Handbook of Numerical Analysis

Article · January 2000

3 authors:

Robert Eymard
University Gustave Eiffel
304 PUBLICATIONS **9,793** CITATIONS
SEE PROFILE

Raphaèle Herbin
Aix-Marseille Université
283 PUBLICATIONS **9,343** CITATIONS
SEE PROFILE

Gallouet Thierry
288 PUBLICATIONS **14,353** CITATIONS
SEE PROFILE

Finite Volume Methods

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

January 2003. 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

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.

1.1 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

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

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 $F_{K,\sigma} = -F_{L,\sigma}$.

1.2 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

$$q_t(x,t) + \mathrm{div}\mathbf{F}(x,t) = f(x,t), \tag{1.12}$$

at each point x and each time t where the conservation of q is to be written. In equation (1.12), $(\cdot)_t$ 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 $(x,t)\ldots$. The transport equation of Example 1.1 is a particular case of (1.12) 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 1.3 (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 (1.12), 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 (1.12) may be seen as the expression of the conservation of q in an infinitesimal domain; it is formally equivalent to the equation

$$\begin{aligned} \int_K q(x,t_2)dx - \int_K q(x,t_1)dx + \int_{t_1}^{t_2} \int_{\partial K} \mathbf{F}(x,t) \cdot \boldsymbol{n}_K(x) d\gamma(x) dt \\ = \int_{t_1}^{t_2} \int_K f(x,t) dx dt, \end{aligned} \tag{1.13}$$

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 (1.13) 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.

1.2.1 Time discretization

The time discretization of Equation (1.12) 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 (1.12) could be written with the use of a space-time divergence. Hence, Equation (1.12) 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 (1.12). 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.

1.2.2 Space discretization

In order to perform a space finite volume discretization of equation (1.12), 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 (1.12) over each cell K of the mesh $\mathcal{T}$. One obtains a conservation law under a nonlocal form (related to equation (1.13)) written for the volume K. Using the Euler time discretization, this yields

$$\int_K \frac{q^{(n+1)}(x) - q^{(n)}(x)}{k} dx + \int_{\partial K} \mathbf{F}(x,t_n)\cdot \boldsymbol{n}_K(x) d\gamma(x) = \int_K f(x,t_n)dx, \tag{1.14}$$

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.

1.3 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 2.3 and 3.3, 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.

1.4 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.

2.1 A finite volume method for the Dirichlet problem

2.1.1 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:

$$\begin{array}{lcl} -u_{xx}(x) & = & f(x), \quad x \in (0,1), \\ u(0) & = & 0, \\ u(1) & = & 0. \end{array} \tag{2.1}$$

If $f \in C([0,1],\mathbb{R})$, there exists a unique solution $u \in C^2([0,1],\mathbb{R})$ to Problem (2.1). 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 2.1 (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 (2.1) is integrated over each cell K_i, as in (1.14) 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

$$|R_{i+\frac{1}{2}}| = |F^\star_{i+\frac{1}{2}} + u_x(x_{i+\frac{1}{2}})| \leq Ch, \text{ where } F^\star_{i+\frac{1}{2}} = -\frac{u(x_{i+1}) - u(x_i)}{h_{i+\frac{1}{2}}}. \tag{2.2}$$

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

Remark 2.1 (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 (2.1). 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 (2.1) reads

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i, \; i = 1, \ldots, N \tag{2.3}$$

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}, \; i = 1, \ldots, N-1, \tag{2.4}$$

$$F_{\frac{1}{2}} = -\frac{u_1}{h_{\frac{1}{2}}}, \tag{2.5}$$

$$F_{N+\frac{1}{2}} = \frac{u_N}{h_{N+\frac{1}{2}}}. \tag{2.6}$$

Note that (2.4), (2.5), (2.6) may also be written

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}, \; i = 0, \ldots, N, \tag{2.7}$$

setting

$$u_0 = u_{N+1} = 0. \tag{2.8}$$

The numerical scheme (2.3)-(2.6) may be written under the following matrix form:

$$AU = b, \tag{2.9}$$

where $U = (u_1, \ldots, u_N)^t$, $b = (b_1, \ldots, b_N)^t$, with (2.8) and with A and b defined by

$$(AU)_i = \frac{1}{h_i}\Big(-\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}}\Big),\ i = 1, \ldots, N, \tag{2.10}$$

$$b_i = \frac{1}{h_i}\int_{K_i} f(x)dx,\ i = 1, \ldots, N, \tag{2.11}$$

Remark 2.2 There are other finite volume schemes for problem (2.1).

1. For instance, it is possible, in Definition 2.1, to take $x_1 \geq 0$, $x_N \leq 1$ and, for the definition of the scheme (that is (2.3)-(2.6)), to write (2.3) only for $i = 2, \ldots, N-1$ and to replace (2.5) and (2.6) by $u_1 = u_N = 0$ (note that (2.4) does not change). For this so-called "modified finite volume" scheme, it is also possible to obtain an error estimate as for the scheme (2.3)-(2.6) (see Remark 2.5). 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.

2.1.2 Comparison with a finite difference scheme

With the same notations as in Section 2.1.1, 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 (2.9)-(2.11), 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 2.1 Let $f \equiv 1$ and consider a mesh of $(0, 1)$, in the sense of Definition 2.1, 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 (2.3)-(2.6) 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 (2.1), the consistent finite difference scheme would be, omitting boundary conditions,

$$\frac{4}{2h_i + h_{i-1} + h_{i+1}}\Big[-\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}}\Big] = f(x_i),\ i = 2, \ldots, N-1, \tag{2.12}$$

Remark 2.3 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 2.2.2.

In the case of Problem (2.1), both the finite volume and finite difference schemes are convergent. The finite difference scheme (2.12) 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 (2.3)-(2.6) needs some more work and is described in Section 2.2.1.

2.1.3 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 (2.1), the two methods yield two different schemes. For Problem (2.1), 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 (2.1) 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:

$$q \in H^1((0,1)),\ u \in L^2((0,1)), \tag{2.13}$$

$$\int_0^1 q(x)p(x)dx = \int_0^1 u(x)p_x(x)dx,\ \forall\, p \in H^1((0,1)), \tag{2.14}$$

$$\int_0^1 q_x(x)v(x)dx = \int_0^1 f(x)v(x)dx,\ \forall\, v \in L^2((0,1)). \tag{2.15}$$

Considering an admissible mesh of $(0,1)$ (see Definition 2.1), 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 (2.14) and $v = \psi_k, k = 1, \ldots, N$ in (2.15). 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 (2.3)-(2.6). 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 (2.14):

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

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].

2.2 Convergence and error analysis for the Dirichlet problem

2.2.1 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 2.1
Let $f \in C([0,1], \mathbb{R})$ and let $u \in C^2([0,1], \mathbb{R})$ be the (unique) solution of Problem (2.1). Let $\mathcal{T} = (K_i)_{i=1,\ldots,N}$ be an admissible mesh in the sense of Definition 2.1. Then, there exists a unique vector $U = (u_1, \ldots, u_N)^t \in \mathbb{R}^N$ solution to (2.3) -(2.6) and there exists $C \geq 0$, only depending on u, such that

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}} \leq C^2 h^2, \tag{2.16}$$

and

$$|e_i| \leq Ch,\ \forall i \in \{1, \ldots, N\}, \tag{2.17}$$

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 (2.3)-(2.6). Indeed, multiplying (2.3) 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 (2.3) 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 (2.3) (with (2.4)-(2.6)).

One now proves (2.16). 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 (2.3), 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 (2.2), there exists $C > 0$, only depending on u, such that

$$F^{\star}_{i+\frac{1}{2}} = \overline{F}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}} \text{ and } |R_{i+\frac{1}{2}}|. \leq Ch, \tag{2.18}$$

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

$$-\frac{e_{i+1} - e_i}{h_{i+\frac{1}{2}}} - R_{i+\frac{1}{2}} + \frac{e_i - e_{i-1}}{h_{i-\frac{1}{2}}} + R_{i-\frac{1}{2}} = 0,\ \forall i \in \{1, \ldots, N\}. \tag{2.19}$$

Multiplying (2.19) 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 (2.18))

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}} \leq Ch \sum_{i=0}^{N} |e_{i+1} - e_i|. \tag{2.20}$$

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

$$\sum_{i=0}^{N} |e_{i+1} - e_i| \leq \Big(\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+\frac{1}{2}}}\Big)^{\frac{1}{2}} \Big(\sum_{i=0}^{N} h_{i+\frac{1}{2}}\Big)^{\frac{1}{2}}. \tag{2.21}$$

Since $\displaystyle\sum_{i=0}^{N} h_{i+\frac{1}{2}} = 1$ in (2.21) and from (2.20), one deduces (2.16).

Since, for all $i \in \{1, \ldots, N\}$, $e_i = \displaystyle\sum_{j=1}^{i} (e_j - e_{j-1})$, one can deduce, from (2.21) and (2.16) that (2.17) holds. ■

Remark 2.4 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 2.5

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 2.2 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 2.1 gives (2.16) 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 2.6

1. If the assumption $f \in C([0,1],\mathbb{R})$ is replaced by the assumption $f \in L^2((0,1))$ in Theorem 2.1, then $u \in H^2((0,1))$ instead of $C^2([0,1],\mathbb{R})$, but the estimates of Theorem 2.1 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 3.4 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 2.1, 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 (2.18) 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 2.7 Estimate (2.16) can be interpreted as a "discrete H_0^1" estimate on the error. A theoretical result which underlies the L^∞ estimate (2.17) 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 3.1 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 (2.17), under slightly different conditions, which uses finite difference techniques.

2.2.2 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 (2.1). Using the notations of Section 2.1.2 (recall that $\overline{U} = (u(x_1), \dots, u(x_N))^t$, 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

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

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 2.1. In order to prove the convergence of the scheme, a stability property is established in Lemma 2.2.

Lemma 2.1 *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $(0,1)$, in the sense of Definition 2.1 page 12, such that x_i is the center of K_i for all $i = 1, \dots, N$. Let $\alpha_\mathcal{T} > 0$ be such that $h_i > \alpha_\mathcal{T} h$ for all $i = 1, \dots, N$ (recall that $h = \max\{h_1, \dots, h_N\}$). Let $\overline{U} = (u(x_1), \dots, u(x_N))^t \in \mathbb{R}^N$, where u is the solution to (2.1), and assume $u \in C^3([0,1], \mathbb{R})$. Let A be the matrix defining the numerical scheme, given in (2.10) page 14. Then there exists a unique $U = (u_1, \dots, u_N)$ solution of (2.3)-(2.6) and there exists $\overline{\overline{r}}$ and $V \in \mathbb{R}^N$ such that*

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

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

Proof of Lemma 2.1
The existence and uniqueness of U is classical (it is also proved in Theorem 2.1).
For $i = 0, \dots N$, 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

$$r_i = \frac{1}{h_i}(R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}}), \text{ for } i = 0, \dots, N, \tag{2.22}$$

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

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

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

$$-\frac{v_{i+1}-v_i}{h_{i+\frac{1}{2}}} = R_{i+\frac{1}{2}} + 0(h^2),\ i = 1,\ldots,N-1,$$

$$-\frac{2v_1}{h_1} = R_{\frac{1}{2}} + 0(h^2),$$

$$\frac{2v_N}{h_N} = R_{N+\frac{1}{2}} + 0(h^2).$$

Since $h_i \geq \alpha_{\mathcal{T}} h$, for $i = 1,\ldots,N$, replacing $R_{i+\frac{1}{2}}$ in (2.22) gives that $r_i = (AV)_i + 0(h)$, for $i = 1,\ldots,N$, and $\|V\|_\infty = 0(h^2)$. Hence the lemma is proved. ■

Lemma 2.2 (Stability) *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $[0,1]$ in the sense of Definition 2.1. Let A be the matrix defining the finite volume scheme given in (2.10). Then A is invertible and*

$$\|A\|_\infty^{-1} \leq \frac{1}{4}. \tag{2.23}$$

Proof of Lemma 2.2

First we prove a discrete maximum principle; indeed if $b_i \geq 0$, for all $i = 1,\ldots,N$, and if U is solution of $AU = b$ then we prove that $u_i \geq 0$ for all $i = 1,\ldots,N$.
Let $a = \min\{u_i, i = 0,\ldots,N+1\}$ (recall that $u_0 = u_{N+1} = 0$) and $i_0 = \min\{i \in \{0,\ldots,N+1\};\ u_i = a\}$.
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 $i = 1,\ldots,N$.
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 $\Phi = (\phi_1,\ldots,\phi_N)$ 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 (2.3)-(2.6), 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 $W = (w_1,\ldots,w_N) \in \mathbb{R}^N$ such that $W_i = -\frac{h_i^2}{8}$ (see proof of Lemma 2.1). 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 2.1, yields the convergence of the finite volume scheme, formulated in the next theorem.

Theorem 2.2 *Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh of $[0,1]$ in the sense of Definition 2.1 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 (2.1)). Let $U = (u_1,\ldots,u_N)$ be the solution given by the numerical scheme (2.3)-(2.6). Then there exists $C > 0$, only depending on $\alpha_{\mathcal{T}}$ and u, such that $\|U - \overline{U}\|_\infty \leq Ch$.*

Remark 2.8 In the proof of Lemma 2.2, 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 2.9 With the notations of Lemma 2.1, 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.

2.3 General 1D elliptic equations

2.3.1 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 3.1.2 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:

$$-(\lambda u_x)_x(x) + au_x(x) + bu(x) = f(x), \; x \in [0,1], \tag{2.24}$$

$$u(0) = c, \; u(1) = d. \tag{2.25}$$

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 2.10 Problem (2.24)-(2.25) 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 $g = f - au_x - bu \in L^1((0,1))$, 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 2.1 page 12, such that the discontinuities of λ coincide with the interfaces of the mesh.
The notations being the same as in section 2.1, integrating Equation (2.24) 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 $i = 1, \ldots, N-1$):

$$u_{i+\frac{1}{2}} = \frac{u_{i+1}\dfrac{\lambda_{i+1}}{h_{i+1}^-} + u_i\dfrac{\lambda_i}{h_i^+}}{\dfrac{\lambda_{i+1}}{h_{i+1}^-} + \dfrac{\lambda_i}{h_i^+}} \tag{2.26}$$

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

$$\begin{aligned} H_{i+\frac{1}{2}} &= -\tau_{i+\frac{1}{2}}(u_{i+1} - u_i),\ i = 1, \ldots, N-1, \\ H_{\frac{1}{2}} &= -\frac{\lambda_1}{h_1^-}(u_1 - c), \\ H_{N+\frac{1}{2}} &= -\frac{\lambda_N}{h_N^+}(d - u_N) \end{aligned} \tag{2.27}$$

with

$$\tau_{i+\frac{1}{2}} = \frac{\lambda_i\lambda_{i+1}}{h_i^+\lambda_{i+1} + h_{i+1}^-\lambda_i},\ i = 1, \ldots, N-1. \tag{2.28}$$

Example 2.2 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 (2.24)-(2.25) is therefore,

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + bh_i u_i = h_i f_i,\ \forall i \in \{1,\dots,N\}, \tag{2.29}$$

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

$$F_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u_{i+1} - u_i) + au_i,\ \forall i \in \{1,\dots,N-1\}, \tag{2.30}$$

$$F_{\frac{1}{2}} = -\frac{\lambda_1}{h_1^-}(u_1 - c) + ac,\ F_{N+\frac{1}{2}} = -\frac{\lambda_N}{h_N^+}(d - u_N) + au_N. \tag{2.31}$$

Remark 2.11 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 (2.28), (2.30) and (2.31) into account, the numerical scheme (2.29) yields a system of N equations with N unknowns $u_1,\dots,u_N$.

