Convergence of finite volume schemes for semilinear convection diffusion equations

Robert Eymard*, Thierry Gallouët†and Raphaèle Herbin‡

October 29, 1998

Abstract. The topic of this work is the discretization of semilinear elliptic problems in two space dimensions by the cell centered finite volume method. Dirichlet boundary conditions are considered here. A discrete Poincaré inequality is used, and estimates on the approximate solutions are proven. The convergence of the scheme without any assumption on the regularity of the exact solution is proven using some compactness results which are shown to hold for the approximate solutions.

1 Introduction

The aim of this work is to study the discretization by the finite volume method of convection diffusion problems on general structured or non structured grids; these grids may consist of polygonal control volumes satisfying adequate geometrical conditions (which are stated in the sequel) and not necessarily ordered in a cartesian grid.

Finite volume methods have been used for over twenty years in several engineering fields (see e.g. [26], but its mathematical analysis has only recently been undertaken. We are interested here by the so called "cell centered" approach, i.e. the discrete unknowns are located at some point in the control volumes. For the "finite volume element" and "control volume finite element" approaches where the unknowns are located at the vertices, see [5], [4], [15], [16] and [10], see also [25] for other types of finite volume methods.

Error estimates and convergence results for cell centered finite volume schemes for the discretization of linear elliptic equations have recently been studied. The one dimensional case and multi-dimensional rectangular cases were studied in [23], [17] using a finite difference technique.

Error estimates for unstructured meshes were also obtained in [2] in the case where the control volumes are constructed using a dual mesh of a triangular finite element mesh. Error estimates for the finite volume method applied to diffusion operators may also be obtained for non structured triangular meshes via finite element technique methods by writing the finite volume scheme as a mixed finite element scheme with numerical integration [3], [1], [27].

A first order estimate for triangular meshes was obtained in [18] for a convection diffusion where the diffusion operator is the Laplacian under C^2 regularity assumptions of the solution. It generalizes easily to the case of Voronoï meshes, see [11], to the case of a diffusion operator involving discontinuous tensor diffusion coefficients and the time dependent case [19]. Error estimates assuming H^2 regularity of the solution may also be obtained for linear convection diffusion equations for Dirichlet boundary conditions [21], [22], [24], [11] and Neumann or Fourier boundary conditions [11]. Note also that the finite volume scheme is well adapted to the discretization of hyperbolic systems (see e.g. [11] and references therein) and is therefore a good candidate for the discretization of systems of equations of different types, see e.g. [20], [28], [29].

*Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée; eymard@enpc.fr
†Université de Provence, Marseille; gallouet@gyptis.univ-mrs.fr
‡Université de Provence, Marseille; herbin@armstrong.univ-mrs.fr

We consider here a semilinear convection diffusion equation with non homogeneous Dirichlet boundary conditions, for which no regularity of the solution is known. The case of a semilinear time dependent convection diffusion equation is addressed in [13] where the convergence of a coupled finite element-finite volume scheme is proven (see also [11] for error estimates in the case of time-dependent linear equations and convergence results in the case of nonlinear equations). Here we consider a pure finite volume scheme and prove the existence and convergence of the approximate solutions in the stationary case. The proof uses the property of consistency of the numerical fluxes on regular test functions. This, together with a compactness result on the set of approximate solutions allows to prove a subsequence of approximate solutions (for which an *a priori* estimate is obtained) is a weak solution of the semilinear equation.

Let us consider the following semilinear elliptic equation:

$$-\Delta u(x) + \mathrm{div}(\mathbf{v} q(u))(x) = f(x, u(x)), \quad x \in \Omega, \tag{1}$$

with Dirichlet boundary condition:

$$u(x) = g(x), \quad x \in \partial\Omega, \tag{2}$$

where

Assumption 1

(i) Ω *is an open bounded polygonal subset of* $\mathbb{R}^2$,

(ii) $f : \Omega \to \mathbb{R}$ *is such that*

$$\begin{array}{c} f(x,s) \text{ is measurable with respect to } x \in \Omega \text{ for all } s \in \mathbb{R} \\ \text{and continuous with respect to } s \in \mathbb{R} \text{ for a.e. } x \in \Omega, \end{array} \tag{3}$$

$$\begin{array}{l} \text{There exist } \alpha < \dfrac{1}{\mathrm{diam}(\Omega)^2} \text{ and } \beta \in \mathbb{R} \text{ such that } f(x,s)s \le \alpha s^2 + \beta \\ \text{and } |f(x,s)| \le \beta |s|, \text{ for all } s \in \mathbb{R}, \text{ for a.e. } x \in \Omega. \end{array} \tag{4}$$

(iii) $\mathbf{v} \in C^1(\overline{\Omega}, \mathbb{R}^d), \mathrm{div}\mathbf{v} = 0$, *and let* $V = \max_{x \in \overline{\Omega}} |\mathbf{v}(x)|$.

(iv) $q \in C^1(\mathbb{R}, \mathbb{R})$ *is such that* $q' \geq 0$ *and there exists* $c_q \in \mathbb{R}_+$ *such that* $|q(s)| \leq c_q |s|$.

(v) $g \in H^{1/2}(\partial\Omega, \mathbb{R})$ *is such that there exists* $\tilde{g} \in H^1(\Omega)$ *such that* $\overline{\gamma}(\tilde{g}) = g$ *a.e. on* $\partial\Omega$.

Remark 1

(i) The bound $\alpha < \frac{1}{\mathrm{diam}(\Omega)^2}$ *in assumption (ii) on* f *may be somewhat relaxed if more restrictive assumptions on the mesh than the ones given below hold for the meshes. For instance using discrete Sobolev inequalities, as in [11], it is sufficient to assume* $\alpha < \frac{1}{\mathrm{m}(\Omega)}$.

(ii) The Laplace operator is considered here for the sake of simplicity, but more general elliptic operators are possible to handle, for instance operators of the form $-div(a(u)\nabla u)$ *with adequate assumptions on* u.

(iii) Since $\mathrm{div}\mathbf{v} = 0$, *we can assume, without loss of generality, that* $q(0) = 0$. *This assumption will simplify the presentation of some proofs.*

Here, and in the sequel, $\overline{\gamma}$ 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 one-dimensional Lebesgue measure on $\partial\Omega$.

Let us introduce the weak formulation of problem (1),(2). A weak solution of (1),(2) under Assumption 1 is a function $u = \tilde{u} + \tilde{g} \in H^1(\Omega)$ satisfying

$$
\begin{cases}
u = \tilde{u} + \tilde{g} \text{ where } \tilde{u} \in H_0^1(\Omega) \text{ and} \\
\displaystyle\int_\Omega (\nabla u(x) \nabla \varphi(x) + \operatorname{div}(\mathbf{v}(x) q(u(x))) \varphi(x) dx = \\
\displaystyle\int_\Omega f(x, u(x)) \varphi(x) dx, \ \forall \varphi \in H_0^1(\Omega).
\end{cases}
\tag{5}
$$

Using Schauder's fixed point theorem (see e.g. [8]) or the convergence theorem 2 which is proved below, it is possible to prove that there exists at least one solution to (5).

2 The finite volume schemes

The finite volume scheme is found by integrating equation (1) on a given control volume of a discretization mesh and finding an approximation of the fluxes on the control volume boundary in terms of the discrete unknowns. Let us first give the assumptions which are needed on the mesh.

2.1 Meshes

Assume K and L to be two neighbouring control volumes of the mesh. A consistent discretization of the normal flux $-\nabla u \cdot \mathbf{n}$ 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.

