An introduction to finite volume schemes for elliptic equations

Pascal Omnes

October 21, 2022

CEA, Université Paris-Saclay, DM2S-STMF, F-91191 Gif-sur-Yvette Cedex, France.

Université Sorbonne Paris Nord, LAGA, CNRS UMR 7539, Institut Galilée, 99, Avenue J.-B. Clément F-93430 Villetaneuse Cedex, France.

1 Introduction

Elliptic equations are encountered in a wide range of physical models (fluid flows, electrostatic equilibrium, etc). Their theoretical analysis are usually performed using variational formulations, which are in this context powerful tools. Based on these tools and on the theory of polynomial interpolation, a natural, flexible and very efficient way to discretize elliptic equations is with the help of finite element methods. On the other hand, hyperbolic equations (like the Euler equations of gas dynamics) have been efficiently discretized by finite volume techniques. Then, the question to be answered when one decides to discretize models that couple hyperbolic and elliptic equations is how to adapt finite volume techniques to elliptic equations (or, alternatively, how to adapt finite element methods to hyperbolic equations, an option which we will not consider in this course). We shall see that the answer to this question is far from trivial, especially in the multi-dimensional case, where anisotropic diffusion and/or distorted meshes will cause severe problems. For this reason, we shall perform the numerical analysis of finite volume methods essentially on the one dimensional Laplace equation, in order to keep technicalities at a minimum, while still giving an idea as complete as possible of the various tools that can be used to perform such an analysis.

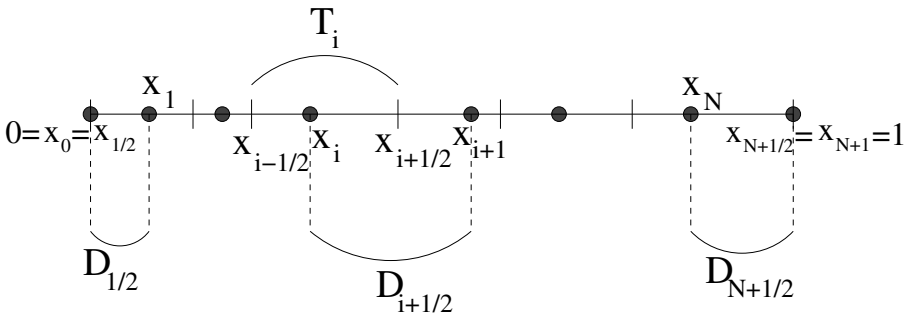

Figure 1: A one dimensional mesh and associated notations

2 Approximation of the Laplace equation in dimension one

The domain of computation will be $\Omega =]0;1[$. A function f is given in $L^2(\Omega)$, and we look for an approximation of the following problem

$$-u'' = f \quad \text{over} \quad \Omega, \tag{1}$$

$$u(0) = u(1) = 0 \tag{2}$$

by a cell-centered finite volume scheme.

2.1 Construction of the finite volume scheme

2.1.1 The mesh

We fix $N \in \mathbb{N}$ and choose $N+1$ points $(x_{i+1/2})_{i\in[0,N]}$ with $0 = x_{1/2} < x_{3/2} < \cdots < x_{N-1/2} < x_{N+1/2} = 1$. We subdivide Ω into N segments $T_i = [x_{i-1/2}; x_{i+1/2}]$, with $i \in [1, N]$ and we associate to each T_i a point $x_i \in T_i$ which is not necessarily the midpoint of T_i (although it is advised to perform such a choice). We shall also set $x_0 = 0$ and $x_{N+1} = 1$ and denote by $|T_i| = x_{i+1/2} - x_{i-1/2}$ the length of T_i. Let us denote the mesh size by $h := \sup_{i\in[1,N]} |T_i|$.

Dual cells are defined according to $D_{i+1/2} = [x_i; x_{i+1}]$, with $i \in [0, N]$. Note that the point $x_{i+1/2}$ is not necessarily the midpoint of $D_{i+1/2}$. In particular, we note that this is never the case for $D_{1/2}$ and $D_{N+1/2}$ which are "half" dual cells. We denote by $|D_{i+1/2}| = x_{i+1} - x_i$, the length of the dual cells.

2.1.2 General principle of the finite volume method

We associate with any control volume T_i a degree of freedom u_i, which is supposed to represent some approximation of the value of u at the point x_i (rather

than an approximation of the mean value of u over T_i, as we shall see later). Then, we integrate equation (1) over the control volumes:

$$\frac{1}{|T_i|}\int_{T_i}(-u'')(x)dx = \frac{1}{|T_i|}\int_{T_i} f(x)dx. \tag{3}$$

The left-hand side of (3) may be integrated exactly and yields

$$\frac{1}{|T_i|}\int_{T_i}(-u'')(x)dx = \frac{1}{|T_i|}[-u'(x_{i+1/2}) + u'(x_{i-1/2})].$$

The right-hand side of (3) equals the mean-value of f over T_i, denoted by f_i. Thus, we obtain

$$\frac{1}{|T_i|}[-u'(x_{i+1/2}) + u'(x_{i-1/2})] = f_i. \tag{4}$$

Up to now, no approximation has been introduced. The construction of the scheme relies on the approximation of u' at the different points $x_{j+1/2}$ with the help of the unknowns (u_j) of the scheme. A simple and sensible way to perform such an approximation is to use the finite difference

$$u'(x_{i+1/2}) \approx \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i} \approx \frac{u_{i+1} - u_i}{|D_{i+1/2}|} \ , \ \forall i \in [0, N]. \tag{5}$$

Indeed, if u is the exact solution of the problem and if u is regular enough, then we may write

$$\begin{aligned} u(x_{i+1}) &= u(x_{i+1/2}) + (x_{i+1} - x_{i+1/2})\, u'(x_{i+1/2}) \\ &+ \frac{1}{2}(x_{i+1} - x_{i+1/2})^2\, u''(x_{i+1/2}) + O(h^3), \end{aligned}$$

$$\begin{aligned} u(x_i) &= u(x_{i+1/2}) + (x_i - x_{i+1/2})\, u'(x_{i+1/2}) \\ &+ \frac{1}{2}(x_i - x_{i+1/2})^2\, u''(x_{i+1/2}) + O(h^3). \end{aligned}$$

Thus,

$$\begin{aligned} u(x_{i+1}) - u(x_i) &= (x_{i+1} - x_i)\, u'(x_{i+1/2}) \\ &+ \frac{1}{2}[(x_{i+1} - x_{i+1/2})^2 - (x_i - x_{i+1/2})^2]\, u''(x_{i+1/2}) + O(h^3). \end{aligned}$$

We may consider two cases: either $x_{i+1/2}$ is the midpoint of $D_{i+1/2}$ and then

$$u'(x_{i+1/2}) \approx \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i} + O(h^2), \tag{6}$$

or, if this is not the case

$$u'(x_{i+1/2}) \approx \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i} + O(h). \tag{7}$$

In any case, the approximation is at least of order one in $O(h)$.

Finally, inserting (5) (written for i and $i-1$) into (4), the finite volume scheme may be written

$$\frac{1}{|T_i|}\left[-\frac{u_{i+1}-u_i}{|D_{i+1/2}|}+\frac{u_i-u_{i-1}}{|D_{i-1/2}|}\right]=f_i\ ,\ \forall i\in[1,N]. \tag{8}$$

Remark: For $i=1$ (first cell of the mesh), we need in (8) a value for u_0. In the same way, for $i=N$ (last cell of the mesh), we need a value for u_{N+1}. These values are given by the boundary conditions (2):

$$u_0=u_{N+1}=0. \tag{9}$$

Remark: In the case of a uniform mesh, and if x_i is chosen as the midpoint of T_i, there holds $|T_i|=\frac{1}{N}:=\Delta x$ and $|D_{i+1/2}|=\Delta x$ for all $i\in[1,N-1]$ and $|D_{1/2}|=|D_{N+1/2}|=\frac{\Delta x}{2}$. We thus have

$$\begin{aligned}\frac{-u_2+3u_1-2u_0}{\Delta x^2} &= f_1\\ \frac{-u_{i+1}+2u_i-u_{i-1}}{\Delta x^2} &= f_i\ ,\ \forall i\in[2,N-1]\\ \frac{-2u_{N+1}+3u_N-u_{N-1}}{\Delta x^2} &= f_N.\end{aligned}$$

In the case of a uniform mesh, the equations obtained for $i\in[2,N-1]$ are thus very much like those obtained by a standard finite difference scheme or by a standard P_1 finite element scheme; only the right-hand side slightly differs.

2.2 Properties of the scheme

2.2.1 Conservativity

Scheme (8) may be written under the form

$$F_{i,i+1/2}+F_{i,i-1/2}=|T_i|f_i,$$

where $F_{i,i+1/2}:=-\dfrac{u_{i+1}-u_i}{|D_{i+1/2}|}$ is called the flux on the edge $i+1/2$ of cell T_i and $F_{i,i-1/2}:=\dfrac{u_i-u_{i-1}}{|D_{i-1/2}|}$ is called the flux on the edge $i-1/2$ of cell T_i.

The conservativity principle simply states that $F_{i,i+1/2}+F_{i+1,i+1/2}=0$ for all internal interface $i+1/2$ (with $i\in[1,N-1]$).

It is easily checked that this is the case here.

2.2.2 Definition of discrete derivatives and discrete scalar products

We introduce the following operator d (discrete divergence - in one dimension it is a discrete derivative on the primal cells)

$$\begin{aligned}d:\mathbb{R}^{N+1} &\longrightarrow \mathbb{R}^N\\ (v_{i+1/2})_{i\in[0,N]} &\mapsto (dv)_i:=\frac{v_{i+1/2}-v_{i-1/2}}{|T_i|}\end{aligned} \tag{10}$$

We also introduce the operator g (discrete gradient - in one dimension it is also a discrete derivative, but on the dual cells)

$$\begin{array}{rcl} g : \mathbb{R}^{N+2} & \longrightarrow & \mathbb{R}^{N+1} \\ (u_i)_{i\in[0,N+1]} & \mapsto & (gu)_{i+1/2} := \dfrac{u_{i+1}-u_i}{|D_{i+1/2}|} \end{array} \tag{11}$$

With these definitions, scheme (8) may simply be rewritten

$$-(dgu)_i = f_i \ , \ \forall i \in [1,N]. \tag{12}$$

On the primal mesh, we also define the discrete scalar product $(\cdot,\cdot)_T$ by

$$(u_i)_{i\in[1,N]}, (w_i)_{i\in[1,N]} \mapsto (u,w)_T := \sum_{i\in[1,N]} |T_i| u_i w_i. \tag{13}$$

And on the dual mesh, we define the discrete scalar product $(\cdot,\cdot)_D$ by

$$(a_{i+1/2})_{i\in[0,N]}, (b_{i+1/2})_{i\in[0,N]} \mapsto (a,b)_D := \sum_{i\in[0,N]} |D_{i+1/2}| a_{i+1/2} b_{i+1/2}. \tag{14}$$

Proposition: Let $(w_i)_{i\in[0,N+1]}, (v_{i+1/2})_{i\in[0,N]}$ be given. There holds

$$(dv,w)_T = -(v,gw)_D + v_{N+1/2} w_{N+1} - v_{1/2} w_0 \, , \tag{15}$$

which is the discrete equivalent of

$$(v',w)_{L^2(\Omega)} = -(w',v)_{L^2(\Omega)} + v(1)w(1) - v(0)w(0).$$

Proof: There holds

$$(dv,w)_T = \sum_{i\in[1,N]} |T_i| \, (dv)_i \, w_i = \sum_{i\in[1,N]} (v_{i+1/2} - v_{i-1/2}) \, w_i \tag{16}$$

according to (13) and (10). Now, in the last sum in (16), for a given $i_0 \in [1, N-1]$, the term $v_{i_0+1/2}$ appears twice: once when $i = i_0$ and once when $i = i_0 + 1$. In the first case, the term $v_{i_0+1/2}$ is multiplied by w_{i_0} and in the second case, it is multiplied by $-w_{i_0+1}$. Thus, rearranging the sum in (16), there holds

$$\begin{array}{rcl} (dv,w)_T & = & -v_{1/2}w_1 - \displaystyle\sum_{i\in[1,N-1]} v_{i+1/2} \, (w_{i+1} - w_i) + v_{N+1/2} w_N \\ & = & -v_{1/2}w_0 - v_{1/2}(w_1 - w_0) - \displaystyle\sum_{i\in[1,N-1]} v_{i+1/2} \, (w_{i+1} - w_i) \\ & & -v_{N+1/2}(w_{N+1} - w_N) + v_{N+1/2} w_{N+1} \\ & = & -\displaystyle\sum_{i\in[0,N]} v_{i+1/2} \, (w_{i+1} - w_i) + v_{N+1/2} w_{N+1} - v_{1/2} w_0 \end{array}$$

$$
\begin{aligned}
&= -\sum_{i\in[0,N]} |D_{i+1/2}| v_{i+1/2} \frac{(w_{i+1}-w_i)}{|D_{i+1/2}|} + v_{N+1/2}w_{N+1} - v_{1/2}w_0 \\
&= -\sum_{i\in[0,N]} |D_{i+1/2}| v_{i+1/2} (gw)_{i+1/2} + v_{N+1/2}w_{N+1} - v_{1/2}w_0 \\
&= -(v, gw)_D + v_{N+1/2}w_{N+1} - v_{1/2}w_0 \,,
\end{aligned}
$$

the last two lines being obtained thanks to (11) and (14).

2.2.3 A discrete variational formulation for the finite volume scheme

Now, consider (12) and (9), and consider any $(w_i)_{i\in[0,N+1]}$ with $w_0 = w_{N+1} = 0$. Thanks to (15) and to the boundary conditions on w, there holds

$$
(gu, gw)_D = -(dgu, w)_T = (f, w)_T, \tag{17}
$$

where $(f_i)_{i\in[1,N]}$ is the vector of $\mathbb{R}^N$ whose entries are the mean-values f_i. Reciprocally, since (17) holds for a w with arbitrary values w_i for $i \in [1, N]$, (17) implies (12). Thus, the scheme may be written under the discrete variational formulation: Find $(u_i)_{i\in[0,N+1]}$ with $u_0 = u_{N+1} = 0$, such that for all $(w_i)_{i\in[0,N+1]}$ with $w_0 = w_{N+1} = 0$, there holds

$$
(gu, gw)_D = (f, w)_T. \tag{18}
$$

This is a discrete equivalent of the continuous variational formulation: find $u \in H^1_0(\Omega)$ such that for all $w \in H^1_0(\Omega)$

$$
(u', w')_{L^2(\Omega)} = (f, w)_{L^2(\Omega)}.
$$

2.2.4 Existence and uniqueness of the discrete solution

The scheme is set as a system of $N+2$ equations (one per cell T_i, Eq. (12), and two boundary conditions, Eq. (9)) with $N+2$ unknowns $(u_i)_{i\in[0,N+1]}$. In $\mathbb{R}^{N+2}$, existence for all data and uniqueness are equivalent for a square system. Let us prove uniqueness, which, by linearity of the discrete equations, is equivalent to prove that if $f_i = 0$ for all $i \in [1, N]$, then $u_i = 0$ for all $i \in [0, N+1]$. Now, if $f_i = 0$ for all $i \in [1, N]$, the right-hand side of (18) vanishes for all possible w. But since $u_0 = u_{N+1} = 0$, we may consider $w = u$, which leads us to

$$
(gu, gu)_D = 0 = \sum_{i\in[0,N]} |D_{i+1/2}| (gu)^2_{i+1/2}.
$$

Since no $|D_{i+1/2}|$ vanishes, this is equivalent to $(gu)_{i+1/2} = 0$ for all $i \in [0, N]$. According to the definition (11) of $(gu)_{i+1/2}$, there holds $u_i = u_{i+1}$ for all $i \in [0, N]$, which means that there exists $c \in \mathbb{R}$ so that $u_i = c$ for all i in $[0, N+1]$. But since $u_0 = 0$, this constant c is necessarily 0, so that $u_i = 0$ for all i in $[0, N+1]$, which proves uniqueness and thus existence for all data.

2.2.5 The case of Neumann boundary conditions

We consider equation (1), associated with the boundary conditions

$$u'(0) = u'(1) = 0. \tag{19}$$

Remark: Since only the gradient of u is involved in system (1)–(19), then u is determined only up to an additive constant (i.e. if $x \mapsto u(x)$ is solution, then $x \mapsto u(x)+c$, with c an arbitrary constant, is also a solution). Thus, to uniquely determine u, we impose

$$\int_\Omega u(x)\,dx = 0. \tag{20}$$

Remark: A necessary condition over f for the solution of (1)–(19) to exist is obtained by integrating (1) over Ω, which yields

$$\int_\Omega f(x)\,dx = -\int_\Omega u''(x)\,dx = -u'(1) + u'(0) = 0, \tag{21}$$

using (19).

Now, (1) is still discretized by (12), while boundary conditions (19) are discretized by $(gu)_{1/2} = (gu)_{N+1/2} = 0$, which yields

$$u_0 = u_1 \quad \text{and} \quad u_{N+1} = u_N. \tag{22}$$

Moreover, (20) is discretized by

$$\sum_{i\in[1,N]} |T_i| u_i = 0. \tag{23}$$

Thus, there are now $N+3$ equations given by (12), (22) and (23), but only $N+2$ unknowns. However, the set of equations given by (12) and (22) is not independent. Indeed, let us multiply each equation in (12) by $|T_i|$ and sum over $i \in [1, N]$. We obtain

$$\begin{aligned} \sum_{i=1}^{N} [-(gu)_{i+1/2} + (gu)_{i-1/2}] &= \sum_{i=1}^{N} |T_i|\, f_i \\ -(gu)_{N+1/2} + (gu)_{1/2} &= \sum_{i=1}^{N} |T_i| \,\frac{1}{|T_i|} \int_{T_i} f(x)\,dx. \end{aligned} \tag{24}$$

The left-hand side of (24) vanishes because of boundary conditions (22), while the right-hand side of (24) equals $\int_\Omega f(x)\,dx$, which vanishes because of (21). Thus, one non-trivial linear combination of the $N+2$ equations (12) and (22) yields the equality $0 = 0$, which means that there are at most $N+1$ independent equations in this set, which, together with (23) yield at most $N+2$ independent equations. Since there are also $N+2$ unknowns, proving injectivity is enough to prove existence and uniqueness. Now, consider (12) and (19), and consider any

$(w_i)_{i\in[0,N+1]}$. Thanks to (15) and to the boundary conditions on (gu), there still holds

$$(gu, gw)_D = (f, w)_T. \tag{25}$$

If f vanishes, choosing $w = u$ in (25) leads to $(gu, gu)_D = 0$. Then, like in section 2.2.4, this is equivalent to $(gu)_{i+1/2} = 0$ for all $i \in [0, N]$. According to the definition (11) of $(gu)_{i+1/2}$, there holds $u_i = u_{i+1}$ for all $i \in [0, N]$, which means that there exists $c \in \mathbb{R}$ so that $u_i = c$ for all i in $[0, N + 1]$. Then, (23) implies that $(\sum_{i=1}^{N} |T_i|)\, c = 0$, which means that $c = 0$, and, finally, that $u_i = 0$ for all i in $[0, N + 1]$, which proves uniqueness and thus existence for all data.