2.3.2 Error estimate

Theorem 2.3
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 (2.24)-(2.25). Let $\mathcal{T} = (K_i)_{i=1,\cdots,N}$ be an admissible mesh, in the sense of Definition 2.1, 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 $U = (u_1,\dots,u_N)^t \in \mathbb{R}^N$ solution to (2.28)-(2.31),*
2. *there exists C, only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, such that*

$$\sum_{i=0}^{N} \tau_{i+\frac{1}{2}}(e_{i+1} - e_i)^2 \leq Ch^2, \tag{2.32}$$

where $\tau_{i+\frac{1}{2}}$ is defined in (2.28), and

$$|e_i| \leq Ch,\ \forall i \in \{1,\dots,N\}, \tag{2.33}$$

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

Proof of Theorem 2.3
Step 1. Existence and uniqueness of the solution to (2.28)-(2.31).
Multiplying (2.29) by u_i and summing for $i = 1,\dots,N$ yields that if $c = d = 0$ and $f_i = 0$ for any $i \in \{1,\dots,N\}$, then the unique solution to (2.28)-(2.31) is obtained by taking $u_i = 0$, for any $i \in \{1,\dots,N\}$. This yields existence and uniqueness of the solution to (2.28)-(2.31).

Step 2. Consistency of the fluxes.
Recall that $h = \max\{h_1,\dots,h_N\}$. 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 $i = 0,\dots,N$, 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

$$\begin{aligned} &H^\star_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + T_{i+\frac{1}{2}}, \\ &|T_{i+\frac{1}{2}}| \leq C_1 h,\ i = 0,\dots,N. \end{aligned} \tag{2.34}$$

In order to show this, let us introduce

$$H^{\star,-}_{i+\frac{1}{2}} = -\lambda_i \frac{u(x_{i+\frac{1}{2}}) - u(x_i)}{h_i^+} \text{ and } H^{\star,+}_{i+\frac{1}{2}} = -\lambda_{i+1} \frac{u(x_{i+1}) - u(x_{i+\frac{1}{2}})}{h_{i+1}^-}; \tag{2.35}$$

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

$$H^{\star,-}_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + R^-_{i+\frac{1}{2}}, \text{ where } |R^-_{i+\frac{1}{2}}| \leq Ch, \, i = 1, \ldots, N, \tag{2.36}$$

and

$$H^{\star,+}_{i+\frac{1}{2}} = \overline{H}_{i+\frac{1}{2}} + R^+_{i+\frac{1}{2}}, \text{ where } |R^+_{i+\frac{1}{2}}| \leq Ch, \, i = 0, \ldots, N-1. \tag{2.37}$$

This yields (2.34) for $i = 0$ and $i = N$.
The following equality:

$$\overline{H}_{i+\frac{1}{2}} = H^{\star,-}_{i+\frac{1}{2}} - R^-_{i+\frac{1}{2}} = H^{\star,+}_{i+\frac{1}{2}} - R^+_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1, \tag{2.38}$$

yields that

$$u(x_{i+\frac{1}{2}}) = \frac{\dfrac{\lambda_{i+1}}{h_{i+1}^-} u(x_{i+1}) + \dfrac{\lambda_i}{h_i^+} u(x_i)}{\dfrac{\lambda_i}{h_i^+} + \dfrac{\lambda_{i+1}}{h_{i+1}^-}} + S_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1, \tag{2.39}$$

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 (2.39) of $u(x_{i+1/2})$ in $H^{\star,-}_{i+1/2}$ defined by (2.35) (note that the computation is similar to that performed in (2.26)-(2.27)); this yields

$$H^{\star,-}_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u(x_{i+1}) - u(x_i)) - \frac{\lambda_i}{h_i^+} S_{i+\frac{1}{2}}, \, i = 1, \ldots, N-1. \tag{2.40}$$

Using (2.38), 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 (2.36) and (2.37), 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 (2.34) is proved.

Define now the total exact fluxes;

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

and define

$$F^\star_{i+\frac{1}{2}} = -\tau_{i+\frac{1}{2}}(u(x_{i+1}) - u(x_i)) + au(x_i),\ \forall i \in \{1, \dots, N-1\},$$

$$F^\star_{\frac{1}{2}} = -\frac{\lambda_1}{h_1^-}(u(x_1) - c) + ac,\ F^\star_{N+\frac{1}{2}} = -\frac{\lambda_N}{h_N^+}(d - u(x_N)) + au_N.$$

Then, from (2.34) and the regularity of u, there exists C_2, only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, such that

$$F^\star_{i+\frac{1}{2}} = \overline{F}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}},\ \text{with } |R_{i+\frac{1}{2}}| \leq C_2 h,\ i = 0, \dots, N. \tag{2.41}$$

Hence the numerical approximation of the flux is consistent.

Step 3. Error estimate.
Integrating Equation (2.24) over each control volume yields that

$$\overline{F}_{i+\frac{1}{2}} - \overline{F}_{i-\frac{1}{2}} + bh_i(u(x_i) + S_i) = h_i f_i,\ \forall i \in \{1, \dots, N\}, \tag{2.42}$$

where $S_i \in \mathbb{R}$ is such that there exists C_3 only depending on u such that $|S_i| \leq C_3 h$, for $i = 1, \dots, N$. Using (2.41) yields that

$$F^\star_{i+\frac{1}{2}} - F^\star_{i-\frac{1}{2}} + bh_i(u(x_i) + S_i) = h_i f_i + R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}},\ \forall i \in \{1, \dots, N\}. \tag{2.43}$$

Let $e_i = u(x_i) - u_i$, for $i = 1, \dots, N$, and $e_0 = e_{N+1} = 0$. Substracting (2.29) from (2.43) yields

$$-\tau_{i+\frac{1}{2}}(e_{i+1} - e_i) + \tau_{i-\frac{1}{2}}(e_i - e_{i-1}) + a(e_i - e_{i-1}) + bh_i e_i = -bh_i S_i + R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}},\ \forall i \in \{1, \dots, N\}.$$

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

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

and therefore

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

Since $|S_i| \leq C_3 h$ and thanks to (2.41), one has

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

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

$$A^2 \leq \sum_{i=1}^{N} bC_3 h_i hAB + C_2 hAB.$$

Now, since

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

one obtains that $A \leq C_4 h$, with C_4 only depending on $\underline{\lambda}, \overline{\lambda}, \gamma$ and δ, which yields Estimate (2.32). Applying once again the Cauchy-Schwarz inequality yields Estimate (2.33). ■

2.3.3 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 (2.24), (2.25) (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:

$$-(\lambda u_x(x))_x = f(x), \quad x \in (0,\underline{x}) \cup (\underline{x},1), \tag{2.44}$$

$$u(0) = 0, \tag{2.45}$$

$$u(1) = 0, \tag{2.46}$$

$$(\lambda u_x)^+(\underline{x}) - (\lambda u_x)^-(\underline{x}) = -\alpha, \tag{2.47}$$

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

$$-(\lambda u_x)_x = g \text{ in } \mathcal{D}'((0,1)), \tag{2.48}$$

$$u(0) = 0, \tag{2.49}$$

$$u(1) = 0, \tag{2.50}$$

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 $\varphi \in \mathcal{D}((0,1)) = C_c^\infty((0,1), \mathrm{I\!R})$, 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 $F^+_{i+\frac12}$ and $F^-_{i+\frac12}$, must satisfy, following Equation (2.47),

$$F^+_{i+\frac12} - F^-_{i+\frac12} = \alpha. \tag{2.51}$$

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

$$F^+_{i+\frac12} = -\lambda_{i+1}\frac{u_{i+1} - u_{i+\frac12}}{h^-_{i+1}}$$

$$F^-_{i+\frac12} = -\lambda_i\frac{u_{i+\frac12} - u_i}{h^+_i}.$$

Replacing these expressions in (2.51) yields

$$u_{i+\frac12} = \frac{h^+_i h^-_{i+1}}{(h^-_{i+1}\lambda_i + h^+_i\lambda_{i+1})}[\frac{\lambda_{i+1}}{h^-_{i+1}}u_{i+1} + \frac{\lambda_i}{h^+_i}u_i + \alpha].$$

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 2.3 still holds in this case (under adequate assumptions).

2.4 A semilinear elliptic problem

2.4.1 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:

$$-u_{xx}(x) = f(x, u(x)),\ x \in (0,1), \tag{2.52}$$

$$u(0) = u(1) = 0, \tag{2.53}$$

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

$$\begin{array}{c} f(x,s) \text{ is measurable with respect to } x \in (0,1) \text{ for all } s \in \mathbb{R} \\ \text{and continuous with respect to } s \in \mathbb{R} \text{ for a.e. } x \in (0,1), \end{array} \tag{2.54}$$

$$f \in L^\infty((0,1) \times \mathbb{R}). \tag{2.55}$$

It is possible to prove that there exists at least one weak solution to (2.52), (2.53), that is a function u such that

$$u \in H^1_0((0,1)),\ \int_0^1 u_x(x) v_x(x) dx = \int_0^1 f(x, u(x)) v(x) dx,\ \forall v \in H^1_0((0,1)). \tag{2.56}$$

Note that (2.56) 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 2.4 which is proved in the sequel.

Let $\mathcal{T}$ be an admissible mesh of $[0,1]$ in the sense of Definition 2.1. In order to discretize (2.52), (2.53), let us consider the following (finite volume) scheme

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i(u_i),\ i = 1, \ldots, N, \tag{2.57}$$

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}},\ i = 0, \ldots, N, \tag{2.58}$$

$$u_0 = u_{N+1} = 0, \tag{2.59}$$

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 2.4), one first proves the existence of a solution to (2.57)-(2.59), a stability result, that is, an estimate on the solution of (2.57)-(2.59) (Lemma 2.3) and a compactness lemma (Lemma 2.4).

Lemma 2.3 (Existence and stability result) *Let* $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ *satisfying (2.54), (2.55) and* $\mathcal{T}$ *be an admissible mesh of* $(0,1)$ *in the sense of Definition 2.1. Then, there exists* $(u_1, \ldots, u_N)^t \in \mathbb{R}^N$ *solution of (2.57)-(2.59) and which satisfies:*

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq C, \tag{2.60}$$

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

Proof of Lemma 2.3

Define $M = \|f\|_{L^\infty((0,1)\times\mathbb{R})}$. The proof of estimate (2.60) is given in a first step, and the existence of a solution to (2.57)-(2.59) 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 (2.57)-(2.59) with $f_i(v_i)$ instead of $f_i(u_i)$ in the right hand-side (see Theorem 2.1 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 (2.57)-(2.59) if and only if $U = (u_1, \ldots, u_N)^t$ is a fixed point to F.
Multiplying (2.57) by u_i and summing over i yields

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq M \sum_{i=1}^{N} h_i |u_i|, \tag{2.61}$$

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 (2.61) yields, with $C = M^2$,

$$\sum_{i=0}^{N} \frac{(u_{i+1} - u_i)^2}{h_{i+\frac{1}{2}}} \leq C. \tag{2.62}$$

This gives, in particular, Estimate (2.60) if $(u_1, \ldots, u_N)^t \in \mathbb{R}^N$ is a solution of (2.57)-(2.59) (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 (2.57)-(2.59). ∎

2.4.2 Compactness results

.

Lemma 2.4 (Compactness)
For an admissible mesh $\mathcal{T}$ of $(0,1)$ (see definition 2.1), let $(u_1,\ldots,u_N)^t \in \mathbb{R}^N$ satisfy (2.60) 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 $x \in K_i$, $i = 1,\ldots,N$. 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, $u \in H^1_0((0,1))$.

Proof of Lemma 2.4

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 $i = 1,\ldots,N$, 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 $H^1_0((0,1))$, see Remark 3.9 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 3.1.2 page 37 and Section 3.6 page 93) when the construction of such a function, $\overline{u}_\mathcal{T}$, “close” to $u_\mathcal{T}$ and bounded in $H^1_0((0,1))$ (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 (2.60), since one has, for all $x \in [0,1]$ (since $u_0 = 0$ and by the Cauchy-Schwarz inequality),

$$|u_\mathcal{T}(x)| \le \sum_{i=0}^N |u_{i+1} - u_i| \le (\sum_{i=0}^N \frac{(u_{i+1}-u_i)^2}{h_{i+\frac12}})^{\frac12} \le C.$$

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

$$\|u_\mathcal{T}(\cdot + \eta) - u_\mathcal{T}\|^2_{L^2(\mathbb{R})} \le C\eta(\eta + 2h). \tag{2.63}$$

(Recall that $h = \mathrm{size}(\mathcal{T})$.)
Indeed, for $i \in \{0,\ldots,N\}$ 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 $x \in \mathbb{R}$,

$$\begin{aligned}(u_\mathcal{T}(x+\eta) - u_\mathcal{T}(x))^2 &\le \Big(\sum_{i=0}^N |u_{i+1} - u_i|\chi_{i+\frac12}(x)\Big)^2 \\ &\le \Big(\sum_{i=0}^N \frac{(u_{i+1}-u_i)^2}{h_{i+\frac12}}\chi_{i+\frac12}(x)\Big)\Big(\sum_{i=0}^N \chi_{i+\frac12}(x) h_{i+\frac12}\Big).\end{aligned} \tag{2.64}$$

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 $i \in \{0,\ldots,N\}$, integrating (2.64) over $\mathbb{R}$ yields (2.63).

Step 3. For $0 < \eta < 1$, Estimate (2.63) 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 3.6, see Theorem 3.9 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 2.4, one may use Theorem 3.10 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 (2.63), with $u_{\mathcal{T}_n}$ instead of $u_{\mathcal{T}}$ (and size($\mathcal{T}_n$) instead of h). One obtains

$$\| \frac{u(\cdot + \eta) - u}{\eta} \|^2_{L^2(\mathbb{R})} \le C. \tag{2.65}$$

Since $(u(\cdot + \eta) - u)/\eta$ tends to Du (the distribution derivative of u) in the distribution sense, as $\eta \to 0$, Estimate (2.65) 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 2.4 is complete. ■

.

2.4.3 Convergence

The following convergence result follows from lemmata 2.3 and 2.4.

