heintzé E. (1998), A rough finite volume scheme for modeling two-phase flow in a pipeline Computers and Fluids 28, 2, 213–241.

[61] Feistauer M., J. Felcman and M. Lukacova-Medvidova (1995), Combined finite element-finite volume solution of compressible flow, *J. Comp. Appl. Math.* **63**, 179-199.

[62] Feistauer M., J. Felcman and M. Lukacova-Medvidova (1997)], On the convergence of a combined finite volume-finite element method for nonlinear convection- diffusion problems, *Numer. Meth. for P.D.E.'s*, **13**, 163-190.

[63] Fernandez G. (1989), Simulation numérique d'écoulements réactifs à petit nombre de Mach, Thèse de Doctorat, Université de Nice.

[64] Fezoui L., S. Lanteri, B. Larrouturou and C. Olivier (1989), Résolution numérique des équations de Navier-Stokes pour un Fluide Compressible en Maillage Triangulaire, INRIA report 1033.

[65] Fiard, J.M. (1994), Modélisation mathématique et simulation numérique des piles au gaz naturel à oxyde solide, thèse de Doctorat, Université de Chambéry.

[66] Fiard, J.M., R. Herbin (1994), Comparison between finite volume finite element methods for the numerical simulation of an elliptic problem arising in electrochemical engineering, Comput. Meth. Appl. Mech. Engin., 115, 315-338.

[67] Forsyth, P.A. (1989), A control volume finite element method for local mesh refinement, SPE 18415, 85-96.

[68] Forsyth, P.A. (1991), A control volume finite element approach to NAPL groundwater contamination, SIAM J. Sci. Stat. Comput., 12, 5, 1029-1057.

[69] Forsyth, P.A. and P.H. Sammon (1988), Quadratic Convergence for Cell-Centered Grids, Appl. Num. Math. 4, 377-394.

[70] Gallouët, T. (1996), Rough schemes for systems for complex hyperbolic systems, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 10-21.

[71] Gallouët, T. and R. Herbin (1994), A Uniqueness Result for Measure Valued Solutions of a Nonlinear Hyperbolic Equations. Int. Diff. Int. Equations., 6,6, 1383-1394.

[72] Gallouët, T., R. Herbin and M.H. Vignal, (1999), Error estimate for the approximate finite volume solutions of convection diffusion equations with general boundary conditions, SIAM J. Numer. Anal., 37, 6, 1935-1972, 2000.

[73] Girault, V. and P. A. Raviart(1986), *Finite Element Approximation of the Navier-Stokes Equations* (Springer-Verlag).

[74] Ghidaglia J.M., A. Kumbaro and G. Le Coq (1996), Une méthode "volumes finis" à flux caractéristiques pour la résolution numérique de lois de conservation, *C.R. Acad. Sci. Paris Sér. I* **332**, 981-988.

[75] Godlewski E. and P. A. Raviart (1991), *Hyperbolic systems of conservation laws*, Ellipses.

[76] Godlewski E. and P. A. Raviart (1996), *Numerical approximation of hyperbolic systems of conservation laws*, *Applied Mathematical Sciences* **118** (Springer, New York).

[77] Godunov S. (1976), *Résolution numérique des problèmes multidimensionnels de la dynamique des gaz* (Editions de Moscou).

[78] Guedda, M., D. Hilhorst and M.A. Peletier (1997), Disappearing interfaces in nonlinear diffusion, *Adv. Math. Sciences and Applications*, 7, 695-710.

[79] Guo, W. and M. Stynes (1997), An analysis of a cell-vertex finite volume method for a parabolic convection-diffusion problem, *Math. Comput.* **66**, 217, 105-124.

[80] Harten A. (1983), On a class of high resolution total-variation-stable finite-difference schemes, *J. Comput. Phys.* **49**, 357-393.

[81] Harten A., P. D. Lax and B. Van Leer (1983), On upstream differencing and Godunov-type schemes for hyperbolic conservations laws, *SIAM review* **25**, 35-61.