Remark: We have seen that the condition $\sum_{i=1}^{N} |T_i|\, f_i = 0$ is a necessary condition for the existence of a solution and that this equality is true if f_i is computed, in an exact way, as the mean-value of f over T_i. However, in practice, it is more than likely that f_i will not be computed exactly, for example if it is computed by a quadrature formula. In order to ensure existence and uniqueness of the discrete solution, it is necessary to project f over the set of discrete functions with vanishing integral over Ω. This will be performed, for example, by replacing the data $(f_i)_{i\in[1,N]}$ by the slightly modified data $(\tilde{f}_i)_{i\in[1,N]}$ defined by the formula

$$\tilde{f}_i = f_i - \frac{\sum_{j=1}^{N} |T_j| f_j}{\sum_{j=1}^{N} |T_j|}, \tag{26}$$

which ensures that $\sum_{i=1}^{N} |T_i|\, \tilde{f}_i = 0$. Note that (26) is indeed a slight modification of f_i, since if each f_j is computed as the mean-value of f over T_j with an error of order $O(h^\alpha)$ for some $\alpha > 0$, then $\frac{\sum_{j=1}^{N} |T_j| f_j}{\sum_{j=1}^{N} |T_j|}$ is also of order $O(h^\alpha)$.

2.2.6 The case of Robin boundary conditions

We consider equation (1), associated with the boundary condition

$$u'(0) - \lambda_0 u(0) = u'(1) + \lambda_1 u(1) = 0, \tag{27}$$

with $\lambda_0 > 0$ and $\lambda_1 > 0$. These boundary conditions are discretized by

$$(gu)_{1/2} = \lambda_0 u_0 \quad \text{and} \quad (gu)_{N+1/2} = -\lambda_1 u_{N+1}. \tag{28}$$

Existence and uniqueness are proved like before: there are $N + 2$ equations and $N + 2$ unknowns, so that it is enough to prove injectivity. Using (12), (15) and (28), there holds, for any $(w_i)_{i\in[1,N]}$

$$(gu, gw)_D + \lambda_1 u_{N+1} w_{N+1} + \lambda_0 u_0 w_0 = (f, w)_T. \tag{29}$$

So that if $f = 0$, then choosing $w = u$ in (29) leads to $u_0 = u_{N+1} = 0$ and $(gu)_{i+1/2} = 0$ for all $i \in [0, N]$. Like in section 2.2.4, this leads to $u_i = 0$ for all $i \in [0, N + 1]$, which proves uniqueness and thus existence for all data.

2.2.7 Discrete maximum principle

Proposition: We suppose here that f is positive on Ω and that $u_0 = u_{N+1} = 0$. Then, the discrete solution remains positive on Ω, i.e. $u_i \geq 0$. Actually, we even have a stronger property, which is the absence of local minimum: Let i be in $[1, N]$; if $f \geq 0$ on T_i, then $u_i \geq \min(u_{i-1}, u_{i+1})$, and $u_i = \min(u_{i-1}, u_{i+1})$ if and only if $u_i = u_{i-1} = u_{i+1}$ and $f_i = 0$.

Proof: if $f \geq 0$ on T_i, then $f_i \geq 0$. Moreover, there holds, according to (8)

$$\frac{1}{|T_i|}\left[-\frac{u_{i+1}-u_i}{|D_{i+1/2}|}+\frac{u_i-u_{i-1}}{|D_{i-1/2}|}\right]=f_i \tag{30}$$

Now if $u_i < \min(u_{i-1}, u_{i+1})$, then $u_i - u_{i-1} < 0$ and $u_i - u_{i+1} < 0$, so that the left-hand side in Eq. (30) is strictly negative, while the right-hand side is non-negative, which is a contradiction. Additionally, if $u_i = \min(u_{i-1}, u_{i+1})$, then $u_i - u_{i-1} < 0$ or $u_i - u_{i+1} < 0$ is also a contradiction to the positivity of f_i, so that $u_i = u_{i-1} = u_{i+1}$ and $f_i = 0$. We have thus proved the absence of local minimum.

Now, if f is positive on Ω, then let us suppose that, for a given i in $[1, N]$, there holds $u_i < 0$. Then, we may consider an integer $i_0 \in [1, N]$ for which $u_{i_0} = \min_{i\in[1,N]}(u_i)$. If i_0 is not unique, we shall choose the smallest in $[1, N]$. Then, of course, $u_{i_0} < 0$ and $u_{i_0} \leq \min(u_{i_0-1}, u_{i_0+1})$ by definition. But since f is positive on Ω, it is positive on T_{i_0}. Thus, according to the absence of local minimum $u_{i_0} \geq \min(u_{i_0-1}, u_{i_0+1})$, so that there necessarily holds $u_{i_0-1} = u_{i_0} = u_{i_0+1}$. Since we have chosen i_0 to be the smallest integer in $[1, N]$ where the minimum of $(u_i)_{i\in[1,N]}$ is reached, and since u_{i_0-1} is also equal to this minimum, then $i_0 = 1$ and $i_0 - 1 = 0$. Thus, there would hold $u_0 = u_{i_0-1} = u_{i_0} = \min_{i\in[1,N]}(u_i) < 0$, while the boundary condition imposed on u_0 is $u_0 = 0$, which is a contradiction. Thus, the hypothesis that there exists a strictly negative u_i is false, which proves the discrete maximum principle.

Remark: of course, if f is negative on Ω, the scheme respects the absence of local maximum and a discrete minimum principle.

2.2.8 Equivalence with a finite element like method

We first recall that the variational formulation of the problem (1)-(2) is to find $u \in H_0^1(\Omega)$, such that for all $w \in H_0^1(\Omega)$

$$(u', w')_{L^2(\Omega)} = (f, w)_{L^2(\Omega)}. \tag{31}$$

The approximation of this problem by conforming finite elements amounts to find a function u_h in a finite dimensional space $V_h \subset H_0^1$ such that $\forall v_h \in V_h$,

$$(u_h', w_h')_{L^2(\Omega)} = (f, w_h)_{L^2(\Omega)}.$$

But we have seen with (18) that the finite volume scheme we consider can be set under a discrete variational form, where we recall that the discrete gradient is given by the formula

$$(gu)_{i+1/2} = \frac{u_{i+1}-u_i}{|D_{i+1/2}|}.$$

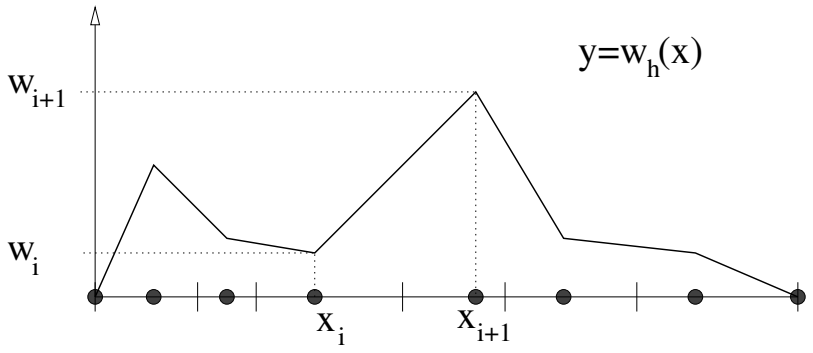

Figure 2: The function w_h associated to the values $(w_i)_{i\in[0,N+1]}$.

We remark that $(gu)_{i+1/2}$ is thus the derivative, on the cell $D_{j+1/2}$, of the function which is P^1 on $D_{j+1/2}$, and which is determined by its values u_i in x_i and u_{i+1} in x_{i+1}. We thus define

$$V_{h0} = \Big\{ w_h \in C_0(\Omega),\ \text{s.t. } (w_h)_{|D_{i+1/2}} \in P^1(D_{j+1/2}) \text{ and } w_h(0) = w_h(1) = 0. \Big\} \tag{32}$$

Note that it suffices to know the values $w_h(x_i)$ to completely determine a function w_h in V_{h0}. With each element $(w_i)_{i\in[0,N+1]}$ with $w_0 = w_{N+1} = 0$, we associate a function w_h which belongs to V_{h0} and which is uniquely defined by $w_h(x_i) = w_i$ for all $i \in [0, N+1]$. The derivative of w_h over each $D_{i+1/2}$ is thus a constant whose value is

$$(w_h)'_{|D_{i+1/2}} = \frac{w_{i+1} - w_i}{|D_{i+1/2}|} = (gw)_{i+1/2}.$$

Thus, in the discrete variational formulation of the scheme, the term $(gu, gw)_D$ is exactly equal to

$$\begin{aligned} (gu, gw)_D &= \sum_{i=0}^{N} |D_{i+1/2}|\, (gu)_{i+1/2}\, (gw)_{i+1/2} \\ &= \sum_{i=0}^{N} |D_{i+1/2}|\, (u_h)'_{|D_{i+1/2}}\, (w_h)'_{|D_{i+1/2}}. \end{aligned}$$

But since $(u_h)'_{|D_{i+1/2}}$ and $(w_h)'_{|D_{i+1/2}}$ are constants, we may write

$$|D_{i+1/2}|\, (u_h)'_{|D_{i+1/2}}\, (w_h)'_{|D_{i+1/2}} = \int_{D_{i+1/2}} u'_h\, w'_h\, (x)\, dx.$$

So that, finally,

$$(gu, gw)_D = \sum_{i=0}^{N} \int_{D_{i+1/2}} u'_h\, w'_h\, (x)\, dx = \int_{\Omega} u'_h\, w'_h\, (x)\, dx.$$

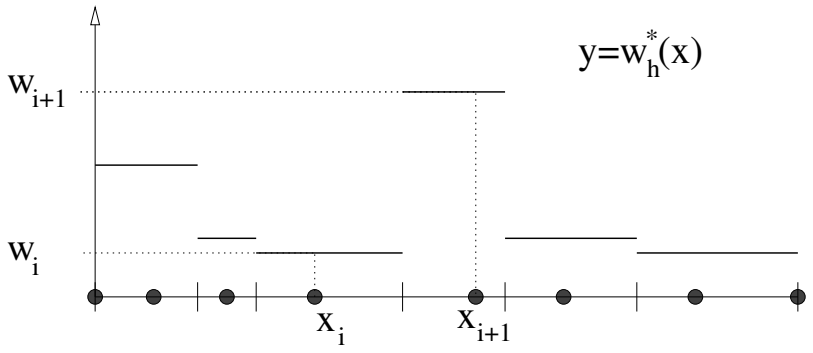

Figure 3: The function w_h^* associated to the values $(w_i)_{i\in[1,N]}$.

As far as the term $(f,w)_T$ is concerned, there holds

$$(f,w)_T = \sum_{i=1}^{N} |T_i|\, f_i w_i = \sum_{i=1}^{N} w_i \int_{T_i} f(x)dx = \sum_{i=1}^{N} \int_{T_i} f(x)\, w_i dx.$$

If we define the function w_h^* through $w_h(x) = w_i$ for all $x \in T_i$, then there holds

$$(f,w)_T = \sum_{i=1}^{N} \int_{T_i} f(x)\, w_h^*(x)\, dx = \int_{\Omega} f(x)\, w_h^*(x)\, dx.$$

The finite volume method is thus equivalent to the following finite element like method:

Find $u_h \in V_{h0}$ such that, for all $w_h \in V_{h0}$,

$$\int_{\Omega} u_h'\, w_h'\,(x)\, dx = \int_{\Omega} f(x)\, w_h^*(x)\, dx. \tag{33}$$

Up to now, we have only proved that the finite volume solution may be transformed into a function that is solution of the finite element scheme, but since both of these solutions are unique, the two problems are indeed equivalent.

3 Error estimation and convergence to the solution of the continuous problem

The aim of this section is to prove that the finite volume scheme solution converges to the solution of the continuous problem and to find some error estimations. We shall give different proofs of the same results: this is in order to give the reader as many tools as possible that could be used to analyze more difficult, possibly higher dimensional related problems.

In what follows, we shall denote by $h := \sup_{i\in[1,N]} |T_i|$ the mesh-step size. Note that this definition implies that $|D_{i+1/2}| \leq 2h$.

We shall suppose in general that the function f is in $L^2(\Omega)$, and in some of the subsections, we shall suppose that it belongs to $H^1(\Omega)$. We recall that if $f \in H^m(\Omega)$ (with the convention $H^0(\Omega) = L^2(\Omega)$), then the solution of (1)-(2), which we will denote by $\hat{u}$ in what follows, belongs to $H^{m+2}(\Omega) \subset C^{m+1}(\bar{\Omega})$ *because Ω is a one-dimensional domain.*

3.1 Estimations in the energy norm

3.1.1 In the finite volume sense

Since the exact solution $\hat{u}$ is at least in $C^1(\bar{\Omega})$, we may consider its values at points x_i. We thus define the projection $\Pi\hat{u}$ by $(\Pi\hat{u})_i = \hat{u}(x_i)\,\forall\, i \in [0, N+1]$. Note that since $\hat{u}(0) = \hat{u}(1) = 0$, this implies that $(\Pi\hat{u})_0 = (\Pi\hat{u})_{N+1} = 0$. Moreover, since the derivative $\hat{u}'$ of the exact solution is at least in $C^0(\bar{\Omega})$, we may consider its values at points $x_{i+1/2}$. We thus define the projection $P\hat{u}'$ by $(P\hat{u}')_{i+1/2} = \hat{u}'(x_{i+1/2})\,\forall\, i \in [0, N]$.

We shall estimate the discrete H_0^1 (energy) norm of the difference $u - \Pi\hat{u}$ (where u is the solution of the finite volume method) defined by

$$|u - \Pi\hat{u}|_{1,D} := (g(u - \Pi\hat{u}), g(u - \Pi\hat{u}))_D^{1/2}.$$

We shall start with a useful lemma:

Let $(w_i)_{i\in[0,N+1]}$ with $w_0 = w_{N+1} = 0$, then, if u is the solution of the finite volume scheme, there holds

$$(gu, gw)_D = (P\hat{u}', gw)_D. \tag{34}$$

Proof: since u is the solution of the finite volume scheme and since $w_0 = w_{N+1} = 0$, we know that, thanks to (18),

$$(gu, gw)_D = (f, w)_T = \sum_{i\in[1,N]} |T_i| f_i w_i = \sum_{i\in[1,N]} w_i \int_{T_i} f(x)dx.$$

But since $f(x) = -\hat{u}''(x)$, we have

$$\begin{aligned} \int_{T_i} f(x)dx &= -\int_{T_i} \hat{u}''(x)dx = -\hat{u}'(x_{i+1/2}) + \hat{u}'(x_{i-1/2}) \\ &= -(P\hat{u}')_{i+1/2} + (P\hat{u}')_{i-1/2} \\ &= -|T_i| \frac{(P\hat{u}')_{i+1/2} - (P\hat{u}')_{i-1/2}}{|T_i|}. \end{aligned} \tag{35}$$

Thus,

$$(gu, gw)_D = -\sum_{i\in[1,N]} |T_i| w_i d(P\hat{u}') = -(w_i, d(P\hat{u}'))_T = (P\hat{u}', gw)_D$$

by the discrete Green formula (15).

With the help of this lemma, setting $w = u - \Pi\hat{u}$, we shall write, since $w_0 = w_{N+1} = 0$,

$$\begin{aligned} |u - \Pi\hat{u}|^2_{1,D} &= (gu, gw)_D - (g(\Pi\hat{u}), gw)_D = (P\hat{u}', gw)_D - (g(\Pi\hat{u}), gw)_D \\ &= (P\hat{u}' - g(\Pi\hat{u}), gw)_D \leq ||P\hat{u}' - g(\Pi\hat{u})||_{0,D}\, |w|_{1,D} \end{aligned} \tag{36}$$

thanks to the discrete Cauchy-Schwarz inequality, with the notation

$$||a||_{0,D} := (a, a)^{1/2}_D.$$

Thus, since $w = u - \Pi\hat{u}$, (36) results in

$$|u - \Pi\hat{u}|_{1,D} \leq ||P\hat{u}' - g(\Pi\hat{u})||_{0,D}. \tag{37}$$

Remark: the evaluation of $|u - \Pi\hat{u}|_{1,D}$ can thus be performed only in terms of the exact solution $\hat{u}$, which is always a good starting point to perform numerical analysis.

There holds

$$||P\hat{u}' - g(\Pi\hat{u})||^2_{0,D} = \sum_{i=0}^{N} |D_{i+1/2}|\, \varepsilon^2_{i+1/2} \tag{38}$$

where $\varepsilon_{i+1/2}$ is the difference in the approximation of $\hat{u}'(x_{i+1/2})$ by the finite difference $\frac{\hat{u}(x_{i+1}) - \hat{u}(x_i)}{|D_{i+1/2}|}$:

$$\varepsilon_{i+1/2} := \hat{u}'(x_{i+1/2}) - \frac{\hat{u}(x_{i+1}) - \hat{u}(x_i)}{|D_{i+1/2}|}.$$

Since $\hat{u}(x_{i+1}) - \hat{u}(x_i) = \int_{D_{i+1/2}} \hat{u}'(x)dx$, and since $\hat{u}'(x_{i+1/2})$ is a constant, there holds

$$\begin{aligned} |D_{i+1/2}|\varepsilon_{i+1/2} &= \int_{D_{i+1/2}} [\hat{u}'(x_{i+1/2}) - \hat{u}'(x)]dx \\ &= \int_{x_i}^{x_{i+1/2}} [\hat{u}'(x_{i+1/2}) - \hat{u}'(x)]dx \\ &+ \int_{x_{i+1/2}}^{x_{i+1}} [\hat{u}'(x_{i+1/2}) - \hat{u}'(x)]dx. \end{aligned} \tag{39}$$

Let us now consider $x \in [x_{i+1/2}, x_{i+1}]$ (the case $x \in [x_i, x_{i+1/2}]$ is treated similarly). Then,

$$\hat{u}'(x_{i+1/2}) - \hat{u}'(x) = -\int_{x_{i+1/2}}^{x} \hat{u}''(t)dt = \int_{x_{i+1/2}}^{x} f(t)dt.$$

Since f belongs to $L^2(\Omega)$, the Cauchy-Schwarz inequality may be used to bound $\int_{x_{i+1/2}}^{x} f(t)dt$. This results in

$$\begin{aligned} |\hat{u}'(x_{i+1/2}) - \hat{u}'(x)| &\leq \left(\int_{x_{i+1/2}}^{x} dt\right)^{1/2} \left(\int_{x_{i+1/2}}^{x} f^2(t)dt\right)^{1/2} \\ &\leq (x - x_{i+1/2})^{1/2} \left(\int_{x_{i+1/2}}^{x_{i+1}} f^2(t)dt\right)^{1/2}. \end{aligned}$$

Thus, there holds

$$\left|\int_{x_{i+1/2}}^{x_{i+1}} [\hat{u}'(x_{i+1/2}) - \hat{u}'(x)]dx\right| \leq \frac{2}{3}(x_{i+1} - x_{i+1/2})^{3/2} \left(\int_{x_{i+1/2}}^{x_{i+1}} f^2(t)dt\right)^{1/2}$$

and then

$$\begin{aligned} |D_{i+1/2}||\varepsilon_{i+1/2}| &\leq \frac{2}{3}(x_{i+1/2} - x_i)^{3/2} \left(\int_{x_i}^{x_{i+1/2}} f^2(t)dt\right)^{1/2} \\ &+ \frac{2}{3}(x_{i+1} - x_{i+1/2})^{3/2} \left(\int_{x_{i+1/2}}^{x_{i+1}} f^2(t)dt\right)^{1/2} \\ &\leq \frac{2}{3}\left((x_{i+1} - x_{i+1/2})^3 + (x_{i+1/2} - x_i)^3\right)^{1/2} \left(\int_{D_{i+1/2}} f^2(t)dt\right)^{1/2} \\ &\leq \frac{2}{3}|D_{i+1/2}|^{3/2} \left(\int_{D_{i+1/2}} f^2(t)dt\right)^{1/2}. \end{aligned}$$

Thus,

$$|\varepsilon_{i+1/2}| \leq \frac{2}{3}|D_{i+1/2}|^{1/2} \left(\int_{D_{i+1/2}} f^2(t)dt\right)^{1/2}$$

and then

$$|D_{i+1/2}|\varepsilon_{i+1/2}^2 \leq \left(\frac{2}{3}\right)^2 |D_{i+1/2}|^2 \int_{D_{i+1/2}} f^2(t)dt \leq \left(\frac{2}{3}\right)^2 4h^2 \int_{D_{i+1/2}} f^2(t)dt, \tag{40}$$

since $|D_{i+1/2}| \leq 2h$. According to (38), this results in

$$||P\hat{u}' - g(\Pi\hat{u})||_{0,D}^2 \leq \left(\frac{2}{3}\right)^2 4h^2 \sum_{i=0}^{N} \int_{D_{i+1/2}} f^2(t)dt = \left(\frac{2}{3}\right)^2 4h^2 \int_{\Omega} f^2(t)dt.$$

Finally, according to (37), there holds

$$|u - \Pi\hat{u}|_{1,D} \leq \frac{4}{3}h\,||f||_{L^2(\Omega)}. \tag{41}$$

The scheme exhibits a first-order convergence, as was expected from (7).