Theorem 2.4 *Let $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ satisfying (2.54), (2.55). For an admissible mesh, $\mathcal{T}$, of $(0,1)$ (see Definition 2.1), let $(u_1, \dots, u_N)^t \in \mathbb{R}^N$ be a solution to (2.57)-(2.59) (the existence of which is given by Lemma 2.3), and let $u_{\mathcal{T}} : (0,1) \to \mathbb{R}$ by $u_{\mathcal{T}}(x) = u_i$, if $x \in K_i$, $i = 1, \dots, N$.*
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 (2.52), (2.53) (that is, a solution to (2.56)).*

PROOF of Theorem 2.4
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 2.3 and 2.4, 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 $\varphi_i = \varphi(x_i)$, $i = 1, \dots, N$, and $\varphi_0 = \varphi_{N+1} = 0$. If $(u_1, \dots, u_N)$ is a solution to (2.57)-(2.59), multiplying (2.57) by φ_i and summing over $i = 1, \dots, N$ yields

$$\int_0^1 u_{\mathcal{T}}(x) \psi_{\mathcal{T}}(x) dx = \int_0^1 f_{\mathcal{T}}(x) \varphi_{\mathcal{T}}(x) dx, \tag{2.66}$$

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,

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

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 (2.66) 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 (2.60),

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

and therefore $-u_{xx} = f(\cdot, u)$ in the distribution sense in $(0, 1)$. This concludes the proof of Theorem 2.4. 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 2.12 It is possible to give some extensions of the results of this section. For instance, Theorem 2.4 is true with an assumption of "sublinearity" on f instead of (2.55). Furthermore, in order to have both existence and uniqueness of the solution to (2.56) and a rate of convergence (of order h) in Theorem 2.4, it is sufficient to assume, instead of (2.54) and (2.55), that $f \in C^1([0, 1] \times \mathbb{R}, \mathbb{R})$ and that there exists $\gamma < 1$, such that $(f(x, s) - f(x, t))(s - t) \le \gamma(s - t)^2$, 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 3.4 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 3.5 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 3.6 is devoted to the proof of compactness results which are used in the proofs of convergence of the schemes.

3.1 Dirichlet boundary conditions

Let us consider here the following elliptic equation

$$-\Delta u(x) + \mathrm{div}(\mathbf{v}u)(x) + bu(x) = f(x), \quad x \in \Omega, \tag{3.1}$$

with Dirichlet boundary condition:

$$u(x) = g(x), \quad x \in \partial\Omega, \tag{3.2}$$

where

Assumption 3.1

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 3.1, by the Lax-Milgram theorem, there exists a unique variational solution $u \in H^1(\Omega)$ of Problem (3.1)-(3.2). (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 $H^1_0(\Omega)$ satisfying

$$\begin{aligned} &\int_\Omega \Big(\nabla w(x) \cdot \nabla\psi(x) + \mathrm{div}(\mathbf{v}w)(x)\psi(x) + bw(x)\psi(x)\Big)dx = \\ &\int_\Omega \Big(-\nabla\tilde{g}(x) \cdot \nabla\psi(x) - \mathrm{div}(\mathbf{v}\tilde{g})(x)\psi(x) - b\tilde{g}(x)\psi(x) + f(x)\psi(x)\Big)dx, \; \forall\psi \in H^1_0(\Omega). \end{aligned} \tag{3.3}$$

3.1.1 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 (3.1)-(3.2) 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:

$$\begin{aligned} &-\Delta u(x,y) = f(x,y), \; (x,y) \in \Omega, \\ &u(x,y) = 0, \; (x,y) \in \partial\Omega. \end{aligned} \tag{3.4}$$

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 2.1)

Assumption 3.2 *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*

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

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

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 (3.4) over each control volume $K_{i,j}$, which yields

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

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

$$F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j} + F_{i,j+\frac{1}{2}} - F_{i,j-\frac{1}{2}} = h_{i,j} f_{i,j},\ \forall\ (i, j) \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}, \tag{3.5}$$

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

$$\begin{aligned} F_{i+\frac{1}{2},j} &= -\frac{k_j}{h_{i+\frac{1}{2}}}(u_{i+1,j} - u_{i,j}),\ \text{for } i = 0, \cdots, N_1, j = 1, \cdots, N_2, \\ F_{i,j+\frac{1}{2}} &= -\frac{h_i}{k_{j+\frac{1}{2}}}(u_{i,j+1} - u_{i,j}),\ \text{for } i = 1, \cdots, N_1, j = 0, \cdots, N_2, \end{aligned} \tag{3.6}$$

$$u_{0,j} = u_{N_1+1,j} = u_{i,0} = u_{i,N_2+1} = 0,\ \text{for } i = 1, \ldots, N_1, j = 1, \ldots, N_2. \tag{3.7}$$

The numerical scheme (3.5)-(3.7) is therefore clearly conservative and the numerical approximations of the fluxes can easily be shown to be consistent.

Proposition 3.1 (Error estimate) *Let* $\Omega = (0, 1) \times (0, 1)$ *and* $f \in L^2(\Omega)$. *Let u be the unique variational solution to (3.4). Under Assumptions 3.2, let* $\zeta > 0$ *be such that* $h_i \geq \zeta h$ *for* $i = 1, \ldots, N_1$ *and* $k_j \geq \zeta h$ *for* $j = 1, \ldots, N_2$. *Then, there exists a unique solution* $(u_{i,j})_{i=1,\cdots,N_1,j=1,\cdots,N_2}$ *to (3.5)-(3.7). Moreover, there exists* $C > 0$ *only depending on* u, Ω *and* ζ *such that*

$$\sum_{i,j} \frac{(e_{i+1,j} - e_{i,j})^2}{h_{i+\frac{1}{2}}} k_j + \sum_{i,j} \frac{(e_{i,j+1} - e_{i,j})^2}{k_{j+\frac{1}{2}}} h_i \leq Ch^2 \tag{3.8}$$

and

$$\sum_{i,j} (e_{i,j})^2 h_i k_j \leq Ch^2, \tag{3.9}$$

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 (3.4) belongs to $H^2(\Omega)$. We do not give here the proof of this proposition since it is in fact included in Theorem 3.4 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 (3.8) and (3.9) 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 3.3 page 52; in this case, the value C in (3.8) and (3.9) independent of ζ, and therefore the assumption $h_i \geq \zeta h$ for $i = 1, \ldots, N_1$ and $k_j \geq \zeta h$ for $j = 1, \ldots, N_2$ is no longer needed.

Relation (3.8) can be seen as an estimate of a “discrete H_0^1 norm” of the error, while relation (3.9) gives an estimate of the L^2 norm of the error.

Remark 3.1 Some slight modifications of the scheme (3.5)-(3.7) are possible, as in the first item of Remark 2.2 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 (3.9)), exactly as in the 1D case, see Remark 2.5 page 18. In the case equivalent to the second case of Remark 2.5, the point $x_{i,j}$ is not necessarily the center of $K_{i,j}$.

When the mesh is no longer rectangular, the scheme (3.5)-(3.6) 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 3.4, turning back to the usual notation x for the current point of $\mathbb{R}^2$,

$$\begin{array}{l} -\Delta u(x) = f(x),\ x \in \Omega, \\ u(x) = 0,\ x \in \partial\Omega. \end{array} \tag{3.10}$$

Let $\mathcal{T}$ be a mesh defined over Ω; then, integrating the first equation of (3.10) 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 3.2 page 37 and 3.8 page 48

Several computer codes use the following "natural" extension of (3.6) 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 $C_{i,j}$).

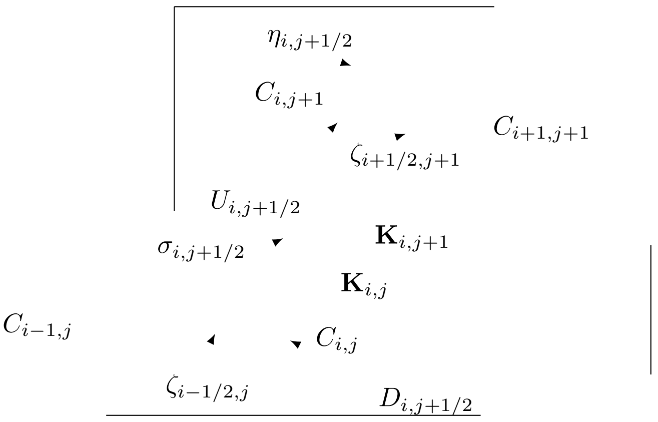

Figure 3.1: FV9 scheme

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

$$\phi_{i,j+1/2} = -\frac{u^U_{i,j+1/2} - u^D_{i,j+1/2}}{d(U_{i,j+1/2}, D_{i,j+1/2})}, \tag{3.11}$$

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 (3.11);

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

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 2.9 page 21 of Section 2.3), the exact solution of (3.10) 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 $-\text{div}(\ \Lambda\ \mathbf{grad}.)$, 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 2.3). 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 3.2 The proof of convergence of this scheme is hindered by the lack of consistency for the discrete adjoint operator (see Section 3.1.4). 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.

3.1.2 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 3.1 page 37 and 3.4 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 3.1 (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.

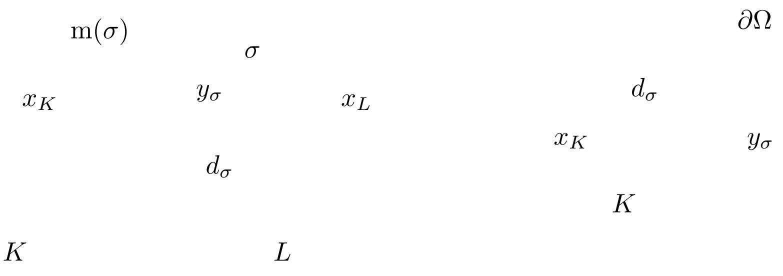

Figure 3.2: Admissible meshes

Remark 3.3 (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 3.1 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 3.1 (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 (3.17)-(3.19) is convenient since (3.18) yields a consistent approximation of the diffusion fluxes and since the transmissibilities (denoted by $\tau_{K|L}$) are positive.

Example 3.2 (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 3.1 if the assumption "on the boundary", namely part (v) of Definition 3.1, 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 3.1 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 (3.20)-(3.23) page 42 allows us to obtain convergence and error estimates results (as in theorems 3.1 page 45 and 3.3 page 52) for all Voronoï meshes. This modification is the obvious generalization of the scheme described in the first item of Remark 2.2 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 (3.20), 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 3.1 (then, the scheme (3.20)-(3.23) 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 3.2 (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 3.1 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

$$\|u\|_{1,\mathcal{T}} = \Big(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma u)^2\Big)^{\frac{1}{2}}, \tag{3.12}$$

where $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$ and $D_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$, $D_\sigma u = |u_K|$ if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$, and where u_K denotes the value taken by u on the control volume K and the sets $\mathcal{E}$, $\mathcal{E}_{\mathrm{int}}$, $\mathcal{E}_{\mathrm{ext}}$ and $\mathcal{E}_K$ are defined in Definition 3.1 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 3.1. One of the tools used below is the following “discrete Poincaré inequality” which may also be found in Temam [141]:

Lemma 3.1 (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 3.1 and $u \in X(\mathcal{T})$ (see Definition 3.2), then*

$$\|u\|_{L^2(\Omega)} \le \mathrm{diam}(\Omega)\|u\|_{1,\mathcal{T}}, \tag{3.13}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 3.2 page 39.

Remark 3.4 (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 3.1 page 40 proven in the sequel.*

Proof of Lemma 3.1
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 3.2 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:

$$|u_K|^2 \le \sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma u)^2}{d_\sigma c_\sigma} \chi_\sigma(x, y(x)) \sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x, y(x)), \text{ for a.e. } x \in K. \tag{3.14}$$

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

$$\sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x, y(x)) \le \mathrm{diam}(\Omega). \tag{3.15}$$

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 (3.15) for such points x leads to (3.15) a.e. on Ω, since $\boldsymbol{d}$ is fixed).

There exists $\sigma_x \in \mathcal{E}_{\text{ext}}$ 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 3.1), this gives (3.15).

Let us integrate (3.14) over Ω; (3.15) gives

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

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

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

Hence the result. ■

Let $\mathcal{T}$ be an admissible mesh. Let us now define a finite volume scheme to discretize (3.1), (3.2) page 32. Let

$$f_K = \frac{1}{\text{m}(K)} \int_K f(x) dx, \forall K \in \mathcal{T}. \tag{3.16}$$

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 2.3), 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 $\text{div}(\mathbf{v}(x)u(x))$ in a stable way (see Section 2.3 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 $v_{K,\sigma} \geq 0$, and $u_{\sigma,+} = u_L$ otherwise; if $\sigma \subset K \cap \partial\Omega$, then $u_{\sigma,+} = u_K$ if $v_{K,\sigma} \geq 0$ 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:

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + b\text{m}(K) u_K = \text{m}(K) f_K, \ \forall K \in \mathcal{T}, \tag{3.17}$$

$$F_{K,\sigma} = -\tau_{K|L}(u_L - u_K), \ \forall \sigma \in \mathcal{E}_{\text{int}}, \text{ if } \sigma = K|L, \tag{3.18}$$

$$F_{K,\sigma} = -\tau_\sigma(g(y_\sigma) - u_K), \ \forall \sigma \in \mathcal{E}_{\text{ext}} \text{ such that } \sigma \in \mathcal{E}_K. \tag{3.19}$$

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:

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + b\text{m}(K) u_K = \text{m}(K) f_K, \ \forall K \in \mathcal{T}, \tag{3.20}$$

$$F_{K,\sigma} = -F_{L,\sigma}, \ \forall \sigma \in \mathcal{E}_{\text{int}}, \text{ if } \sigma = K|L, \tag{3.21}$$

$$F_{K,\sigma} d_{K,\sigma} = -\mathrm{m}(\sigma)(u_\sigma - u_K),\ \forall \sigma \in \mathcal{E}_K,\ \forall K \in \mathcal{T}, \tag{3.22}$$

$$u_\sigma = g(y_\sigma),\ \forall \sigma \in \mathcal{E}_{\mathrm{ext}}. \tag{3.23}$$

Note that (3.20)-(3.23) 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 3.5

1. Note that one may have, for some $\sigma \in \mathcal{E}_K$, $x_K \in \sigma$, and therefore, thanks to (3.22), $u_\sigma = u_K$.
2. The choice $u_\sigma = g(y_\sigma)$ in (3.23) 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 (3.1), (3.2) 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 3.2.

For the sake of simplicity, it is assumed in Definition 3.1 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 $\sigma = K|L$.

3.1.3 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 (3.13) and will yield convergence thanks to a compactness theorem given in Section 3.6 page 93.

Lemma 3.2 (Existence and estimate) *Under Assumptions 3.1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 3.1 page 37; there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to equations (3.20)-(3.23). Furthermore, assuming $g = 0$ and defining $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 3.2) by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, and for any $K \in \mathcal{T}$, the following estimate holds:*

$$\|u_\mathcal{T}\|_{1,\mathcal{T}} \leq \mathrm{diam}(\Omega) \|f\|_{L^2(\Omega)}, \tag{3.24}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 3.2.

Proof of Lemma 3.2

Equations (3.20)-(3.23) 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 (3.20) by u_K and sum over K; from (3.21) and (3.22) one deduces

$$b \sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K^2 + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} u_K + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} u_K = 0, \tag{3.25}$$

which gives, reordering the summation over the set of edges

$$b\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^2+\sum_{\sigma\in\mathcal{E}}\tau_\sigma(D_\sigma u)^2+\sum_{\sigma\in\mathcal{E}}v_\sigma\Big(u_{\sigma,+}-u_{\sigma,-}\Big)u_{\sigma,+}=0, \tag{3.26}$$

where
$|D_\sigma u|=|u_K-u_L|$, if $\sigma=K|L$ and $|D_\sigma u|=|u_K|$, if $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}$;
$v_\sigma=|\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}d\gamma(x)|$, $\mathbf{n}$ 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 $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}$, 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 $\sigma\in\mathcal{E}_{\mathrm{ext}}$.

