A Finite Volume Scheme for Diffusion Problems on General Meshes Applying Monotony Constraints

O. Angélini, C. Chavant, Eric Chénier, R. Eymard

A finite volume scheme for diffusion problems on general meshes applying monotony constraints

O. Angelini[1&2], C. Chavant[1], E. Chénier[2] and R. Eymard[2]

[1]Laboratoire de Mécanique des Structures Industrielles Durables, UMR EDF/CNRS 2832, France
[2]Université Paris-Est, Laboratoire Modélisation et Simulation Multi Echelle,
MSME FRE3160 CNRS, 5 bd Descartes, 77454 Marne-la-Vallée, France

Abstract

In order to increase the accuracy and the stability of a scheme dedicated to the approximation of diffusion operators on any type of grids, we propose a method which reduces the curvature of the discrete solution where the loss of monotony is observed. The discrete solution is shown to fulfill a variational formulation thanks to the use of Lagrange multipliers. We can then show its convergence to the solution of the continuous problem, and an error estimate is derived. A numerical method, based on Uzawa's algorithm, is shown to provide accurate and stable approximate solutions to various problems. Numerical results show the increase of precision due to the application of the method.

1 Introduction

A few pioneering works address the problem of monotony of numerical schemes for handling anisotropic diffusion operators on general grids. Let us quote [15; 16; 13; 14; 17; 6; 3], who attempted various methods in order to circumvent this problem. Recall that all of these methods are nonlinear, and rely on coefficients computed from the solution itself. Among them, let us recall a natural one, which consists in writing the equations associated with each unknown in a way such that each unknown can be expressed as a convex combination of neighboring ones; this is achieved in finite volume methods, writing adapted two-point expressions for the fluxes at the edges of a control volume, and ensuring the conservation of the fluxes. One difficulty which must be overcome using such methods is to achieve the convergence of the fixpoint method, necessary for solving these nonlinear schemes (see [14] for a discussion on this problem).
We have therefore investigated another way of increasing the numerical stability of numerical schemes, using new nonlinear scheme, which modifies the curvature of the discrete solution. In order to describe the method, let us first give the classical Dirichlet problem, with a heterogeneous and anisotropic diffusion matrix.
Our aim is to provide an approximation to the following problem:

$$\left\{\begin{array}{rl} -\mathrm{div}\left(\overline{\overline{\Lambda}}(x)\nabla u\right) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{array}\right. \tag{1}$$

with the following hypotheses on the data:

$$\begin{array}{l} \Omega \text{ is a polyhedral open bounded connected subset of } \mathbb{R}^d \text{ with } d \in \mathbb{N}^\star, \\ \overline{\overline{\Lambda}} \text{ is a measurable function from } \Omega \text{ to } M_d(\mathbb{R}), \\ \text{There exist } \underline{\lambda} \text{ and } \overline{\lambda} \text{ such that } 0 < \underline{\lambda} \leq \overline{\lambda} \text{ and } \mathrm{Sp}(\overline{\overline{\Lambda}}(x)) \subset [\underline{\lambda}, \overline{\lambda}] \text{ for a.e. } x \in \Omega, \\ f \in L^2(\Omega), \end{array} \tag{2}$$

where we denote by $M_d(\mathbb{R})$ the set of symmetric positive definite $d \times d$ real matrices, and, for $A \in M_d(\mathbb{R})$, $\mathrm{Sp}(A)$ is the set of the eigenvalues of A. Under these hypotheses, the weak solution of (1) is the unique function u satisfying:

$$\left\{\begin{array}{l} u \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \overline{\overline{\Lambda}}(x)\nabla u(x).\nabla v(x)dx = \int_\Omega f(x)v(x)dx, \ \ \forall v \in H_0^1(\Omega) \end{array}\right. \tag{3}$$

The approach that we propose in this paper is the following. We first consider the use of the hybrid finite volume scheme (see [8] or [9]) for finding an approximate solution to Problem (3). This scheme, which can be related to the mimetic finite difference method [4; 5], has the property to degenerate on the well-known two-point flux finite volume method with harmonic averaging of the diffusion [10] in the particular case of triangles or rectangles and isotropic diffusion, which has the advantage to be cheap and easily implementable. In such a case, accounting from the above discussion, monotony properties are fully satisfied.
The hybrid finite volume method consists in the general case in finding an element $u \in V$ which satisfies the minimum value of a real α-elliptic function J on some finite dimensional Euclidean space V. The main mathematical results, concerning this scheme, are recalled in Section 2.
We then introduce the subset $K \subset V$ of all $v \in V$ such that $G_i(v) \leq 0$ for $i = 1, \ldots, p$. These functions G_i are convex regular functions, such that the interpolation in V of any smooth function belongs to K, at least for sufficiently fine discretizations. The new problem to solve is then to find $u \in K$ which satisfies the minimum on K of the function J. We show in Section 3 that this problem is well-posed, and that the discrete solution thus provided converges to the exact solution. An error estimate is also proposed.
We finally show, on a numerical example proposed in Section 4, the efficiency of the method (the discrete solution being approximated thanks to Uzawa's algorithm, which has the advantage in this framework to be easily implemented in a code dedicated to the computation of the non-constrained scheme). We recall in an appendix the classical theorems concerning the existence of Lagrange multipliers and the convergence of Uzawa's algorithm.