Particular case when $x_{i+1/2}$ is the midpoint of $D_{i+1/2}$. In accordance to (6), we shall now see that we may get a better local estimate if $x_{i+1/2}$ is the midpoint of $D_{i+1/2}$, and if $\hat{u}$ is regular enough (for this we shall suppose that f belongs to $H^1(\Omega)$, so that $\hat{u}$ belongs to $H^3(\Omega)$ as explained above). However, we shall not get a global second-order convergence, since, at least for $i = 0$ and $i = N$, the point $x_{i+1/2}$ is not the midpoint of $D_{i+1/2}$. Let us consider the right-hand side of Eq. (39); in the first integral, let us perform the change of variable $x = x_{i+1/2} - t$, and in the second, let us set $x = x_{i+1/2} + t$. Since $x_{i+1/2} - x_i = x_{i+1} - x_i = \frac{1}{2}|D_{i+1/2}|$, there holds

$$\begin{aligned} |D_{i+1/2}|\varepsilon_{i+1/2} &= -\int_{\frac{|D_{i+1/2}|}{2}}^{0} \left[\hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} - t)\right] dt \\ &\quad + \int_{0}^{\frac{|D_{i+1/2}|}{2}} \left[\hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} + t)\right] dt \\ &= \int_{0}^{\frac{|D_{i+1/2}|}{2}} A(t)dt, \end{aligned} \tag{42}$$

with $A(t) = \hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} - t) + \hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} + t)$. We may write

$$\begin{aligned} \hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} - t) &= \int_0^t \hat{u}''(x_{i+1/2} - s)ds \\ \hat{u}'(x_{i+1/2}) - \hat{u}'(x_{i+1/2} + t) &= -\int_0^t \hat{u}''(x_{i+1/2} + s)ds, \end{aligned}$$

so that, setting $B(s) = \hat{u}''(x_{i+1/2} - s) - \hat{u}''(x_{i+1/2} + s)$, there holds

$$A(t) = \int_0^t B(s)ds. \tag{43}$$

Since $-\hat{u}'' = f$, then $-\hat{u}''' = f'$ and we may write

$$B(s) = -\int_{-s}^{s} \hat{u}'''(x_{i+1/2} + \tau)d\tau = \int_{-s}^{s} f'(\tau)d\tau.$$

Since we have supposed that f is in $H^1(\Omega)$, then f' is in $L^2(\Omega)$ and we may use the Cauchy-Schwarz inequality to obtain

$$|B(s)| \leq (2s)^{1/2} \left(\int_{-s}^{s} (f')^2(\tau)d\tau\right)^{1/2} \leq (2s)^{1/2} ||f'||_{L^2(D_{i+1/2})}.$$

Thus, Eq. (43) leads to

$$|A(t)| \leq \sqrt{2}\frac{2}{3}t^{3/2} ||f'||_{L^2(D_{i+1/2})},$$

which, according to (42), leads in turn to

$$\begin{aligned} |D_{i+1/2}|\,|\varepsilon_{i+1/2}| &\leq \sqrt{2}\frac{2}{3}\frac{2}{5}\left(\frac{|D_{i+1/2}|}{2}\right)^{5/2}||f'||_{L^2(D_{i+1/2})} \\ &\leq \frac{1}{15}|D_{i+1/2}|^{5/2}||f'||_{L^2(D_{i+1/2})}. \end{aligned}$$

Thus,

$$|\varepsilon_{i+1/2}| \leq \frac{1}{15}|D_{i+1/2}|^{3/2}||f'||_{L^2(D_{i+1/2})}$$

and,

$$|D_{i+1/2}|\,|\varepsilon_{i+1/2}|^2 \leq \left(\frac{1}{15}\right)^2 |D_{i+1/2}|^4||f'||^2_{L^2(D_{i+1/2})} \leq \left(\frac{4}{15}\right)^2 h^4||f'||^2_{L^2(D_{i+1/2})} \tag{44}$$

since $|D_{i+1/2}| \leq 2h$. Now, in order to find a bound for $\sum_{i=0}^{N}|D_{i+1/2}|\,|\varepsilon_{i+1/2}|^2$, we have to distinguish the cells that are such that $x_{i+1/2}$ is the midpoint of $D_{i+1/2}$ from those for which this is not the case, since, as noticed above, this is never the case for $i = 0$ and $i = N$. We may thus consider the general case in which there is a (bounded) number P of cells $D_{i+1/2}$ for which $x_{i+1/2}$ is not the midpoint of $D_{i+1/2}$. This is for example the case when a family of meshes is obtained from an arbitrary coarse initial mesh by dividing recursively each cell into two identical sub-cells, and when one chooses as x_i the midpoints of the cells T_i. In this family of meshes, it is very easy to show that all $x_{i+1/2}$ are the midpoints of the cells $D_{i+1/2}$ except for the points $x_{i+1/2}$ associated with the interfaces of the initial coarse cells.

There holds

$$\begin{aligned} \sum_{i=0}^{N}|D_{i+1/2}|\,|\varepsilon_{i+1/2}|^2 &= \sum_{i \text{ s.t. } x_{i+1/2}=\frac{x_i+x_{i+1}}{2}}|D_{i+1/2}|\,|\varepsilon_{i+1/2}|^2 \\ &+ \sum_{i \text{ s.t. } x_{i+1/2}\neq\frac{x_i+x_{i+1}}{2}}|D_{i+1/2}|\,|\varepsilon_{i+1/2}|^2. \end{aligned} \tag{45}$$

According to (44), the first sum in the right-hand side of (45) is bounded by

$$\left(\frac{4}{15}\right)^2 h^4 \sum_{i \text{ s.t. } x_{i+1/2}=\frac{x_i+x_{i+1}}{2}}||f'||^2_{L^2(D_{i+1/2})} \leq \left(\frac{4}{15}\right)^2 h^4||f'||^2_{L^2(\Omega)}.$$

According to (40), the second sum in the right-hand side of (45) is bounded by

$$\left(\frac{2}{3}\right)^2 4h^2 \sum_{i \text{ s.t. } x_{i+1/2}\neq\frac{x_i+x_{i+1}}{2}}\int_{D_{i+1/2}} f^2(t)dt.$$

Now, since we have supposed that $f \in H^1(\Omega)$, then f is continuous on $\bar{\Omega}$, and there holds (since $|\Omega| = 1$)

$$||f||_{L^\infty(\Omega)} \leq ||f||_{L^2(\Omega)} + ||f'||_{L^2(\Omega)},$$

which shows that

$$\int_{D_{i+1/2}} f^2(t)dt \leq (2h)(||f||_{L^2(\Omega)} + ||f'||_{L^2(\Omega)})^2.$$

Finally, the second sum in the right-hand side of (45) is bounded by

$$\left(\frac{2}{3}\right)^2 8Ph^3(||f||_{L^2(\Omega)} + ||f'||_{L^2(\Omega)})^2.$$

According to (37), (38), (45), and the above considerations, there holds

$$|u - \Pi\hat{u}|_{1,D} \leq \frac{4}{15}h^2||f'||_{L^2(\Omega)} + \frac{4}{3}\sqrt{2}P^{1/2}h^{3/2}(||f||_{L^2(\Omega)} + ||f'||_{L^2(\Omega)}),$$

whose leading term behaves like $O(P^{1/2}h^{3/2})||f||_{H^1(\Omega)}$. Thus, if P is indeed bounded when h goes to zero, then the convergence is at least of order 1.5.

Particular case when x_i is the midpoint of T_i. Moreover, we can prove the second order convergence of the scheme by a different technique when one chooses the points x_i as the midpoints of the cell T_i. Note that this proof is valid only in dimension one, while the previous computations can be performed in higher dimensions.
We first note that $[(gu)_{i+1/2}]_{i\in[0,N]}$ and $[(P\hat{u}')_{i+1/2}]_{i\in[0,N]}$ solve the same difference equations, according to (12) and (35), so that

$$[d(gu - P\hat{u}')] = 0.$$

According to the definition of d (see (10)), this implies that there exists a constant $c \in \mathbb{R}$ such that $(gu - P\hat{u}')_{i+1/2} = c$ for all $i \in [0, N]$. Moreover, since $u_0 = u_{N+1} = 0$, there holds $\sum_{i=0}^{N} |D_{i+1/2}|(gu)_{i+1/2} = 0$; this implies that

$$c\sum_{i=0}^{N} |D_{i+1/2}| = \sum_{i=0}^{N} |D_{i+1/2}|(gu - P\hat{u}')_{i+1/2} = -\sum_{i=0}^{N} |D_{i+1/2}|(P\hat{u}')_{i+1/2},$$

which means finally (since $\sum_{i=0}^{N} |D_{i+1/2}| = |\Omega| = 1$) that

$$(gu - P\hat{u}')_{i+1/2} = -\sum_{j=0}^{N} |D_{j+1/2}|(P\hat{u}')_{j+1/2} \ \forall\, i \in [0, N]. \tag{46}$$

Since x_i is the midpoint of T_i, there holds $|D_{i+1/2}| = \frac{1}{2}(|T_i| + |T_{i+1}|)$ for $i \in [1, N-1]$, while $|D_{1/2}| = \frac{1}{2}|T_1|$ and $|D_{N+1/2}| = \frac{1}{2}|T_N|$. Therefore,

$$\begin{aligned}\sum_{i=0}^{N} |D_{i+1/2}|(P\hat{u}')_{i+1/2} &= \frac{1}{2}|T_1|(P\hat{u}')_{1/2} + \frac{1}{2}|T_N|(P\hat{u}')_{N+1/2} \\ &+ \sum_{i=1}^{N-1} \frac{1}{2}(|T_i| + |T_{i+1}|)(P\hat{u}')_{i+1/2}.\end{aligned} \tag{47}$$

Rearranging the sum in the right-hand side of (47), we obtain

$$\sum_{i=0}^{N} |D_{i+1/2}|(P\hat{u}')_{i+1/2} = \sum_{i=1}^{N} |T_i| \frac{1}{2}\left[(P\hat{u}')_{i-1/2} + (P\hat{u}')_{i+1/2}\right]. \tag{48}$$

Let us prove that the right-hand side of (48) provides a second-order approximation of $\int_\Omega \hat{u}'(x)dx$. Since $T_i = [x_{i-1/2}; x_{i+1/2}]$, there holds

$$\begin{aligned}\int_{T_i} \hat{u}'(x)dx &= \int_0^{|T_i|} \hat{u}'(x_{i-1/2} + t)dt \\ &= |T_i|\hat{u}'(x_{i-1/2}) + \int_0^{|T_i|} (\hat{u}'(x_{i-1/2} + t) - \hat{u}'(x_{i-1/2}))dt \\ &= |T_i|(P\hat{u}')_{i-1/2} + \int_0^{|T_i|} \left(\int_0^t \hat{u}''(x_{i-1/2} + \tau)d\tau\right) dt.\end{aligned} \tag{49}$$

In the same way,

$$\begin{aligned}\int_{T_i} \hat{u}'(x)dx &= \int_0^{|T_i|} \hat{u}'(x_{i+1/2} - t)dt \\ &= |T_i|\hat{u}'(x_{i+1/2}) + \int_0^{|T_i|} (\hat{u}'(x_{i+1/2} - t) - \hat{u}'(x_{i+1/2}))dt \\ &= |T_i|(P\hat{u}')_{i+1/2} - \int_0^{|T_i|} \left(\int_0^t \hat{u}''(x_{i+1/2} - \tau)d\tau\right) dt.\end{aligned} \tag{50}$$

It follows from (49) and (50) that

$$\begin{aligned}\int_{T_i} \hat{u}'(x)dx &= |T_i|\frac{1}{2}\left[(P\hat{u}')_{i-1/2} + (P\hat{u}')_{i+1/2}\right] \\ &+ \frac{1}{2}\int_0^{|T_i|} \left(\int_0^t (\hat{u}''(x_{i-1/2} + \tau) - \hat{u}''(x_{i+1/2} - \tau))d\tau\right) dt.\end{aligned} \tag{51}$$

We may estimate

$$\hat{u}''(x_{i-1/2} + \tau) - \hat{u}''(x_{i+1/2} - \tau) = \int_{x_{i+1/2}-\tau}^{x_{i-1/2}+\tau} \hat{u}'''(s)ds = -\int_{x_{i+1/2}-\tau}^{x_{i-1/2}+\tau} f'(s)ds.$$

Since f' belongs to $L^2(\Omega)$, and since $|(x_{i-1/2}+\tau)-(x_{i+1/2}-\tau)| \leq |T_i|$ for $0 \leq \tau \leq t \leq |T_i|$, we obtain by the Cauchy-Schwarz inequality

$$\left|\hat{u}''(x_{i-1/2}+\tau)-\hat{u}''(x_{i+1/2}-\tau)\right| \leq |T_i|^{1/2}||f'||_{L^2(T_i)}. \tag{52}$$

From (51) and (52), we obtain

$$\begin{aligned}\left|\int_{T_i}\hat{u}'(x)dx-|T_i|\frac{1}{2}\left[(P\hat{u}')_{i-1/2}+(P\hat{u}')_{i+1/2}\right]\right| &\leq \frac{1}{2}|T_i|^{5/2}||f'||_{L^2(T_i)}\\ &\leq \frac{1}{2}h^2|T_i|^{1/2}||f'||_{L^2(T_i)}\end{aligned}$$

and therefore

$$\left|\int_{\Omega}\hat{u}'(x)dx-\sum_{i=1}^{N}|T_i|\frac{1}{2}\left[(P\hat{u}')_{i-1/2}+(P\hat{u}')_{i+1/2}\right]\right| \leq \frac{1}{2}h^2\sum_{i=1}^{N}|T_i|^{1/2}||f'||_{L^2(T_i)}. \tag{53}$$

The right-hand side of (53) may be bounded, through the discrete Cauchy-Schwarz inequality by $(h^2/2)\,||f'||_{L^2(\Omega)}$. Now, since $\int_\Omega \hat{u}'(x)dx$ equals $\hat{u}(1)-\hat{u}(0)$ which vanishes due to boundary conditions (2), there holds

$$\left|\sum_{i=1}^{N}|T_i|\frac{1}{2}\left[(P\hat{u}')_{i-1/2}+(P\hat{u}')_{i+1/2}\right]\right| \leq \frac{1}{2}h^2||f'||_{L^2(\Omega)},$$

which taking into account (46) and (48) yields

$$|(gu-P\hat{u}')_{i+1/2}| \leq \frac{1}{2}h^2||f||_{H^1(\Omega)}\,\forall i \in [0,N],$$

which means second-order convergence in the max norm for the discrete gradient gu.

3.1.2 In the finite element sense

We have seen in section 2.2.4 that there is an equivalence of the finite volume scheme with a finite element like technique. We shall therefore use this equivalence and tools that are standard in the numerical analysis of finite element schemes in order to obtain an error estimate for the solution u_h of (33) in the (continuous) H_0^1 semi-norm which shall be denoted in the following by $|\cdot|_{H_0^1(\Omega)}$. By a triangular inequality, there holds, for all w_h in V_{h0} (see definition (32))

$$|\hat{u}-u_h|_{H_0^1(\Omega)} \leq |\hat{u}-w_h|_{H_0^1(\Omega)}+|w_h-u_h|_{H_0^1(\Omega)}. \tag{54}$$

The second term in this sum may be estimated as follows

$$\begin{aligned}|w_h-u_h|^2_{H_0^1(\Omega)} &= (w_h'-u_h',w_h'-u_h')_{L^2(\Omega)}\\ &= (w_h'-\hat{u}',w_h'-u_h')_{L^2(\Omega)}+(\hat{u}'-u_h',w_h'-u_h')_{L^2(\Omega)}.\end{aligned} \tag{55}$$

Let us set $v_h = w_h - u_h$; since V_{h0} is included in $H_0^1(\Omega)$, Eq. (31) implies

$$(\hat{u}', v_h')_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)}. \tag{56}$$

Moreover, thanks to (33), there holds

$$(u_h', v_h')_{L^2(\Omega)} = (f, v_h^*)_{L^2(\Omega)}. \tag{57}$$

Combining (55), (56) and (57) and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |w_h - u_h|^2_{H_0^1(\Omega)} &= (w_h' - \hat{u}', w_h' - u_h')_{L^2(\Omega)} + (f, v_h - v_h^*)_{L^2(\Omega)} \\ &\leq |w_h - \hat{u}|_{H_0^1(\Omega)} |w_h - u_h|_{H_0^1(\Omega)} + |(f, v_h - v_h^*)_{L^2(\Omega)}| \end{aligned}$$

which, if $|w_h - u_h|_{H_0^1(\Omega)} = |v_h|_{H_0^1(\Omega)}$ does not vanish, implies

$$|w_h - u_h|_{H_0^1(\Omega)} \leq |w_h - \hat{u}|_{H_0^1(\Omega)} + \frac{|(f, v_h - v_h^*)_{L^2(\Omega)}|}{|v_h|_{H_0^1(\Omega)}}. \tag{58}$$

Finally, for all w_h in V_{h0}, we infer from (54) and (58)

$$\begin{aligned} |\hat{u} - u_h|_{H_0^1(\Omega)} &\leq 2|\hat{u} - w_h|_{H_0^1(\Omega)} + \frac{|(f, v_h - v_h^*)_{L^2(\Omega)}|}{|v_h|_{H_0^1(\Omega)}} \\ &\leq 2|\hat{u} - w_h|_{H_0^1(\Omega)} + \sup_{v_h \neq 0 \in V_{h0}} \frac{|(f, v_h - v_h^*)_{L^2(\Omega)}|}{|v_h|_{H_0^1(\Omega)}} \\ &\leq 2 \inf_{w_h \in V_{h0}} |\hat{u} - w_h|_{H_0^1(\Omega)} + \sup_{v_h \neq 0 \in V_{h0}} \frac{|(f, v_h - v_h^*)_{L^2(\Omega)}|}{|v_h|_{H_0^1(\Omega)}} \end{aligned} \tag{59}$$

The first term in the right-hand side of (59) is called the interpolation error and measures the way in which the space V_{h0} approaches $H_0^1(\Omega)$; the second term in the right-hand side of (59) is called the consistency error and is due to the fact that (33) contains the term $(f, v_h^*)_{L^2(\Omega)}$ instead of the term $(f, v_h)_{L^2(\Omega)}$ which is usual in finite element techniques.