[82] Harten A., J. M. Hyman and P. D. Lax (1976), On finite difference approximations and entropy conditions, *Comm. Pure Appl. Math.* **29**, 297-322.

[83] Heinrich B. (1986), Finite diference methods on irregular networks, I.S.N.M. 82(Birkhauser).

[84] Herbin R. (1995), An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh, *Num. Meth. P.D.E.* **11**, 165-173.

[85] Herbin R. (1996), Finite volume methods for diffusion convection equations on general meshes, *in Finite volumes for complex applications, Problems and Perspectives*, F. Benkhaldoun and R. Vilsmeier eds, Hermes, 153-160.

[86] Herbin R. and O. Labergerie (1997), Finite volume schemes for elliptic and elliptic-hyperbolic problems on triangular meshes, *Comp. Meth. Appl. Mech. Engin.* **147**,85-103.

[87] Herbin R. and E. Marchand (1997), *Numerical approximation of a nonlinear problem with a Signorini condition* , Third IMACS International Symposium On Iterative Methods In Scientific Computation, Wyoming, USA, July 97.

[88] Kagan, A.M., Linnik, Y.V., Rao, C.R. (1973): Characterization Problems in Mathematical Statistics. Wiley, New York

[89] Kamenomostskaja, S.L. (1995), On the Stefan problem, *Mat. Sb.* 53, 489-514 (1961 in Russian).

[90] Keller H. B. (1971), A new difference sheme for parabolic problems, *Numerical solutions of partial differential equations, II*, B. Hubbard ed., (Academic Press, New-York) 327-350.

[91] Kröner D. (1997) *Numerical schemes for conservation laws in two dimensions*, Wiley-Teubner Series Advances in Numerical Mathematics. (John Wiley and Sons, Ltd., Chichester; B. G. Teubner, Stuttgart).

[92] Kröner D. and M. Rokyta (1994), Convergence of upwind finite volume schemes on unstructured grids for scalar conservation laws in two dimensions, *SIAM J. Numer. Anal.* **31**, 2, 324-343.

[93] Kröner D., S. Noelle and M. Rokyta (1995), Convergence of higher order upwind finite volume schemes on unstructured grids for scalar conservation laws in several space dimensions,*Numer. Math.* **71**, 527-560,.

[94] Krushkov S.N. (1970), First Order quasilinear equations with several space variables, *Math. USSR. Sb.* **10**, 217-243.

[95] Kumbaro A. (1992), Modélisation, analyse mathématique et numérique des modèles bi-fluides d'écoulement diphasique, Thesis, Université Paris XI Orsay.

[96] Kuznetsov (1976), Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation, *USSR Comput. Math. and Math. Phys.* **16**, 105-119.

[97] Ladyženskaja, O.A., V.A. Solonnikov and Ural'ceva, N.N. (1968), Linear and Quasilinear Equations of Parabolic Type, *Transl. of Math. Monographs* **23**.

[98] Lazarov R.D. and I.D. Mishev (1996), Finite volume methods for reaction diffusion problems in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 233-240.

[99] Lazarov R.D., I.D. Mishev and P.S Vassilevski (1996), Finite volume methods for convection-diffusion problems, SIAM J. Numer. Anal., 33, 1996, 31-55.

[100] LeVeque, R. J. (1990), *Numerical methods for conservation laws* (Birkhauser verlag).

[101] Li, R. Z. Chen and W. Wu(2000), *Generalized difference methods for partial differential equations* Marcel Dekker, New York.

[102] Lions, P. L. (1996), *Mathematical Topics in Fluid Mechanics; v.1: Incompressible Models*, Oxford Lecture Series in Mathematics & Its Applications, No.3, Oxford Univ. Press.

[103] Mackenzie, J.A., and K.W. Morton (1992), Finite volume solutions of convection-diffusion test problems, *Math. Comput.* **60**, 201, 189-220.