Definition 1 (Admissible meshes) *Let Ω be an open bounded polygonal subset of $\mathbb{R}^2$. 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 (with positive measure), a family of subsets of $\overline{\Omega}$ contained in hyperplanes of $\mathbb{R}^2$, denoted by $\mathcal{E}$ (these are the edges of the control volumes), with strictly positive one-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 $\partial K = \overline{K} \setminus K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$. Let $\mathcal{E} = \cup_{K \in \mathcal{T}} \mathcal{E}_K$.*
- *(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 $K|L$.*

In the sequel, the following notations are used. The mesh size is defined by: $\operatorname{size}(\mathcal{T}) = \sup\{\operatorname{diam}(K), K \in \mathcal{T}\}$. *For any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$, $\mathrm{m}(K)$ is the area of K and $\mathrm{m}(\sigma)$ the length 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 $\mathcal{N}(K) = \{L \in \mathcal{T};\ \exists \sigma \in \mathcal{E}_K,\ \overline{\sigma} = \overline{K} \cap \overline{L}\}$. For any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$ we denote by $d_{K,\sigma}$ the Euclidean distance between x_K and σ. For any $\sigma in \mathcal{E}$, we define $d_\sigma = d_{K,\sigma} + d_{L,\sigma}$ if $\sigma = K|L \in \mathcal{E}_{\text{int}}$ (in which case d_σ is the Euclidean distance between x_K and x_L) and $d_\sigma = d_{K,\sigma}$ if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$.*
For any $\sigma \in \mathcal{E}$; the "transmissibility" through σ is defined by $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$ if $d_\sigma \neq 0$ and $\tau_\sigma = 0$ if $d_\sigma = 0$. 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.

Example 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 use of such a grid for the finite volume scheme using differential quotients for the approximation of the normal flux yields a 4-point scheme [18]. 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.

Example 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 $\mathcal{P} = \{(k_1 h, \ldots, k_2 h),\ (k_1, k_2) \in \mathbb{Z}^2\} \cap \Omega$, for a given $h > 0$. The control volumes of the Voronoï mesh are defined with respect to each point x of $\mathcal{P}$ by

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

Recall that $|x - y|$ denotes the euclidean distance between x and y.
Voronoï meshes are admissible in the sense of Definition 1 if the assumption "on the boundary", namely part (v) of Definition 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 1 may be replaced by the weaker assumption "$d(y_\sigma, \sigma) \leq \text{size}(\mathcal{T})$ for any $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$" which is much easier to satisfy. Note also that a slight modification of the treatment of the boundary conditions in the finite volume scheme (8)-(12) allows us to obtain convergence results (as in Theorem 2) for all Voronoï meshes. This modification consists in replacing, for $K \in \mathcal{T}$ such that $\mathcal{E}_K \cap \mathcal{E}_{\text{ext}} \neq \emptyset$, the equation (8), associated with 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 1 (then, the scheme (8)-(12) page 5 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.

Note that cell centered finite volume schemes may also be defined on meshes which are not admissible in the sense of the above definition [12], [7]. In this case, however, some more technical assumptions are needed on the mesh to show the convergence of the scheme.

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

Definition 2 *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, $d = 2$ or 3, and $\mathcal{T}$ an admissible mesh. Define $X(\mathcal{T})$ as the set of functions from Ω to $\mathbb{R}$ which are constant over each control volume of the mesh.*

Definition 3 (Discrete norms) *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 1.*
For $u \in X(\mathcal{T})$, define the discrete H_0^1 norm by

$$\|u\|_{1,\mathcal{T}} = \Big(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma u)^2\Big)^{\frac{1}{2}} \tag{6}$$

where, for any $\sigma \in \mathcal{T}$, $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$ *and*
$D_\sigma u = |u_K - u_L|$ *if* $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$,
$D_\sigma u = |u_K|$ *if* $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$,
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 1.*

2.2 The schemes

Let $\mathcal{T}$ be an admissible mesh. Let us now define a finite volume scheme to discretize (1)-(2).
Let $(u_K)_{K\in\mathcal{T}}$ denote the discrete unknowns and let

$$f_K(u_K) = \frac{1}{\mathrm{m}(K)} \int_K f(x, u_K)dx, \forall K \in \mathcal{T}. \tag{7}$$

In order to describe the scheme in the most general way, one introduces 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 on an edge 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.

We may now write the finite volume scheme for the discretization of Problem (1)-(2) under assumptions 1 as the following set of equations:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(u_{\sigma,+}) = \mathrm{m}(K) f_K(u_K),\ \forall K \in \mathcal{T}, \tag{8}$$

where $u_{\sigma,+}$ is defined by

$$\begin{array}{l} \text{if } \sigma = K|L, \text{ then } u_{\sigma,+} = u_K \text{ if } v_{K,\sigma} \geq 0, \text{ and } u_{\sigma,+} = u_L \text{ otherwise;} \\ \text{if } \sigma \subset K \cap \partial\Omega, \text{ then } u_{\sigma,+} = u_K \text{ if } v_{K,\sigma} \geq 0 \text{ and } u_{\sigma,+} = u_\sigma \text{ otherwise,} \end{array} \tag{9}$$

and $F_{K,\sigma}$ is defined by

$$F_{K,\sigma} = -F_{L,\sigma}, \forall \sigma \in \mathcal{E}_{\mathrm{int}}, \text{ if } \sigma = K|L, \tag{10}$$

$$F_{K,\sigma} d_{K,\sigma} = -\mathrm{m}(\sigma)(u_\sigma - u_K), \forall \sigma \in \mathcal{E}_K, \forall K \in \mathcal{T}, \tag{11}$$

and

$$u_\sigma = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma g(y) d\gamma(y),\ \forall \sigma \in \mathcal{E}_{\mathrm{ext}}. \tag{12}$$

Note that the values u_σ for $\sigma \in \mathcal{E}_{\mathrm{int}}$ are auxiliary values which may be eliminated so that (8)-(12) leads to a nonlinear system of N equations with N unknowns, namely the $(u_K)_{K\in\mathcal{T}}$, with $N = \mathrm{card}(\mathcal{T})$. This nonlinear system can be written, using some ordering of the unknowns and equations, as

$$AU + B(U) = C(U) + D(g), \tag{13}$$

where:
$U \in \mathbb{R}^N$ is the vector of discrete unknowns (that is the u_K, $K \in \mathcal{T}$), N being the number of cells of the mesh $\mathcal{T}$,
A is a linear application from $\mathbb{R}^N$ to $\mathbb{R}^N$ and AU corresponds to the discretization of $-\Delta u(x)$,
B is a continuous application from $\mathbb{R}^N$ to $\mathbb{R}^N$ and $B(U)$ corresponds to the discretization of $\mathrm{div}(\mathbf{v}q(u))(x)$,

C is a continuous application from $\mathbb{R}^N$ to $\mathbb{R}^N$ and $C(U)$ corresponds to the discretization of $f(x, u(x))$ and $D(g)$ is a vector of $\mathbb{R}^N$ which contains all the terms depending on g (note that D is and application from $L^1(\partial\Omega)$ into $\mathbb{R}^N$.

3 Discrete Poincaré inequalities and trace inequalities

We give in this section some inequalities for piecewise constant functions.

Let us first prove discrete Poincaré inequality for the discrete H_0^1 norm of a piecewise constant function. Note that a "discrete mean value Poincaré inequality" may also be proven in order to deal with Neumann boundary conditions (see [11])

Lemma 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 1 and $u \in X(\mathcal{T})$ (see Definition 2), then*

$$||u||_{L^2(\Omega)} \leq \mathrm{diam}(\Omega)||u||_{1,\mathcal{T}}, \tag{14}$$

where $||\cdot||_{1,\mathcal{T}}$ is the discrete H_0^1 norm defined in Definition 3.