The interpolation error of $H_0^1(\Omega)$ by V_{h0} is easily bounded (see any finite element course) by $Ch||f||_{L^2(\Omega)}$, where the constant C does not depend on the mesh. The consistency error may be estimated in the following way

$$\begin{aligned} |(f, v_h - v_h^*)_{L^2(\Omega)}| &\leq \sum_{i=0}^{N} \left| \int_{D_{i+1/2}} f(x) \left[v_h(x) - v_h^*(x) \right] dx \right| \\ &\leq \sum_{i=0}^{N} ||f||_{L^2(D_{i+1/2})} ||v_h - v_h^*||_{L^2(D_{i+1/2})}. \end{aligned} \tag{60}$$

We recall here that $D_{i+1/2} = [x_i; x_{i+1/2}] \cup [x_{i+1/2}; x_{i+1}]$ and that v_h^* is a constant on each of these subintervals, equal to $v_i = v_h(x_i)$ on $[x_i; x_{i+1/2}]$

and equal to $v_{i+1} = v_h(x_{i+1})$ on $[x_{i+1/2}; x_{i+1}]$. We shall therefore estimate $||v_h - v_h^*||_{L^2(D_{i+1/2})}$ in the following way

$$||v_h - v_h^*||^2_{L^2(D_{i+1/2})} = \int_{x_i}^{x_{i+1/2}} (v_h(x) - v_h(x_i))^2 \, dx + \int_{x_{i+1/2}}^{x_{i+1}} (v_h(x) - v_h(x_{i+1}))^2 \, dx.$$

Now we may write, by a Cauchy-Schwarz inequality

$$\begin{aligned} v_h(x) - v_h(x_i) &= \int_{x_i}^{x} v_h'(\tau) d\tau \\ &\leq (x - x_i)^{1/2} \left(\int_{x_i}^{x} (v_h')^2(\tau) d\tau \right)^{1/2} \\ &\leq (x - x_i)^{1/2} \left(\int_{x_i}^{x_{i+1/2}} (v_h')^2(\tau) d\tau \right)^{1/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_{x_i}^{x_{i+1/2}} (v_h(x) - v_h(x_i))^2 \, dx &\leq \frac{1}{2}(x_{i+1/2} - x_i)^2 \int_{x_i}^{x_{i+1/2}} (v_h')^2(\tau) d\tau \\ &\leq \frac{1}{2}|D_{i+1/2}|^2 \int_{x_i}^{x_{i+1/2}} (v_h')^2(\tau) d\tau. \end{aligned}$$

With the same kind of computation on $[x_{i+1/2}; x_{i+1}]$, we finally obtain

$$\begin{aligned} ||v_h - v_h^*||^2_{L^2(D_{i+1/2})} &\leq \frac{1}{2}|D_{i+1/2}|^2 |v_h|^2_{H_0^1(D_{i+1/2})} \\ &\leq 2h^2 |v_h|^2_{H_0^1(D_{i+1/2})}. \end{aligned}$$

With this result and (60), we obtain by the discrete Cauchy-Schwarz inequality

$$|(f, v_h - v_h^*)_{L^2(\Omega)}|^2 \leq 2h^2 |v_h|^2_{H_0^1(\Omega)} ||f||^2_{L^2(\Omega)},$$

from which

$$\frac{|(f, v_h - v_h^*)_{L^2(\Omega)}|}{|v_h|_{H_0^1(\Omega)}} \leq \sqrt{2} h ||f||_{L^2(\Omega)}. \tag{61}$$

Finally, from (59), (61) and the $O(h)$ estimation on the interpolation error, we obtain a first-order error estimate, like in the general case of section 3.1.1: there exists a constant C, which does not depend on the mesh, such that

$$|\hat{u} - u_h|_{H_0^1(\Omega)} \leq Ch ||f||_{L^2(\Omega)}. \tag{62}$$

3.2 Estimations in the L^2 norm

3.2.1 Through a discrete Poincaré inequality

Lemma: Let $(w_i)_{i \in [0, N+1]}$ such that $w_0 = 0$; then $||w||_{0,T} \leq |w|_{1,D}$.
Proof: For all $0 \leq i \leq N$, there holds, since $w_0 = 0$

$$|w_i| = \left| w_0 + \sum_{j=0}^{i-1} (w_{j+1} - w_j) \right|$$

$$
\begin{aligned}
&= \left| \sum_{j=0}^{i-1} |D_{j+1/2}| \frac{(w_{j+1} - w_j)}{|D_{j+1/2}|} \right| \\
&= \left| \sum_{j=0}^{i-1} |D_{j+1/2}| (gw)_{j+1/2} \right| \\
&\leq \left(\sum_{j=0}^{i-1} |D_{j+1/2}| \right)^{1/2} \left(\sum_{j=0}^{i-1} |D_{j+1/2}| (gw)_{j+1/2}^2 \right)^{1/2}
\end{aligned}
$$

through the discrete Cauchy-Schwarz inequality. Now

$$
\left(\sum_{j=0}^{i-1} |D_{j+1/2}| \right)^{1/2} \leq \left(\sum_{j=0}^{N} |D_{j+1/2}| \right)^{1/2} = |\Omega|^{1/2} = 1
$$

and

$$
\left(\sum_{j=0}^{i-1} |D_{j+1/2}| (gw)_{j+1/2}^2 \right)^{1/2} \leq \left(\sum_{j=0}^{N} |D_{j+1/2}| (gw)_{j+1/2}^2 \right)^{1/2} = |w|_{1,D}.
$$

Therefore, for all $0 \leq i \leq N$, there holds $|w_i| \leq |w|_{1,D}$, and thus

$$
||w||_{0,T} = \left(\sum_{i=1}^{N} |T_i| w_i^2 \right)^{1/2} \leq \left(\sum_{i=1}^{N} |T_i| \right)^{1/2} |w|_{1,D} = |\Omega| \, |w|_{1,D} = |w|_{1,D}.
$$

The application of this result to $w = u - \Pi\hat{u}$ proves the first-order convergence of u to $\Pi\hat{u}$ in the general case, an order of convergence of order 1.5 if all but a bounded number of cells $D_{i+1/2}$ have the point $x_{i+1/2}$ as midpoint and the second-order convergence if x_i is the midpoint of T_i for all i, if f is regular enough.

3.2.2 By a duality argument

This is an argument which is known in the finite element theory as the Aubin-Nitsche lemma. We measure the L^2 norm of the difference $\hat{u} - u_h$, where u_h is the function in V_{h0} which solves the equivalent finite element like formulation (33). Since $L^2(\Omega)$ is its own dual, there holds

$$
||\hat{u} - u_h||_{L^2(\Omega)} = \sup_{v \in L^2(\Omega), v \neq 0} \frac{(\hat{u} - u_h, v)_{L^2(\Omega)}}{||v||_{L^2(\Omega)}}. \tag{63}
$$

For any given $v \in L^2(\Omega)$, there exists a unique $w \in H^2(\Omega) \cap H_0^1(\Omega)$ which solves

$$
\begin{cases} -w'' = v \quad \text{in } \Omega \\ w(0) = w(1) = 0 \end{cases}.
$$

Moreover, there exists a constant C (the constant in the Poincaré inequality over Ω) such that

$$|w|_{H_0^1(\Omega)} \leq C||v||_{L^2(\Omega)}. \tag{64}$$

The following equalities hold

$$\begin{aligned}(\hat{u}-u_h,v)_{L^2(\Omega)} &= -(\hat{u}-u_h,w'')_{L^2(\Omega)} = ((\hat{u}-u_h)',w')_{L^2(\Omega)} \\ &= ((\hat{u}-u_h)',(w-w_h)')_{L^2(\Omega)} + ((\hat{u}-u_h)',w_h')_{L^2(\Omega)} \end{aligned} \tag{65}$$

for any $w_h \in V_{h0}$. The last term in the right-hand side of (65) may be transformed into

$$((\hat{u}-u_h)',w_h')_{L^2(\Omega)} = (\hat{u}',w_h')_{L^2(\Omega)} - (u_h',w_h')_{L^2(\Omega)} = (f,w_h)_{L^2(\Omega)} - (f,w_h^*)_{L^2(\Omega)} \tag{66}$$

thanks to (31) and (33). We have seen from (61) that

$$|(f,w_h-w_h^*)_{L^2(\Omega)}| \leq \sqrt{2}h||f||_{L^2(\Omega)}|w_h|_{H_0^1(\Omega)}.$$

This, (66) and (64) imply that

$$\begin{aligned}\left|((\hat{u}-u_h)',w_h')_{L^2(\Omega)}\right| &\leq \sqrt{2}h||f||_{L^2(\Omega)}(|w_h-w|_{H_0^1(\Omega)}+|w|_{H_0^1(\Omega)}) \\ &\leq Ch||f||_{L^2(\Omega)}(|w_h-w|_{H_0^1(\Omega)}+||v||_{L^2(\Omega)}). \end{aligned} \tag{67}$$

On the other hand, with a Cauchy-Schwarz inequality and (62), we obtain

$$\begin{aligned}\left|((\hat{u}-u_h)',(w-w_h)')_{L^2(\Omega)}\right| &\leq |\hat{u}-u_h|_{H_0^1(\Omega)}|w-w_h|_{H_0^1(\Omega)} \\ &\leq Ch||f||_{L^2(\Omega)}|w-w_h|_{H_0^1(\Omega)}. \end{aligned} \tag{68}$$

Finally, (65), (67) and (68) lead to

$$\left|(\hat{u}-u_h,v)_{L^2(\Omega)}\right| \leq Ch||f||_{L^2(\Omega)}(|w_h-w|_{H_0^1(\Omega)}+||v||_{L^2(\Omega)}) \tag{69}$$

for all w_h in V_{h0}. We may choose w_h as the function of V_{h0} which interpolates w at the points x_i: $w_h(x_i)=w(x_i)$. Standard interpolation theory provides the following inequality

$$|w_h-w|_{H_0^1(\Omega)} \leq Ch||w''||_{L^2(\Omega)} = Ch||v||_{L^2(\Omega)}.$$

This, together with (69) and (63) imply that there exists a constant C such that

$$||\hat{u}-u_h||_{L^2(\Omega)} \leq Ch||f||_{L^2(\Omega)}. \tag{70}$$

Remark: there is here a major difference with finite element theory: indeed, the P^1 finite element approximation of the Laplace equation converges to the exact solution with second-order accuracy, while (70) indicates (at least in the general case) first order convergence of the finite volume solution to the exact solution. This is due to the second term in the right-hand side of (65) which is of order one as shown by (67) whereas it vanishes due to Galerkin orthogonality in the finite element approximation. The next section will help us understand that (70) is the optimal estimation we may have in the general case, but that it may be improved in particular cases.

3.2.3 Through the use of Green functions

Let $i \in [1, N]$, we define the Green function g^i associated to the point x_i in T_i in the following way

$$\left\{ \begin{array}{l} -(g^i)'' = \delta_{x_i} \quad \text{in } \mathcal{D}'(\Omega) \\ g^i(0) = g^i(1) = 0 \end{array} \right. , \tag{71}$$

where δ_{x_i} is the Dirac distribution associated with the point x_i: $\langle \delta_{x_i}, \phi \rangle = \phi(x_i)$ for all ϕ in $\mathcal{D}(\Omega)$. This is an easy task to show that the function g^i is given by

$$g^i(x) = \left\{ \begin{array}{ll} (1-x_i)x & \text{if} \quad x \leq x_i \\ x_i(1-x) & \text{if} \quad x \geq x_i \end{array} \right. . \tag{72}$$

Note that this function belongs to $H_0^1(\Omega)$. The fundamental property verified by g^i is the following

$$\hat{u}(x_i) = \int_0^1 g^i(x) f(x) dx. \tag{73}$$

This equality would be straightforward to prove if $\hat{u}$ were in $\mathcal{D}(\Omega)$; however, since this is not the case, we use a direct computation:

$$\int_0^1 g^i(x) f(x) dx = \int_0^{x_i} (1-x_i) x f(x) dx + \int_{x_i}^1 x_i (1-x) f(x) dx.$$

Since $f = -\hat{u}''$, we may use two integrations by parts and the continuity of $\hat{u}'$ in x_i to obtain

$$\begin{array}{rcl} \displaystyle\int_0^1 g^i(x) f(x) dx & = & (1-x_i) \displaystyle\int_0^{x_i} \hat{u}'(x) dx - x_i \int_{x_i}^1 \hat{u}'(x) dx \\ & = & (1-x_i)[\hat{u}(x_i) - \hat{u}(0)] - x_i[\hat{u}(1) - \hat{u}(x_i)] \\ & = & \hat{u}(x_i) \end{array}$$

since $\hat{u}(0) = \hat{u}(1) = 0$.

We may now define the discrete Green function $(G_j^i)_{j \in [0, N+1]}$ associated with the cell T_i by

$$\left\{ \begin{array}{l} -[dg(G^i)]_j = \frac{\delta_j^i}{|T_j|} \ , \forall\, j \ \text{ in } [1, N] \\ G_0^i = G_{N+1}^i = 0 \end{array} \right. , \tag{74}$$

where δ_j^i is the standard Kronecker symbol: $\delta_j^i = 1$ if $i = j$ and $\delta_j^i = 0$ otherwise. System (74) is exactly of the form (12) and (9), for which we have proved existence and uniqueness; therefore (74) really defines a unique set of values $(G_j^i)_{j \in [0, N+1]}$.

Proposition: there holds

$$G_j^i = g^i(x_j) \ \forall\, j \in [0, N+1]. \tag{75}$$

Proof: by uniqueness of the solution of (74), it is enough to prove that the set $(g^i(x_j))_{j \in [0, N+1]}$ verifies the set of equations given by (74). First, the boundary

conditions are obvious since $x_0 = 0$ and $x_{N+1} = 1$, and since $g^i(0) = g^i(1) = 0$ by (71). Moreover, for $0 \leq j \leq i$, there holds $x_j \leq x_i$, therefore $g^i(x_j) = (1 - x_i)x_j$ and

$$[g(g^i)]_{j+1/2} = (1 - x_i)\frac{x_{j+1} - x_j}{|D_{j+1/2}|} = (1 - x_i)\ , \forall j \text{ s. t. } 0 \leq j < i,$$

which implies that

$$[dg(g^i)]_j = \frac{1}{|T_j|}\left\{[g(g^i)]_{j+1/2} - [g(g^i)]_{j-1/2}\right\} = 0\ , \forall j \text{ s. t. } 1 \leq j \leq i-1. \quad (76)$$

In the same way, for $i \leq j \leq N+1$, there holds $x_i \leq x_j$, therefore $g^i(x_j) = (1 - x_j)x_i$ and

$$[g(g^i)]_{j+1/2} = x_i\frac{(1 - x_{j+1}) - (1 - x_j)}{|D_{j+1/2}|} = -x_i\ , \forall j \text{ s. t. } i \leq j \leq N,$$

which implies that

$$[dg(g^i)]_j = \frac{1}{|T_j|}\left\{[g(g^i)]_{j+1/2} - [g(g^i)]_{j-1/2}\right\} = 0\ , \forall j \text{ s. t. } i+1 \leq j \leq N. \quad (77)$$

Finally, when $j = i$, we have on the one hand $[g(g^i)]_{i+1/2} = -x_i$ and, on the other hand $[g(g^i)]_{i-1/2} = (1 - x_i)$. Therefore

$$[dg(g^i)]_i = \frac{1}{|T_i|}\left\{[g(g^i)]_{i+1/2} - [g(g^i)]_{i-1/2}\right\} = -\frac{1}{|T_i|}. \quad (78)$$

Therefore, (76), (77) and (78) show that $-[dg(g^i)]_j = \delta^i_j$ for all j in $[1, N]$, which is exactly the first equation in (74).

The fundamental property verified by G^i is the following

$$u_i = \int_0^1 (g^i)^*(x)\, f(x)dx, \quad (79)$$

with $(g^i)^*(x) = g^i(x_j)$ for all $x \in T_j$ and all j in $[1, N]$.
Proof: by definition of G^i, see (74), there holds, by double application of the discrete Green formula (15), by property (12) and thanks to the boundary conditions on G^i and u

$$\begin{aligned} u_i = \sum_{j=1}^N u_j|T_j|\frac{\delta^i_j}{|T_j|} &= -(u, [dg(G^i)])_T \\ &= -(dg(u), G^i)_T \\ &= (f, G^i)_T \\ &= \sum_{j=1}^N |T_j| f_j G^i_j \\ &= \sum_{j=1}^N g^i(x_j) \int_{T_j} f(x)dx \end{aligned} \quad (80)$$

through (75) and the definition of the mean-value f_j. Finally, by definition of $(g^i)^*$, the right-hand side of (80) exactly equals $\sum_{j=1}^N \int_{T_j} (g^i)^*(x)\, f(x)dx = \int_0^1 (g^i)^*(x)\, f(x)dx$, which proves the result.

Pointwise second-order convergence We shall now investigate two cases in which we may prove second-order convergence; in both cases we shall suppose that f belongs to $H^1(\Omega)$.

Suppose that x_i is the midpoint of T_i for all i in $[1, N]$. Let us define the following piecewise constant function

$$\begin{array}{rccl} \Pi f: & \Omega & \longrightarrow & \mathbb{R} \\ & x & \mapsto & \Pi f(x) = f_j \text{ if } x \in T_j \end{array}$$

According to (73) and (79), there holds

$$\hat{u}(x_i) - u_i = (g^i - (g^i)^*, f)_{L^2(\Omega)} = (g^i - (g^i)^*, f - \Pi f)_{L^2(\Omega)} + (g^i - (g^i)^*, \Pi f)_{L^2(\Omega)}. \tag{81}$$

Since f is in $H^1(\Omega)$ (it actually suffices that f is H^1 on each cell T_j), there exists a constant C which does not depend on the mesh such that

$$||f - \Pi f||_{L^2(T_j)} \leq C|T_j|\, ||f'||_{L^2(T_j)} \leq Ch||f||_{H^1(T_j)}.$$

Therefore,

$$||f - \Pi f||_{L^2(\Omega)} \leq Ch||f||_{H^1(\Omega)}. \tag{82}$$

Moreover

$$\left|\left[g^i - (g^i)^*\right](x)\right| \leq |x - x_j| \sup(x_i, 1 - x_i),\ \forall\, x \in T_j.$$

Since $|x - x_j| \leq h$ and $\sup(x_i, 1 - x_i) \leq 1$, there holds

$$\left|\left[g^i - (g^i)^*\right](x)\right| \leq h\ , \forall\, x \in \Omega. \tag{83}$$

By the Cauchy-Schwarz inequality, we obtain from (82) and (83)

$$\left|(g^i - (g^i)^*, f - \Pi f)_{L^2(\Omega)}\right| \leq Ch^2||f||_{H^1(\Omega)}, \tag{84}$$

which is a bound for the first term in the right-hand side of (81). As far as the second term is concerned, we note that, for $j \neq i$, the function $\left[g^i - (g^i)^*\right] \Pi f$ is affine on T_j; its integral over T_j can then be evaluated exactly by the midpoint rule:

$$\int_{T_j} \left[g^i - (g^i)^*\right](x) \Pi f(x)\, dx = |T_j| \left[g^i - (g^i)^*\right](x_j) \Pi f(x_j)$$

which vanishes thanks to the definition of $(g^i)^*$ given after (79). Therefore, the only remaining contribution to $(g^i - (g^i)^*, \Pi f)_{L^2(\Omega)}$ is that for $j = i$:

$$(g^i - (g^i)^*, \Pi f)_{L^2(\Omega)} = \int_{T_i} \left[g^i - (g^i)^*\right](x) \Pi f(x)\, dx.$$

Since g^i is affine on $[x_{i-1/2}, x_i]$ and also on $[x_i, x_{i+1/2}]$, this integral can be computed by applying the trapezoidal rule on the two subintervals whose lengths are $\frac{|T_i|}{2}$. This results in

$$\begin{aligned}(g^i-(g^i)^*, \Pi f)_{L^2(\Omega)} &= f_i \frac{|T_i|}{4}\left\{\left[g^i-(g^i)^*\right](x_{i-1/2}) + \left[g^i-(g^i)^*\right](x_i)\right\} \\ &+ f_i \frac{|T_i|}{4}\left\{\left[g^i-(g^i)^*\right](x_i) + \left[g^i-(g^i)^*\right](x_{i+1/2})\right\}.\end{aligned}$$

Since $\left[g^i-(g^i)^*\right](x_i) = 0$ and $\left[g^i-(g^i)^*\right](x_{i-1/2}) = -(1-x_i)(x_i - x_{i-1/2})$ and $\left[g^i-(g^i)^*\right](x_{i+1/2}) = -x_i(x_{i+1/2}-x_i)$, there holds

$$\left|(g^i-(g^i)^*, \Pi f)_{L^2(\Omega)}\right| = |f_i|\,\frac{|T_i|^2}{8} \leq \frac{h^2}{8}|f_i|.$$

Since f is in $H^1(\Omega)$ and since Ω is a one dimensional domain, f is bounded by $C||f||_{H^1(\Omega)}$, and so is f_i. Therefore, there holds

$$\left|(g^i-(g^i)^*, \Pi f)_{L^2(\Omega)}\right| \leq \frac{h^2}{8}||f||_{H^1(\Omega)}. \tag{85}$$

Finally, through (81), (84) and (85), there holds

$$|\hat{u}(x_i) - u_i| \leq Ch^2||f||_{H^1(\Omega)},$$

which is a second-order L^∞ bound on the pointwise error due to the finite volume approximation. From this, of course, we obtain a second-order convergence in the (discrete L^2) $||\cdot||_{0,T}$ norm.