Now, remark that

$$\sum_{\sigma\in\mathcal{E}}v_\sigma u_{\sigma,+}(u_{\sigma,+}-u_{\sigma,-})=\frac{1}{2}\sum_{\sigma\in\mathcal{E}}v_\sigma\Big((u_{\sigma,+}-u_{\sigma,-})^2+(u_{\sigma,+}^2-u_{\sigma,-}^2)\Big) \tag{3.27}$$

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

$$\sum_{\sigma\in\mathcal{E}}v_\sigma(u_{\sigma,+}^2-u_{\sigma,-}^2)=\sum_{K\in\mathcal{T}}\Big(\int_{\partial K}\mathbf{v}(x)\cdot\mathbf{n}_K d\gamma(x)\Big)u_K^2=\int_\Omega(\mathrm{div}\mathbf{v}(x))u_{\mathcal{T}}^2(x)dx\ge 0. \tag{3.28}$$

Hence,

$$b\|u_{\mathcal{T}}\|_{L^2(\Omega)}^2+\|u_{\mathcal{T}}\|_{1,\mathcal{T}}^2=b\sum_{K\in\mathcal{T}}\mathrm{m}(K)u_K^2+\sum_{\sigma\in\mathcal{E}}\tau_\sigma(D_\sigma u)^2\le 0, \tag{3.29}$$

One deduces, from (3.29), 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 (3.20)-(3.23), for any $\{g(y_\sigma),\ \sigma\in\mathcal{E}_{\mathrm{ext}}\}$ and $\{f_K,\ K\in\mathcal{T}\}$.

Step 2 (estimate)
Assume $g=0$. Multiply (3.20) by u_K, sum over K; then, thanks to (3.21), (3.22), (3.27) and (3.28) one has

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

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 (3.13), 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 (3.20)-(3.23); this is an interesting stability property, even though it will not be used in the proofs of the convergence and error estimate.

Proposition 3.2 *Under Assumption 3.1 page 32, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 3.1 page 37, let $(f_K)_{K\in\mathcal{T}}$ be defined by (3.16). If $f_K\ge 0$ for all $K\in\mathcal{T}$, and $g(y_\sigma)\ge 0$, for all $\sigma\in\mathcal{E}_{\mathrm{ext}}$, then the solution $(u_K)_{K\in\mathcal{T}}$ of (3.20)-(3.23) satisfies $u_K\ge 0$ for all $K\in\mathcal{T}$.*

PROOF of Proposition 3.2

Assume that $f_K \geq 0$ for all $K \in \mathcal{T}$ and $g(y_\sigma) \geq 0$ for all $\sigma \in \mathcal{E}_{\text{ext}}$. 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 $\mathcal{E}_K \subset \mathcal{E}_{\text{int}}$, and that $u_{K_0} \leq 0$. Then, from (3.20),

$$\sum_{\sigma \in \mathcal{E}_{K_0}} F_{K_0,\sigma} + \sum_{\sigma \in \mathcal{E}_{K_0}} v_{K_0,\sigma} u_{\sigma,+} \geq 0; \tag{3.30}$$

since for any neighbour L of K_0 one has $u_L \geq u_{K_0}$, then, noting that $\text{div}\mathbf{v} \geq 0$, 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 $\mathcal{E}_K \not\subset \mathcal{E}_{\text{int}}$, 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 $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, relations (3.22) and (3.23) yield $g(y_\sigma) < 0$, which is in contradiction with the assumption. Hence Proposition 3.2 is proved. ■

Remark 3.6 The maximum principle immediately yields the existence and uniqueness of the solution of the numerical scheme (3.20)-(3.23), which was proved directly in Lemma 3.2.

3.1.4 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 3.2 together with the compactness theorem 3.10 given at the end of this chapter to prove the convergence result. In order to use Theorem 3.10, one needs the following lemma.

Lemma 3.3 *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 3.1 page 37 and $u \in X(\mathcal{T})$ (see Definition 3.2). 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*

$$\|\tilde{u}(\cdot + \eta) - \tilde{u}\|^2_{L^2(\mathbb{R}^d)} \leq \|u\|^2_{1,\mathcal{T}} |\eta| (|\eta| + C\,\text{size}(\mathcal{T})), \forall \eta \in \mathbb{R}^d. \tag{3.31}$$

PROOF of Lemma 3.3

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 3.2 page 39 for the definition of $D_\sigma u$).

This gives, using the Cauchy-Schwarz inequality,

$$|\tilde{u}(x+\eta) - \tilde{u}(x)|^2 \leq \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) \frac{|D_\sigma u|^2}{d_\sigma c_\sigma} \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) d_\sigma c_\sigma, \quad \text{for a.e. } x \in \mathbb{R}^d, \tag{3.32}$$

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

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, x+\eta) d_\sigma c_\sigma \leq |\eta| + C\,\text{size}(\mathcal{T}), \tag{3.33}$$

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 (3.33) for such points x gives (3.33) 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 $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$ and $z_1 = x_L$ or $y_{\tilde{\sigma}}$ with $\tilde{\sigma} \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_L$, depending on the position of y and z in $\overline{K}$ or $\overline{L}$ respectively.
Since $y_1 = y + y_2$, with $|y_2| \leq \text{size}(\mathcal{T})$, and $z_1 = z + z_2$, with $|z_2| \leq \text{size}(\mathcal{T})$, one has

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

and

$$\sum_{\sigma \in \mathcal{E}} \chi_\sigma(y, z) d_\sigma c_\sigma \leq |y - z| + 2\,\text{size}(\mathcal{T}). \tag{3.34}$$

Note that this yields (3.33) 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

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

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 (3.34) with $y = x_i$ and $z = x_{i+1}$, for $i = 1, \ldots, n-1$, one has (3.33) with $C = 2(N-1)$ (in particular, if Ω is convex, $C = 2$ is convenient for (3.33) and therefore for (3.31) as we shall see below).

In order to conclude the proof of Lemma 3.3, remark that, for all $\sigma \in \mathcal{E}$,

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

Therefore, integrating (3.32) over $\mathbb{R}^d$ yields, with (3.33),

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

■

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 3.2 page 51), following Eymard, Gallouët and Herbin [55].

Theorem 3.1 (Convergence, homogeneous Dirichlet boundary conditions) *Under Assumption 3.1 page 32 with $g = 0$, let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 3.1 page 37). Let$(u_K)_{K \in \mathcal{T}}$ be the solution of the system given by equations (3.20)-(3.23) (existence and uniqueness of $(u_K)_{K \in \mathcal{T}}$ are given in Lemma 3.2). 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 (3.1), (3.2) as* $\text{size}(\mathcal{T}) \to 0$. *Furthermore $\|u_\mathcal{T}\|_{1,\mathcal{T}}$ converges to $\|u\|_{H^1_0(\Omega)}$ as* $\text{size}(\mathcal{T}) \to 0$.

Remark 3.7

1. In Theorem 3.1, 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 3.1, 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 3.5 page 60).

   It is also possible to obtain convergence results, towards a "very weak solution" of Problem (3.1), (3.2), 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 3.1, 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 3.1

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 3.1 page 37. First, we want to prove that $u_{\mathcal{T}}$ tends to the unique solution (in $H_0^1(\Omega)$) to (3.3) as size$(\mathcal{T}) \to 0$.

Thanks to Lemma 3.2 and to the discrete Poincaré inequality (3.13), 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 3.3 and to the compactness result given in Theorem 3.10 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 (3.3), 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 (3.3). 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 (3.20) by $\psi(x_K)$, and summing the result over $K \in \mathcal{T}$ yields

$$T_1 + T_2 + T_3 = T_4, \tag{3.35}$$

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

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

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

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

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

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

from Estimate (3.24), 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

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

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

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

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

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

which yields, with the Cauchy-Schwarz inequality,

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

from which one deduces, with Estimate (3.24), that $r_3 \to 0$ as 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 size$(\mathcal{T}) \to 0$ in (3.35) 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 (3.3) 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 size$(\mathcal{T}) \to 0$, where u is the unique solution (in $H_0^1(\Omega)$) to (3.3).

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 3.1. ■

Remark 3.8 (Consistency for the adjoint operator) The proof of Theorem 3.1 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 3.2 page 37). On the contrary, the error estimates proved in sections 3.1.5 and 3.1.6 directly use the consistency for the discretized Dirichlet operator itself.

Remark 3.9 (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 3.10 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 3.4 *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 3.1 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

$$\mathcal{N}(\tilde{g}, \mathcal{T}) = \Big(\sum_{\sigma = K|L \in \mathcal{E}_{\mathrm{int}}} \tau_{K|L} (\tilde{g}_K - \tilde{g}_L)^2 + \sum_{\sigma \in \mathcal{E}_{\mathrm{ext}}} \tau_\sigma (\tilde{g}_{K(\sigma)} - \tilde{g}_\sigma)^2 \Big)^{\frac{1}{2}}, \tag{3.36}$$

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

$$\mathcal{N}(\tilde{g}, \mathcal{T}) \leq C \|\tilde{g}\|_{H^1(\Omega)}. \tag{3.37}$$

Proof of Lemma 3.4

Lemma 3.4 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 3.4. 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

$$(\tilde{g}_K - \tilde{g}_\sigma)^2 \leq C_1 \frac{\mathrm{diam}(K)}{\mathrm{m}(\sigma)} \int_K |\nabla \tilde{g}(x)|^2 dx, \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K, \tag{3.38}$$

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 $J(x_1) = \{x_2 \in \mathbb{R}$, 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,

$$(\tilde{g}_K - \tilde{g}_\sigma)^2 \leq \frac{2}{\mathrm{m}(K)\mathrm{m}(\sigma)}(A + B), \tag{3.39}$$

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,

$$A = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \big(\int_{z_2(x_1,\overline{y})}^{x_2} D_2\tilde{g}(x_1, s) ds\big)^2 d\overline{y} dx_2 dx_1.$$

The Cauchy-Schwarz inequality yields

$$A \le \mathrm{diam}(K) \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \int_{J(x_1)} \big(D_2\tilde{g}(x_1, s)\big)^2 ds d\overline{y} dx_2 dx_1$$

and therefore

$$A \le \mathrm{diam}(K)^3 \int_K \big(D_2\tilde{g}(x)\big)^2 dx. \tag{3.40}$$

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

$$B = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \big(\int_0^{x_1} [D_1\tilde{g}(s, z_2(s,\overline{y})) + \frac{\beta - \overline{y}}{\alpha} D_2\tilde{g}(s, z_2(s,\overline{y}))] ds\big)^2 d\overline{y} dx_2 dx_1.$$

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

$$B \le 2\mathrm{diam}(K)(B_1 + \frac{1}{\zeta^2} B_2), \tag{3.41}$$

with

$$B_i = \int_0^\alpha \int_{J(x_1)} \int_{J(0)} \int_0^{x_1} \big(D_i\tilde{g}(s, z_2(s,\overline{y}))\big)^2 ds d\overline{y} dx_2 dx_1,\ i = 1,\ 2.$$

First, using Fubini's theorem, one has

$$B_i = \int_{J(0)} \int_0^\alpha \big(D_i\tilde{g}(s, z_2(s,\overline{y}))\big)^2 \int_s^\alpha \int_{J(x_1)} dx_2 dx_1 ds d\overline{y}.$$

Therefore

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

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

$$B_i \le \mathrm{diam}(K) \int_0^\alpha \int_{J(s)} \big(D_i\tilde{g}(s, z_2)\big)^2 \frac{\alpha - s}{1 - \frac{s}{\alpha}} dz_2 ds.$$

Hence

$$B_i \le \mathrm{diam}(K)^2 \int_K \big(D_i\tilde{g}(x)\big)^2 dx. \tag{3.42}$$

Using the fact that $\mathrm{m}(K) \ge \pi\zeta^2\big(\mathrm{diam}(K)\big)^2$, (3.39), (3.40), (3.41) and (3.42), one concludes (3.38).

In order to conclude the proof of (3.37), one remarks that

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

Because, for all $K\in\mathcal{T}$ and $\sigma\in\mathcal{E}_K$, $d_\sigma\ge\zeta\mathrm{diam}(K)$, one gets thanks to (3.38), that

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

The above inequality shows that

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

which implies (3.37). ■

Theorem 3.2 (Convergence, non homogeneous Dirichlet boundary condition)
Assume items 1, 2, 3 and 4 of Assumption 3.1 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 3.1 page 37) such that $d_{K,\sigma}\ge\zeta\mathrm{diam}(K)$ for all control volumes $K\in\mathcal{T}$ and for all $\sigma\in\mathcal{E}_K$, and $\mathrm{card}(\mathcal{E}_K)\le M$ for all $K\in\mathcal{T}$. Let $(u_K)_{K\in\mathcal{T}}$ be the solution of the system given by equations (3.20)-(3.22) and

$$u_\sigma=\frac{1}{\mathrm{m}(\sigma)}\int_\sigma g(x)d\gamma(x),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}. \tag{3.43}$$

(note that the proofs of existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ which were given in Lemma 3.2 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 (3.1), (3.2) as $\mathrm{size}(\mathcal{T})\to 0$.

Proof of Theorem 3.2
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 3.1 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

$$\sum_{\sigma\in\mathcal{E}_K}\tilde{F}_{K,\sigma}=\mathrm{m}(K)f_K-\sum_{\sigma\in\mathcal{E}_K}G_{K,\sigma},\ \forall K\in\mathcal{T}, \tag{3.44}$$

$$\tilde{F}_{K,\sigma}=-\tau_{K|L}(\tilde{u}_L-\tilde{u}_K),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if }\sigma=K|L, \tag{3.45}$$

$$\tilde{F}_{K,\sigma}=\tau_\sigma(\tilde{u}_K),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}\text{ such that }\sigma\in\mathcal{E}_K. \tag{3.46}$$

$$G_{K,\sigma}=-\tau_{K|L}(\tilde{g}_L-\tilde{g}_L),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if }\sigma=K|L, \tag{3.47}$$

$$G_{K,\sigma}=-\tau_\sigma(\tilde{g}_\sigma-\tilde{g}_L),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}\text{ such that }\sigma\in\mathcal{E}_K, \tag{3.48}$$

where $\tilde{g}_\sigma=\frac{1}{\mathrm{m}(\sigma)}\int_\sigma g(x)d\gamma(x)$ Multiplying (3.44) 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

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

from the definition (3.36) page 49 of $\mathcal{N}(\tilde{g},\mathcal{T})$ and Definition 3.2 page 39 of $\|\cdot\|_{1,\mathcal{T}}$. Therefore, thanks to Lemma 3.4 page 49 and the discrete Poincaré inequality (3.13), there exists $C_1\in\mathbb{R}$, only depending on Ω, $\|\tilde{g}\|_{H^1(\Omega)}$, ζ, M and f, such that $\|\tilde{u}_{\mathcal{T}}\|_{1,\mathcal{T}} \le C_1$ and $\|\tilde{u}_{\mathcal{T}}\|_{L^2(\Omega)} \le C_1$. 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 $H_0^1(\Omega)$ to (3.3). We proceed as in Theorem 3.1 page 45. Using Lemma 3.3, the compactness result given in Theorem 3.10 page 94 and the uniqueness of the solution (in $H_0^1(\Omega)$) of (3.3), it is sufficient to prove that if $\tilde{u}_{\mathcal{T}}$ converges towards some $\tilde{u} \in H_0^1(\Omega)$, in $L^2(\Omega)$ as size$(\mathcal{T}) \to 0$, then $\tilde{u}$ is the solution to (3.3). 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 (3.3) is identical to the corresponding part in the proof of Theorem 3.1 page 45. This completes the proof of Theorem 3.2. ■

Remark 3.10 (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 3.4 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.