Proof of Lemma 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 $\mathbf{d}$ 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 $\mathbf{d}$. 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

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

where the notations $D_\sigma u$ and u_K are defined in Definition 3. Let $c_\sigma = |\mathbf{d}\cdot\mathbf{n}_\sigma|$ (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:

$$|u_K|^2 \leq \sum_{\sigma\in\mathcal{E}} \frac{(D_\sigma u)^2}{d_\sigma c_\sigma}\chi_\sigma(x,y(x)) \sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x,y(x)), \text{ for a.e. } x \in K. \tag{15}$$

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

$$\sum_{\sigma\in\mathcal{E}} d_\sigma c_\sigma \chi_\sigma(x, y(x)) \leq \mathrm{diam}(\Omega). \tag{16}$$

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 (16) for such points x leads to (16) a.e. on Ω, since $\mathbf{d}$ is fixed). There exists $\sigma \in \mathcal{E}_{\mathrm{ext}}$ such that $y(x) \in \sigma$. Then,

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

Since x_K, $y_\sigma \in \Omega$, this gives (16).

Integrating (15) over Ω and using (16) gives

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

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

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

Hence the result. ■

The following result will also be useful for getting estimates on the approximate solutions in the case of non homogeneous Dirichlet boundary conditions.

Lemma 2 *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 $H^1_0(\Omega)$ to $H^{1/2}(\partial\Omega)$). Let $\mathcal{T}$ be an admissible mesh (in the sense of Definition 1) and let :*

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

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

and

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

Then there exists $C\in\mathbb{R}_+$, only depending on $\zeta = \min\{\dfrac{d_{K,\sigma}}{\mathrm{diam}(K)}, K\in\mathcal{T}, \sigma\in\mathcal{E}_K\}$ and $M = \max\{\mathrm{card}(\mathcal{E}_K), K\in\mathcal{T}\}$, such that

$$|D\tilde{g}|_{\mathcal{T}} \le C\|\tilde{g}\|_{H^1(\Omega)}. \tag{17}$$

Proof of Lemma 2

Lemma 2 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 2. 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

$$(\tilde{g}_K-\tilde{g}_\sigma)^2 \le C_1\frac{\mathrm{diam}(K)}{\mathrm{m}(\sigma)}\int_K |\nabla\tilde{g}(x)|^2dx, \forall K\in\mathcal{T}, \forall\sigma\in\mathcal{E}_K, \tag{18}$$

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,

$$(\tilde{g}_K-\tilde{g}_\sigma)^2 \le \frac{2}{\mathrm{m}(K)\mathrm{m}(\sigma)}(A+B), \tag{19}$$

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

$$A \le \mathrm{diam}(K)^3 \int_K \big(D_2\tilde{g}(x)\big)^2 dx. \tag{20}$$

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

$$B \le 2\mathrm{diam}(K)(B_1 + \frac{1}{\zeta^2} B_2), \tag{21}$$

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, using 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

$$B_i \le \mathrm{diam}(K)^2 \int_K \big(D_i\tilde{g}(x)\big)^2 dx. \tag{22}$$

Using the fact that $\mathrm{m}(K) \ge \pi\zeta^2\big(\mathrm{diam}(K)\big)^2$, (19), (20), (21) and (22), one concludes (18).

In order to conclude the proof of (17), one remarks that

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

Since $d_\sigma \geq \zeta \mathrm{diam}(K)$ for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, one gets, using (18), that

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

The above inequality shows that

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

which implies (17). ∎

4 Existence and estimates for the approximate solution

Let us first prove the existence of the approximate solution and an estimate on this solution. This estimate will be obtained by using the discrete inequalities which were proved in the previous sections, and will yield convergence thanks to a compactness theorem given in the appendix.

Lemma 3 (Existence and estimate) *Under Assumptions 1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 1, and let:*

$$\zeta = \min\Big(\min_{K \in \mathcal{T}} \min_{\sigma \in K} \frac{d_{K,\sigma}}{\mathrm{diam}}(K), \min_{K \in \mathcal{T}} \min_{\sigma \in K} \frac{d_{K,\sigma}}{d_\sigma} \Big), \tag{23}$$

then there exists a solution $(u_K)_{K \in \mathcal{T}}$ to the system of equations (8)-(12).
Furthermore, let $u_{\mathcal{T}} \in X(\mathcal{T})$ (see Definition 2) be defined by $u_{\mathcal{T}}(x) = u_K$ for a.e. $x \in K$, and for any $K \in \mathcal{T}$; there exists $C \in \mathbb{R}$, only depending on Ω, $\|\tilde{g}\|_{H^1(\Omega)}$, ζ, $M = \max_{K \in \mathcal{T}} \mathrm{card}(\mathcal{E}_K)$, f and q, such that

$$\|\tilde{u}_{\mathcal{T}}\|_{1,\mathcal{T}} \leq C \text{ and } \|\tilde{u}_{\mathcal{T}}\|_{L^2(\Omega)} \leq C, \tag{24}$$

where

$$\tilde{u}_{\mathcal{T}}(x) = \tilde{u}_K = u_K - \frac{1}{\mathrm{m}(K)} \int_K \tilde{g}(y) dy \text{ for all } x \in K \text{ and all } K \in \mathcal{T}. \tag{25}$$

Remark 2 *In the case of homogeneous Dirichlet boundary conditions, the additional assumption (23) is not required. This technical assumption is essentially needed for the proof of Lemma 2 whereas the proof of the estimate (24) for homogeneous Dirichlet boundary conditions only requires the use of the discrete Poincaré inequality (1) (see [11]). Similarly, if there is no convection (i.e. $q = 0$ or $\mathbf{v} = 0$), then (23) is not needed. This latter assumption is used to obtain a bound on the convection terms in the estimate of the approximate solution.*

Proof of Lemma 3

Equations (8)-(12) lead, after an easy elimination of the auxiliary unknowns, to a nonlinear system of N equations with N unknowns, namely the $(u_K)_{K \in \mathcal{T}}$, with $N = \mathrm{card}(\mathcal{T})$.
We shall first prove the existence and uniqueness of the solution to the linearized system which is obtained from the numerical scheme. We shall then prove an estimate on any possible function $u_{\mathcal{T}}$ of (8)-(12). These two steps will allow to prove the existence of the solution to the numerical scheme by a topological degree argument.
We assume (without loss of generality) that $q(0) = 0$.

Step 1 (existence and uniqueness of the solution to the linear system)
Let $(r_K)_{K \in \mathcal{T}}$ be a given vector of $\mathbb{R}^N$ (with $N = \mathrm{card}(\mathcal{T})$). Let us introduce the linear systems of equations with unknowns $(u_K)_{K \in \mathcal{T}}$ consisting of the following equation:

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = r_K \;\; \forall K \in \mathcal{T}, \tag{26}$$

where $F_{K,\sigma}$, for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$ is defined with respect to the unknowns $(u_K)_{K\in\mathcal{T}}$ by (10) and (11), and equation (12).
First assume that $(u_K)_{K\in\mathcal{T}}$ satisfies the linear system (26), (10), (11), (12) with $r_K = 0$ for all $K \in \mathcal{T}$, and $\int_\sigma g(y)d\gamma(y) = 0$ for any $\sigma \in \mathcal{E}_{\text{ext}}$. Let us prove that in this case $(u_K) = 0$ for all $K \in \mathcal{T}$. This yields the uniqueness (and thus the existence) of the solution to the linear system (26), (10), (11), (12).
Multiplying (26) by u_K, summing over K, and using (10) and (11) leads to

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}u_K = 0. \tag{27}$$

Let us now perform a "discrete integration by parts", that is a reordering of the summations over the edges of the mesh. We obtain:

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_\sigma u)^2 = 0, \tag{28}$$

where $|D_\sigma u| = |u_K - u_L|$, if $\sigma = K|L$ and $|D_\sigma u| = |u_K|$, if $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}$.
Hence,

$$\|u_\mathcal{T}\|^2_{1,\mathcal{T}} = 0, \tag{29}$$

which yields $u_K = 0$ for all $K \in \mathcal{T}$.
This proves the existence and the uniqueness of the solution to the linear system (26), (10), (11), (9), (12) for any $(w_K)_{K\in\mathcal{T}} \in \mathbb{R}^N$, and for any $\{\int_\sigma g(y)d\gamma(y),\ \sigma \in \mathcal{E}_{\text{ext}}\}$.