Suppose that $x_{i+1/2}$ is the midpoint of $D_{i+1/2}$ for all i in $[1, N-1]$. Let us define the following piecewise constant function

$$\begin{array}{rccl} Pf: & \Omega & \longrightarrow & \mathbb{R} \\ & x & \mapsto & Pf(x) = f_{j+1/2} := \frac{1}{|D_{j+1/2}|}\int_{D_{j+1/2}} f(y)dy \text{ if } x \in D_{j+1/2}. \end{array}$$

According to (73) and (79), there holds

$$\hat{u}(x_i) - u_i = (g^i-(g^i)^*, f)_{L^2(\Omega)} = (g^i-(g^i)^*, f-Pf)_{L^2(\Omega)} + (g^i-(g^i)^*, Pf)_{L^2(\Omega)}. \tag{86}$$

Since f is in $H^1(\Omega)$ (it actually suffices that f is H^1 on each cell $D_{j+1/2}$), we can proceed in the same way as in the previous paragraph to obtain

$$\left|(g^i-(g^i)^*, f-Pf)_{L^2(\Omega)}\right| \leq Ch^2||f||_{H^1(\Omega)}. \tag{87}$$

Moreover, g^i is affine on every $D_{j+1/2}$, thus

$$\int_{D_{j+1/2}} g^i(x)dx = \frac{|D_{j+1/2}|}{2}(g^i(x_j) + g^i(x_{j+1})).$$

On the other hand, by definition, $(g^i)^*$ is a constant on $[x_j, x_{j+1/2}[$ (respectively on $]x_{j+1/2}, x_{j+1}]$) and equals $g^i(x_j)$ (resp. $g^i(x_{j+1})$); and since, for $j \in [1, N-1]$, the point $x_{j+1/2}$ is the midpoint of $D_{j+1/2}$, the lengths of the two subintervals are both equal to $\frac{|D_{j+1/2}|}{2}$. Therefore,

$$\int_{D_{j+1/2}} (g^i)^*(x)dx = \frac{|D_{j+1/2}|}{2}(g^i(x_j) + g^i(x_{j+1})).$$

Thus, for $j \neq 0$ and $j \neq N$, there holds

$$\int_{D_{j+1/2}} \left[g^i - (g^i)^*\right](x) Pf(x)\, dx = 0$$

so that

$$\begin{aligned}(g^i - (g^i)^*, Pf)_{L^2(\Omega)} &= f_{1/2} \int_{D_{1/2}} \left[g^i - (g^i)^*\right](x)dx \\ &+ f_{N+1/2} \int_{D_{N+1/2}} \left[g^i - (g^i)^*\right](x)dx.\end{aligned}$$

Now, it follows from the expressions of g^i and $(g^i)^*$ that $\left|\left[g^i - (g^i)^*\right](x)\right| \leq |D_{1/2}| \leq h$ for all x in $D_{1/2}$ and that $\left|\left[g^i - (g^i)^*\right](x)\right| \leq |D_{N+1/2}| \leq h$ for all x in $D_{N+1/2}$. So that, finally

$$\begin{aligned}\left|(g^i - (g^i)^*, Pf)_{L^2(\Omega)}\right| &\leq h^2(|f_{1/2}| + |f_{N+1/2}|) \\ &\leq Ch^2||f||_{H^1(\Omega)} \qquad (88)\end{aligned}$$

for the same reason as before. Through (86), (87) and (88), we thus conclude that

$$|\hat{u}(x_i) - u_i| \leq Ch^2||f||_{H^1(\Omega)},$$

which is a second-order L^∞ bound on the pointwise error due to the finite volume approximation. From this, of course, we obtain a second-order convergence in the (discrete L^2) $||\cdot||_{0,T}$ norm. This way of proceeding may easily be extended to the case in which there is a bounded (with respect to h) number of cells $D_{j+1/2}$ for which $x_{j+1/2}$ is not the midpoint of $D_{j+1/2}$.

4 Approximation of the Laplace equation in dimension two on admissible meshes

Let $\Gamma_D \neq \emptyset$ and Γ_N by two subsets of $\partial\Omega$ such that $\bar{\Gamma}_D \cup \bar{\Gamma}_N = \partial\Omega$ and $\Gamma_D \cap \Gamma_N = \emptyset$. We are interested in the numerical approximation of the following system

$$\left\{\begin{array}{lll} -\Delta u = f & \text{in} & \Omega \\ u = u_d & \text{on} & \Gamma_D \\ \nabla u \cdot n = G & \text{on} & \Gamma_N \end{array}\right., \qquad (89)$$

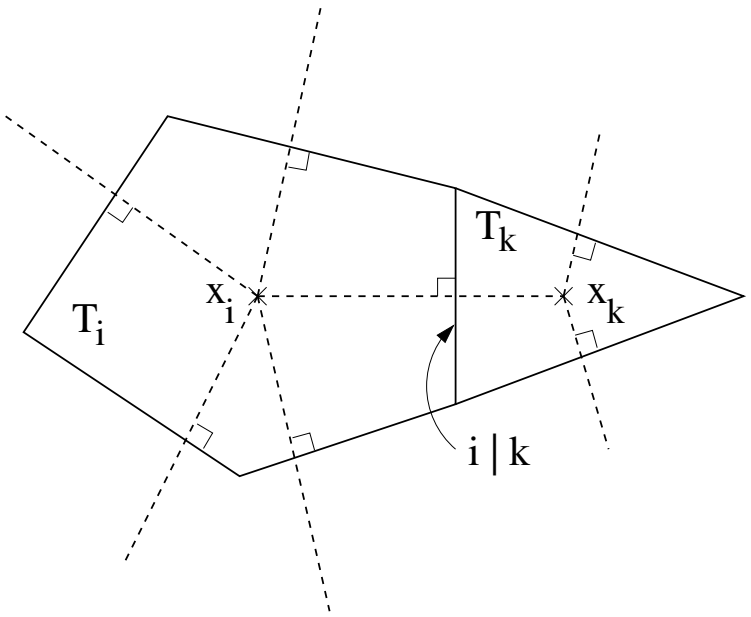

Figure 4: Two neighboring cells of an admissible mesh

where f is a given function in $L^2(\Omega)$, u_d is a given function in $H^{3/2}(\Gamma_D)$, G is a given function in $H^{1/2}(\Gamma_N)$ and n is the outgoing unit vector orthogonal to Γ_N.

4.1 Construction of the finite volume scheme

4.1.1 Admissible meshes

The name "admissible meshes" is borrowed to [1].

Let Ω be a polygonal domain covered by the elements $(T_i)_{i\in[1,N]}$ of a mesh. With each element T_i, we associate a point $x_i \in \overset{\circ}{T}_i$. We shall denote by $i|k$ the common edge of T_i and T_k when these two elements are neighbors. The mesh is said to be admissible if $[x_i x_k]$ is orthogonal to $i|k$ for any couple (T_i, T_k) of neighboring elements, and if, for any element T_i which has an edge on $\partial\Omega$, the orthogonal projection of the associated point x_i on the straight line going over the considered edge belongs to this edge. In that case, the orthogonal projection on the edge is still denoted by x_k, with $k \in [N+1, N+N_b]$, where N_b denotes the number of boundary edges of the mesh, and the edge is still denoted by $i|k$.

For $i \in [1, N]$, we denote by $V(i) \subset [1, N+N_b]$ the set of neighboring indexes of the element T_i. When T_i has one (or more) edge(s) that are on $\partial\Omega$, the set $V(i)$ will contain elements that are strictly greater than N. We shall denote by

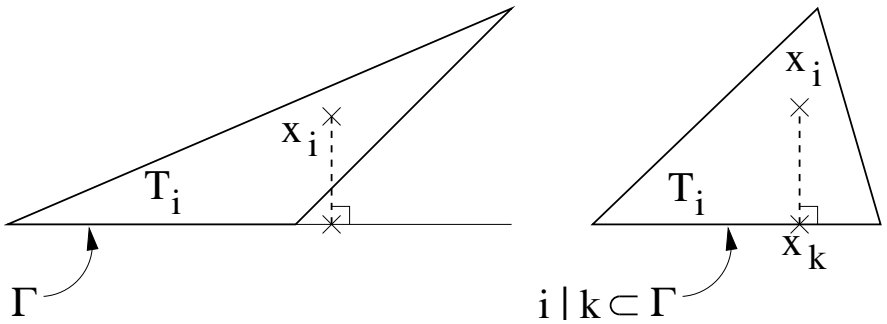

Figure 5: Non-admissibility (left) and admissibility (right) at the boundary

$|T_i|$ the area of T_i, by ℓ_{ik} the length of $i|k$ (note that $\ell_{ik} = \ell_{ki}$), and by n_{ik} the unit vector orthogonal to $i|k$ pointing from T_i to T_k (note that $n_{ik} = -n_{ki}$). Finally, we shall denote by $d_{ik} = d_{ki} = ||\overrightarrow{x_i x_k}||$.

Examples of admissible meshes The first example is a triangular mesh, where the point x_i in T_i is chosen as the center of the circumscribed circle to the vertices of T_i. Since in that case $[x_i x_k]$ is a subset of the orthogonal bisector of the edge $i|k$, this ensures the orthogonality of $[x_i x_k]$ and $i|k$. A necessary and sufficient condition for all the points x_i to be in the interiors of the triangles T_i is that all the angles of all the triangles of the mesh should be strictly lower than $\pi/2$ (see however the remark at the end of subsection 4.1.2).

As a second example of admissible meshes, we may cite the Voronoi mesh associated to a set of points $(x_i)_{i\in[1,N]}$. The element T_i is then defined as the set of points of Ω that are closer to x_i than to any other point x_k, with $k \neq i$:

$$T_i = \{x \in \Omega \text{ s.t. } ||\overrightarrow{xx_i}|| \leq ||\overrightarrow{xx_k}|| \, \forall k \neq i\} .$$

By definition, the common edge of two such neighboring elements T_i and T_k is a subset of the orthogonal bisector of the segment $[x_i x_k]$, which ensures admissibility for interior edges. Additional (but not detailed here) constraints on the boundary Γ and on the set of points $(x_i)_{i\in[1,N]}$ will ensure that the resulting mesh is admissible.

4.1.2 Principle of the scheme

We associate with any finite volume T_i of the mesh an unknown denoted by u_i, which will approach the value $u(x_i)$, and we integrate the first equation in (89) over T_i. Setting f_i as the mean-value of f over T_i, there holds

$$-\int_{T_i} \Delta u(x)\, dx = \int_{T_i} f(x)\, dx = |T_i|\, f_i. \tag{90}$$

The left-hand side of (90) may be evaluated thanks to the Green formula

$$-\int_{T_i} \Delta u(x)\,dx = -\int_{\partial T_i} \nabla u \cdot n(\sigma)\,d\sigma,$$

where n is the unit vector normal to ∂T_i pointing outward T_i. Since ∂T_i is composed of the edges $i|k$ when k runs over $V(i)$, we may write

$$-\int_{T_i} \Delta u(x)\,dx = -\sum_{k\in V(i)} \int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma.$$

Thus, the following formula contains no approximation

$$-\sum_{k\in V(i)} \int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma = |T_i|\, f_i. \tag{91}$$

The construction of the finite volume scheme amounts to the approximation of the flux $\int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma$ as a function of the unknowns of the scheme. When the mesh is admissible, a reasonable approximation of $\nabla u \cdot n_{ik}$ on the edge $i|k$ is given by

$$\nabla u \cdot n_{ik} \approx \frac{u(x_k) - u(x_i)}{d_{ik}}. \tag{92}$$

Indeed, there always holds $u(x_k)-u(x_i) = \int_{x_i}^{x_k} \nabla u \cdot \frac{\overrightarrow{x_i x_k}}{||\overrightarrow{x_i x_k}||}\,(\sigma)\,d\sigma$ and since $\overrightarrow{x_i x_k}$ is orthogonal to $i|k$, then $\frac{\overrightarrow{x_i x_k}}{||\overrightarrow{x_i x_k}||} = n_{ik}$ and $u(x_k)-u(x_i) \approx ||\overrightarrow{x_i x_k}||\nabla u \cdot n_{ik}(M_{ik})$ with second-order accuracy, where M_{ik} is the midpoint of $[x_i x_k]$. Thus, as soon as $k \in [1, N]$ (i.e. if $i|k$ is an interior edge), we may approach

$$\int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma \approx \frac{\ell_{ik}}{d_{ik}}(u_k - u_i).$$

Now, if $i|k \subset \Gamma$, there are two cases

The case of a Dirichlet boundary edge: The formula (92) is still a good approximation of the gradient in the normal direction. However, in that case, $u(x_k)$ is known since $x_k \in \Gamma_D$, according to the second equation in (89). We may thus approach

$$\int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma \approx \frac{\ell_{ik}}{d_{ik}}(u_d(x_k) - u_i).$$

The case of a Neumann boundary edge: This case is very simple since, according to the third equation in (89), the quantity $\nabla u \cdot n_{ik}$ is equal to G. We may thus write

$$\int_{i|k} \nabla u \cdot n_{ik}(\sigma)\,d\sigma = \int_{i|k} G(\sigma)\,d\sigma = \ell_{ik} G_{ik},$$

where we have defined G_{ik} as the mean-value of G over the edge $i|k$. Finally, the ith equation of the scheme thus writes

$$-\sum_{i|k\not\subset\Gamma}\frac{\ell_{ik}}{d_{ik}}(u_k-u_i)-\sum_{i|k\subset\Gamma_D}\frac{\ell_{ik}}{d_{ik}}(u_d(x_k)-u_i)-\sum_{i|k\subset\Gamma_N}\ell_{ik}G_{ik}=|T_i|f_i. \quad (93)$$

Leaving in the left-hand side the unknown terms and in the right-hand side the data, there holds

$$-\sum_{i|k\not\subset\Gamma}\frac{\ell_{ik}}{d_{ik}}(u_k-u_i)+\sum_{i|k\subset\Gamma_D}\frac{\ell_{ik}}{d_{ik}}u_i=$$
$$|T_i|f_i+\sum_{i|k\subset\Gamma_D}\frac{\ell_{ik}}{d_{ik}}u_d(x_k)+\sum_{i|k\subset\Gamma_N}\ell_{ik}G_{ik}. \quad (94)$$

Remark: if the mesh is not admissible, then for an edge $i|k$, the direction $\overrightarrow{x_ix_k}$ is not orthogonal to $i|k$. Let us denote by ν_{ik} the unit vector pointing from x_i to x_k and τ_{ik} the unit vector such that (ν_{ik},τ_{ik}) is a positively-oriented orthonormal basis of $\mathbb{R}^2$. Then $\overrightarrow{x_ix_k}=d_{ik}\nu_{ik}$ and a reasonable approximation of ∇u in the direction $\overrightarrow{x_ix_k}$ is still given by

$$\nabla u\cdot\nu_{ik}\approx\frac{u(x_k)-u(x_i)}{d_{ik}}.$$

But what we are looking for is an approximation of $\nabla u\cdot n_{ik}$; if we decompose ∇u in the basis (ν_{ik},τ_{ik}), we obtain

$$\nabla u=(\nabla u\cdot\nu_{ik})\nu_{ik}+(\nabla u\cdot\tau_{ik})\tau_{ik}$$

and thus

$$\nabla u\cdot n_{ik}\approx\frac{u_k-u_i}{d_{ik}}\,\nu_{ik}\cdot n_{ik}+(\nabla u\cdot\tau_{ik})\,\tau_{ik}\cdot n_{ik}. \quad (95)$$

When ν_{ik} and n_{ik} are not the same direction, then $\tau_{ik}\cdot n_{ik}\neq 0$ and a good approximation of $\nabla u\cdot\tau_{ik}$ is thus necessary to get a good approximation of $\nabla u\cdot n_{ik}$; this is not an obvious task, and we postpone this question to section 6.

Remark: Up to now, we have supposed that the point x_i belongs to $\overset{\circ}{T}_i$, so that d_{ik} can never vanish, and the division by d_{ik} is totally licit. However, in practice, it may happen that a mesh generator which generates Delaunay triangulations may produce right-angled triangles. Two cases may occur that will lead to a situation where $x_i=x_k$, and thus to a vanishing length d_{ik}, for which one can no longer write an equation such as (94). The first case is if two such right-angled triangles T_i and T_k share their longest edge. The remedy in this case is to merge T_i and T_k into a single element $T_j=T_i\cup T_k$, whose associated point will be $x_j=x_i=x_k$. It is an easy matter to verify that the resulting mesh is admissible (see Fig. 6). The second case occurs if the longest edge of a right-angled triangle T_i is a subset of Γ_D. Since x_i in that case is itself on Γ_D, where u is known to be equal to u_d, then the remedy is simply to replace the ith equation by the equation $u_i=u_d(x_i)$ (see Fig 7).

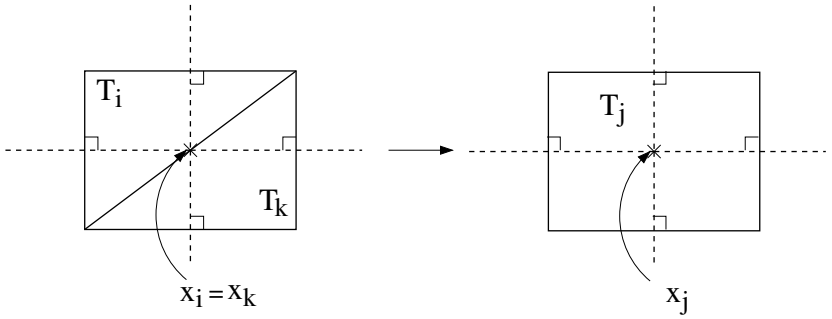

Figure 6: Merging of two right-angled triangles into a single rectangle

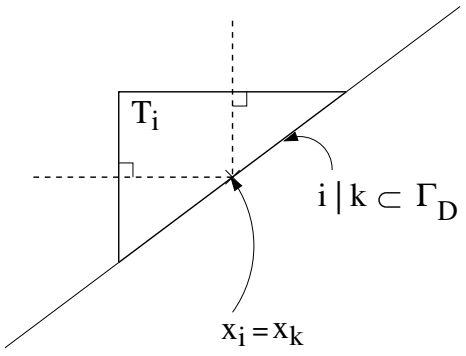

Figure 7: A right-angled triangle with its longest edge on the Dirichlet boundary

4.2 Properties of the scheme

4.2.1 Conservativity

Formula (93) may be rewritten as

$$\sum_{k \in V(i)} F_{ik} = |T_i| f_i.$$

The conservativity principle states that if $i|k$ is an internal edge (not on the boundary), then $F_{ik} + F_{ki} = 0$.