3.1.5 C^2 error estimate

Under adequate regularity assumptions on the solution of Problem (3.1)-(3.2), one may prove that the error between the exact solution and the approximate solution given by the finite volume scheme (3.20)-(3.23) 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 3.3 *Under Assumption 3.1 page 32, let $\mathcal{T}$ be an admissible mesh as defined in Definition 3.1 page 37 and $u_{\mathcal{T}} \in X(\mathcal{T})$ (see Definition 3.2 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 (3.20)-(3.23). Assume that the unique variational solution u of Problem (3.1)-(3.2) 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

$$\|e_{\mathcal{T}}\|_{1,\mathcal{T}} \le C\mathrm{size}(\mathcal{T}), \tag{3.49}$$

where $\|\cdot\|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 3.2,

$$\|e_{\mathcal{T}}\|_{L^2(\Omega)} \le C\mathrm{size}(\mathcal{T}) \tag{3.50}$$

and

$$\begin{aligned}
&\sum_{\substack{\sigma \in \mathcal{E}_{\mathrm{int}} \\ \sigma = K|L}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u_L - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x) \Big)^2 + \\
&\sum_{\substack{\sigma \in \mathcal{E}_{\mathrm{ext}} \\ \sigma \in \overline{K} \cap \partial\Omega}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{g(y_\sigma) - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x) \Big)^2 \le C\mathrm{size}(\mathcal{T})^2.
\end{aligned} \tag{3.51}$$

Remark 3.11

1. Inequality (3.49) (resp. (3.50)) yields an estimate of order 1 for the discrete H_0^1 norm (resp. L^2 norm) of the error on the solution. Note also that, since $u \in C^1(\overline{\Omega})$, one deduces, from (3.50), the existence of C only depending on u and Ω such that $\|u - u_{\mathcal{T}}\|_{L^2(\Omega)} \le C\mathrm{size}(\mathcal{T})$. Inequality (3.51) 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 3.4 page 85). Note that this result is quite different of those of the remarks 2.5 page 18 and 3.1 page 35, which are obtained by using a higher order approximation of the flux.

3. The proof of Theorem 3.3 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 3.2 page 43 (which is used for error estimates with finite difference schemes).

Proof of Theorem 3.3

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 (3.20)-(3.23). Let us write the flux balance for any $K \in \mathcal{T}$;

$$\sum_{\sigma\in\mathcal{E}_K}\left(\overline{F}_{K,\sigma} + \overline{V}_{K,\sigma}\right) + b\int_K u(x)dx = \int_K f(x)dx, \tag{3.52}$$

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 $v_{K,\sigma}\ge 0$ (resp. $v_{K,\sigma}\le 0$) and $x_{\sigma,+} = x_K$ (resp. y_σ) if $\sigma = \mathcal{E}_K\cap\mathcal{E}_{\rm ext}$ and $v_{K,\sigma}\ge 0$ (resp. $v_{K,\sigma}\le 0$). Then, the consistency error on the diffusion and convection fluxes may be defined as

$$R_{K,\sigma} = \frac{1}{{\rm m}(\sigma)}(\overline{F}_{K,\sigma} - F^\star_{K,\sigma}), \tag{3.53}$$

$$r_{K,\sigma} = \frac{1}{{\rm m}(\sigma)}(\overline{V}_{K,\sigma} - V^\star_{K,\sigma}), \tag{3.54}$$

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 $|R_{K,\sigma}| + |r_{K,\sigma}| \le C_1{\rm size}(\mathcal{T})$ 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 $|\rho_K| \le C_2{\rm size}(\mathcal{T})$ with some $C_2\in\mathbb{R}_+$ only depending on u.

Substract (3.20) to (3.52); thanks to (3.53) and (3.54), one has

$$\sum_{\sigma\in\mathcal{E}_K}\left(G_{K,\sigma} + W_{K,\sigma}\right) + b{\rm m}(K)e_K = b{\rm m}(K)\rho_K - \sum_{\sigma\in\mathcal{E}_K}{\rm m}(\sigma)(R_{K,\sigma} + r_{K,\sigma}), \tag{3.55}$$

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

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

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

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 (3.55) 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

$$\|e_{\mathcal{T}}\|^2_{1,\mathcal{T}} + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} e_{\sigma,+} e_K + b\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} \le b\sum_{K\in\mathcal{T}} \mathrm{m}(K)\rho_K e_K - \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)(R_{K,\sigma} + r_{K,\sigma}) e_K, \quad (3.56)$$

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 $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$, $|D_\sigma e| = |e_K|$, if $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}$,
$e_{\sigma,+} = u(x_{\sigma,+}) - u_{\sigma,+}$.
By Young's inequality, the first term of the left hand side satisfies:

$$|\sum_{K\in\mathcal{T}} \mathrm{m}(K)\rho_K e_K| \le \frac{1}{2}\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} + \frac{1}{2} C_2^2 (\mathrm{size}(\mathcal{T}))^2 \mathrm{m}(\Omega). \quad (3.57)$$

Thanks to the assumption div$\mathbf{v} \ge 0$, one obtains, through a computation similar to (3.27)-(3.28) page 43 that

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

Hence, (3.56) and (3.57) yield that there exists C_3 only depending on u, b and Ω such that

$$\|e_{\mathcal{T}}\|^2_{1,\mathcal{T}} + \frac{1}{2} b\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} \le C_3 (\mathrm{size}(\mathcal{T}))^2 - \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)(R_{K,\sigma} + r_{K,\sigma}) e_K, \quad (3.58)$$

Thanks to the property of conservativity, one has $R_{K,\sigma} = -R_{L,\sigma}$ and $r_{K,\sigma} = -r_{L,\sigma}$ for $\sigma \in \mathcal{E}_{\mathrm{int}}$ 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

$$\begin{aligned} |\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)(R_{K,\sigma} + r_{K,\sigma}) e_K| \le \sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma)(D_\sigma e)(R_\sigma + r_\sigma) \le \\ \Big(\sum_{\sigma\in\mathcal{E}} \frac{\mathrm{m}(\sigma)}{d_\sigma}(D_\sigma e)^2\Big)^{\frac{1}{2}} \Big(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma (R_\sigma + r_\sigma)^2\Big)^{\frac{1}{2}}. \end{aligned} \quad (3.59)$$

Now, since $|R_\sigma + r_\sigma| \le C_1 \mathrm{size}(\mathcal{T})$ and since $\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma = d\,\mathrm{m}(\Omega)$, (3.58) and (3.59) 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

$$\|e_{\mathcal{T}}\|^2_{1,\mathcal{T}} + b\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} \le C_5 (\mathrm{size}(\mathcal{T}))^2. \quad (3.60)$$

This inequality yields Estimate (3.49) and, in the case $b > 0$, Estimate (3.50). In the case where $b = 0$, one uses the discrete Poincaré inequality (3.13) and the inequality (3.60) to obtain

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

which yields (3.50).

Remark now that (3.49) can be written

$$\sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}}\\ \sigma=K|L}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u_L - u_K}{d_\sigma} - \frac{u(x_L)-u(x_K)}{d_\sigma}\Big)^2 + \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{ext}}\\ \sigma\in\overline{K}\cap\partial\Omega}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{g(y_\sigma) - u_K}{d_\sigma} - \frac{u(y_\sigma)-u(x_K)}{d_\sigma}\Big)^2 \leq (C\mathrm{size}(\mathcal{T}))^2. \tag{3.61}$$

From Definition (3.53) and the consistency of the fluxes, one has

$$\sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}}\\ \sigma=K|L}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u(x_L)-u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\Big)^2 + \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{ext}}\\ \sigma\in\overline{K}\cap\partial\Omega}} \mathrm{m}(\sigma) d_\sigma \Big(\frac{u(y_\sigma)-u(x_K)}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\Big)^2 = \sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma R_\sigma^2 \leq d\mathrm{m}(\Omega) C_1^2 (\mathrm{size}(\mathcal{T}))^2. \tag{3.62}$$

Then (3.61) and (3.62) give (3.51). ■

3.1.6 H^2 error estimate

In Theorem 3.3, the hypothesis $u \in C^2(\overline{\Omega})$ was used. In the following theorem (Theorem 3.4), one obtains Estimates (3.49) and (3.50), in the case $b = \mathbf{v} = 0$ and assuming some additional assumption on the mesh (see Definition 3.3 below), under the weaker assumption $u \in H^2(\Omega)$. This additional assumption on the mesh is not completely necessary (see Remark 3.13 and Gallouët, Herbin and Vignal [72]). It is also possible to obtain Estimates (3.49) and (3.50) in the cases $b \neq 0$ or $\mathbf{v} \neq 0$ assuming $u \in H^2(\Omega)$ (see Remark 3.13 and Gallouët, Herbin and Vignal [72]). Some similar results are also in Lazarov, Mishev and Vassilevski [99] and Coudière, Vila and Villedieu [41].

Definition 3.3 (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 3.1 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 3.4 (H^2 regularity) *Under Assumption 3.1 page 32 with $b = \mathbf{v} = 0$, let $\mathcal{T}$ be a restricted admissible mesh in the sense of Definition 3.3 and $u_{\mathcal{T}} \in X(\mathcal{T})$ (see Definition 3.2 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 (3.20)-(3.23) (existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ are given by Lemma 3.2). Assume that the unique solution, u, of (3.3) (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 (3.49), (3.50) and (3.51) hold.

Remark 3.12

1. In Theorem 3.4, 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 3.1 with $b = \mathbf{v} = g = 0$ the (unique) solution of (3.3) is necessarily in $H^2(\Omega)$ provided that Ω is convex.

PROOF of Theorem 3.4

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 $\sigma \in \mathcal{E}_{\text{int}}$, 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 $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, 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 3.3), such that, for all control volumes K and for all $\sigma \in \mathcal{E}_K$,

$$|R_{K,\sigma}|^2 \leq C \frac{(\text{size}(\mathcal{T}))^2}{\text{m}(\sigma) d_\sigma} \int_{\mathcal{V}_\sigma} |H(u)(z)|^2 dz, \tag{3.63}$$

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 (3.53)), that is:

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

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

Note that $R_{K,\sigma}$ is well defined, thanks to $u \in H^2(\Omega)$, see Remark 3.12.

In Step 1, one proves (3.63), and, in Step 2, we conclude the proof of Estimates (3.49) and (3.50).

Step 1. Proof of (3.63).

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 (3.63) 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 (3.63) if $\sigma \in \mathcal{E}_{\text{int}}$. 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 $\alpha > \zeta \text{diam}(K)$, $\beta > \zeta \text{diam}(L)$ (ζ is defined in Definition 3.3 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 2.6, 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

$$B_{K,\sigma} = \frac{C_1}{\mathrm{m}(\sigma) d_\sigma} \int_\sigma \int_0^1 |H(u)(tx + (1-t)x_K)||x_K - x|^2 t dt d\gamma(x), \tag{3.64}$$

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 (3.64). 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 $d = 2$, (3.1.6) 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 (3.63) 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 (3.1.6)):

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

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 (3.1.6) yields that:

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

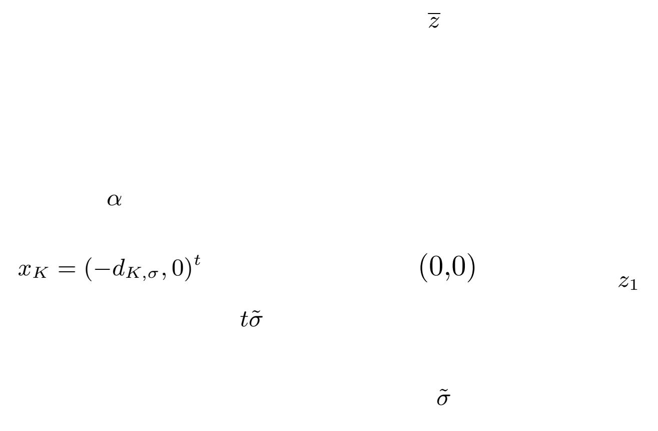

Figure 3.3: Consistency error, $d = 3$

and therefore (3.1.6) gives:

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

and then

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

With a similar estimate on $B_{L,\sigma}$, this yields (3.63) for some C only depending on d and ζ.

Now, one proves (3.63) 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 $\alpha \ge \frac{1}{2}\zeta\mathrm{diam}(K)$. The above proof gives (see Definition 3.1 page 37 for the definition of y_σ), with some C_2 only depending on d,

$$|\frac{u(y_\sigma) - u(x_K)}{2\alpha} - \frac{1}{\mathrm{m}(\hat{\sigma})}\int_{\hat{\sigma}} \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)|^2 \le C_2 \frac{(\mathrm{size}(\mathcal{T}))^2}{\mathrm{m}(\sigma)d_\sigma}\int_{\mathcal{V}_{\hat{\sigma}}} |H(u)(z)|^2 dz, \tag{3.65}$$

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 $t \ge \frac{1}{2}$),

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

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

$$|I_\sigma - I_{\hat\sigma}|^2 \leq \frac{C_4(\mathrm{diam}(K))^2}{\mathrm{m}(\sigma)d_\sigma}\int_{E_\sigma}|H(u)(z)|^2dz, \tag{3.66}$$

with some C_4 only depending on d and ζ.
Inequalities (3.65) and (3.66) yield (3.63) for some C only depending on d and ζ.

One may therefore choose $C \in \mathbb{R}_+$ such that (3.63) holds for $\sigma \in \mathcal{E}_{\mathrm{int}}$ or $\sigma \in \mathcal{E}_{\mathrm{ext}}$. This concludes Step 1.

Step 2. Proof of Estimates (3.49), (3.50) and (3.51).
In order to obtain Estimate (3.49) (and therefore (3.50) from the discrete Poincaré inequality (3.13)), one proceeds as in Theorem 3.3. Inequality (3.56) reads here, since $R_{K,\sigma} = -R_{L,\sigma}$, if $\sigma = K|L$,

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

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

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

The main consequence of (3.63) is that

$$\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma)d_\sigma R^2_\sigma \leq C(\mathrm{size}(\mathcal{T}))^2 \sum_{\sigma\in\mathcal{E}} \int_{\mathcal{V}_\sigma} |H(u)(z)|^2dz = C(\mathrm{size}(\mathcal{T}))^2 \int_\Omega |H(u)(z)|^2dz. \tag{3.67}$$

Then, one obtains

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

This concludes the proof of (3.49) since $u \in H^2(\Omega)$ implies $\int_\Omega |H(u)(z)|^2dz < \infty$.
Estimate (3.51) follows from (3.67) in a similar manner as in the proof of Theorem 3.3. This concludes the proof of Theorem 3.4. ■

Remark 3.13 (Generalizations)

1. By developping the method used to bound the consistency error on the flux on the elements of $\mathcal{E}_{\mathrm{ext}}$, it is possible to replace, in Theorem 3.4, the hypothesis $d_{K,\sigma} \geq \zeta\mathrm{diam}(K)$ in Definition 3.3 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 $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$ (for $\sigma \in \mathcal{E}_{\mathrm{ext}}$, one always needs $y_\sigma - x_K = d_\sigma \mathbf{n}_{K,\sigma}$).