Step 2 (Existence of a solution) Using the formulation (13) of the numerical scheme, let us prove the existence of U solution to (13). From step 1, A is invertible and (13) is therefore equivalent to:

$$U = -A^{-1}B(U) + A^{-1}C(U) + A^{-1}D(g). \tag{30}$$

In order to show that (30) admits at least one solution in $\mathbb{R}^N$, and therefore that (8)-(12) admits at least one solution, we are going to use a topological degree argument (see also Remark 3).
For $t \in [0,1]$ and $U \in \mathbb{R}^N$, let $F(t,U) = -A^{-1}B(tU) + tA^{-1}C(U) + A^{-1}D(tg)$, so that F is continuous from $[0,1] \times \mathbb{R}^N$ in $\mathbb{R}^N$.

Let us endow the space $\mathbb{R}^N$ with some norm; let us choose for instance the norm defined by $|U|^2 = \sum_{K\in\mathcal{T}} \mathrm{m}(K)u_K^2$, where the u_K, $K \in \mathcal{T}$, are the components of U. We shall show in step 3 below that

$$\exists R > 0 \text{ such that if } (t,U) \in [0,1]\times\mathbb{R}^N \text{ and } U = F(t,U) \text{ then } |U| \neq R. \tag{31}$$

Assuming that (31) is satisfied, it is possible to define for $t \in [0,1]$, the (Brouwer) topological degree of the application $Id - F(t,.)$ with respect to $B_R = \{U \in \mathbb{R}^N, |U| < R\}$ and 0, which is denoted by $d(Id - F(t,.), B_R, 0)$ (see e.g. [8] for the definition of the topological degree and its properties). Then, thanks to the homotopy invariance of the degree and since $F(0,U) = 0$ (for all $U \in \mathbb{R}^N$), one has

$$d(Id - F(1,.), B_R, 0) = d(Id, B_R, 0).$$

Since $d(Id, B_R, 0) = 1$, this leads to $d(Id - F(1,.), B_R, 0) \neq 0$ which proves the existence of $U \in B_R$ such that $U - F(1,U) = 0$ i.e. that U is a solution of (30). This proves the existence of a solution to (8)-(12).

Step 3 (Proof of (31) and estimate on a possible solution) To conclude the proof of the lemma, it remains to prove (31) and an estimate on the solutions.
Let $U \in \mathbb{R}^N$ and $t \in [0,1]$ such that $U = F(t,U)$ (where F is defined in step 2). Let $(u_K)_{K\in\mathcal{T}}$ be the components of U, one has

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+}) = \mathrm{m}(K) t f_K(u_K),\ \forall K\in\mathcal{T}, \tag{32}$$

$$F_{K,\sigma} = -F_{L,\sigma},\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if } \sigma = K|L, \tag{33}$$

$$F_{K,\sigma} d_{K,\sigma} = -\mathrm{m}(\sigma)(u_\sigma - u_K),\ \forall\sigma\in\mathcal{E}_K,\ \forall K\in\mathcal{T}, \tag{34}$$

$$\begin{array}{l}\text{If } \sigma = K|L,\ \text{then } u_{\sigma,+} = u_K \text{ if } v_{K,\sigma}\geq 0,\ \text{and } u_{\sigma,+} = u_L \text{ otherwise;}\\ \text{if } \sigma\subset K\cap\partial\Omega,\ \text{then } u_{\sigma,+} = u_K \text{ if } v_{K,\sigma}\geq 0 \text{ and } u_{\sigma,+} = u_\sigma \text{ otherwise.}\end{array} \tag{35}$$

$$u_\sigma = \frac{t}{\mathrm{m}(\sigma)} \int_\sigma g(y) d\gamma(y),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}}. \tag{36}$$

Let $\tilde{g}\in H^1(\Omega)$ be such that the trace of $\tilde{g}$ on $\partial\Omega$ is equal to g. One defines $\tilde{u}_{\mathcal{T}}^{(t)}\in X(\mathcal{T})$ by:

$$\tilde{u}_{\mathcal{T}}^{(t)}(x) = \tilde{u}_K^{(t)} = u_K - \frac{t}{\mathrm{m}(K)}\int_K \tilde{g}(y) dy \text{ for all } x\in K \text{ and all } K\in\mathcal{T}. \tag{37}$$

Then $(\tilde{u}_K^{(t)})_{K\in\mathcal{T}}$ satisfies

$$\sum_{\sigma\in\mathcal{E}_K} \tilde{F}_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+}) = \mathrm{m}(K) t f_K(u_K) - \sum_{\sigma\in\mathcal{E}_K} tG_{K,\sigma},\ \forall K\in\mathcal{T}, \tag{38}$$

$$\tilde{F}_{K,\sigma} = -\tau_{K|L}(\tilde{u}_L^{(t)} - \tilde{u}_K^{(t)}),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if } \sigma = K|L, \tag{39}$$

$$\tilde{F}_{K,\sigma} = \tau_\sigma(\tilde{u}_K^{(t)}),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}} \text{ such that } \sigma\in\mathcal{E}_K. \tag{40}$$

$$G_{K,\sigma} = -\tau_{K|L}\Big(\frac{1}{\mathrm{m}(L)}\int_L \tilde{g}(y)dy - \frac{1}{\mathrm{m}(K)}\int_K \tilde{g}(y)dy\Big),\ \forall\sigma\in\mathcal{E}_{\mathrm{int}},\ \text{if } \sigma = K|L, \tag{41}$$

$$G_{K,\sigma} = -\tau_\sigma\Big(\frac{1}{\mathrm{m}(\sigma)}\int_\sigma g(x)d\gamma(x) - \frac{1}{\mathrm{m}(K)}\int_K \tilde{g}(y)dy\Big),\ \forall\sigma\in\mathcal{E}_{\mathrm{ext}} \text{ such that } \sigma\in\mathcal{E}_K. \tag{42}$$

Multiplying (38) by $\tilde{u}_K^{(t)}$, summing over $K\in\mathcal{T}$, gathering by edges in the left hand side, using item (ii) Assumption 1 and remarking that for any $\varepsilon > 0$,

$$\|u_{\mathcal{T}}\|^2_{L^2(\Omega)} \leq (\|\tilde{u}_{\mathcal{T}}^{(t)}\|_{L^2(\Omega)} + t\|\tilde{g}\|_{L^2(\Omega)})^2 \leq (1+\varepsilon)\|\tilde{u}_{\mathcal{T}}^{(t)}\|^2_{L^2(\Omega)} + t^2(1+\frac{4}{\varepsilon})\|\tilde{g}\|^2_{L^2(\Omega)}$$

yields that, for any $\varepsilon > 0$, there exists $C_\varepsilon \geq 0$ depending only on ε such that

$$\begin{array}{l}\|\tilde{u}_{\mathcal{T}}^{(t)}\|^2_{1,\mathcal{T}} + \displaystyle\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+})\tilde{u}_K^{(t)} \leq \\ \alpha(1+\varepsilon)\|\tilde{u}_{\mathcal{T}}^{(t)}\|^2_{L^2(\Omega)} + C_\varepsilon t^2\|\tilde{g}\|^2_{L^2(\Omega)} + \beta\mathrm{m}(\Omega) + \beta\|\tilde{u}_{\mathcal{T}}^{(t)}\|_{L^2(\Omega)}\|\tilde{g}\|_{L^2(\Omega)} + \beta\|\tilde{g}\|^2_{L^2(\Omega)} + t|D\tilde{g}|_{\mathcal{T}}\|\tilde{u}_{\mathcal{T}}\|_{1,\mathcal{T}},\end{array} \tag{43}$$