[104] Manteuffel, T., and A.B.White (1986), The numerical solution of second order boundary value problem on non uniform meshes,*Math. Comput.* **47**, 511-536.

[105] Masella, J. M. (1997), Quelques méthodes numériques pour les écoulements diphasiques bi-fluide en conduites petrolières, Thesis, Université Paris VI.

[106] Masella, J. M., I. Faille, and T. Gallouët (1996), On a rough Godunov scheme, accepted for publication in Intl. J. Computational Fluid Dynamics.

[107] Meirmanov, A.M. (1992), *The Stefan Problem* (Walter de Gruyter Ed, New York).

[108] Meyer, G.H. (1973), Multidimensional Stefan Problems, *SIAM J. Num. Anal.* **10**, 522-538.

[109] Mishev, I.D. (1998), Finite volume methods on Voronoï meshes *Num. Meth. P.D.E.*, 14, 2, 193-212.

[110] Morton, K.W. (1996), *Numerical Solutions of Convection-Diffusion problems* (Chapman and Hall, London).

[111] Morton, K.W. and E. Süli (1991), Finite volume methods and their analysis, IMA J. Numer. Anal. **11**, 241-260.

[112] Morton, K.W., Stynes M. and E. Süli (1997), Analysis of a cell-vertex finite volume method for a convection-diffusion problems, *Math. Comput.* **66**, 220, 1389-1406.

[113] Nečas, J. (1967), *Les méthodes directes en théorie des é quations elliptiques* (Masson, Paris).

[114] Nicolaides, R.A. (1992) Analysis and convergence of the MAC scheme I. The linear problem, *SIAM J. Numer. Anal.***29**, 6, 1579-1591.

[115] Nicolaides, R.A. (1993) The Covolume Approach to Computing Incompressible Flows, in: *Incompressible computational fluid dynamics* M.D Gunzburger, R. A. Nicolaides eds, 295–333.

[116] Nicolaides, R.A. and X. Wu (1996) Analysis and convergence of the MAC scheme II. Navier-Stokes equations ,*Math. Comput.***65**, 213, 29-44.

[117] Noëlle, S. (1996) A note on entropy inequalities and error estimates for higher-order accurate finite volume schemes on irregular grids, *Math. Comput.* **65**, 1155-1163.

[118] Oden, J.T. (1991), *Finite elements: An Introduction* in: *Handbook of Numerical Analysis* **II** (North-Holland,Amsterdam) 3-15.

[119] Oleinik, O.A. (1960), A method of solution of the general Stefan Problem, *Sov. Math. Dokl.* **1**, 1350-1354.

[120] Oleinik, O. A. (1963), On discontinuous solutions of nonlinear differential equations,*Am. Math. Soc. Transl., Ser. 2* **26**, 95-172.

[121] Osher S. (1984), Riemann Solvers, the entropy Condition, and difference approximations, *SIAM J. Numer. Anal.* **21**, 217-235.

[122] Otto, F. (1996): Initial-boundary value problem for a scalar conservation law. C. R. Acad. Sci. Paris Sér. I Math. **8**, 729–734

[123] Patankar, S.V. (1980), *Numerical Heat Transfer and Fluid Flow*, Series in Computational Methods in Mechanics and Thermal Sciences, Minkowycz and Sparrow Eds. (Mc Graw Hill).

[124] Patault, S., Q.-H. Tran, Q.H. (1996): Modèle et schéma numérique du code TACITE-NPW, tech. report, rapport IFP 42415

[125] Pfertzel, A. (1987), Sur quelques schémas numériques pour la résolution des écoulements diphasiques en milieux poreux, Thesis, Université de Paris 6.

[126] Roberts J.E. and J.M. Thomas (1991), Mixed and hybrids methods, in: *Handbook of Numerical Analysis* **II** (North-Holland,Amsterdam) 523-640.