2. It is also possible to prove Theorem 3.4 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 (3.3) 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 3.3. 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 3.4 (see also Eymard, Gallouët and Herbin [55] or Gallouët, Herbin and Vignal [72]).

It is also possible, essentially under Assumption 3.1 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 3.5 (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 3.4 page 63, and let $\zeta > 0$ be such that*

$$\forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K, \qquad d_{K,\sigma} \geq \zeta d_\sigma, \tag{3.68}$$

Let be $u \in X(\mathcal{T})$ (see definition 3.2 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 $d = 3$,

$$\|u\|_{L^q(\Omega)} \leq Cq\|u\|_{1,\mathcal{T}}, \tag{3.69}$$

where $\| \cdot \|_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in definition 3.2 page 39.

Proof of Lemma 3.5

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

$$v_K^2 \leq \sum_{\sigma\in\mathcal{E}} D_\sigma v \, \chi_\sigma^{(1)}(x) \sum_{\sigma\in\mathcal{E}} D_\sigma v \, \chi_\sigma^{(2)}(x), \tag{3.70}$$

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 (3.70) 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

$$\int_\Omega v^2(x)dx \leq \Big(\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) D_\sigma v \Big)^2. \tag{3.71}$$

Applying the inequality (3.71) 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 $\sum\limits_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)d_{K,\sigma} = 2\mathrm{m}(K)$, 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 $p' = \frac{2\alpha}{\alpha+1}$,

$$\|u\|_{L^q(\Omega)} = \Big(\int_\Omega |u(x)|^{2\alpha}dx\Big)^{\frac{1}{2\alpha}} \le \alpha 2^{\frac{1}{p}}\big(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{|D_\sigma u|^{p'}}{{d_{K,\sigma}}^{p'}}\mathrm{m}(\sigma)d_{K,\sigma}\big)^{\frac{1}{p'}}, \tag{3.72}$$

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 (3.72) gives

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

and then (3.69) 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 $\mathcal{D}^1_x$, $\mathcal{D}^2_x$ and $\mathcal{D}^3_x$ 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 $\chi^{(1)}_\sigma$, $\chi^{(2)}_\sigma$ and $\chi^{(3)}_\sigma$ by

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

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:

$$\begin{aligned}&\int_{-A}^{A}\int_{-A}^{A}|v(x_1,x_2,x_3)|^{\frac{3}{2}}dx_1dx_2\\ &\le \Big(\int_{-A}^{A}\int_{-A}^{A}|v(x_1,x_2,x_3)|dx_1dx_2\Big)^{\frac{1}{2}}\Big(\int_{-A}^{A}\int_{-A}^{A}|v(x_1,x_2,x_3)|^2dx_1dx_2\Big)^{\frac{1}{2}}.\end{aligned} \tag{3.73}$$

Now remark that

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

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

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

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 (3.73) in the third direction yields:

$$\int_\Omega |v(x)|^{\frac{3}{2}}dx \le \Big(\sum_{\sigma\in\mathcal{E}}\mathrm{m}(\sigma)D_\sigma v\Big)^{\frac{3}{2}}. \tag{3.74}$$

Now let $v = |u|^4\mathrm{sign}(u)$, since $|v_K - v_L| \leq 4(|u_K|^3 + |u_L|^3)|u_K - u_L|$, Inequality (3.74) yields:

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

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

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

and since $d_{K,\sigma} \geq \zeta d_\sigma$, this yields (3.69) with, for instance, $C = \frac{4\sqrt{3}}{\sqrt{\zeta}}$.

■

Remark 3.14 (Discrete Poincaré Inequality) In the above proof, Inequality (3.71) leads to another proof of some discrete Poincaré inequality (as in Lemma 3.1 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 3.1 page 37 (but more general meshes are possible). Let $v \in X(\mathcal{T})$. Then, (3.71), the Cauchy-Schwarz inequality and the fact that $\displaystyle\sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma) d_\sigma = 2\,\mathrm{m}(\Omega)$ yield

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

A similar result holds in the three-dimensional case.

Corollary 3.1 (Error estimate) *Under the same assumptions and with the same notations as in Theorem 3.3 page 52, or as in Theorem 3.4 page 55, and assuming that the mesh satisfies, for some $\zeta > 0$, $d_{K,\sigma} \geq \zeta d_\sigma$, 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*

$$\|e_\mathcal{T}\|_{L^q(\Omega)} \leq Cq\mathrm{size}(\mathcal{T});\ \textit{for any } q \in \begin{cases} [1, 6] & \textit{if } d = 3, \\ [1, +\infty) & \textit{if } d = 2; \end{cases} \tag{3.75}$$

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

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} \leq C\mathrm{size}(\mathcal{T})(|\ln(\mathrm{size}(\mathcal{T}))| + 1), \qquad \text{if } d = 2. \tag{3.76a}$$

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} \leq C\mathrm{size}(\mathcal{T})^{2/3}, \qquad \text{if } d = 3. \tag{3.76b}$$

Proof of Corollary 3.1

Estimate (3.49) of Theorem 3.3 (or Theorem 3.4) and Inequality (3.69) of Lemma 3.5 immediately yield Estimate (3.75) in the case $d = 2$. Let us now prove (3.76). Remark that

$$\|e_\mathcal{T}\|_{L^\infty(\Omega)} = \max\{|e_K|, K \in \mathcal{T}\} \leq \Big(\frac{1}{\zeta_\mathcal{T}\mathrm{size}(\mathcal{T})^2}\Big)^{\frac{1}{q}} \|e_\mathcal{T}\|_{L^q}. \tag{3.77}$$

For $d = 2$, a study of the real function defined, for $q \geq 2$, 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 $\ln h \leq -\frac{1}{2}$. Therefore (3.75) and (3.77) yield(3.76). The 3 dimensional case is an immediate consequence of(3.75) with $q = 6$. ■

3.2 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 3.3 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:

$$-\Delta u(x) = f(x),\ x \in \Omega, \tag{3.78}$$

with Neumann boundary conditions:

$$\nabla u(x)\cdot \boldsymbol{n}(x) = g(x),\ x \in \partial\Omega, \tag{3.79}$$

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 3.3

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 3.3, Problem (3.78), (3.79) 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:

$$u \in H^1(\Omega),\ \int_\Omega u(x)dx = 0, \tag{3.80}$$

$$\int_\Omega \nabla u(x)\nabla\psi(x) = \int_\Omega f(x)\psi(x)dx + \int_{\partial\Omega} g(x)\overline{\gamma}(\psi)(x)d\gamma(x),\ \forall \psi \in H^1(\Omega). \tag{3.81}$$

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

3.2.1 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 3.1 page 37, thanks to the Neumann boundary conditions and the absence of convection term.

Definition 3.4 (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 (3.78), (3.79), 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 3.1 page 37.

The same notations as in Definition 3.1 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 3.2 page 39.

Definition 3.5 (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 3.4.
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 $\tau_\sigma = \frac{{\rm m}(\sigma)}{d_\sigma}$ and $\mathcal{E}_{\rm int}$ are defined in Definition 3.1 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 3.4 . For $K \in \mathcal{T}$, let us define:

$$f_K = \frac{1}{{\rm m}(K)} \int_K f(x)dx, \tag{3.82}$$

$$\begin{array}{l} g_K = \dfrac{1}{{\rm m}(\partial K \cap \partial\Omega)} \displaystyle\int_{\partial K \cap \partial\Omega} g(x) d\gamma(x) \text{ if } {\rm m}(\partial K \cap \partial\Omega) \neq 0, \\ g_K = 0 \text{ if } {\rm m}(\partial K \cap \partial\Omega) = 0. \end{array} \tag{3.83}$$

Recall that, in formula (3.82), ${\rm m}(K)$ denotes the d-dimensional Lebesgue measure of K, and, in (3.83), ${\rm m}(\partial K \cap \partial\Omega)$ 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 (3.20)-(3.22) page 42, with $b = 0$ and $\mathbf{v} = 0$. This yields:

$$-\sum_{L \in \mathcal{N}(K)} \tau_{K|L}\Big(u_L - u_K\Big) = {\rm m}(K) f_K + {\rm m}(\partial K \cap \partial\Omega) g_K, \ \forall K \in \mathcal{T}, \tag{3.84}$$

(see the notations in Definitions 3.1 page 37 and 3.4 page 63). The condition (3.80) is discretized by:

$$\sum_{K \in \mathcal{T}} {\rm m}(K) u_K = 0. \tag{3.85}$$

Then, the approximate solution, $u_\mathcal{T}$, belongs to $X(\mathcal{T})$ (see Definition 3.2 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 (3.84) and (3.85).

Lemma 3.6 *Under Assumption 3.3. let $\mathcal{T}$ be an admissible mesh (see Definition 3.4) and $\{f_K, K \in \mathcal{T}\}$, $\{g_K, K \in \mathcal{T}\}$ defined by (3.82), (3.83). Then, there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to (3.84)-(3.85).*

Proof of lemma 3.6

Let $N = {\rm card}(\mathcal{T})$. The equations (3.84) 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 (3.84), multiplying (3.84) (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 (3.84) has a solution, is that

$$\sum_{K\in\mathcal{T}}(\mathrm{m}(K)f_K + \mathrm{m}(\partial K\cap\partial\Omega)g_K) = 0. \tag{3.86}$$

Since the dimension of the null space of A is one, this condition is also a sufficient condition. Therefore, System (3.84) has a solution if and only if (3.86) holds, and this solution is unique up to an additive constant. Adding condition (3.85) yields uniqueness. Note that (3.86) holds thanks to the second item of Assumption 3.3; this concludes the proof of Lemma 3.6. ■

3.2.2 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 3.3), requires a “discrete Poincaré” inequality as in the case of the Dirichlet problem.

Lemma 3.7 (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 3.4 page 63), $\mathcal{T}$, and for all $u\in X(\mathcal{T})$ (see Definition 3.2 page 39), the following inequality holds:*

$$\|u\|^2_{L^2(\Omega)} \le C|u|^2_{1,\mathcal{T}} + 2(\mathrm{m}(\Omega))^{-1}(\int_\Omega u(x)dx)^2, \tag{3.87}$$

where $|\cdot|_{1,\mathcal{T}}$ is the discrete H^1 seminorm defined in Definition 3.5.

PROOF of Lemma 3.7

The proof given here is a “direct proof”; another proof, by contradiction, is possible (see Remark 3.16). 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 3.7 amounts to proving the existence of $D\ge 0$, only depending on Ω, such that

$$\|u - m_\Omega(u)\|^2_{L^2(\Omega)} \le D|u|^2_{1,\mathcal{T}}. \tag{3.88}$$

The proof of (3.88) 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, $\omega\ne\emptyset$ 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

$$\|u(x) - m_\omega(u)\|^2_{L^2(\omega)} \le \frac{1}{\mathrm{m}(\omega)}C_0|u|^2_{1,\mathcal{T}}. \tag{3.89}$$

(Taking $\omega=\Omega$, this proves (3.88) and Lemma 3.7 in the case where Ω is convex.)

Noting that

$$\int_\omega (u(x) - m_\omega(u))^2dx \le \frac{1}{\mathrm{m}(\omega)}\int_\omega\big(\int_\omega (u(x)-u(y))^2dy\big)dx,$$

(3.89) is proved provided that there exists $C_0\in\mathbb{R}_+$, only depending on Ω, such that

$$\int_\omega \int_\omega (u(x) - u(y))^2 dxdy \le C_0 |u|^2_{1,\mathcal{T}}. \tag{3.90}$$

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,

$$(u(x) - u(y))^2 \le \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_{\sigma,y-x}} \chi_\sigma(x, y) \sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_{\sigma,y-x} \chi_\sigma(x, y), \tag{3.91}$$

with

$$c_{\sigma,y-x} = |\frac{y - x}{|y - x|} \cdot \mathbf{n}_\sigma|,$$

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

$$\sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma c_{\sigma,y-x} \chi_\sigma(x, y) = |(x_K - x_L) \cdot \frac{y - x}{|y - x|}|,$$

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 (3.91) 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,

$$\int_\omega \int_\omega (u(x) - u(y))^2 dxdy \le \mathrm{diam}(\Omega) \int_{\mathbb{R}^d} \big(\sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{|D_\sigma u|^2}{d_\sigma c_{\sigma,z}} \int_\omega \chi_\sigma(x, x + z) dx \big) dz. \tag{3.92}$$

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 (3.92):

$$\int_\omega \int_\omega (u(x) - u(y))^2 dxdy \le (\mathrm{diam}(\Omega))^2 \mathrm{m}(B_\Omega) \sum_{\sigma \in \mathcal{E}_{\rm int}} \frac{\mathrm{m}(\sigma) |D_\sigma u|^2}{d_\sigma},$$

where B_Ω denotes the ball of $\mathbb{R}^d$ of center 0 and radius $\mathrm{diam}(\Omega)$.

This inequality proves (3.90) and then (3.89) 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 3.7 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 (3.89) 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

$$\|u - m_I(u)\|^2_{L^2(\omega)} \leq C_1 |u|^2_{1,\mathcal{T}}. \tag{3.93}$$

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, $C_7 = C_6 ({\rm diam}(\omega))^2$. The bounds on A and B yield (3.93).

Step 3 (proof of (3.88))

Let us now prove that there exists $D \in \mathbb{R}_+$, only depending on Ω such that (3.88) 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

$$\|u - m_i\|^2_{L^2(\Omega_i)} \leq C_i |u|^2_{1,\mathcal{T}}, \ \forall i \in \{1, \ldots, n\}, \tag{3.94}$$

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 $(m_i - m_{i,j})^2 {\rm m}(\Omega_i) \leq 2(C_i + C_{i,j}) |u|^2_{1,\mathcal{T}}$, 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, (3.94) yields the existence of D, only depending on Ω, such that (3.88) holds. This completes the proof of Lemma 3.7. ■

An easy consequence of the proof of Lemma 3.7 is the following lemma. Although this lemma is not used in the sequel, it is interesting in its own sake.

Lemma 3.8 (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 3.4 page 63) $\mathcal{T}$ and for all $u \in X(\mathcal{T})$ (see Definition 3.2 page 39), the following inequality holds:*

$$\|u - m_I(u)\|^2_{L^2(\Omega)} \le C|u|^2_{1,\mathcal{T}}$$

where $|\cdot|_{1,\mathcal{T}}$ is the discrete H^1 seminorm defined in Definition 3.5 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 3.4.

Finally, let us point out that a continuous version of lemmata 3.7 (known as the Poincaré-Wirtinger inequality) and 3.8 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 3.4 page 49. The advantage of the latter is that it gives a more explicit bound.

Lemma 3.9 *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:

$$\|u\|^2_{L^2(\Omega)} \le C|u|^2_{H^1(\Omega)} + 2(\mathrm{m}(\Omega))^{-1}(\int_\Omega u(x)dx)^2$$

and

$$\|u - m_I(u)\|^2_{L^2(\Omega)} \le \tilde{C}|u|^2_{H^1(\Omega)},$$

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 $H^{1/2}(\partial\Omega)$.