Let us give an estimate on $\displaystyle\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+})\tilde{u}_K^{(t)}$. Reordering the summation, we may write:

$$\begin{array}{l}\displaystyle\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+})\tilde{u}_K^{(t)} = \sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(\tilde{u}_{\sigma,+}^{(t)} - \tilde{u}_{\sigma,-}^{(t)}) \\ \displaystyle = \sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(u_{\sigma,+} - u_{\sigma,-}) + \sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(t\tilde{g}_{\sigma,+} - t\tilde{g}_{\sigma,-})\end{array} \tag{44}$$

where the indexes $\sigma,+$ (resp $\sigma,-$) denote the upstream (resp. downstream) choice of $\tilde{u}$ or $\tilde{g}$, in the same way as for u (see (9)). Let G_t be a primitive of $q(t\cdot)$. Then:

$$\sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(u_{\sigma,+}-u_{\sigma,-}) = \sum_{\sigma\in\mathcal{E}} v_\sigma (G_t(u_{\sigma,+})-G_t(u_{\sigma,-})) - \int_{u_{\sigma,-}}^{u_{\sigma,+}} v_\sigma(q(ts)-q(tu_{\sigma,+}))ds.$$

Reordering the summation and using the fact that q is non decreasing yields that

$$\sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(u_{\sigma,+}-u_{\sigma,-}) \geq \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} G_t(u_K). \tag{45}$$

Hence, from (44), (45) and the fact that $\mathrm{div}\mathbf{v}=0$, one has:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+})\tilde{u}_K \geq \sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(t\tilde{g}_{\sigma,+}-t\tilde{g}_{\sigma,-}) \tag{46}$$

Since q satisfies the conditions (iv) in Assumption 1, using the Cauchy Schwarz inequality and the fact that the mesh satisfies condition (23), one has:

$$\left\{\begin{array}{l} |\displaystyle\sum_{\sigma\in\mathcal{E}} v_\sigma q(tu_{\sigma,+})(t\tilde{g}_{\sigma,+}-t\tilde{g}_{\sigma,-})| \leq \\ \displaystyle\sum_{\sigma\in\mathcal{E}} V c_q \mathrm{m}(\sigma)|u_{\sigma,+}(t\tilde{g}_{\sigma,+}-t\tilde{g}_{\sigma,-})| \leq \\ 2t(\dfrac{V}{\zeta})^{1/2}\|u_\mathcal{T}\|_{L^2(\Omega)}|D\tilde{g}|_\mathcal{T} \end{array}\right. \tag{47}$$

From (46) (choosing ε small enough) and (47), one has:

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(tu_{\sigma,+})\tilde{u}_K \leq Ct\|u_\mathcal{T}\|_{L^2(\Omega)}|D\tilde{g}|_\mathcal{T} \tag{48}$$

where C depends only on ζ,Ω, $\mathbf{v}$ and f. Therefore, from (43) (choosing ε small enough and (48), using Lemma 2 and the discrete Poincaré inequality (14), there exists $C_1\in\mathbb{R}$, only depending on Ω, $\|\tilde{g}\|_{H^1(\Omega)}$, ζ, $\mathbf{v}$ and f, such that $\|\tilde{u}_\mathcal{T}\|_{1,\mathcal{T}}\leq C_1$ and $\|\tilde{u}_\mathcal{T}\|_{L^2(\Omega)}\leq C_1$.
This gives (31) with $R > C_1+\|\tilde{g}\|_{L^2(\Omega)}$ and the estimates (24) on the solutions of (8)-(12). This concludes the proof of the lemma.

■

Remark 3 *Note that the topological degree argument which was used in the above proof may also be used to show the existence of at least a solution to problem (1)-(2). In fact, the existence of a solution to problem (1)-(2) is also an immediate consequence of the convergence theorem 2 given hereafter.*

We now state a discrete maximum property of the scheme. Even though this property is not used in the sequel, it is a very important feature when dealing with problems where positivity of the solution must be ensured. This stability property is valid for the scheme (8)-(12) for any admissible mesh thanks to the upwind approximation of the convection term; note that a centered approximation of the convection term would still yield a convergent scheme, but that the scheme might become unstable for (that is the maximum principle will not hold) if the convection term is large with respect to the size of the mesh.

Proposition 1 *Under Assumption 1, let $\mathcal{T}$ be an admissible mesh in the sense of Definition 1. If $f\geq 0$ for all $K\in\mathcal{T}$, and if Dirichlet boundary conditions (2) hold with $g\geq 0$, then the solution $(u_K)_{K\in\mathcal{T}}$ of (8)-(12) satisfies $u_K\geq 0$ for all $K\in\mathcal{T}$.*

Proof A proof of this result is given in [18]. (see also [11]) in the linear case. It uses the strong formulation. We use here the weak formulation. For any $K \in \mathcal{T}$, denote by u_K^- the negative part of u_K, that is $u_K^- = \frac{1}{2}(|u_K| - u_K)$ (note that contrary to its name, the negative part of u_K is positive...) Multiplying (8) by u_K^-, summing over $K \in \mathcal{T}$ and using the fact that $f \geq 0$ and $g \geq 0$ yields that:

$$\sum_{\sigma \in \mathcal{E}} \tau_\sigma (u_{\sigma,+} - u_{\sigma,-})(u_{\sigma,+}^- - u_{\sigma,-}^-) + v_\sigma q(u_{\sigma,+})(u_{\sigma,+}^- - u_{\sigma,-}^-) \geq 0$$

Noting that $|u_{\sigma,+}^- - u_{\sigma,-}^-| \leq |u_{\sigma,+} - u_{\sigma,-}|$ and that $(u_{\sigma,+} - u_{\sigma,-})(u_{\sigma,+}^- - u_{\sigma,-}^-) \leq 0$ yields that

$$-\sum_{\sigma \in \mathcal{E}} \tau_\sigma (u_{\sigma,+}^- - u_{\sigma,-}^-)^2 + v_\sigma q(u_{\sigma,+})(u_{\sigma,+}^- - u_{\sigma,-}^-) \geq 0. \tag{49}$$

Let G be a primitive of q such that $G(0) = 0$, calculations similar to those of the proof of Lemma 3 (see (45), (46)) yield that

$$-\sum_{\sigma \in \mathcal{E}} v_\sigma q(u_{\sigma,+})(u_{\sigma,+}^- - u_{\sigma,-}^-) = \sum_{K \in \mathcal{T}} G(-u_K^-) \int_{\partial K} \mathbf{v}(x).\mathbf{n}_K(x) d\gamma(x) + \sum_{\sigma \in \mathcal{E}} v_\sigma \int_{-u_{\sigma,-}^-}^{-u_{\sigma,+}^-} (q(u_{\sigma,+}) - q(s)) ds \geq 0.$$

From (49), one obtains:

$$u_{\sigma,+}^- - u_{\sigma,-}^- = 0 \text{ for any } \sigma \in \mathcal{E}. \tag{50}$$

Hence $u_K \geq 0$ for any $K \in \mathcal{T}$. Indeed, for any $a \in \mathbb{R}$, let Ω_a be the interior of the union of the closures of the control volumes such that $u_K^- = a$. Let $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega_a$. Then, from (50), for any $b \neq a$, $\sigma \not\subset \partial\Omega_b$, so that $\sigma \subset \partial\Omega$, which proves that $a = 0$ and concludes the proof of Proposition 1. ∎

5 Convergence of the scheme

In order to show the convergence of the scheme, we shall first show that for any $C \geq 0$, the set $\{u_\mathcal{T} \in X(\mathcal{T})$ where $\mathcal{T}$ is an admissible mesh of $\Omega\}$ is relatively compact in $L^2(\Omega)$. In order to do so, we shall use the following compactness result.

Theorem 1 *Let Ω be an open bounded set of $\mathbb{R}^d$, $d \geq 1$, and $\{u_n,\, n \in \mathbb{N}\}$ a bounded sequence of $L^2(\Omega)$. For $n \in \mathbb{N}$, one defines $\overline{u}_n$ by $\overline{u}_n = u_n$ a.e. on Ω and $\overline{u}_n = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$. Assume that there exist $C \in \mathbb{R}$ and $\{h_n,\, n \in \mathbb{N}\} \subset \mathbb{R}_+$ such that $h_n \to 0$ as $n \to \infty$ and*

$$\|\overline{u}_n(\cdot + \eta) - \overline{u}_n\|_{L^2(\mathbb{R}^d)}^2 \leq C|\eta|(|\eta| + h_n), \forall n \in \mathbb{N},\ \forall \eta \in \mathbb{R}^d. \tag{51}$$

Then, $\{u_n,\, n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$. Furthermore, if $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$, then $u \in H_0^1(\Omega)$.