[127] Roe P. L. (1980), The use of Riemann problem in finite difference schemes, *Lectures notes of physics* **141**, 354-359.

[128] Roe P. L. (1981), Approximate Riemann solvers, parameter vectors, and difference schemes, *J. Comp. Phys.* **43**, 357-372.

[129] Rudin W. (1987), *Real and Complex analysis* (McGraw Hill).

[130] Samarski A.A. (1965), On monotone difference schemes for elliptic and parabolic equations in the case of a non-selfadjoint elliptic operator, *Zh. Vychisl. Mat. i. Mat. Fiz.* **5**, 548-551 (Russian).

[131] Samarskii A.A. (1971), *Introduction to the Theory of Difference Schemes*, Nauka, Moscow (Russian).

[132] Samarskii A.A., R.D. Lazarov and V.L. Makarov (1987), *Difference Schemes for differential equations having generalized solutions*, Vysshaya Shkola Publishers, Moscow (Russian).

[133] Sanders R. (1983), On the Convergence of Monotone Finite Difference Schemes with Variable Spatial Differencing, *Math. Comput.* **40**, 161, 91-106.

[134] Selmin V. (1993), The node-centered finite volume approach: Bridge between finite differences and finite elements, *Comp. Meth. in Appl. Mech. Engin.* **102**, 107- 138.

[135] Serre D. (1996), *Systèmes de lois de conservation* (Diderot).

[136] Shashkov M. (1996), *Conservative Finite-difference Methods on general grids*, CRC Press, New York.

[137] Sod G. A. (1978), A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws, *J. Comput. Phys.*, **27**, 1-31.

[138] Sonier F. and R. Eymard (1993), Mathematical and numerical properties of control-volume finite-element scheme for reservoir simulation, *paper SPE 25267, 12th symposium on reservoir simulation, New Orleans.*

[139] Süli E. (1992), The accuracy of cell vertex finite volume methods on quadrilateral meshes, *Math. Comp.* **59**, 200, 359-382.

[140] Szepessy A. (1989), An existence result for scalar conservation laws using measure valued solutions, *Comm. P.D.E.* **14**, 10, 1329-1350.

[141] Temam R. (1977), *Navier-Stokes Equations* (North Holland, Amsterdam).

[142] Tichonov A.N. and A.A. Samarskii (1962), Homogeneous difference schemes on nonuniform nets *Zh. Vychisl. Mat. i. Mat. Fiz.* **2**, 812-832 (Russian).

[143] Thomas J.M. (1977), Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes, Thesis, Université Pierre et Marie Curie.

[144] Thomée V. (1991), Finite differences for linear parabolic equations, in: *Handbook of Numerical Analysis* **I** (North-Holland,Amsterdam) 5-196.

[145] Turkel E. (1987), Preconditioned methods for solving the imcompressible and low speed compressible equations, *J. Comput. Phys.* **72**, 277-298.

[146] Van Leer B. (1974), Towards the ultimate conservative difference scheme, II. Monotonicity and conservation combined in a second-order scheme *J. Comput. Phys.* **14**, 361-370.

[147] Van Leer B. (1977), Towards the ultimate conservative difference scheme, IV. a new approach to numerical convection. *J. Comput. Phys.* **23**, 276-299.

[148] Van Leer B. (1979), Towards the ultimate conservative difference scheme, V, *J. Comput. Phys.* **32**, 101-136.

[149] Vanselow R. (1996), Relations between FEM and FVM, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 217-223.

[150] Vassileski, P.S., S.I. Petrova and R.D. Lazarov (1992), Finite difference schemes on triangular cell-centered grids with local refinement, *SIAM J. Sci. Stat. Comput.* **13**, 6, 1287-1313.

[151] Vignal M.H. (1996), Convergence of a finite volume scheme for a system of an elliptic equation and a hyperbolic equation *Modél. Math. Anal. Numér.* **30**, 7, 841-872.