4.2.2 Definitions of discrete differential operators and discrete scalar products

We shall denote by E the set of all the edges in the mesh and by N_e the number of edges in the set E (including the N_b boundary edges). We recall that N is the number of elements in the mesh. We define the following discrete divergence operator

$$\begin{array}{rcl} d : \mathbb{R}^{N_e} & \longrightarrow & \mathbb{R}^N \\ (v_{ik})_{ik \in E} & \mapsto & (dv)_i := \dfrac{1}{|T_i|} \displaystyle\sum_{k \in V(i)} \ell_{ik} v_{ik}. \end{array} \tag{96}$$

Note that in this definition, the following convention has been used

$$v_{ki} = -v_{ik}. \tag{97}$$

We also introduce the operator g (discrete normal gradient)

$$\begin{array}{rcl} g : \mathbb{R}^{N+N_b} & \longrightarrow & \mathbb{R}^{N_e} \\ (u_i) & \mapsto & (gu)_{ik} := \dfrac{u_k - u_i}{d_{ik}} \end{array} \tag{98}$$

We note that $(gu)_{ki} = -(gu)_{ik}$. With these definitions, the scheme (94) may simply be rewritten

$$-(dgu)_i = f_i \ , \ \forall i \in [1, N], \tag{99}$$

with boundary values u_k, $k \in [N+1, N+N_b]$ given by

$$u_k = u_d(x_k) \text{ if } x_k \in \Gamma_D \tag{100}$$

$$u_k = u_i + d_{ik} G_{ik} \text{ if } x_k \in \Gamma_N. \tag{101}$$

On the primal mesh, we also define the discrete scalar product $(\cdot, \cdot)_T$ by

$$(u_i)_{i \in [1,N]}, (w_i)_{i \in [1,N]} \mapsto (u, w)_T := \sum_{i \in [1,N]} |T_i| u_i w_i. \tag{102}$$

And on the dual mesh, associated with the edges, we define the discrete scalar product $(\cdot,\cdot)_D$ by

$$(a_{ik})_{ik\in E}, (b_{ik})_{ik\in E} \mapsto (a,b)_D := \sum_{ik\in E} \frac{d_{ik}\ell_{ik}}{2} a_{ik}b_{ik}. \tag{103}$$

Remark: Note that $\frac{d_{ik}\ell_{ik}}{2}$ is the area of the quadrilateral $x_i S_{ik1} S_{ik2} x_k$, where S_{ik1} and S_{ik2} are the two vertices of the edge $i|k$. Therefore, $\sum_{ik\in E} \frac{d_{ik}\ell_{ik}}{2} = |\Omega|$.

Finally, we define boundary scalar product on Γ, on Γ_D and Γ_N by

$$(a_{ik})_{i|k\subset\Gamma}, (b_{ik})_{i|k\subset\Gamma} \mapsto (a,b)_{\Gamma_h} := \sum_{i|k\subset\Gamma} \ell_{ik} a_{ik} b_{ik} \tag{104}$$

and by similar expressions by replacing Γ by Γ_D or Γ_N.

Proposition: Let $(u_i)_{i\in[1,N+N_b]}, (v_{ik})_{ik\in E}$ be given. There holds

$$(dv,u)_T = -2(v,gu)_D + (v,\gamma u)_{\Gamma_h}, \tag{105}$$

where the discrete trace operator γ is defined in the following way

$$\begin{array}{rcl} \gamma : \mathbb{R}^{N+N_b} & \longrightarrow & \mathbb{R}^{N_b} \\ (u_i)_{i\in[1,N+N_b]} & \mapsto & (\gamma u)_{ik} := u_k \text{ when } x_k \in \Gamma. \end{array} \tag{106}$$

Proof: There holds

$$(dv,u)_T = \sum_{i\in[1,N]} |T_i|\,(dv)_i\,u_i = \sum_{i\in[1,N]} \sum_{k\in V(i)} \ell_{ik} v_{ik}\,u_i \tag{107}$$

according to (102) and (96). Now, in the right-hand side of (107), for a given interface $i_0|k_0 \in E$, the term $v_{i_0k_0}$ appears twice if $i_0|k_0$ is an interior edge: once when $i = i_0$ and once when $i = k_0$. In the first case, the term $v_{i_0k_0}$ is multiplied by $\ell_{i_0k_0}u_{i_0}$ and in the second case, it is multiplied by $-\ell_{i_0k_0}u_{k_0}$, since, by (97), $v_{k_0i_0} = -v_{i_0k_0}$ and $\ell_{i_0k_0} = \ell_{k_0i_0}$. On the other hand, the term $v_{i_0k_0}$ appears only once if $i_0|k_0$ is a boundary edge, and it is multiplied by $\ell_{i_0k_0}u_{i_0} = \ell_{i_0k_0}(u_{i_0} - u_{k_0}) + \ell_{i_0k_0}u_{k_0}$. Thus, rearranging the sum in (107), there holds

$$\begin{aligned} (dv,u)_T &= \sum_{ik\in E} \ell_{ik}v_{ik}(u_i - u_k) + \sum_{i|k\subset\Gamma} \ell_{ik}v_{ik}u_k \\ &= -2\sum_{ik\in E} \frac{d_{ik}\ell_{ik}}{2} v_{ik} \frac{(u_k - u_i)}{d_{ik}} + \sum_{i|k\subset\Gamma} \ell_{ik}v_{ik}(\gamma u)_{ik} \\ &= -2(v,gu)_D + (v,\gamma u)_{\Gamma_h}. \end{aligned}$$

Remark: The discrete Green formula (105) is the discrete equivalent of

$$(\nabla\cdot v, u)_{L^2(\Omega)} = -(v,\nabla u)_{L^2(\Omega)} + (v\cdot n, u)_{L^2(\Gamma)}.$$

The factor 2 in front of the scalar product $(v, gu)_D$ may be interpreted by the fact that the continuous scalar product $(v, \nabla u)_{L^2(\Omega)}$ involves vector fields v and ∇u, the dot product of which is composed of two terms: the product of their normal components and the product of their tangential components. Since in the discrete scalar product only the normal components are involved, the factor 2 may be interpreted as "a compensation" for this "half" scalar product.

4.2.3 A discrete variational formulation for the finite volume scheme

Consider any $(w_i)_{i\in[1,N+N_b]}$, with $w_k = 0$ for all k such that $x_k \in \Gamma_D$. There holds

$$2(gu, gw)_D = (f, w)_T + (G, \gamma w)_{\Gamma_{N\,h}}. \tag{108}$$

Proof: Let us start from (99), then multiply by $|T_i|w_i$ and sum over $i \in [1, N]$. We obtain

$$-(dgu, w)_T = (f, w)_T.$$

Combining this with the discrete Green formula (105), we obtain

$$2(gu, gw)_D - (gu, \gamma w)_{\Gamma_h} = (f, w)_T. \tag{109}$$

Now, on Γ_D, γw vanishes, while on Γ_N, Eq. (98) and (101) imply that $(gu)_{ik} = G_{ik}$, and therefore (109) implies (108).

4.2.4 Existence and uniqueness of the discrete solution

The finite volume scheme may be written as a system of $N + N_b$ unknowns (one per cell T_i and one per boundary edge) and $N + N_b$ equations given by (100), (101) and (99). Therefore, existence for all data (f, u_d, G) and uniqueness are equivalent. Moreover, uniqueness may be proved by injectivity: if $f_i = 0$ for all $i \in [1, N]$, if $u_d(x_k) = 0$ for all $x_k \in \Gamma_D$ and if $G_{ik} = 0$ for all $x_k \in \Gamma_N$, then we may choose $u = w$ in (108) and there holds, thanks to (103)

$$(gu, gu)_D = 0 = \sum_{ik\in E} \frac{d_{ik}\ell_{ik}}{2}(gu)_{ik}^2.$$

Since $d_{ik} \neq 0$ and $\ell_{ik} \neq 0$, this implies that $(gu)_{ik} = 0$ for all $ik \in E$. By definition of $(gu)_{ik}$, this means that $u_i = u_k$ for all $i|k$ (including the boundary edges). Therefore, there exists a constant c such that $u_i = c$ for all $i \in [1, N + N_b]$, and, by the homogeneous Dirichlet boundary conditions (since $u_d(x_k)$ vanishes), this constant is equal to 0. Thus, $u_i = 0$ for all $i \in [1, N + N_b]$, which means injectivity and thus existence for all data (f, u_d, G) and uniqueness.

4.2.5 Discrete maximum principle

We suppose that f is positive on Ω and that $u_k = 0$, for all $k \in [N + 1, N + N_b]$ (homogeneous Dirichlet boundary conditions). We wish to show that the discrete solution is positive on Ω, i.e. that $u_i \geq 0$ for all $i \in [1, N]$. Actually,

we shall prove a stronger property, which is the absence of local minimum: Let i be in $[1, N]$; if $f \geq 0$ on T_i, then $u_i \geq \min_{k \in V(i)}(u_k)$, and $u_i = \min_{k \in V(i)}(u_k)$ if and only if $u_i = u_k$ for all $k \in V(i)$ and $f_i = 0$.

Proof: if $f \geq 0$ on T_i, then $f_i \geq 0$, which implies that, according to (94)

$$- \sum_{k \in V(i)} \frac{\ell_{ik}}{d_{ik}}(u_k - u_i) = |T_i| f_i \geq 0. \tag{110}$$

Now $u_i < \min_{k \in V(i)}(u_k)$ is impossible since the left-hand side of (110) would be the sum of non positive terms, one of which would be strictly negative. Moreover, if $u_i = \min_{k \in V(i)}(u_k)$, then the left-hand side of (110) is non strictly negative if and only if $u_i = u_k$ for all $k \in V(i)$ and thus $f_i = 0$.

Now, for $f \geq 0$, if we suppose that there exists an $i \in [1, N]$ such that $u_i < 0$, then we may consider an index i_0 in which the minimum value is reached: $u_{i_0} = \min_{i \in [1,N]} u_i < 0$. Since $u_k = 0$ for all $k \in [N+1, N+N_b]$, there also holds $u_{i_0} = \min_{i \in [1, N+N_b]} u_i < 0$. By definition, since u_{i_0} is the minimum, there holds $u_{i_0} \leq \min_{k \in V(i_0)}(u_k)$; moreover, by absence of local minimum, $u_{i_0} \geq \min_{k \in V(i_0)}(u_k)$; therefore

$$u_{i_0} = \min_{k \in V(i_0)}(u_k)$$

and then $u_{i_0} = u_k = \min_{i \in [1,N]} u_i$ for all $k \in V(i_0)$. This means that the minimum of u is reached in all neighboring elements of T_{i_0}, in which we may repeat the same reasoning to prove that the minimum of u is also reached in the neighbors of the neighbors of T_{i_0}. We may easily infer that all u_i, for all $i \in [1, N]$, would be equal for to the strictly negative minimum of u. If we repeat again the same reasoning for a T_i which has at least one of its edges $i|k$ on the boundary, then we prove that the corresponding u_k is also equal to this strictly negative minimum of u, which is impossible since $u_k = 0$ on the boundary.

5 Error estimation for the approximation of the Laplace equation on admissible meshes

We shall give error estimates in both a discrete H_0^1 norm (actually a norm related to the discrete normal gradient) and in a discrete L^2 norm. We shall denote by $h := \sup_{i \in [1,N]} diam(T_i)$ the mesh step size. We shall suppose that $\hat{u}$, the exact solution of (89) with homogeneous Dirichlet boundary conditions ($u_d = 0$ and $\Gamma_N = \emptyset$) belongs to $H^2(\Omega)$. Note that the regularity of $\hat{u}$ depends both on that of f and on that of the boundary Γ. It is for example sufficient that Ω is a convex polygon and that f belongs to $L^2(\Omega)$. Since, in dimension two, the functions of $H^2(\Omega)$ also belong to $C_0(\bar{\Omega})$, we may consider the pointwise projection of $\hat{u}$ defined as follows

$$(\Pi \hat{u})_i = \hat{u}(x_i) \ \forall i \in [1, N + N_b].$$

Note that this implies that

$$(\Pi\hat{u})_k = 0 \ \forall k \in [N+1, N+N_b]. \tag{111}$$

We shall estimate the difference between the projection $((\Pi\hat{u})_i)_{i\in[1,N+N_b]}$ and the values $(u_i)_{i\in[1,N+N_b]}$ obtained from the finite volume scheme (99) with boundary conditions

$$u_k = 0 \ \forall k \in [N+1, N+N_b]. \tag{112}$$

5.1 Estimation in the energy norm

We wish to estimate the norm of the discrete normal gradient of the error

$$|u - \Pi\hat{u}|_{1,D} := (g(u - \Pi\hat{u}), g(u - \Pi\hat{u}))_D^{1/2}. \tag{113}$$

For this, let us first define the following averaged normal gradient on each edge $i|k$:

$$(\delta\hat{u})_{ik} := \frac{1}{\ell_{ik}} \int_{i|k} \nabla\hat{u} \cdot n_{ik}(\sigma) d\sigma.$$

Note that this implies that $(\delta\hat{u})_{ik} = -(\delta\hat{u})_{ki}$.

Lemma: there holds, for any $(w_i)_{i\in[1,N+N_b]}$ with $w_k = 0$ for all $k \in [N+1, N+N_b]$

$$(f, w)_T = 2(\delta\hat{u}, gw)_D. \tag{114}$$

Proof: By definition of f_i, there holds

$$\begin{aligned} |T_i| f_i &= \int_{T_i} f(x)dx = -\int_{T_i} \Delta\hat{u}(x)dx = -\int_{\partial T_i} \nabla\hat{u} \cdot n(\sigma) d\sigma \\ &= -\sum_{k\in V(i)} \int_{i|k} \nabla\hat{u} \cdot n_{ik}(\sigma) d\sigma = -\sum_{k\in V(i)} \ell_{ik} (\delta\hat{u})_{ik} \\ &= -|T_i| (d\delta\hat{u})_i \end{aligned} \tag{115}$$

according to definition (96). The fact that $(\delta\hat{u})_{ik} = -(\delta\hat{u})_{ki}$ is consistent with the convention (97). Now (115) and (105) imply that

$$(f, w)_T = -(d(\delta\hat{u}), w)_T = 2(\delta\hat{u}, gw)_D$$

since w vanishes on the boundary.

Now, thanks to (108) and (114), there holds

$$(\delta\hat{u}, gw)_D = (gu, gw)_D \tag{116}$$

for all w vanishing on the boundary. Setting $w = u - \Pi u$ in (113) and since w vanishes on the boundary thanks to (111) and (112), we may write, thanks to (116) and to the Cauchy-Schwarz inequality

$$\begin{aligned} |w|_{1,D}^2 &= |u - \Pi\hat{u}|_{1,D}^2 = (g(u - \Pi\hat{u}), gw)_D = (\delta\hat{u} - g(\Pi\hat{u}), gw)_D \\ &\leq |\delta\hat{u} - g(\Pi\hat{u})|_{0,D} |w|_{1,D} \end{aligned}$$

and thus,

$$|w|_{1,D} \leq |\delta\hat{u} - g(\Pi\hat{u})|_{0,D}. \tag{117}$$

The right-hand side of (117) contains terms which are only related to the exact solution of (89) and not to the discrete solution of the scheme any more, which is always a good starting point to perform numerical analysis. In order to evaluate this right-hand side, we shall use a technique which is often summarized in the numerical analysis literature as "the Bramble-Hilbert lemma and a scaling argument". Let us detail how this technique may be adapted in the present case.

First, the "scaling argument". This is nothing but a change of variables. Let us consider the quadrilateral $x_i S_{ik1} S_{ik2} x_k$, where S_{ik1} and S_{ik2} are the two vertices of the edge $i|k$. Since the segments $[x_i x_k]$ and $[S_{ik1}S_{ik2}]$ are mutually orthogonal, we may suppose, without loss of generality, that they are the axis of an orthogonal basis, and that their intersection is the origin, so that $[S_{ik1}S_{ik2}] = \left\{(x,0) \text{ s.t. } x \in [-\frac{\ell_{ik}}{2}, \frac{\ell_{ik}}{2}]\right\}$ and $[x_i x_k] = \left\{(0,y) \text{ s.t. } y \in [-d_{ik}^-, d_{ik}^+]\right\}$ with $d_{ik}^+ + d_{ik}^- = d_{ik}$. Then, we define the affine transform $\mathcal{T}$ by

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \mathcal{T}\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} \frac{2x}{\ell_{ik}} \\ \frac{2y - d_{ik}^+ + d_{ik}^-}{d_{ik}} \end{pmatrix} \tag{118}$$

so that the diamond-cell $\mathcal{D}_{ik} = x_i S_{ik1} S_{ik2} x_k$ is mapped onto the reference quadrilateral $\hat{\mathcal{D}} = (0,-1),(1,0),(0,1),(-1,0)$. Moreover, let us set

$$v(x,y) = \nabla\hat{u} \cdot n_{ik}(x,y) = \frac{\partial \hat{u}}{\partial y}(x,y)$$

and

$$\hat{v}(\hat{x},\hat{y}) := v(x,y) = v \circ \mathcal{T}^{-1}(\hat{x},\hat{y}).$$

If $\hat{u} \in H^2(\Omega)$, then $v \in H^1(\mathcal{D}_{ik})$ and $\hat{v} \in H^1(\hat{\mathcal{D}})$ and moreover, by (118), there holds

$$\nabla\hat{v}(\hat{x},\hat{y}) = \frac{1}{2}\begin{pmatrix} \ell_{ik} & 0 \\ 0 & d_{ik} \end{pmatrix} \nabla v(x,y)$$

so that

$$\left|\nabla\hat{v}(\hat{x},\hat{y})\right|^2 \leq \frac{1}{4}\max(\ell_{ik}^2, d_{ik}^2)\left|\nabla v(x,y)\right|^2$$

and then

$$||\nabla\hat{v}||_{L^2(\hat{D})} \leq \frac{1}{2}\max(\ell_{ik}, d_{ik})\left|\det\begin{pmatrix} 2\ell_{ik}^{-1} & 0 \\ 0 & 2d_{ik}^{-1} \end{pmatrix}\right|^{1/2} ||\nabla v||_{L^2(\mathcal{D}_{ik})}$$

which means

$$|\hat{v}|_{1,\hat{D}} \leq \max\left(\sqrt{\frac{\ell_{ik}}{d_{ik}}}, \sqrt{\frac{d_{ik}}{\ell_{ik}}}\right)|v|_{1,\mathcal{D}_{ik}}. \tag{119}$$

Then, let us define on each quadrilateral cell

$$\begin{aligned} e_{ik}(v) &= (\delta\hat{u})_{ik} - (g(\Pi\hat{u}))_{ik} \\ &= \frac{1}{\ell_{ik}} \int_{i|k} \nabla\hat{u} \cdot n_{ik}(\sigma) d\sigma - \frac{1}{d_{ik}} \int_{[S_{ik2}S_{ik1}]} \nabla\hat{u} \cdot n_{ik}(\sigma) d\sigma \\ &= \frac{1}{\ell_{ik}} \int_{-\frac{\ell_{ik}}{2}}^{\frac{\ell_{ik}}{2}} v(x,0) dx - \frac{1}{d_{ik}} \int_{-d_{ik}^-}^{d_{ik}^+} v(0,y) dy. \end{aligned} \tag{120}$$

Through the change of variables given by (118) in the integrals in the right-hand side of (120), there holds

$$e_{ik}(v) = \hat{e}_{ik}(\hat{v}) := \frac{1}{2} \int_{-1}^{1} \hat{v}(\hat{x}, \hat{y}_{ik}) d\hat{x} - \frac{1}{2} \int_{-1}^{1} \hat{v}(0,\hat{y}) d\hat{y}. \tag{121}$$

with $\hat{y}_{ik} := \frac{-d_{ik}^+ + d_{ik}^-}{d_{ik}} \in [-1,1]$. Now comes the Bramble-Hilbert lemma: $\hat{e}_{ik}$ is a linear application which is continuous on $H^1(\hat{D})$. Moreover, it vanishes over the set of constant functions. Then, there holds

$$|\hat{e}_{ik}(\hat{v})| \leq \|\hat{e}_{ik}\| \, |\hat{v}|_{1,\hat{D}} \tag{122}$$

for all $\hat{v}$ in $H^1(\hat{D})$, where $\|\hat{e}_{ik}\|$ is the continuity norm of $\hat{e}_{ik}$. Note that the definition (121) implies that $\|\hat{e}_{ik}\|$ only depends on $\hat{y}_{ik}$, but it can be shown (and we admit it here) that $\|\hat{e}_{ik}\|$ may be bounded uniformly by a constant $\hat{C}$ which does not depend on $\hat{y}_{ik}$, which means that $\hat{C}$ does not depend on the particular diamond-cell $\mathcal{D}_{ik}$ at all. Finally, through (122) and (119), there holds

$$|e_{ik}(v)| \leq \hat{C} |\hat{v}|_{1,\hat{D}} \leq \hat{C} \max\left(\sqrt{\frac{\ell_{ik}}{d_{ik}}}, \sqrt{\frac{d_{ik}}{\ell_{ik}}} \right) |v|_{1,\mathcal{D}_{ik}}.$$

Then, by (117) and the definition of $e_{ik}(v)$, there holds

$$\begin{aligned} |w|_{1,D}^2 &\leq \hat{C}^2 \sum_{i|k} \frac{\ell_{ik} d_{ik}}{2} \max\left(\frac{\ell_{ik}}{d_{ik}}, \frac{d_{ik}}{\ell_{ik}} \right) |v|_{1,\mathcal{D}_{ik}}^2 \\ &\leq \hat{C}^2 \max_{i|k} \left(\ell_{ik}^2, d_{ik}^2 \right) \sum_{i|k} |v|_{1,\mathcal{D}_{ik}}^2. \end{aligned}$$

But of course $|v|_{1,\mathcal{D}_{ik}} \leq |\hat{u}|_{H^2(D_{ik})}$, and $\ell_{ik} \leq h$ and $d_{ik} \leq 2h$, so that finally

$$|u - \Pi\hat{u}|_{1,D} \leq Ch \, |\hat{u}|_{H^2(\Omega)}, \tag{123}$$

where the constant C does not depend on the mesh. This is a first-order convergence in the discrete gradient.