3.2.3 Error estimate

Under Assumption 3.3, let $\mathcal{T}$ be an admissible mesh (see Definition 3.4) and $\{f_K, K \in \mathcal{T}\}$, $\{g_K, K \in \mathcal{T}\}$ defined by (3.82), (3.83). By Lemma 3.6, there exists a unique solution $(u_K)_{K\in\mathcal{T}}$ to (3.84)-(3.85). Under an additional regularity assumption on the exact solution, the following error estimate holds:

Theorem 3.5 *Under Assumption 3.3 page 63, let $\mathcal{T}$ be an admissible mesh (see Definition 3.4 page 63) and $h = \mathrm{size}(\mathcal{T})$. Let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (3.84) and (3.85) (thanks to (3.82) and (3.83), existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 3.6). Let $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 3.2 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 (3.80), (3.81) satisfies $u \in C^2(\overline{\Omega})$.*
Then there exists $C \in \mathbb{R}_+$ which only depends on u and Ω such that

$$\|u_\mathcal{T} - u\|_{L^2(\Omega)} \le Ch, \tag{3.95}$$

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)d_\sigma(\frac{u_L - u_K}{d_\sigma} - \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x))^2 \le Ch^2. \tag{3.96}$$

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 3.5
Let $C_{\mathcal{T}} \in {\rm I\!R}$ be such that

$$\sum_{K\in\mathcal{T}} \overline{u}(x_K){\rm m}(K) = 0,$$

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

$$|e_{\mathcal{T}}|_{1,\mathcal{T}} \le Ch \;\text{ and }\; \|e_{\mathcal{T}}\|_{L^2(\Omega)} \le Ch. \tag{3.97}$$

Integrating (3.78) page 63 over $K \in \mathcal{T}$, and taking (3.79) page 63 into account yields:

$$\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) = \int_K f(x)dx + \int_{\partial K\cap\partial\Omega} g(x)d\gamma(x). \tag{3.98}$$

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:

$$R_{K,\sigma} = \frac{1}{{\rm m}(\sigma)} \int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) - \frac{u(x_L) - u(x_K)}{d_\sigma}. \tag{3.99}$$

Note that the definition of $R_{K,\sigma}$ remains with $\overline{u}$ instead of u in (3.99).
Thanks to the regularity of the solution u, there exists $C_1 \in {\rm I\!R}_+$, only depending on u, such that $|R_{K,L}| \le C_1 h$. Using (3.98), (3.99) and (3.84) yields

$$\sum_{K|L\in\mathcal{E}_{\rm int}} \tau_{K|L}(e_L - e_K)^2 \le d{\rm m}(\Omega)(C_1 h)^2,$$

which gives the first part of (3.97).
Thanks to the discrete Poincaré inequality (3.87) applied to the function $e_{\mathcal{T}}$, and since

$$\sum_{K\in\mathcal{T}} {\rm m}(K) e_K = 0$$

(which is the reason why $e_{\mathcal{T}}$ was defined with $\overline{u}$ instead of u) one obtains the second part of (3.97), that is the existence of C_2 only depending on u and Ω such that

$$\sum_{K\in\mathcal{T}} {\rm m}(K)(e_K)^2 \le C_2 h^2.$$

From (3.97), one deduces (3.95) 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

$$\sum_{K\in\mathcal{T}} \int_K (u(x_K) - u(x))^2 dx \le \sum_{K\in\mathcal{T}} {\rm m}(K)(C_2 h)^2 = {\rm m}(\Omega)(C_2 h)^2.$$

Then, noting that

$$\begin{aligned}\|u_{\mathcal{T}} - u\|^2_{L^2(\Omega)} &= \sum_{K\in\mathcal{T}} \int_K (u_K - u(x))^2 dx \\ &\le 3\sum_{K\in\mathcal{T}} {\rm m}(K)(e_K)^2 + 3(C_{\mathcal{T}})^2{\rm m}(\Omega) + 3\sum_{K\in\mathcal{T}} \int_K (u(x_K) - u(x))^2 dx\end{aligned}$$

yields (3.95).

The proof of Estimate (3.96) 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 7.3.6). ■

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 3.6), Estimates (3.97) and (3.96) hold under the weaker assumption $u \in H^2(\Omega)$ (see Theorem 3.6 below). It is therefore also possible to obtain (3.95) under the additional assumption that u is Lipschitz continuous.

Definition 3.6 (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 3.4 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 3.6 (H^2 regularity, Neumann problem) *Under Assumption 3.3 page 63, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 3.6 and $h = \mathrm{size}(\mathcal{T})$. Let $u_{\mathcal{T}} \in X(\mathcal{T})$ (see Definition 3.2 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 (3.84) and (3.85) (thanks to (3.82) and (3.83), existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 3.6). Assume that the unique solution, u, of (3.80), (3.81) 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 (3.97) and (3.96) hold.

Note that, in Theorem 3.6, 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 3.12).

Proof of Theorem 3.6

The proof is very similar to that of Theorem 3.4 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 3.6), such that, for all $\sigma \in \mathcal{E}_{\mathrm{int}}$,

$$|R_\sigma|^2 \leq C \frac{h^2}{\mathrm{m}(\sigma) d_\sigma} \int_{\mathcal{V}_\sigma} |(H(u)(z)|^2 dz, \tag{3.100}$$

and therefore

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \mathrm{m}(\sigma) d_\sigma R_\sigma^2 \leq C h^2 \int_\Omega |H(u)(z)|^2 dz. \tag{3.101}$$

The proof of (3.100) (from which (3.101) is an easy consequence) was already done in the proof of Theorem 3.4 (note that, here, there is no need to consider the case of $\sigma \in \mathcal{E}_{\mathrm{ext}}$). In order to obtain Estimate (3.97), one proceeds as in Theorem 3.4. 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

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

Then, one obtains, with (3.101),

$$|e_{\mathcal{T}}|_{1,\mathcal{T}} \leq \sqrt{C} h\Big(\int_\Omega |H(u)(z)|^2 dz\Big)^{\frac{1}{2}}.$$

This concludes the proof of the first part of (3.97). The second part of (3.97) is a consequence of the discrete Poincaré inequality (3.87). Using (3.101) also easily leads (3.96).
Note also that, if u is Lipschitz continuous, Inequality (3.95) follows from the second part of (3.97) and the definition of $\overline{u}$ as in Theorem 3.5.
This concludes the proof of Theorem 3.6. ■

Some generalizations of Theorem 3.6 are possible, as for the Dirichlet case, see Remark 3.13 page 59.

3.2.4 Convergence

A convergence result, under Assumption 3.3, 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 3.10 (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 3.4 page 63, and $u \in X(\mathcal{T})$ (see Definition 3.2 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*

$$\|\overline{\gamma}(u)\|_{L^2(\partial\Omega)} \leq C(|u|_{1,\mathcal{T}} + \|u\|_{L^2(\Omega)}). \tag{3.102}$$

Remark 3.15 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 3.4 page 63) so as to take into account non connected subsets.

Proof of Lemma 3.10
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

$$\|\alpha_i \overline{\gamma}(u)\|_{L^2(\Gamma_i)} \leq C_i\big(|u|_{1,\mathcal{T}} + \|u\|_{L^2(\Omega)}\big). \tag{3.103}$$

The existence of C, only depending on Ω, such that (3.102) 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 (3.103).

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 $i \in \{1,\ldots,N\}$ 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,

$$|\alpha_i \overline{\gamma}(u)(x)|^2 \leq A(x) + B(x) \tag{3.104}$$

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 (3.104) over Γ_i, the bounds on A and B lead (3.103) for some convenient C_i and it concludes the proof of Lemma 3.10. ■

Remark 3.16 Using this "trace inequality" (3.102) and the Kolmogorov theorem (see Theorem 3.9 page 94, it is possible to prove Lemma 3.7 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 3.4) 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 3.7 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 3.11 page 95, yields that $D_i u = 0$, for all $i \in \{1, \ldots, n\}$ (even if $\mathrm{size}(\mathcal{T}_n) \not\to 0$, 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 (3.84) and (3.85), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (3.82) and (3.83) is stable: the approximate solution given by the scheme is bounded independently of the mesh, as we proceed to show.

Lemma 3.11 (Estimate for the Neumann problem) *Under Assumption 3.3 page 63, let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 3.4 page 63). Let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (3.84) and (3.85), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (3.82) and (3.83); the existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 3.6. Let $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 3.2) 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*

$$|u_\mathcal{T}|_{1,\mathcal{T}} \leq C, \tag{3.105}$$

where $|\cdot|_{1,\mathcal{T}}$ is defined in Definition 3.5 page 64.

Proof of Lemma 3.11

Multiplying (3.84) by u_K and summing over $K \in \mathcal{T}$ yields

$$\sum_{K|L\in\mathcal{E}_{\mathrm{int}}} \tau_{K|L}(u_L - u_K)^2 = \sum_{K\in\mathcal{T}} \mathrm{m}(K) f_K u_K + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} u_{K_\sigma} g_{K_\sigma} \mathrm{m}(\sigma), \tag{3.106}$$

where, for $\sigma \in \mathcal{E}_{\mathrm{ext}}$, $K_\sigma \in \mathcal{T}$ is such that $\sigma \in \mathcal{E}_{K_\sigma}$.

We get (3.105) from (3.106) using (3.102), (3.87) and the Cauchy-Schwarz inequality. ■

Using the estimate (3.105) on the approximate solution, a convergence result is given in the following theorem.

Theorem 3.7 (Convergence in the case of the Neumann problem)

Under Assumption 3.3 page 63, let u be the unique solution to (3.80),(3.81). For an admissible mesh (in the sense of Definition 3.4 page 63) $\mathcal{T}$, let $(u_K)_{K\in\mathcal{T}}$ be the unique solution to (3.84) and (3.85) (where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (3.82) and (3.83), the existence and uniqueness of $(u_K)_{K\in\mathcal{T}}$ is given in Lemma 3.6) and define $u_\mathcal{T} \in X(\mathcal{T})$ (see Definition 3.2) by $u_\mathcal{T}(x) = u_K$ for a.e. $x \in K$, for all $K \in \mathcal{T}$. Then,

$$u_{\mathcal{T}} \to u \textit{ in } L^2(\Omega) \textit{ as } \mathrm{size}(\mathcal{T}) \to 0,$$

$$|u_{\mathcal{T}}|^2_{1,\mathcal{T}} \to \int_{\Omega} |\nabla u(x)|^2 dx \textit{ as } \mathrm{size}(\mathcal{T}) \to 0$$

and

$$\overline{\gamma}(u_{\mathcal{T}}) \to \overline{\gamma}(u) \textit{ in } L^2(\partial\Omega) \textit{ for the weak topology as } \mathrm{size}(\mathcal{T}) \to 0,$$

where the function $\overline{\gamma}(u)$ *stands for the trace of* u *on* $\partial\Omega$ *in the sense given in Lemma 3.10 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 3.7

Step 1 (Compactness)
Denote by Y the set of approximate solutions $u_{\mathcal{T}}$ for all admisible meshes $\mathcal{T}$. Thanks to Lemma 3.11 and to the discrete Poincaré inequality (3.87), 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 3.9 and 3.11 page 95, provided that there exists a real positive number C only depending on Ω, f and g such that

$$\|\tilde{u}_{\mathcal{T}}(\cdot+\eta) - \tilde{u}_{\mathcal{T}}\|^2_{L^2(\mathbb{R}^d)} \le C|\eta|, \text{ for any admissible mesh } \mathcal{T} \text{ and for any } \eta \in \mathbb{R}^d,\ |\eta| \le 1, \tag{3.107}$$

and that, for any compact subset $\bar{\omega}$ of Ω,

$$\begin{aligned}&\|u_{\mathcal{T}}(\cdot+\eta) - u_{\mathcal{T}}\|^2_{L^2(\bar{\omega})} \le C|\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})), \text{ for any admissible mesh } \mathcal{T}\\ &\text{and for any } \eta \in \mathbb{R}^d \text{ such that } |\eta| < d(\bar{\omega}, \Omega^c).\end{aligned} \tag{3.108}$$

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 (3.107) and (3.108), 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:

$$|\tilde{u}(x+\eta) - \tilde{u}(x)| \le \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \chi_\sigma(x,x+\eta)|D_\sigma u| + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} \chi_\sigma(x,x+\eta)|u_\sigma|, \text{ for a.e. } x \in \Omega, \tag{3.109}$$

where, for $\sigma \in \mathcal{E}_{\mathrm{ext}}$, $u_\sigma = u_K$, 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 (3.108). 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 (3.109) is 0, and the same proof as in Lemma 3.3 page 44 gives, from an integration over $\bar{\omega}$ instead of Ω and from (3.33) with $C = 2$ since $[x, x+\eta] \subset \Omega$ for $x \in \bar{\omega}$,

$$\|u_{\mathcal{T}}(\cdot+\eta) - u_{\mathcal{T}}\|^2_{L^2(\bar{\omega})} \le |u|^2_{1,\mathcal{T}}|\eta|(|\eta| + 2\,\mathrm{size}(\mathcal{T})). \tag{3.110}$$

In order to prove (3.107), remark that the number of non zero terms in the second term of the right hand side of (3.109) 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

$$|\tilde{u}(x+\eta) - \tilde{u}(x)|^2 \le C_1(\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \chi_\sigma(x,x+\eta)|D_\sigma u|)^2 + C_1 \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} \chi_\sigma(x,x+\eta)u_\sigma^2, \text{ for a.e. } x \in \Omega. \tag{3.111}$$

Let us integrate this inequality over $\mathbb{R}^d$. As seen in the proof of Lemma 3.3 page 44,

$$\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm int}} \chi_\sigma(x,x+\eta)|D_\sigma u|\Big)^2 dx \le |u|_{1,\mathcal{T}}^2 |\eta|(|\eta| + 2(N-1){\rm size}(\mathcal{T}));$$

hence, by Lemma 3.11 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 (3.111) integrated over $\mathbb{R}^d$;

$$\begin{aligned}\int_{\mathbb{R}^d} \Big(\sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi_\sigma(x,x+\eta)u_\sigma^2\Big) dx \quad &\le \quad \sum_{\sigma\in\mathcal{E}_{\rm ext}} {\rm m}(\sigma)|\eta| u_\sigma^2 \\ &\le \quad \|\overline{\gamma}(u_\mathcal{T})\|_{L^2(\partial\Omega)}^2 |\eta|;\end{aligned}$$

therefore, thanks to Lemma 3.10, Lemma 3.11 and to the discrete Poincaré inequality (3.87), 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 (3.107) 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 (3.80), (3.81) (in $L^2(\Omega)$ as ${\rm size}(\mathcal{T}) \to 0$) is first proved.
Since the solution to (3.80), (3.81) 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 ${\rm size}(\mathcal{T}_n) \to 0$ as $n \to 0$, then u is a solution to (3.80)-(3.81).
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 (3.84)-(3.85) page 64 with $\mathcal{T} = \mathcal{T}_n$. Assume $u_{\mathcal{T}_n} \to u$ in $L^2(\Omega)$ and ${\rm size}(\mathcal{T}_n) \to 0$ 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 (3.80). It remains to show that u satisfies (3.81). 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

$$\begin{aligned}-\int_\Omega u(x)\Delta\varphi(x)dx + \int_{\partial\Omega} \nabla\varphi(x)\cdot\mathbf{n}(x)v(x)d\gamma(x) = \int_\Omega f(x)\varphi(x)dx \\ + \int_{\partial\Omega} g(x)\varphi(x)d\gamma(x),\ \forall\varphi\in C^2(\overline{\Omega}),\end{aligned} \tag{3.112}$$

and then that u satisfies (3.81).