[152] Vignal M.H. (1996), Convergence of Finite Volumes Schemes for an elliptic hyperbolic system with boundary conditions,in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 145-152.

[153] Vignal M.H. and S. Verdière (1998), Numerical and theoretical study of a dual mesh method using finite volume schemes for two phase flow problems in porous media, *Numer. Math.*, 80, 4, 601-639.

[154] Vila J.P. (1986), Sur la théorie et l'approximation numérique de problèmes hyperboliques non linéaires. Application aux équations de saint Venant et à la modélisation des avalanches de neige dense, Thesis, Université Paris VI.

[155] Vila J.P. (1994), Convergence and error estimate in finite volume schemes for general multidimensional conservation laws, I. explicit monotone schemes, *Modél. Math. Anal. Numér.* **28**, 3, 267-285.

[156] Vol'pert A. I. (1967), The spaces BV and quasilinear equations, *Math. USSR Sb. 2*, 225-267.

[157] Vovelle, J. (2002): Convergence of finite volume monotone schemes for scalar conservation laws on bounded domains. Num. Math., **3**, 563–596

[158] Weiser A. and M.F. Wheeler (1988), On convergence of block-centered finite-differences for elliptic problems, *SIAM J. Numer. Anal.*, **25**, 351-375.

Index

Boundary conditions for hyperbolic equations, 147, 198, 236
BV
– Space, 124, 141, 174, 190–191
Initial condition in –, 170, 174, 189, 192
Strong – estimate, 131, 141, 142
Strong time – estimate, 168–172
Weak – estimate, 129–131, 137–139, 161–164, 167–168, 229

CFL condition, 128, 136
Compactness
– of the approximate solutions in L^1, 145
– of the approximate solutions in L^1_{loc}, 130
– of the approximate solutions in L^2, 46, 74, 75
– results in L^2, 93–96
Discrete – results, 28–30
Helly's – theorem, 141
Kolmogorov's – theorem, 94
Nonlinear weak – in L^∞, 201
Weak – in L^∞, 201
Conservativity, 4, 8, 35, 83, 211
Consistency
– error, 53, 56, 59, 70, 127
– in the finite difference sense, 10, 15, 127, 128
– of the fluxes, 6, 13, 23, 25, 35, 36, 212
– error, 13
Weak –, 21, 37
Control volume, 4, 5, 12, 33, 37, 39
Control volume finite element method, 9, 85, 223
Convection term, 41, 63, 97, 99, 213, 218, 219
Convergence
– for the weak star topology, 21, 116, 131
Nonlinear weak star –, 154, 200
Convergence of the approximate solutions
– towards the entropy process solution, 182
– towards the entropy weak solution, 187
Convergence of the finite volume method
– for a linear hyperbolic equation, 132
– for a nonlinear hyperbolic equation, 139
– for a nonlinear parabolic equation, 113
– for a semilinear elliptic equation, 30–31
– for an elliptic equation, 45, 51, 74

Delaunay condition, 39, 53, 85, 88, 90
Dirichlet boundary condition, 12–14, 16–21, 32–37
Discontinuous coefficients, 8, 78
Discrete
– $L^2(0,T;H^1(\Omega))$ seminorm, 107
– H^1_0 norm, 39
– Poincaré inequality, 42, 62, 65–69, 74, 82
– Sobolev inequality, 60–62
– entropy inequalities, 136, 172–173
– maximum principle, 43, 106
Donald dual cell, 100