2 The initial scheme

We now describe in details the scheme, first proposed in [8] and extended in [9], used to provide the properties that are required to follow the lines given in the introduction. To this purpose, we first define the geometrical elements entering into the definition of a discretization.

Definition 2.1 (Discretization) *Let Ω be a polyhedral open bounded connected subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$, and $\partial\Omega = \overline{\Omega}\backslash\Omega$ its boundary. A discretization of Ω, denoted by $\mathcal{D}$, is defined as the triplet $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where:*

- *$\mathcal{M}$ is a finite family of non empty connected open disjoint subsets of Ω ("'the control volumes"') such that $\overline{\Omega} = \bigcup_{K\in\mathcal{M}} \overline{K}$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K}\backslash K$ be the boundary of K ; $m_K > 0$ denotes the measure of K and h_K denotes the diameter of K.*
- *$\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "'edges"' of the mesh), such that, for all $\sigma \in \mathcal{E}$, σ is a non empty closed subset of a hyperplane of $\mathbb{R}^d$, which has a measure $m_\sigma > 0$ for the $(d-1)$ dimensional measure of σ. We assume that, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. We then denote by $\mathcal{M}_\sigma = \{K \in \mathcal{M}, \sigma \in \mathcal{E}_K\}$. We then assume that, for all $\sigma \in \mathcal{E}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \subset \partial\Omega$ (the set of these interfaces, called boundary interfaces, is denoted by $\mathcal{E}_{ext}$) or $\mathcal{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, is denoted by $\mathcal{E}_{int}$). For all $\sigma \in \mathcal{E}$, we denote by x_σ the barycenter of σ. For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}_K$, we denote by $\overline{n}_{K,\sigma}$ the unit vector normal to σ outward to K.*
- *$\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (x_K)_{K\in\mathcal{M}}$, such that for all $K \in \mathcal{M}$, $x_K \in K$ and K is assumed to be x_k-star-shaped, which means that for all $x \in K$, the property $[x_K, x] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between x_K and the hyperplane including σ, one assumes that $d_{K,\sigma} > 0$.*

The following notations are used. The size of the discretization is defined by: $h_\mathcal{D} = \sup\{h_K, K \in \mathcal{M}\}$. The regularity of the mesh is measured through the parameter

$$\theta_\mathcal{D} = \max\left(\max_{\sigma\in\mathcal{E}_{\text{int}},K,L\in\mathcal{M}_\sigma} \frac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{\sigma\in\mathcal{E}_K,K\in\mathcal{M}_\sigma} \frac{h_K}{d_{K,\sigma}}\right) \tag{4}$$

We have, thanks to the assumption that K is x_K-star-shaped, the property:

$$\sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} = d\, m_K \tag{5}$$

Let $X_\mathcal{D} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{E}}), v_K \in \mathbb{R}, v_\sigma \in \mathbb{R}\} = \mathbb{R}^\mathcal{M} \times \mathbb{R}^\mathcal{E}$ be the discrete space into which we look for the solution of the schemes.
Let $X_{\mathcal{D},0} \subset X_\mathcal{D}$ be defined by $X_{\mathcal{D},0} = \{u \in X_\mathcal{D}, u_\sigma = 0, \sigma \in \mathcal{E}_{ext}\}$, taking into account the homogeneous Dirichlet boundary conditions.

For all $\phi \in C^0(\overline{\Omega})$, we denote by $P_{\mathcal{D}}\phi$ the element of $X_{\mathcal{D}}$ defined by $\left((\phi(x_K))_{K\in\mathcal{M}}, (\phi(x_\sigma))_{\sigma\in\mathcal{E}}\right)$.
The space $X_{\mathcal{D},0}$ is equipped with a Euclidean structure, defined by the following inner product:

$$\forall (u,v) \in (X_{\mathcal{D},0})^2,\ [u,v]_{\mathcal{D}} = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d_{K,\sigma}}(u_\sigma - u_K)(v_\sigma - v_K), \tag{6}$$

and the associated norm: $\|u\|_{1,\mathcal{D}} = ([u,u]_{\mathcal{D}})^{\frac{1}{2}}$.
Let $H_{\mathcal{M}}(\Omega) \subset L^2(\Omega)$ be the set of piecewise constant functions on the control volumes on the mesh $\mathcal{M}$. For any $u \in X_{\mathcal{D}}$, we define

$$\nabla_K u = \frac{1}{m_K} \sum_{\sigma\in\mathcal{E}_K} m_\sigma (u_\sigma - u_K)\overline{n}_{K,\sigma}\ \ , \forall K \in \mathcal{M}, \tag{7}$$

and we denote by $\nabla_{\mathcal{D}} u \in H_{\mathcal{M}}(\Omega)^d$ the function defined by

$$\nabla_{\mathcal{D}} u(x) = \nabla_K u, \ \text{ for a.e. } x \in K,\ \forall K \in \mathcal{M}. \tag{8}$$