5.2 Estimation in the L^2 norm

We shall use the following discrete Poincaré inequality: Let $(w_i)_{i\in[1,N+N_b]}$ such that $w_k = 0$ for all $x_k \in \Gamma$. Then, there exists a constant $C = 2|\Omega|^{1/2}$ such that

$$||w||_{0,T} = \left(\sum_{i=1}^{N} |T_i| w_i^2\right)^{1/2} \leq C|w|_{1,D}. \tag{124}$$

Proof: Let us define the following function

$$\begin{array}{rcl} \omega : \Omega & \longrightarrow & \mathbb{R} \\ x & \mapsto & \omega(x) = w_i \ \text{ if } x \in T_i. \end{array} \tag{125}$$

Let $x \in \Omega$ be given, we define by D_x^1 and D_x^2 the two straight lines going through x with direction $(1,0)$ and $(0,1)$ respectively. For a given $i|k$ in the set of edges E and a given $x \in \Omega$, we also define the functions χ_{ik}^j, with $j = 1$ and $j = 2$ by

$$\begin{array}{rcl} \chi_{ik}^j : \Omega & \longrightarrow & \mathbb{R} \\ x & \mapsto & \chi_{ik}^j(x) = \left\{ \begin{array}{lll} 1 & \text{if} & i|k \cap D_x^j \neq \emptyset \\ 0 & \text{if} & i|k \cap D_x^j = \emptyset \end{array} \right. . \end{array} \tag{126}$$

Then, for a given T_i and all $x \in T_i$, there holds

$$\omega(x) = w_i = (w_i - w_{k_1}) + (w_{k_1} - w_{k_2}) + \cdots + (w_{k_{q-1}} - w_{k_q}) + (w_{k_q} - w_k),$$

where $i|k_1$, $k_1|k_2$, $\cdots$, $k_{q-1}|k_q$, and $k_q|k$ are the edges of the mesh which are successively intersected by D_x^1 and where the index k is such that x_k belongs to an edge of the mesh which is included in Γ, so that $w_k = 0$. In general such an x_k is not unique, so that one chooses any such x_k. Since

$$w_{k_\ell} - w_{k_{\ell+1}} = d_{k_\ell k_{\ell+1}} (gw)_{k_\ell k_{\ell+1}},$$

there holds

$$|\omega(x)| \leq \sum_{k_1|k_2 \in E} d_{k_1 k_2} |(gw)_{k_1 k_2}| \chi_{k_1 k_2}^1(x).$$

Performing the same calculation with $j = 2$ instead of $j = 1$ and multiplying the two inequalities, there holds

$$|\omega(x)|^2 \leq \left(\sum_{k_1|k_2 \in E} d_{k_1 k_2} |(gw)_{k_1 k_2}| \chi_{k_1 k_2}^1(x)\right) \left(\sum_{k_1|k_2 \in E} d_{k_1 k_2} |(gw)_{k_1 k_2}| \chi_{k_1 k_2}^2(x)\right).$$

Integrating over Ω, and taking into account that ω is a constant over each T_i, there holds

$$\begin{array}{rcl} \displaystyle\sum_{i=1}^{N} |T_i|\, w_i^2 & \leq & \displaystyle\int_\Omega \left(\sum_{k_1|k_2 \in E} d_{k_1 k_2} |(gw)_{k_1 k_2}| \chi_{k_1 k_2}^1(x)\right) \\ & & \displaystyle\left(\sum_{k_1|k_2 \in E} d_{k_1 k_2} |(gw)_{k_1 k_2}| \chi_{k_1 k_2}^2(x)\right) dx_1 dx_2. \end{array}$$

Now, it is easily seen that the function $\chi^1_{k_1k_2}$ only depends on x_2 and that $\chi^2_{k_1k_2}$ only depends on x_1, so that, setting

$$a := \min\{x_1 \text{ s.t. } (x_1, x_2) \in \Omega\}, b := \max\{x_1 \text{ s.t. } (x_1, x_2) \in \Omega\}$$

$$\alpha := \min\{x_2 \text{ s.t. } (x_1, x_2) \in \Omega\}, \beta := \max\{x_2 \text{ s.t. } (x_1, x_2) \in \Omega\}$$

we get

$$\begin{aligned}\sum_{i=1}^{N} |T_i|\, w_i^2 \quad \leq \quad & \int_\alpha^\beta \sum_{k_1|k_2 \in E} d_{k_1k_2} |(gw)_{k_1k_2}| \chi^1_{k_1k_2}(x_2)\, dx_2 \\ & \int_a^b \sum_{k_1|k_2 \in E} d_{k_1k_2} |(gw)_{k_1k_2}| \chi^2_{k_1k_2}(x_1)\, dx_1.\end{aligned}$$

It is also easily seen that

$$\int_\alpha^\beta \chi^1_{k_1k_2}(x_2)\, dx_2 \leq \ell_{k_1k_2} \ ; \ \int_a^b \chi^2_{k_1k_2}(x_1)\, dx_1 \leq \ell_{k_1k_2},$$

so that we finally get

$$\sum_{i=1}^{N} |T_i|\, w_i^2 \leq \left(\sum_{k_1|k_2 \in E} d_{k_1k_2} \ell_{k_1k_2} |(gw)_{k_1k_2}| \right)^2.$$

Applying the discrete Cauchy-Schwarz inequality, there holds

$$\sum_{i=1}^{N} |T_i|\, w_i^2 \leq 4 \left(\sum_{k_1|k_2 \in E} \frac{d_{k_1k_2} \ell_{k_1k_2}}{2} \right) \left(\sum_{k_1|k_2 \in E} \frac{d_{k_1k_2} \ell_{k_1k_2}}{2} |(gw)_{k_1k_2}|^2 \right),$$

which is exactly the discrete Poincaré inequality (124).

Having proved this, the following estimation in the discrete L^2 norm follows directly from the estimation in the discrete energy norm (123): there exists a constant C which does not depend on h such that:

$$||u - \Pi\hat{u}||_{0,T} \leq Ch\, |\hat{u}|_{H^2(\Omega)}. \tag{127}$$

Note that, when $\hat{u}$ is in $H^3(\Omega)$, the estimate (127) is suboptimal on admissible meshes made up of rectangular cells in which the associated points are the cell centers, since in that case, second-order convergence has been proved; it is also probably suboptimal on Delaunay triangulations in which the associated points are the cell circumcenters, since numerical computations also indicate second-order convergence, although this fact has never been proved.

6 Approximation of anisotropic diffusion equations on non orthogonal meshes

6.1 Statement of the problem

The construction of the finite volume scheme for the Laplace equation in section 4 and its error analysis in section 5 are based on the notion of "admissible meshes". However, there are a number of circumstances in which it may not be always possible to construct the family of points (x_i) such that the resulting segments $[x_i x_k]$ are all orthogonal to their associated interfaces $i|k$. This is for example the case of nonconforming meshes, where this is impossible. Another problem might come from anisotropic diffusion equations, as we shall see now. Let K be a given symmetric positive definite (SPD) matrix. Consider the diffusion equation with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -\nabla \cdot K\nabla u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{128}$$

Let us also consider a partition of Ω by elements $(T_i)_{i\in[1,N]}$ and integrate both sides of Eq. (128) on T_i. Performing like in section 4.1.2, we end up with a formula which is much alike (91):

$$-\sum_{k\in V(i)} \int_{i|k} (K\nabla u)\cdot n_{ik}(\sigma)\, d\sigma = |T_i|\, f_i. \tag{129}$$

Approaching the integrals in the left-hand side of (129), denoting by $(\nabla_h u)_{ik}$ an approximation of ∇u along $i|k$, and applying the equality $[K(\nabla_h u)_{ik}]\cdot n_{ik} = (\nabla_h u)_{ik}\cdot(K^T n_{ik}) = (\nabla_h u)_{ik}\cdot(K n_{ik})$ (since K has been supposed to be symmetric), we end up with the following equation

$$-\sum_{k\in V(i)} \ell_{ik}(\nabla_h u)_{ik}\cdot(K n_{ik}) = |T_i|\, f_i, \tag{130}$$

which shows that we need an approximation of the gradient of u in the direction Kn_{ik}. It will be in general very difficult, even on triangular meshes, to find a set of points (x_i) associated to the elements (T_i) such that the resulting segments $[x_i x_k]$ will all be collinear to the directions (Kn_{ik}), in particular when K is anisotropic. Note however that since K is positive definite, then $(Kn_{ik})\cdot n_{ik} \geq \alpha||n_{ik}||^2 = \alpha > 0$, which means that n_{ik} and Kn_{ik} are both oriented from T_i to T_k. In order to find the approximation of ∇u in the direction Kn_{ik}, a first possibility is to construct only that direction of the gradient (like we approached only the normal component gu of the gradient on admissible meshes), while a second possibility is first to reconstruct the whole gradient $(\nabla_h u)_{ik}$, then perform the matrix-vector product $K(\nabla_h u)_{ik}$ and finally take the dot product of the resulting vector with n_{ik}.

6.2 Construction of the gradient in the direction Kn_{ik}.

Let us denote by $\eta_{ik} = \frac{1}{||Kn_{ik}||} Kn_{ik}$ the unit vector in the direction Kn_{ik}. Then, of course, $(\nabla_h u)_{ik} \cdot (Kn_{ik}) = ||Kn_{ik}|| (\nabla_h u)_{ik} \cdot \eta_{ik}$. Let us consider two points $\tilde{x}_i$ and $\tilde{x}_k$ located on the straight line D_{ik} defined by the midpoint of the interface $i|k$ and by the vector η_{ik}. Then a reasonable approximation of $\nabla u \cdot \eta_{ik}$ at that midpoint is given by

$$(\nabla_h u)_{ik} \cdot \eta_{ik} = \frac{u(\tilde{x}_k) - u(\tilde{x}_i)}{||\tilde{x}_i \tilde{x}_k||}. \tag{131}$$

This is nothing but a finite difference formula like the one we used in (92). The two questions are: first, how to choose $\tilde{x}_i$ and $\tilde{x}_k$, and, second, how to approach reasonably well the values of u at those points, when the unknowns of the scheme u_j are supposed to be good approximations of the exact values $u(x_j)$. Our main requirements will be that $\tilde{x}_i$ and $\tilde{x}_k$ are not too far away one from the other and from the interface $i|k$, in order for (131) to be a good approximation of the gradient at the interface, and that formula (131) will be exact if u is locally a first-order polynomial: we say that the gradient is consistent; we have seen in the proof of the convergence of the scheme on admissible meshes (see section 5.1) that this allows us to apply the Bramble-Hilbert lemma which is a key ingredient in the proof.

The construction of $\tilde{x}_i$ may be performed in the following way. The straight-line D_{ik} divides the plane into two half-planes. We choose one of the points x_j associated to an element T_j located near T_i (for example T_j will have a common vertex with the interface $i|k$), such that x_j is in the opposite half-plane from x_i. We choose the point $\tilde{x}_i$ as the intersection of $[x_j x_i]$ with D_{ik}. Then, $u(\tilde{x}_i)$ is approached by a linear interpolation of $u(x_i)$ and $u(x_j)$ which will be exact if u is P_1

$$u(\tilde{x}_i) \approx \frac{||\tilde{x}_i x_j||\, u(x_i) + ||x_i \tilde{x}_i||\, u(x_j)}{||x_i x_j||} \approx \frac{||\tilde{x}_i x_j||\, u_i + ||x_i \tilde{x}_i||\, u_j}{||x_i x_j||}. \tag{132}$$

We have chosen x_j such that $\tilde{x}_i$ belongs to the segment $[x_i x_j]$ in order to express $u(\tilde{x}_i)$ as a convex combination of u_i and u_j, meaning that $\frac{||\tilde{x}_i x_j||}{||x_i x_j||} \geq 0$ and $\frac{||x_i \tilde{x}_i||}{||x_i x_j||} \geq 0$ and their sum is 1. If x_j were in the same half-plane as x_i, the approximation of $u(\tilde{x}_i)$ by a linear extrapolation of $u(x_i)$ and $u(x_j)$ would have resulted in a non-convex combination.

Remark: There are often more than only one point x_j that fulfill the above criteria. Choosing the point which is as close to x_i as possible provides for the best approximation of $u(\tilde{x}_i)$ by (132). Choosing the point which is as close to the midpoint of the interface $i|k$ as possible provides for the best approximation of $\nabla u \cdot \eta_{ik}$ by (131). To our knowledge, there is no well established rule for choosing one option or the other.

Remark: Particular case of elements near the boundary

6.3 Construction of the whole gradient at the interface: the diamond-cell method

The idea is to construct the whole gradient by using two finite difference formula in two different directions: Let α_{ik} be the unit vector joining x_i and x_k and β_{ik} be a unit vector in the direction given by $i|k$. We write

$$\nabla u \cdot \alpha_{ik} \approx (\nabla_h u)_{ik} \cdot \alpha_{ik} = \frac{u_k - u_i}{d_{ik}} \tag{133}$$

$$\nabla u \cdot \beta_{ik} \approx (\nabla_h u)_{ik} \cdot \beta_{ik} = \frac{u(N_{ik}) - u(S_{ik})}{\ell_{ik}}, \tag{134}$$

where N_{ik} and S_{ik} are the vertices of the interface $i|k$ oriented so that $\overrightarrow{S_{ik}N_{ik}} \cdot \beta_{ik} > 0$. Note that the two-dimensional vector $(\nabla_h u)_{ik}$ is completely defined by (133) and (134) which are its scalar products with two independent vectors. Denoting by n'_{ik} the unit normal vector orthogonal to α_{ik} oriented so that $n'_{ik} \cdot \beta_{ik} \geq 0$, and since n_{ik} is orthogonal to β_{ik}, there holds

$$(\nabla_h u)_{ik} = \frac{(\nabla_h u)_{ik} \cdot \alpha_{ik}}{\alpha_{ik} \cdot n_{ik}} n_{ik} + \frac{(\nabla_h u)_{ik} \cdot \beta_{ik}}{\beta_{ik} \cdot n'_{ik}} n'_{ik}. \tag{135}$$

Denoting by Δ_{ik} the associated diamond-cell $x_i S_{ik} x_k N_{ik}$, it is a matter of elementary geometry to show that

$$|\Delta_{ik}| = \frac{\ell_{ik} d_{ik}}{2} \alpha_{ik} \cdot n_{ik} = \frac{\ell_{ik} d_{ik}}{2} \beta_{ik} \cdot n'_{ik}. \tag{136}$$

Compiling (133), (134), (135) and (136), we obtain the following formula for the gradient

$$(\nabla_h u)_{ik} = \frac{1}{2|\Delta_{ik}|} [(u_k - u_i)\ell_k n_{ik} + (u(N_{ik}) - u(S_{ik}))d_{ik} n'_{ik}]. \tag{137}$$

In this formula, the values of u at the nodes of the mesh are not known, and, in order to maintain a square linear system of algebraic equations, the diamond-cell method proposes to reconstruct them from a linear combination of the unknowns (u_j). In order to do this for a given node N of the mesh, we shall denote by $T(N)$ the set of mesh elements which share N as a vertex and set

$$u(N) = \sum_{j \in T(N)} \alpha_j u_j. \tag{138}$$

In order that formula (137) remains exact if u is a first-order polynome, we shall constrain the weights (α_j) to be such that formula (138) will be exact if u is a first-order polynome. This is equivalent to the following requirement

$$\forall (a,b,c) \in \mathbb{R}^3, \quad a x_N + b y_N + c = \sum_{j \in T(N)} \alpha (a x_{x_j} + b y_{x_j} + c), \tag{139}$$

where (x_N, y_N) are the coordinates of the point N and (x_{x_j}, y_{x_j}) the coordinates of the points x_j associated to the elements T_j. Condition (139) is equivalent to the following system

$$\left\{ \begin{array}{lcl} \sum_{j\in T(N)} \alpha_j & = & 1 \\ \sum_{j\in T(N)} \alpha_j x_{x_j} & = & x_N \\ \sum_{j\in T(N)} \alpha_j y_{x_j} & = & y_N \end{array} \right. .$$

This is a system of three equations in $card(T(N))$ unknowns. For an interior node, $card(T(N))$ is greater or equal to 3, so that we shall sometimes (often, actually) need more equations to fully determine the set (α_j). If $card(T(N)) = 4$, one might think of requiring that the reconstruction (138) will be exact on Q_1 polynomials; if $card(T(N)) = 6$, the weights could be chosen so that the reconstruction (138) will be exact on P_2 polynomials. But this does not cover the general case, and the following least-square reconstruction has been proposed: find coefficients $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ so that the values at the points x_j of the first-order polynomial function $\alpha x + \beta y + \gamma$ are as close as possible to the values u_j in the following sense: Let (w_j) be given strictly positive weights, find the set $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ that minimizes

$$\sum_{j\in T(N)} w_j(\alpha x_{x_j} + \beta y_{x_j} + \gamma - u_j)^2. \tag{140}$$

This system has a unique solution which solves

$$\left\{ \begin{array}{lcl} \sum_{j\in T(N)} x_{x_j} w_j(\alpha x_{x_j} + \beta y_{x_j} + \gamma - u_j) & = & 0 \\ \sum_{j\in T(N)} y_{x_j} w_j(\alpha x_{x_j} + \beta y_{x_j} + \gamma - u_j) & = & 0 \\ \sum_{j\in T(N)} w_j(\alpha x_{x_j} + \beta y_{x_j} + \gamma - u_j) & = & 0 \end{array} \right. ,$$

which is equivalent to

$$\left(\begin{array}{ccc} \sum_{j\in T(N)} x_{x_j}^2 w_j & \sum_{j\in T(N)} x_{x_j} y_{x_j} w_j & \sum_{j\in T(N)} x_{x_j} w_j \\ \sum_{j\in T(N)} x_{x_j} y_{x_j} w_j & \sum_{j\in T(N)} y_{x_j}^2 w_j & \sum_{j\in T(N)} y_{x_j} w_j \\ \sum_{j\in T(N)} x_{x_j} w_j & \sum_{j\in T(N)} y_{x_j} w_j & \sum_{j\in T(N)} w_j \end{array} \right) \left(\begin{array}{c} \alpha \\ \beta \\ \gamma \end{array} \right) = \left(\begin{array}{c} \sum_{j\in T(N)} x_{x_j} w_j u_j \\ \sum_{j\in T(N)} y_{x_j} w_j u_j \\ \sum_{j\in T(N)} w_j u_j \end{array} \right). \tag{141}$$

Once the set (α, β, γ) has been found, then $u(N)$ is simply set equal to $\alpha x_N + \beta y_N + \gamma$. One may easily verify that this procedure leads to the exact value of $u(N)$ if u is locally P_1 and if $u_j = u(x_j)$. Indeed, let us consider $u(x, y) = ax + by + c$ and $u_j = ax_{x_j} + by_{x_j} + c$. Then the set $\alpha = a$, $\beta = b$ and $\gamma = c$ minimizes (140) since this expression is always non-negative and vanishes for $\alpha = a$, $\beta = b$ and $\gamma = c$:

$$\sum_{j\in T(N)} w_j(\alpha x_{x_j} + \beta y_{x_j} + \gamma - u_j)^2 = \sum_{j\in T(N)} w_j((\alpha - a)x_{x_j} + (\beta - b)y_{x_j} + (\gamma - c))^2 = 0.$$

Then the value $ax_N + by_N + c$ is indeed the value of $u(x_N, y_N)$.