Edge, 5, 38
Entropy
– function, 160
– process solution, 140, 155, 156, 181–187
– weak solution, 123–125, 154, 181
Discrete – inequalities, 172–173
Continuous – inequalities, 123
Discrete – inequalities, 136
Equation
Conservation –, 6
Convection-diffusion –, 18, 32, 63, 78, 97, 101, 206
Diffusion –, 5, 33, 104
Hyperbolic –, 122, 153, 198
Transport –, 4, 125
Error estimate
– for a hyperbolic equation
– in the general case, 188, 189
– in the one dimensional case, 127
– for a parabolic equation, 101
– for an elliptic equation
– in the general case, 52
– for an elliptic equation
– in the general case, 34, 55, 62, 69, 71, 81, 93
– in the one dimensional case, 16, 21, 23
Estimate on the approximate solution
– for a hyperbolic equation, 128, 131, 136, 142, 160, 164, 168, 229
– for a parabolic equation, 101, 106–113
– for an elliptic equation, 28, 42, 74

Euler equations, 7, 11, 210–212, 214

Finite difference method, 6, 8, 9, 14, 15, 19, 21, 100
Finite element method, 9, - Mixed15, 16, 37, 53, - primal mesh87, 84–90, 100, 203–204, 207, 219, 223
Finite volume scheme
– for multiphase flow problems, 228
Finite volume finite element method, 89–91
Finite volume principles, 4, 6
Finite volume scheme
– for elliptic problems
– in the one dimensional case, 27
– in one space dimension, 13–14, 23, 26
– in two or three space dimensions, 34–37, 41–42, 64, 79–80, 82–84
– for hyperbolic problems
– in one space dimension, 128, 133
– in two or three space dimensions, 5, 157, 158
– for hyperbolic systems, 210–217
– for multiphase flow problems, 222–223, 226
– for parabolic problems, 99–100, 105–106
– for the Stokes system, 219

Galerkin expansion, 9, 16, 89, 100, 219

Helly's theorem, 141

Kolmogorov's theorem, 94
Krushkov's entropies, 124

Lax-Friedrichs scheme, 135, 189
Lax-Wendroff theorem, 143

Mass lumping, 100, 204
Mesh
– Refinement, 93
Admissible –
–for Dirichlet boundary conditions, 37
–for Neumann boundary conditions, 63
–for a general diffusion operator, 79
–for hyperbolic equations, 128, 156
–for regular domains, 116
–in the one-dimensional elliptic case, 12
Restricted – for Dirichlet boundary conditions, 55
Restricted – for Neumann boundary conditions, 71
Dual –, 84
Moving –, 204–206
Rectangular –, 33
Structured –, 33–35
Triangular –, 39
Voronoï –, 39, 86

Navier-Stokes equations, 11, 207, 217, 218
Neumann boundary conditions, 63–78
Newton's algorithm, 209, 214–216, 224

Poincaré
Discrete – inequality, 62, 65–69, 74, 82
Poincaré inequality
– Neumann boundary conditions, 65
Poincaré inequality
– Dirichlet boundary conditions, 40, 69
Poincaré inequality
- Neumann boundary conditions, 69

Roe scheme, 210

Scheme
Explicit Euler –, 5, 7, 146, 204
Higher order –, 146–147
Implicit – for hyperbolic equations, 155, 164–172
Implicit – for parabolic equations, 99, 106, 110, 113
Implicit Euler –, 8, 100
Lax-Friedrichs, 135
Lax-Friedrichs –, 189
Monotone flux –, 134–135, 142
MUSCL –, 146
Roe –, 210
Van Leer –, 146
VFRoe –, 211–213, 215
Singular source terms in elliptic equations, 91–93
Sobolev
Discrete – inequality, 60, 62
Stability, 15, 20, 22, 28, 43, 74, 98, 106, 125, 136, 142, 157, 160, 164
Stabilization of a finite element method, 203
Staggered grid, 11, 207, 217–219
Stokes equations, 219

Transmissibility, 38, 39, 85, 86, 88–90
Two phase flow, 222

Uniqueness
– of the entropy process solution to a hyperbolic equation, 183
– of the solution to a nonlinear diffusion equation, 118
Upstream, 5, 22, 41, 80, 99, 126, 135, 137, 141, 217, 223, 228

Van Leer scheme, 146
Voronoï, 39, 86