Note that, thanks to the Cauchy-Schwarz inequality, we get

$$\|\nabla_{\mathcal{D}} u\|_{L^2(\Omega)^d} \leq \sqrt{d\, m_\Omega}\|u\|_{1,\mathcal{D}}. \tag{9}$$

We also denote by $P_{\mathcal{M}} u \in H_{\mathcal{M}}(\Omega)$ the function such that $P_{\mathcal{M}} u(x) = u_K$, for a.e. $x \in K$ and all $K \in \mathcal{M}$.

For a given family of strictly positive reals $\alpha = (\alpha_K)_{K\in\mathcal{M}}$, we consider the bilinear form defined by:

$$\langle u, v\rangle_{\mathcal{D},\alpha} = \sum_{K\in\mathcal{M}} \left(m_K \nabla_K u \ \cdot \overline{\overline{\Lambda}}_K \ \nabla_K v + \alpha_K \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} R_{K,\sigma}(u) R_{K,\sigma}(v) \overline{n}_{K,\sigma} \cdot \overline{\overline{\Lambda}}_K \overline{n}_{K,\sigma} \right) \tag{10}$$

where we define

$$R_{K,\sigma}(u) = \frac{u_\sigma - u_K - \nabla_K u \cdot (x_\sigma - x_K)}{d_{K,\sigma}},\ \forall K \in \mathcal{M},\ \forall \sigma \in \mathcal{E}_K, \tag{11}$$

and

$$\overline{\overline{\Lambda}}_K = \frac{1}{m_K} \int_K \overline{\overline{\Lambda}}(x) dx,\ \forall K \in \mathcal{M}. \tag{12}$$

The scheme for the approximation of Problem (3) consists in finding $u \in X_{\mathcal{D},0}$ such that

$$\langle u, v\rangle_{\mathcal{D},\alpha} = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx,\ \forall v \in X_{\mathcal{D},0}. \tag{13}$$

We then have the following results, proven in [8].

Lemma 2.1 (Discrete Poincaré inequality)
Let us assume Hypothesis (2). Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 and let $\theta \geq \theta_{\mathcal{D}}$ be given, where $\theta_{\mathcal{D}}$ is defined by (4). Then there exists C_1, only depending on d, Ω and on θ, such that

$$\|P_{\mathcal{M}} u\|_{L^2(\Omega)} \leq C_1 \|u\|_{1,\mathcal{D}}. \tag{14}$$

Lemma 2.2 (Coerciveness of $\langle\cdot,\cdot\rangle_{\mathcal{D},\alpha}$)
Let us assume Hypothesis (2). Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 and let $\theta \geq \theta_{\mathcal{D}}$ be given, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ α be a family of strictly positive reals and let $\underline{\alpha} \in (0, \min\{\alpha_K, K \in \mathcal{M}\})$. Then there exists $\alpha_0 > 0$, only depending on d, $\underline{\alpha}$, $\underline{\lambda}$ and on θ, such that

$$\alpha_0 [u,u]_{\mathcal{D}} \leq \langle u, u\rangle_{\mathcal{D},\alpha},\ \forall u \in X_{\mathcal{D},0}. \tag{15}$$

The two above Lemmas provide sufficient conditions to ensure the existence and uniqueness of the solution to (13). Thanks to the symmetry of $\langle\cdot,\cdot\rangle_{\mathcal{D},\alpha}$, we get that this solution is also given by

$$u = \operatorname{argmin}_{v\in X_{\mathcal{D},0}} J_{\mathcal{D},\alpha}(v), \tag{16}$$

where the notation $\operatorname{argmin}_{x\in E} f(x)$ denotes an element $y \in E$, assumed to exist and to be unique, such that the minimum value of the function f on E is equal to $f(y)$, and where $J_{\mathcal{D},\alpha}$ is defined by

$$J_{\mathcal{D},\alpha}(v) = \frac{1}{2}\langle v, v\rangle_{\mathcal{D},\alpha} - \int_\Omega f(x) P_{\mathcal{M}} v(x) dx,\ \forall v \in X_{\mathcal{D},0}. \tag{17}$$