Proof of Theorem 1

Since $\{h_n,\, n \in \mathbb{N}\}$ is bounded, the fact that $\{u_n,\, n \in \mathbb{N}\}$ is relatively compact in $L^2(\Omega)$ is an immediate consequence of the Kolmogorov compactness theorem. Then, assuming that $u_n \to u$ in $L^2(\Omega)$ as $n \to \infty$, it is only necessary to prove that $u \in H_0^1(\Omega)$. Let us first remark that $\tilde{u}_n \to \overline{u}$ in $L^2(\mathbb{R}^d)$, as $n \to \infty$, with $\overline{u} = u$ a.e. on Ω and $\overline{u} = 0$ a.e. on $\mathbb{R}^d \setminus \Omega$.
Then, for $\varphi \in C_c^\infty(\mathbb{R}^d)$, one has, for all $\eta \in \mathbb{R}^d$, $\eta \neq 0$ and $n \in \mathbb{N}$, using the Cauchy Schwarz inequality and thanks to (51),

$$\int_{\mathbb{R}^d} \frac{(\overline{u}_n(x+\eta) - \overline{u}_n(x))}{|\eta|} \varphi(x) dx \leq \frac{\sqrt{C|\eta|(|\eta| + h_n)}}{|\eta|} \|\varphi\|_{L^2(\mathbb{R}^d)},$$

which gives, letting $n \to \infty$, since $h_n \to 0$,

$$\int_{\mathbb{R}^d} \frac{(\overline{u}(x+\eta)-\overline{u}(x))}{|\eta|}\varphi(x)dx \le \sqrt{C}||\varphi||_{L^2(\mathbb{R}^d)},$$

and therefore, with a trivial change of variables in the integration,

$$\int_{\mathbb{R}^d} \frac{(\varphi(x-\eta)-\varphi(x))}{|\eta|}\overline{u}(x)dx \le \sqrt{C}||\varphi||_{L^2(\mathbb{R}^d)}. \tag{52}$$

Let $\{e_i,\ i=1,\ldots,d\}$ be the canonical basis of $\mathbb{R}^d$. For $i \in \{1,\ldots,d\}$ fixed, taking $\eta = he_i$ in 52 and letting $h \to 0$ (with $h>0$, for instance) leads to

$$-\int_{\mathbb{R}^d} \frac{\partial \varphi(x)}{\partial x_i}\overline{u}(x)dx \le \sqrt{C}||\varphi||_{L^2(\mathbb{R}^d)},$$

for all $\varphi \in C_c^\infty(\mathbb{R}^d)$.
This proves that $D_i\overline{u}$ (the derivative of $\overline{u}$ with respect to x_i in the sense of distributions) belongs to $L^2(\mathbb{R}^d)$, and therefore that $\overline{u} \in H^1(\mathbb{R}^d)$. Since u is the restriction of $\overline{u}$ on Ω and since $\overline{u}=0$ a.e. on $\mathbb{R}^d \setminus \Omega$, therefore $u \in H_0^1(\Omega)$. This completes the proof of Theorem 1. ■

Lemma 4 *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 1 and $u \in X(\mathcal{T})$ (see Definition 2). One defines $\overline{u}$ by $\overline{u}=u$ a.e. on Ω, and $\overline{u}=0$ a.e. on $\mathbb{R}^d \setminus \Omega$. Then there exists $C>0$, only depending on Ω, such that*

$$||\overline{u}(\cdot+\eta)-\overline{u}||^2_{L^2(\mathbb{R}^d)} \le ||u||^2_{1,\mathcal{T}}|\eta|(|\eta|+C\,\mathrm{size}(\mathcal{T})), \forall \eta \in \mathbb{R}^d. \tag{53}$$

Proof of Lemma 4

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

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

(see Definition 3 for the definition of $D_\sigma u$).
This gives, using the Cauchy Schwarz inequality,

$$|\overline{u}(x+\eta)-\overline{u}(x)|^2 \le \sum_{\sigma\in\mathcal{E}} \chi_\sigma(x,x+\eta)\frac{|D_\sigma u|^2}{d_\sigma c_\sigma}\sum_{\sigma\in\mathcal{E}} \chi_\sigma(x,x+\eta)d_\sigma c_\sigma, \text{ for a.e. } x \in \mathbb{R}^d, \tag{54}$$

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

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

$$\sum_{\sigma\in\mathcal{E}} \chi_\sigma(x,x+\eta)d_\sigma c_\sigma \le |\eta| + C\,\mathrm{size}(\mathcal{T}), \tag{55}$$

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 (56) for such points x gives (56) for a.e. $x \in \mathbb{R}^d$, since η is fixed). Since Ω is not assumed to be convex, it may happen that the line segment $[x,x+\eta]$ is not included in $\overline{\Omega}$. In order to deal with this, let $y,z \in [x,x+\eta]$ such that $y \neq z$ and $[y,z] \subset \overline{\Omega}$; there exist $K, L \in \mathcal{T}$ such that $y \in \overline{K}$ and $z \in \overline{L}$. Hence,

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

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

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

and

$$\sum_{\sigma\in\mathcal{E}}\chi_\sigma(y,z)d_\sigma c_\sigma\le|y-z|+2\,\mathrm{size}(\mathcal{T}). \tag{56}$$

Note that this yields (55) 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 $0\le t_1<t_2<\ldots<t_n\le 1$, $n\le N$, 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

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

with $x_i = x+t_i\eta$, for $i=1,\ldots,n$, and $[x_i,x_{i+1}]\subset\overline{\Omega}$. Indeed, one has $x_i\in\partial\Omega$ if $t_i\notin\{0,1\}$.
Then, using (56) with $y=x_i$ and $z=x_{i+1}$, for $i=1,\ldots,n-1$, yields (55) with $C=2(N-1)$ (in particular, if Ω is convex, $C=2$ is convenient for (55) and therefore for (53) as we shall see below).

In order to conclude the proof of Lemma 4, remark that, for all $\sigma\in\mathcal{E}$,

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

Therefore, integrating (54) over $\mathbb{R}^d$ yields, with (55),

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

■

We may now state the convergence result:

Theorem 2 (Convergence)
Assume items 1, 2, 3 and 4 of Assumption 1 and $g\in H^{1/2}(\partial\Omega)$. Let $\zeta\in\mathbb{R}_+$ and $M\in\mathbb{N}$ be given values. Consider a family of admissible meshes of Ω (in the sense of Definition 1) such that $d_{K,\sigma}\ge\zeta\mathrm{diam}(K)$ for all control volume $K\in\mathcal{T}$ and for all $\sigma\in\mathcal{E}_K$, and $\mathrm{card}(\mathcal{E}_K)\le M$ for all $K\in\mathcal{T}$. Assume that $\mathrm{size}(\mathcal{T})$ tends to 0 as n tends to infinity. Let $(u_K)_{K\in\mathcal{T}}$ be the solution of the system given by equations (8)-(12). For a given mesh $\mathcal{T}$, 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, there exists a subsequence of the sequence $u_\mathcal{T}$ which converges in $L^2(\Omega)$ to a function u as $\mathrm{size}(\mathcal{T})\to 0$, where u satisfies (5). Moreover, if there exists only one solution to (5), then the whole sequence converges to u.