Remark: the matrix in system (141) depends only on the geometry of the cells around N and on the weights, while the right-hand side of the system depends linearly on u_j. Thus, the set (α, β, γ) obtained by solving this system is indeed a linear combination of the values u_j. This is important since if it was not the case, the combination of (130), (137) and a non-linear reconstruction for the values at the vertices of the mesh would lead to a non-linear set of equations, which is, in general, much more difficult to solve than a linear set of equations.

To my knowledge, there is no well-defined rule to choose the weights $w_j > 0$. Standard choices include $w_j = 1$ for all $j \in T(N)$, or $w_j = |T_j|$, or $w_j = \frac{1}{||\overrightarrow{Nx_j}||}$ in order to give more weight to the points x_j which are closer to N.

6.4 Construction of the whole gradient at the interface: the discrete duality finite volume method

The diamond scheme presented in the previous section may be used on a wide variety of meshes, including nonconforming ones, and has proved through experimentation to be robust. However, it presents two drawbacks:
1. The associated linear system has, in general, no particular properties, so that it is, in general, not possible to prove that it has a unique solution. Its lack of symmetry also leads to use more expensive (in comparison to the conjugate gradient method) linear solvers to obtain its solution (when there is one).
2. From the numerical analysis point of view, little is known about its convergence when the mesh step size tends to 0. In particular, and this is linked to the fact that the associated matrix is not positive definite in general, there is a lack of coercivity in this scheme which prevents from performing the analysis which was possible in one dimension or in two dimensions on admissible meshes.

The method we present in this section cures these two drawbacks, and we shall see that, like in one dimension and in two dimensions on admissible meshes, it may be expressed under a discrete variational formulation, providing a comfortable setting for the numerical analysis of the scheme.

6.4.1 Construction of the scheme

The main idea of the scheme is to use the discrete gradient formula (137) and to consider the values of u at the interior nodes of the mesh as new unknowns of the scheme, instead of expressing them as linear combinations of the unknowns associated to the elements of the mesh. Since this new scheme has more unknowns, we also need more equations to maintain a square system. These new equations will be obtained by integrating equation (128) over dual cells centered on the interior mesh nodes. The following notations are employed and summarized on figure xxx:

Thus, there will be two sets of unknowns: the set $(u_i^T)_{i\in[1,N+N_b]}$ associated with the elements T_i, with $i \in [1, N]$ and with the midpoints of the boundary edges ($i \in [N+1, N+N_b]$), and the set $(u_j^P)_{j\in[1,Q]}$ associated with the dual

cells P_j, with $j \in [1, Q]$, where Q is the number of nodes in the mesh. We shall suppose for the sake of convenience that the set of nodes is ordered so that indexes of the nodes located on the boundary are within $[Q - N_b + 1, Q]$ and that indexes of the inner nodes are within $[1, Q - N_b]$. We write down the following equations

$$-\sum_{k \in V(i)} \ell_{ik} (\nabla_h u)_{ik} \cdot (K n_{ik}) = |T_i| \, f_i \, , \ \forall i \in [1, N] \tag{142}$$

$$-\sum_{i|k \in E(j)} d_{ik} (\nabla_h u)_{ik} \cdot (K n'_{ik}) s_{ik,j} = |P_j| \, f_j \, , \ \forall j \in [1, Q - N_b] \tag{143}$$

and the boundary conditions are taken into account by

$$u_i^T = 0 \, , \ \forall i \in [N + 1, N + N_b] \tag{144}$$

$$u_j^P = 0 \, , \ \forall j \in [Q - N_b + 1, Q]. \tag{145}$$

6.4.2 Definitions of discrete differential operators and discrete scalar products

We recall that we denote by E the set of all the edges in the mesh and by N_e the number of edges in the set E (including the N_b boundary edges). We recall that N is the number of elements in the mesh and Q the number of nodes. We define the following discrete divergence operator on the primal mesh

$$\begin{aligned} \nabla^T \cdot : \left(\mathbb{R}^{N_e}\right)^2 &\longrightarrow \mathbb{R}^N \\ (v_{ik})_{ik \in E} &\mapsto (\nabla^T \cdot v)_i := \frac{1}{|T_i|} \sum_{k \in V(i)} \ell_{ik} v_{ik} \cdot n_{ik}. \end{aligned} \tag{146}$$

We also introduce a discrete divergence operator on the dual mesh

$$\begin{aligned} \nabla^P \cdot : \left(\mathbb{R}^{N_e}\right)^2 &\longrightarrow \mathbb{R}^Q \\ (v_{ik})_{ik \in E} &\mapsto (\nabla^P \cdot v)_j := \frac{1}{|P_j|} \Bigg(\sum_{i|k \in E(j)} d_{ik} v_{ik} \cdot n'_{ik} s_{ik,j} \\ &\qquad + \sum_{i|k \in E(j), i|k \subset \Gamma} \frac{\ell_{ik}}{2} v_{ik} \cdot n_{ik} \Bigg). \end{aligned} \tag{147}$$

The expression of the discrete gradient is given by

$$\begin{aligned} \nabla_h : \mathbb{R}^{N+N_b} \times \mathbb{R}^Q &\longrightarrow \left(\mathbb{R}^{N_e}\right)^2 \\ ((u_i^T), (u_j^P))_{i \in [1, N+N_b] \, , \, j \in [1, Q]} &\mapsto (\nabla_h u)_{ik} := \frac{1}{2|\Delta_{ik}|} \big[(u_k^T - u_i^T) \ell_k n_{ik} \\ &\qquad + (u^P_{j_2,ik} - u^P_{j_1,ik}) d_{ik} n'_{ik} \big] \end{aligned} \tag{148}$$

With these definitions, scheme (142)–(143) may simply be rewritten

$$\begin{aligned} -(\nabla^T \cdot (K\nabla_h u))_i &= f_i^T \ , \ \forall i \in [1,N], \quad (149)\\ -(\nabla^P \cdot (K\nabla_h u))_j &= f_j^P \ , \ \forall j \in [1,Q-N_b]. \quad (150) \end{aligned}$$

with boundary values u_k^T, $k \in [N+1, N+N_b]$ given by (144) and boundary values u_j^P, $j \in [Q-N_b+1, Q]$ given by (145). On the primal and dual meshes, we also define the discrete scalar product $(\cdot,\cdot)_{T,P}$ by

$$\begin{aligned} &(((u_i^T)_{i\in[1,N]},(u_j^P)_{j\in[1,Q]}),((w_i^T)_{i\in[1,N]},(w_j^P)_{j\in[1,Q]})) \mapsto (u,w)_{T,P}\\ &\mapsto (u,w)_{T,P} := \frac{1}{2}\left(\sum_{i\in[1,N]} |T_i| u_i^T w_i^T + \sum_{j\in[1,Q]} |P_j| u_j^P w_j^P\right). \quad (151) \end{aligned}$$

And on the diamond mesh, associated with the edges, we define the discrete scalar product $(\cdot,\cdot)_D$ by

$$(a_{ik})_{ik\in E},(b_{ik})_{ik\in E} \mapsto (a,b)_D := \sum_{ik\in E} |\Delta_{ik}|\, a_{ik}\cdot b_{ik}. \quad (152)$$

Finally, we define a boundary scalar product on Γ by

$$(a_{ik})_{i|k\subset\Gamma},(b_{ik})_{i|k\subset\Gamma} \mapsto (a,b)_{\Gamma_h} := \sum_{i|k\subset\Gamma} \ell_{ik} a_{ik} b_{ik} \quad (153)$$

Proposition: Let $((u_i^T)_{i\in[1,N+N_b]},(u_j^P)_{j\in[1,Q]})$ and $(v_{ik})_{ik\in E}$ be given. There holds

$$(\nabla^{T,P}\cdot v, u)_{T,P} = -(v,\nabla_h u)_D + (v\cdot n, \gamma u)_{\Gamma_h}, \quad (154)$$

where the discrete trace operator γ is defined in the following way

$$\begin{aligned} \gamma : \mathbb{R}^{N+N_b}\times\mathbb{R}^Q &\longrightarrow \mathbb{R}^{N_b}\\ ((u_i^T)_{i\in[1,N+N_b]},(u_j^P)_{j\in[1,Q]}) &\mapsto (\gamma u)_{ik} := \frac{1}{4}(u_{j_1,ik}^P + 2u_k^T + u_{j_2,ik}^P) \quad (155)\\ &\qquad \text{when } x_k\in\Gamma, \end{aligned}$$

where j_1, ik and j_2, ik are the two indexes of the vertices of the edge $i|k$.
Proof: There holds

$$\begin{aligned} 2(\nabla^{T,P}\cdot v, u)_{T,P} &= \sum_{i\in[1,N]} |T_i|\,(\nabla^T v)_i\, u_i^T + \sum_{j\in[1,Q]} |P_j|\,(\nabla^P v)_j u_j^P\\ &= \sum_{i\in[1,N]} u_i^T \sum_{k\in V(i)} \ell_{ik} v_{ik}\cdot n_{ik} \quad (156)\\ &+ \sum_{j\in[1,Q]} u_j^P \sum_{i|k\in E(j)} d_{ik} v_{ik}\cdot n'_{ik} s_{ik,j} \end{aligned}$$

according to (151), (146) and (147). Now, in the right-hand side of (156), for a given interface $i_0|k_0 \in E$, the term $v_{i_0k_0}$ appears four times if $i_0|k_0$ is an interior

edge: once when $i = i_0$, and once when $i = k_0$ in the first sum, and once for each of the two vertices j_{1,i_0k_0} and j_{2,i_0k_0} of the edge $i_0|k_0$ in the second sum. In the first contribution, the term $v_{i_0k_0}$ is multiplied by $\ell_{i_0k_0}u_{i_0}^T n_{i_0k_0}$ while in the second contribution, it is multiplied by $-\ell_{i_0k_0}u_{k_0}^T n_{i_0k_0}$, since $n_{k_0i_0} = -n_{i_0k_0}$ and $\ell_{i_0k_0} = \ell_{k_0i_0}$. As far as the vertices are concerned, the term $v_{i_0k_0}$ is multiplied by $(u^P_{j_1,i_0k_0} - u^P_{j_2,i_0k_0})d_{i_0k_0}n'_{i_0k_0}$ since $s_{i_0k_0,j} = 1$ if $j = j_{1,i_0k_0}$ and $s_{i_0k_0,j} = -1$ if $j = j_{2,i_0k_0}$. Thus, for an interior edge $i_0|k_0$, the term $v_{i_0k_0}$ is multiplied by $-\ell_{i_0k_0}(u^T_{k_0} - u^T_{i_0})n_{i_0k_0} - d_{i_0k_0}(u^P_{j_2,i_0k_0} - u^P_{j_1,i_0k_0})n'_{i_0k_0} = -2|\Delta_{i_0k_0}|(\nabla_h u)_{i_0k_0}$. On the other hand, the term $v_{i_0k_0}$ appears only once in the first sum in the right-hand side of (156) if $i_0|k_0$ is a boundary edge, and it is multiplied by $\ell_{i_0k_0}u^T_{i_0}n_{i_0k_0} = \ell_{i_0k_0}(u^T_{i_0} - u^T_{k_0})n_{i_0k_0} + \ell_{i_0k_0}u^T_{k_0}n_{i_0k_0}$. Moreover, since $i_0|k_0 \subset \Gamma$, then the discrete divergences $(\nabla^P \cdot v)_{j_1,i_0k_0}$ and $(\nabla^P \cdot v)_{j_2,i_0k_0}$ have the extra contributions $\frac{1}{2}\ell_{i_0k_0}v_{i_0k_0} \cdot n_{i_0k_0}$. Thus, for a boundary edge $i_0|k_0$, the term $v_{i_0k_0}$ is multiplied by

$$\begin{aligned} -\ell_{i_0k_0}(u^T_{k_0} - u^T_{i_0})n_{i_0k_0} - d_{i_0k_0}(u^P_{j_2,i_0k_0} - u^P_{j_1,i_0k_0})n'_{i_0k_0} \\ +\ell_{i_0k_0}u^T_{k_0}n_{i_0k_0} + \frac{1}{2}\ell_{i_0k_0}(u^P_{j_2,i_0k_0} + u^P_{j_1,i_0k_0})n_{i_0k_0} \quad = \\ -2|\Delta_{i_0k_0}|(\nabla_h u)_{i_0k_0} + 2\ell_{i_0k_0}(\gamma u)_{i_0k_0}n_{i_0k_0} \end{aligned}$$

Thus, rearranging the sum in (156), there holds

$$\begin{aligned} 2(\nabla^{T,P} \cdot v, u)_{T,P} &= -2\sum_{ik \in E} |\Delta_{ik}|(\nabla_h u)_{ik} \cdot v_{ik} + 2\sum_{i|k \subset \Gamma} \ell_{ik}(\gamma u)_{ik}v_{ik} \cdot n_{ik} \\ &= -2(v, \nabla_h u)_D + 2(v \cdot n, \gamma u)_{\Gamma_h}. \end{aligned}$$

Remark: The discrete Green formula (154) is the discrete equivalent of

$$(\nabla \cdot v, u)_{L^2(\Omega)} = -(v, \nabla u)_{L^2(\Omega)} + (v \cdot n, u)_{L^2(\Gamma)},$$

better than was the associated formula on (105) on admissible meshes. Indeed, in the present case, the dot product $(v, \nabla_h u)_D$ involves both components of the vector fields v and $\nabla_h u$, and not only their tangential components, as was the case with admissible meshes (see the interpretation at the end of paragraph 4.2.2).

6.4.3 A discrete variational formulation for the discrete duality finite volume scheme

Consider any $((w_i^T), (w_j^P))_{i \in [1, N+N_b],\, j \in [1,Q]}$ with $w_k^T = 0$ for all $k \in [N+1, N+N_b]$ and $w_j^P = 0$ for all $j \in [Q - N_b + 1, Q]$, which means that w vanishes on the boundary. Then the solution $((u_i^T), (u_j^P))_{i \in [1, N+N_b],\, j \in [1,Q]}$ of the scheme is such that

$$(K\nabla_h u, \nabla_h w)_D = (f, w)_{T,P}. \tag{157}$$

Proof: Start from (149) and multiply it by $|T_i|w_i^T$, we obtain for all $i \in [1, N]$

$$-|T_i|(\nabla^T \cdot (K\nabla_h u))_i w_i^T = |T_i| f_i^T w_i^T. \tag{158}$$

In the same way, we have

$$-|P_j| f_j^P w_j^P (\nabla^P \cdot (K \nabla_h u))_j w_j^P = |P_j| f_j^P w_j^P \tag{159}$$

for all $j \in [1, Q]$. For $j \in [1, Q - N_b]$, this comes from (150), and for $j \in [Q - N_b + 1, Q]$, this comes from the fact that $w_j^P = 0$. Therefore, summing (158) over $i \in [1, N]$ and (159) over $j \in [1, Q]$, and dividing the result by two, we obtain, from the definition of the discrete scalar product on the primal and dual meshes (151)

$$-(\nabla^{T,P} \cdot (K \nabla_h u), w)_{T,P} = (f, w)_{T,P}. \tag{160}$$

By the discrete Green formula (154), the left-hand side of (160) may be transformed to yield

$$-(\nabla^{T,P} \cdot (K \nabla_h u), w)_{T,P} = (K \nabla_h u, \nabla_h w)_D - ((K \nabla_h u) \cdot n, \gamma w)_{\Gamma_h},$$

which, together with (160) and since $\gamma w = 0$ identically, proves (157).

6.4.4 Existence and uniqueness of the solution of the discrete duality finite volume scheme

Scheme (142)–(145) has $N + N_b + Q$ unknowns. There are N equations in (142), $Q - N_b$ equations in (143), N_b equations in (144) and also in (145), which implies that the scheme may be expressed as a square linear system. Thus, existence for all data and uniqueness are equivalent, and we shall prove uniqueness by injectivity. Supposing that f vanishes and applying (157) with $w = u$, which is possible because u vanishes on the boundary, we get

$$(K \nabla_h u, \nabla_h u)_D = \sum_{ik \in E} |\Delta_{ik}| (K(\nabla_h u)_{ik}) \cdot (\nabla_h u)_{ik} = 0. \tag{161}$$

In (161), all the terms in the sum are positive since K has been supposed to be a positive matrix. Since the sum vanishes, this implies that $|\Delta_{ik}| (K(\nabla_h u)_{ik}) \cdot (\nabla_h u)_{ik} = 0$ for all $ik \in E$. Since $|\Delta_{ik}| \neq 0$ and since K is definite, this implies that $(\nabla_h u)_{ik} = 0$. Now, the discrete gradient defined by (148) is a linear combination of the two independent vectors n_{ik} and n'_{ik}, so that $(\nabla_h u)_{ik} = 0$ implies that $u_k^T = u_i^T$ and $u_{j_2,ik}^P = u_{j_1,ik}^P$, if S_{j_1} and S_{j_2} are the two vertices of the primal edge $i|k$. It follows by connexity of the domain that all u_i^T, including those at the boundary, are equal to the same value which vanishes because of the boundary conditions. In the same way, all u_j^P are equal to the same value which vanishes.

References

[1] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods, in Ciarlet, P. G. (ed.) et al., Handbook of numerical analysis. Vol. 7. Amsterdam: North-Holland/ Elsevier, pp. 713–1020, 2000.