Let $\mathcal{T}$ be an admissible mesh, $u_\mathcal{T}$ the corresponding approximate solution to the Neumann problem, given by (3.84) and (3.85), where $(f_K)_{K\in\mathcal{T}}$ and $(g_K)_{K\in\mathcal{T}}$ are given by (3.82) and (3.83) 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 (3.84) by φ_K, summing over $K \in \mathcal{T}$ and reordering the terms yields

$$\sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}(K)} \tau_{K|L}(\varphi_L - \varphi_K) = \int_\Omega f(x)\varphi_\mathcal{T}(x)dx + \int_{\partial\Omega} \overline{\gamma}(\varphi_\mathcal{T})(x)g(x)d\gamma(x). \tag{3.113}$$

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 $|R_{K,L}|\le C_4\mathrm{m}(K|L)\mathrm{size}(\mathcal{T})$, where C_4 only depends on φ. Hence (3.113) may be rewritten as

$$\begin{aligned} -\int_\Omega u_{\mathcal{T}}(x)\Delta\varphi(x)dx + \int_{\partial\Omega} \nabla\varphi(x)\cdot\mathbf{n}(x)\overline{\gamma}(u_{\mathcal{T}})(x)d\gamma(x) + r(\varphi,\mathcal{T}) = \\ \int_\Omega f(x)\varphi_{\mathcal{T}}(x)dx + \int_{\partial\Omega}\overline{\gamma}(\varphi_{\mathcal{T}})(x)g(x)d\gamma(x), \end{aligned} \tag{3.114}$$

where

$$\begin{aligned} |r(\varphi,\mathcal{T})| &= C_4\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|D_\sigma u|\mathrm{m}(\sigma)\mathrm{size}(\mathcal{T}) \\ &\le C_4\Big(\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}|D_\sigma u|^2\frac{\mathrm{m}(\sigma)}{d_\sigma}\Big)^{\frac12}\Big(\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}(\sigma)d_\sigma\Big)^{\frac12}\mathrm{size}(\mathcal{T}) \\ &\le C_5\mathrm{size}(\mathcal{T}), \end{aligned}$$

where C_5 is a real positive number only depending on f, g, Ω and φ (thanks to Lemma 3.11).
Writing (3.114) with $\mathcal{T}=\mathcal{T}_n$ and passing to the limit as n tends to infinity yields (3.112).

Let us now prove that u satifies (3.81). Since $u\in H^1(\Omega)$, an integration by parts in (3.112) yields

$$\begin{aligned} \int_\Omega \nabla u(x)\cdot\nabla\varphi(x)dx + \int_{\partial\Omega}\nabla\varphi(x)\cdot\mathbf{n}(x)(v(x)-\overline{\gamma}(u)(x))d\gamma(x) \\ = \int_\Omega f(x)\varphi(x)dx + \int_{\partial\Omega} g(x)\varphi(x)d\gamma(x), \forall\varphi\in C^2(\overline{\Omega}), \end{aligned} \tag{3.115}$$

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 (3.81) (this will conclude the proof of Theorem 3.7), 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 (3.115) 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 (3.115), this yields

$$-\int_{\partial\Omega}\nabla\varphi(x)\cdot\mathbf{n}(x)(v(x)-\overline{\gamma}(u)(x))d\gamma(x)=0,\ \forall\varphi\in C^2(\overline{\Omega}) \text{ such that } \varphi=0 \text{ on } \partial\Omega. \tag{3.116}$$

There remains to show that the wide choice of φ in (3.116) 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 (3.116), 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 (3.80),(3.81).

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 (3.106) together with (3.81) 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 3.7. ■

Note that, with some discrete Sobolev inequality (similar to (3.69)), 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 3.7.

3.3 General elliptic operators

3.3.1 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 3.1 must be adapted. As an illustration, let us consider here the following problem, which was studied in Section 2.3 page 21 in the one-dimensional case:

$$-\text{div}(\Lambda\nabla u)(x) + \text{div}(\mathbf{v}u)(x) + bu(x) = f(x),\ x \in \Omega, \tag{3.117}$$

$$u(x) = g(x),\ x \in \partial\Omega, \tag{3.118}$$

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 3.4

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 $\underline{\lambda}\xi \cdot \xi \leq \Lambda(x)\xi \cdot \xi \leq \overline{\lambda}\xi \cdot \xi$ for any $x \in \Omega$ and any $\xi \in \mathbb{R}^d$.*

2. *$\mathbf{v} \in C^1(\overline{\Omega}, \mathbb{R}^d)$, div$\mathbf{v} \geq 0$ 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 3.1, under Assumption 3.4, there exists a unique variational solution $u \in H^1(\Omega)$ of Problem (3.117), (3.118). 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 (3.117)-(3.118).

Definition 3.7 (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 (3.117)-(3.118) is an admissible mesh $\mathcal{T}$ of Ω in the sense of Definition 3.1 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 $\sigma \in \mathcal{E}_{\mathrm{ext}}$ 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 $\sigma \in \mathcal{E}_{\mathrm{ext}}$, 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 3.1 page 37.

We shall now define the discrete unknowns of the numerical scheme, with the same notations as in Section 3.1.2. 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 $\sigma \in \mathcal{E}_{\mathrm{ext}}$, set $u_\sigma = g(y_\sigma)$. The finite volume scheme for the numerical approximation of the solution to Problem (3.117)-(3.118) is obtained by integrating Equation (3.117) over each control volume K, and approximating the fluxes over each edge σ of K. This yields

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + \mathrm{m}(K) b u_K = f_K, \ \forall K \in \mathcal{T}, \tag{3.119}$$

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 $\sigma \in \mathcal{E}_{\rm ext}$, 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 (2.26) 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:

  $$F_{K,\sigma} = -\tau_\sigma(u_L - u_K), \text{ if } \sigma \in \mathcal{E}_{\rm int},\, \sigma = K|L, \tag{3.120}$$

  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:

  $$F_{K,\sigma} = -\tau_\sigma(g_\sigma - u_K), \text{ if } \sigma \in \mathcal{E}_{\rm ext} \text{ and } x_K \notin \sigma, \tag{3.121}$$

  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 (3.119)) and the numerical flux $F_{K,\sigma}$ is an unknown which may be deduced from (3.119).

Remark 3.17 Note that if $\Lambda = Id$, then the scheme (3.119)-(3.121) is the same scheme than the one described in Section 3.1.2.

Error estimate

Theorem 3.8
Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3. Under Assumption 3.4, let u be the unique variational solution to Problem (3.117)-(3.118). Let $\mathcal{T}$ be an admissible mesh for the discretization of Problem (3.117)-(3.118), in the sense of Definition 3.7. Let ζ_1 and $\zeta_2 \in \mathbb{R}_+$ such that

$$\begin{array}{l}\zeta_1(\mathrm{size}(\mathcal{T}))^2 \le \mathrm{m}(K) \le \zeta_2(\mathrm{size}(\mathcal{T}))^2,\\ \zeta_1\mathrm{size}(\mathcal{T}) \le \mathrm{m}(\sigma) \le \zeta_2\mathrm{size}(\mathcal{T}),\\ \zeta_1\mathrm{size}(\mathcal{T}) \le d_\sigma \le \zeta_2\mathrm{size}(\mathcal{T}).\end{array}$$

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 (3.117)-(3.118)) 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 (3.119)-(3.121); 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

$$\sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma e)^2}{d_\sigma}\mathrm{m}(\sigma) \le C(\mathrm{size}(\mathcal{T}))^2 \tag{3.122}$$

and

$$\sum_{K\in\mathcal{T}} e_K^2\mathrm{m}(K) \le C(\mathrm{size}(\mathcal{T}))^2. \tag{3.123}$$

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 3.8

First, one may use Taylor expansions and the same technique as in the 1D case (see step 2 of the proof of Theorem 2.3, Section 2.3) to show that the expressions (3.120) and (3.121) 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 $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$,

$$\begin{array}{l}F^\star_{K,\sigma} - \int_\sigma -\Lambda(x)\nabla u(x)\cdot \mathbf{n}_{K,\sigma}d\gamma(x) = R_{K,\sigma},\\ \text{with } |R_{K,\sigma}| \le C_1\mathrm{size}(\mathcal{T})\mathrm{m}(\sigma).\end{array}$$

There also exists C_2 only depending on u and $\mathbf{v}$ such that, for all $\sigma \in \mathcal{E}$,

$$\begin{array}{l}v_{K,\sigma}u(x_{K_{\sigma,+}}) - \int_\sigma \mathbf{v}\cdot\mathbf{n}_{K,\sigma}u = r_{K,\sigma},\\ \text{with } |r_{K,\sigma}| \le C_2\mathrm{size}(\mathcal{T})\mathrm{m}(\sigma).\end{array}$$

Let us then integrate Equation (3.117) over each control volume, subtract to (3.119) and use the consistency of the fluxes to obtain the following equation on the error:

$$\left\{\begin{array}{l} -\displaystyle\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}e_{\sigma,+} + \mathrm{m}(K)be_K = \\ \displaystyle\sum_{\sigma\in\mathcal{E}_K}(R_{K,\sigma} + r_{K,\sigma}) + S_K,\ \forall K \in \mathcal{T},\end{array}\right.$$

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

$|S_K| \le \mathrm{m}(K) C_3 h$, where $C_3 \in \mathbb{R}_+$ only depends on u and b. Then, similarly to the proof of Theorem 3.3 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" (3.122). Then, following Herbin [84], one shows the following discrete Poincaré inequality:

$$\sum_{K\in\mathcal{T}} e_K^2 \mathrm{m}(K) \le C_4 \sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma e)^2}{d_\sigma} \mathrm{m}(\sigma), \tag{3.124}$$

where C_4 only depends on Ω, ζ_1 and ζ_2, which in turn yields the L^2 estimate (3.123). ■

Remark 3.18 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 3.1 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 (3.122) and (3.123) on ζ_1 and ζ_2, which arises when proving (3.124) as in Herbin [84].

3.3.2 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):

$$-\mathrm{div}(\lambda_i \nabla u)(x) = f(x),\ x \in \Omega_i,\ i = 1,2, \tag{3.125}$$

$$-\lambda_i \nabla u(x) \cdot \mathbf{n}(x) = \alpha(u(x) - \overline{u}),\ x \in \Gamma_1 \cup \Gamma_3, \tag{3.126}$$

$$\nabla u(x) \cdot \mathbf{n}(x) = 0,\ x \in \Gamma_2, \tag{3.127}$$

$$u(x) = g(x),\ x \in \Gamma_4, \tag{3.128}$$

$$(\lambda_2 \nabla u(x) \cdot \mathbf{n}_I(x))_{|2} = (\lambda_1 \nabla u(x) \cdot \mathbf{n}_I(x))_{|1} + \theta(x),\ x \in I, \tag{3.129}$$

$$u_{|2}(x) - u_{|1}(x) = \Phi(x),\ x \in I, \tag{3.130}$$

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 (3.125)-(3.130) in the sense of Definition 3.7. 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 (3.125) over each control volume K, and approximating the fluxes over each edge σ of K yields the following finite volume scheme:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = f_K,\ \forall K \in \mathcal{T}, \tag{3.131}$$

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 (3.130), 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 (3.131). The expression of $F_{K,\sigma}$ with respect to the unknowns u_K and u_σ is

$$F_{K,\sigma} = -\mathrm{m}(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}}, \forall K \in \mathcal{T}; K \subset \Omega_i \ (i = 1, 2), \forall \sigma \in \mathcal{E}_K; \tag{3.132}$$

which yields $2(N_{\mathcal{E}}^{0} + N_{\mathcal{E}}^{I}) + \sum_{i=1}^{4} N_{\mathcal{E}}^{i}$. (In (3.132), 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 (3.126) yields

$$F_{K,\sigma} = \alpha \mathrm{m}(\sigma)(u_\sigma - \overline{u}), \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K;\ \sigma \subset \Gamma_1 \cup \Gamma_3, \tag{3.133}$$

  that is $N_{\mathcal{E}}^{1} + N_{\mathcal{E}}^{3}$ equations.

- Neumann boundary conditions. Discretizing condition (3.127) yields

$$F_{K,\sigma} = 0, \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K;\ \sigma \subset \Gamma_2, \tag{3.134}$$

  that is $N_{\mathcal{E}}^{2}$ equations.

- Dirichlet boundary conditions. Discretizing condition (3.128) yields

$$u_\sigma = g(y_\sigma), \forall \sigma \in \mathcal{E}; \sigma \subset \Gamma_4, \tag{3.135}$$

  that is $N_{\mathcal{E}}^{4}$ equations.

- Conservativity of the flux. Except at interface I, the flux is continuous, and therefore

$$F_{K,\sigma} = -F_{L,\sigma}, \forall \sigma \in \mathcal{E}; \sigma \not\subset (\bigcup_{i=1}^{4} \Gamma_i \cup I) \text{ and } \sigma = K|L, \tag{3.136}$$

  that is $N_{\mathcal{E}}^{0}$ equations.

- Jump condition on the flux. At interface I, condition (3.129) is discretized into

$$F_{K,\sigma} + F_{L,\sigma} = \int_\sigma \theta(x)ds, \forall \sigma \in \mathcal{E}; \sigma \subset I \text{ and } \sigma = K|L;\ K \subset \Omega_2, \tag{3.137}$$

  that is $N_{\mathcal{E}}^{I}$ equations.

- Jump condition on the unknown. At interface I, condition (3.130) is discretized into

$$u_{\sigma,2} = u_{\sigma,1} + \Phi(y_\sigma), \forall \sigma \in \mathcal{E}; \sigma \subset I \text{ and } \sigma = K|L. \tag{3.138}$$

  that is another $N_{\mathcal{E}}^{I}$ equations.

Hence the total number of equations from (3.131) to (3.138) is N_{tot}, so that the numerical scheme can be expected to be well posed.

The finite volume scheme for the discretization of equations (3.125)-(3.130) is therefore completely defined by (3.131)-(3.138). Particular cases of this scheme are the schemes (3.20)-(3.23) page 42 (written for Dirichlet boundary conditions) and (3.84)-(3.85) page 64 (written for Neumann boundary conditions and no convection term) which were thoroughly studied in the two previous sections.

3.4 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 3.4.4.

3.4.1 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:

$$\left\{ \begin{array}{ll} -\Delta u(x) = f(x), & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega. \end{array} \right. \tag{3.139}$$

Let us show that the "piecewise linear" finite element method for the discretization of (3.139) 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

$$\sum_{K\in\mathcal{V}} \int_\Omega \varphi_K(x)dx = \mathrm{m}(\Omega) \tag{3.140}$$

and

$$\sum_{K\in\mathcal{V}} \nabla\varphi_K(x) = 0, \quad \text{for a.e.} x \in \Omega. \tag{3.141}$$

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 (3.140), 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.

3.4.2 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 3.1 page 37, a