PROOF of Theorem 2
For any mesh $\mathcal{T}$ of the family of admissible meshes which is considered here, let $\tilde u\in X(\mathcal{T})$ be defined by (37) with $t=1$; using Lemma 3 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 $\|\tilde u_\mathcal{T}\|_{L^2(\Omega)}\le C_1$. Furthermore, since $\mathrm{size}(\mathcal{T})\to 0$, from Lemma 4 and Theorem 1, there exists a subsequence, still denoted by $\tilde u_\mathcal{T}$, and $\tilde u\in H^1_0(\Omega)$ such that $\tilde u_\mathcal{T}$ converges to $\tilde u$ in $L^2(\Omega)$.

Let us now prove that $u = \tilde{u} + \tilde{g}$ satisfies (5). Since $\tilde{u} \in H^1_0(\Omega)$, there only remains to show that $\tilde{u}$ satisfies:

$$\begin{aligned}&\int_\Omega (\nabla \tilde{u}(x) \nabla \varphi(x) + \operatorname{div}(\mathbf{v}(x) q(\tilde{u}(x) + \tilde{g}(x)))\varphi(x) = \\ &\int_\Omega f(x, \tilde{u}(x) + \tilde{g}(x))\varphi(x)dx - \int_\Omega (\nabla \tilde{g}(x) \nabla \varphi(x))dx, \ \forall \varphi \in H^1_0(\Omega).\end{aligned} \tag{57}$$

Let $\varphi \in C^\infty_c(\Omega)$ and let size$(\mathcal{T})$ be small enough so that $\varphi(x) = 0$ if $x \in K$ and $K \in \mathcal{T}$ is such that $\partial K \cap \partial\Omega \neq \emptyset$. Taking $t = 1$ in (38), multiplying by $\varphi(x_K)$, and summing the result over $K \in \mathcal{T}$ yields

$$T_1 + T_2 = T_3 + T_4, \tag{58}$$

with

$$T_1 = -\sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \tau_{K|L}(\tilde{u}_L - \tilde{u}_K)\varphi(x_K),$$

$$T_2 = \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma} q(u_{\sigma,+})\varphi(x_K),$$

$$T_3 = \sum_{K \in \mathcal{T}} \mathrm{m}(K)\varphi(x_K) f_K(u_K).$$

$$T_4 = \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \tau_{K|L}(\tilde{g}_L - \tilde{g}_K)\varphi(x_K),$$

where

$$\tilde{g}_K = \frac{1}{\mathrm{m}(K)} \int_K \tilde{g}(x)dx, \forall K \in \mathcal{T}.$$

First remark that, since $u_\mathcal{T}$ tends to u in $L^2(\Omega)$ and thanks to the assumptions on f and q,

$$T_3 \to \int_\Omega f((u(x))\varphi(x)dx \text{ as size}(\mathcal{T}) \to 0.$$

Let us now turn to the study of T_1;

$$T_1 = -\sum_{K|L \in \mathcal{E}_{\text{int}}} \tau_{K|L}(\tilde{u}_L - \tilde{u}_K)(\varphi(x_K) - \varphi(x_L)).$$

Consider the following auxiliary expression:

$$\begin{aligned} T'_1 &= \int_\Omega \tilde{u}_\mathcal{T}(x)\Delta\varphi(x)dx \\ &= \sum_{K \in \mathcal{T}} \tilde{u}_K \int_K \Delta\varphi(x)dx \\ &= \sum_{K|L \in \mathcal{E}_{\text{int}}} (\tilde{u}_K - \tilde{u}_L) \int_{K|L} \nabla\varphi(x) \cdot \mathbf{n}_{K,L} d\gamma(x). \end{aligned}$$

Since $\tilde{u}_\mathcal{T}$ converges to $\tilde{u}$ in $L^2(\Omega)$, it is clear that T'_1 tends to $\displaystyle\int_\Omega \tilde{u}(x)\Delta\varphi(x)\ dx$ as size$(\mathcal{T})$ tends to 0. Define

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

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

$$\begin{aligned} |T_1 + T_1'| &= |\sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L)(\tilde u_K - \tilde u_L) R_{K,L}| \\ &\leq \Big[\sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L)\frac{(\tilde u_K - \tilde u_L)^2}{d_{K|L}} \sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L) d_{K|L} (R_{K,L})^2\Big]^{1/2}, \end{aligned}$$

Regularity properties of the function φ give the existence of $C_1 \in \mathbb{R}$, only depending on φ, such that $|R_{K,L}| \leq C_1 \mathrm{size}(\mathcal{T})$. Therefore, since

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

using Estimate (24), we conclude that $T_1 + T_1' \to 0$ as $\mathrm{size}(\mathcal{T}) \to 0$.

The study of T_4 is similar. Let us introduce the function $\tilde g_{\mathcal{T}} \in X(\mathcal{T})$ 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 $\mathrm{size}(\mathcal{T}) \to 0$. Let

$$T_4' = \int_\Omega \tilde g_{\mathcal{T}}(x) \Delta\varphi(x) dx.$$

With computations similar to those carried out for T_1, we obtain that

$$|T_4 + T_4'| \leq \Big[\sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L)\frac{(\tilde g_K - \tilde g_L)^2}{d_{K|L}} \sum_{K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(K|L) d_{K|L} (R_{K,L})^2\Big]^{1/2}.$$

Hence, thanks to Lemma 2, and the fact that $|R_{K,L}| \leq C_1 \mathrm{size}(\mathcal{T})$, one deduces that $T_4 + T_4' \to 0$ as $\mathrm{size}(\mathcal{T}) \to 0$, and since $\tilde g_{\mathcal{T}} \to \tilde g$ in $L^2(O)$ as $\mathrm{size}(\mathcal{T}) \to 0$,

$$T_4 \to \int_\Omega \nabla \tilde g(x) \nabla\varphi(x) dx \text{ as } \mathrm{size}(\mathcal{T}) \to 0.$$

Let us now show that T_2 tends to $-\int_\Omega \mathbf{v}(x) q(u(x)) \nabla\varphi(x) dx$ as $\mathrm{size}(\mathcal{T}) \to 0$. Let us decompose $T_2 = T_2' + T_2''$ where

$$T_2' = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} (q(u_{\sigma,+}) - q(u_K)) \varphi(x_K)$$

and

$$T_2'' = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} q(u_K) \varphi(x_K) = \int_\Omega \mathrm{div}\mathbf{v}(x) q(u_{\mathcal{T}}(x)) \varphi_{\mathcal{T}}(x) dx,$$

where $\varphi_{\mathcal{T}}$ is defined by $\varphi_{\mathcal{T}}(x) = \varphi(x_K)$ if $x \in K$, $K \in \mathcal{T}$. Since $u_{\mathcal{T}} \to u$ and $\varphi_{\mathcal{T}} \to \varphi$ in $L^2(\Omega)$ as $\mathrm{size}(\mathcal{T}) \to 0$ (indeed, $\varphi_{\mathcal{T}} \to \varphi$ uniformly on Ω as $\mathrm{size}(\mathcal{T}) \to 0$) and thanks to the assumptions on q and the fact that $\mathrm{div}\mathbf{v} \in L^\infty(\Omega)$, one has

$$T_2'' \to \int_\Omega \mathrm{div}\mathbf{v}(x) q(u(x)) \varphi(x) dx \text{ as } \mathrm{size}(\mathcal{T}) \to 0.$$

Let us now rewrite T_2' as $T_2' = T_2''' + r_2$ with

$$T_2''' = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} (q(u_{\sigma,+}) - q(u_K)) \int_\sigma \mathbf{v}(x) \cdot \mathbf{n}_{K,\sigma} \varphi(x) d\gamma(x)$$

and

$$r_2 = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(q(u_{\sigma,+})-q(u_K))\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}(\varphi(x_K)-\varphi(x))d\gamma(x).$$

Thanks to the regularity of $\mathbf{v}$ and φ and the assumptions on q, there exists C_2 only depending on $\mathbf{v}$ and φ such that

$$|r_2| \le C_2\mathrm{size}(\mathcal{T})\sum_{K|L\in\mathcal{E}_{\mathrm{int}}}|u_K-u_L|\mathrm{m}(K|L),$$

which yields, with the Cauchy-Schwarz inequality,

$$|r_2| \le C_2\mathrm{size}(\mathcal{T})(\sum_{K|L\in\mathcal{E}_{\mathrm{int}}}\tau_{K|L}|u_K-u_L|^2)^{\frac{1}{2}}(\sum_{K|L\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}(K|L)d_{K|L})^{\frac{1}{2}},$$

from which one deduces, with Estimate (24), that $r_2\to 0$ as $\mathrm{size}(\mathcal{T})\to 0$.
Next, remark that

$$T_2''' = -\sum_{K\in\mathcal{T}}q(u_K)\sum_{\sigma\in\mathcal{E}_K}\int_\sigma \mathbf{v}(x)\cdot\mathbf{n}_{K,\sigma}\varphi(x)d\gamma(x) = -\sum_{K\in\mathcal{T}}q(u_K)\int_K \mathrm{div}(\mathbf{v}(x)\varphi(x))dx.$$

This implies (since $u_\mathcal{T}\to u$ in $L^2(\Omega)$) that $T_2'''\to -\int_\Omega \mathrm{div}(\mathbf{v}(x)\varphi(x))q(u(x))dx$, so that T_2' has the same limit and $T_2\to -\int_\Omega \mathbf{v}(x)\cdot\nabla\varphi(x)q(u(x))dx$.

Hence, letting $\mathrm{size}(\mathcal{T})\to 0$ in (58) yields that the function $\tilde{u}\in H_0^1(\Omega)$ satisfies

$$\int_\Omega\Big(\nabla\tilde{u}(x)\nabla\varphi(x)+\mathbf{v}(x)q(u(x))\nabla\varphi(x)-f(u(x))\varphi(x)\Big)dx = \int_\Omega\nabla\tilde{g}(x)\nabla\varphi(x)dx,\ \forall\varphi\in C_c^\infty(\Omega),$$

which, in turn, yields (57) thanks to the fact that $\tilde{u}\in H_0^1(\Omega)$, and to the density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$.
Finally, the function $u_\mathcal{T}$ converges in $L^2(\Omega)$, as $\mathrm{size}(\mathcal{T})\to 0$ to $u=\tilde{u}+\tilde{g}\in H^1(\Omega)$.
This concludes the proof of $u_\mathcal{T}\to u$ in $L^2(\Omega)$ as $\mathrm{size}(\mathcal{T})\to 0$, where u satisfies (5).

■

Remark 4 (i) A more simple proof of convergence for the finite volume scheme with non homogeneous Dirichlet boundary condition can be made if g is the trace of a Lipschitz-continuous function $\tilde{g}$. In that case, ζ and M do not have to be introduced and Lemma 2 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.
(ii) Lemma 2 is given in the two dimensional case; a similar result holds in the three dimensional case, but the proof is somewhat longer for technical reasons.

References

[1] Agouzal, A., J. Baranger, J.-F.Maitre and F. Oudin (1995), Connection between finite volume and mixed finite element methods for a diffusion problem with non constant coefficients, with application to Convection Diffusion, *East-West Journal on Numerical Mathematics.*, **3**, 4, 237-254.

[2] Bank R.E. and D.J. Rose (1987), Some error estimates for the box method, *SIAM J. Numer. Anal.*, **24** , 4, 777-787.

[3] J. Baranger, J.-F. Maitre and F. Oudin (1996), Connection between finite volume and mixed finite element methods, *Modél. Math. Anal. Numér.*, 30, 3, 4, 444-465.

[4] Cai, Z. (1991), On the finite volume element method, *Numer. Math.*, **58**, 713-735.

[5] Cai, Z., J. Mandel and S. Mc Cormick (1991), The finite volume element method for diffusion equations on general triangulations, *SIAM J. Numer. Anal.*, **28**, 2, 392-402.

[6] Courbet, B.. and J. P. Croisille, Finite-volume box-schemes on triangular meshes, submitted.

[7] Coudière, Y., J.P. Vila, and P. Villedieu, Convergence of a finite volume scheme for a diffusion problem, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 161-168.

[8] Deimling, K., *Nonlinear Functional Analysis*, (Springer, New York).

[9] Faille, I. (1992), Modélisation bidimensionnelle de la genèse et la migration des hydrocarbures dans un bassin sédimentaire, Thesis, Université de Grenoble.

[10] Eymard, R. and T. Gallouët (1993), Convergence d'un schéma de type eléments finis - volumes finis pour un système couplé elliptique - hyperbolique, *Modél. Math. Anal. Numér.* **27**, 7, 843-861.

[11] Eymard R., T. Gallouët and R. Herbin , The finite volume method, *to appear in Handobook of Numerical Analysis*, P.G. Ciarlet and J.L. Lions eds.

[12] Faille, I. (1992), A control volume method to solve an elliptic equation on a 2D irregular meshing, *Comp. Meth. Appl. Mech. Engrg.*, 100, 275-290.

[13] Feistauer M., J. Felcman and M. Lukacova-Medvidova, On the convergence of a combined finite volume-finite element method for nonlinear convection- diffusion problems, *Numer. Meth. for P.D.E.'s*, to appear.

[14] Fiard, J.M., R. Herbin (1994), Comparison between finite volume finite element methods for the numerical simulation of an elliptic problem arising in electrochemical engineering, Comput. Meth. Appl. Mech. Engin., 115, 315-338.

[15] Forsyth, P.A. (1989), A control volume finite element method for local mesh refinement, SPE 18415, 85-96.

[16] Forsyth, P.A. (1991), A control volume finite element approach to NAPL groundwater contamination, SIAM J. Sci. Stat. Comput., 12, 5, 1029-1057.

[17] Forsyth, P.A. and P.H. Sammon (1988), Quadratic Convergence for Cell-Centered Grids, Appl. Num. Math. 4, 377-394.

[18] Herbin R. (1995), An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh, *Num. Meth. P.D.E.* **11**, 165-173.

[19] Herbin R. (1996), Finite volume methods for diffusion convection equations on general meshes, *in Finite volumes for complex applications, Problems and Perspectives*, F. Benkhaldoun and R. Vilsmeier eds, Hermes, 153-160.

[20] Herbin R. and O. Labergerie (1997), Finite volume schemes for elliptic and elliptic-hyperbolic problems on triangular meshes, *Comp. Meth. Appl. Mech. Engin.* **147**,85-103.

[21] Lazarov R.D. and I.D. Mishev (1996), Finite volume methods for reaction diffusion problems in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 233-240.

[22] Lazarov R.D., I.D. Mishev and P.S Vassilevski, Finite volume methods for convection-diffusion problems, SIAM J. Numer. Anal., 33, 1996, 31-55.

[23] Manteufel, T.A., and A.B.White (1986), The numerical solution of second order boundary value problem on non uniform meshes,*Math. Comput.* **47**, 511-536.

[24] I.D. Mishev, Finite volume methods on Voronoï meshes, to appear.

[25] Morton, K.W. and E. Süli (1991), Finite volume methods and their analysis, IMA J. Numer. Anal. **11**, 241-260.

[26] Patankar, S.V. (1980), *Numerical Heat Transfer and Fluid Flow*, Series in Computational Methods in Mechanics and Thermal Sciences, Minkowycz and Sparrow Eds. (Mc Graw Hill).

[27] Vanselow R. (1996), Relations between FEM and FVM, in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 217-223.

[28] Vignal M.H. (1996), Convergence of a finite volume scheme for a system of an elliptic equation and a hyperbolic equation *Modél. Math. Anal. Numér.* **30**, 7, 841-872.

[29] Vignal M.H. (1996), Convergence of Finite Volumes Schemes for an elliptic hyperbolic system with boundary conditions,in: F. Benkhaldoun and R. Vilsmeier eds, *Finite volumes for complex applications, Problems and Perspectives* (Hermes, Paris), 145-152.