Lemma 2.3 (Relative compactness in $X_{\mathcal{D}^{(m)},0}$)
Let us assume Hypothesis (2). Let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a family of discretizations of Ω in the sense of Definition 2.1, such that there exists $\theta \geq 0$ with $\theta \geq \theta_{\mathcal{D}^{(m)}}$ for all $m \in \mathbb{N}$. Let $(u_m)_{m\in\mathbb{N}}$ be a sequence such that there exists $C_2 > 0$ such that, for all $m \in \mathbb{N}$, $u_m \in X_{\mathcal{D}^{(m)},0}$ and $\|u_m\|_{1,\mathcal{D}^{(m)}} \leq C_2$. Then there exists $u \in H^1_0(\Omega)$ and a sub-sequence of $(u_m)_{m\in\mathbb{N}}$, again denoted by $(u_m)_{m\in\mathbb{N}}$, which converges to u in $L^2(\Omega)$. Moreover, $(\nabla_{\mathcal{D}^{(m)}} u_m)_{m\in\mathbb{N}}$ weakly converges to ∇u in $L^2(\Omega)^d$.

Lemma 2.4 (Consistency of $\nabla_{\mathcal{D}} u$ and $R_{K,\sigma} u$)
Let us assume Hypothesis (2). Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 and let $\theta \geq \theta_{\mathcal{D}}$ be given, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\phi \in C^2(\overline{\Omega})$. Then there exists C_3, only depending on d, θ and ϕ (through its second order partial derivatives) such that

$$|\nabla_K P_{\mathcal{D}}\phi - \nabla\phi(x_K)| \leq C_3 h_{\mathcal{D}},$$

where we denote by $|x|$ the Euclidean norm of any $x \in \mathbb{R}^d$, and

$$|R_{K,\sigma}(P_{\mathcal{D}}\phi)| \leq C_3 h_{\mathcal{D}}.$$

The following lemma is a consequence of the previous ones.

Lemma 2.5 (Consistency of $\langle\cdot,\cdot\rangle_{\mathcal{D},\alpha}$)
Let us assume Hypothesis (2). Let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a family of discretizations of Ω in the sense of Definition 2.1, such that there exists $\theta \geq 0$ with $\theta \geq \theta_{\mathcal{D}^{(m)}}$ for all $m \in \mathbb{N}$. Let $(u_m)_{m\in\mathbb{N}}$ be a sequence such that there exists $C_2 > 0$ such that, for all $m \in \mathbb{N}$, $u_m \in X_{\mathcal{D}^{(m)},0}$ and $\|u_m\|_{1,\mathcal{D}^{(m)}} \leq C_2$. We assume that there exist $\overline{\alpha} \geq \underline{\alpha} > 0$, and, for all $m \in \mathbb{N}$, a family $\alpha^{(m)} = (\alpha_K^{(m)})_{K\in\mathcal{M}^{(m)}}$, such that $\{\alpha_K^{(m)}, K \in \mathcal{M}\} \subset [\underline{\alpha}, \overline{\alpha}]$. We also assume that there exists $u \in H^1_0(\Omega)$ such that $(u_m)_{m\in\mathbb{N}}$ converges to u in $L^2(\Omega)$. Then the following holds

$$\lim_{m\to\infty} \langle u_m, P_{\mathcal{D}^{(m)}}\phi\rangle_{\mathcal{D}^{(m)},\alpha^{(m)}} = \int_\Omega \overline{\overline{\Lambda}}(x)\nabla u(x).\nabla\phi(x)dx,\ \forall\phi \in C_c^\infty(\Omega), \tag{18}$$

where $C_c^\infty(\Omega)$ is the set of elements of $C^\infty(\Omega)$ with compact support. We can finally state the convergence result, which can be easily proven using the previous lemmas (in the next section, we state and prove a convergence result, following similar steps).

Theorem 2.1 (Convergence of the non-constrained scheme)
Let us assume Hypothesis (2). Let $\theta > 0$ and $\overline{\alpha} \geq \underline{\alpha} > 0$ be given. Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 such that $\theta \geq \theta_{\mathcal{D}}$, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ be a family of reals such that $\{\alpha_K, K \in \mathcal{M}\} \subset [\underline{\alpha}, \overline{\alpha}]$. Let $u_{\mathcal{D}}$ be the unique solution of (13). Then $u_{\mathcal{D}}$ tends in $L^2(\Omega)$ to u, the unique solution of (3), as $h_{\mathcal{D}}$ tends to 0. Moreover, $\nabla_{\mathcal{D}} u_{\mathcal{D}}$ tends in $L^2(\Omega)^d$ to ∇u.

3 The scheme with constraints

Let us again assume Hypothesis (2). Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 and let $\theta \geq \theta_{\mathcal{D}}$ be given, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ be a family of strictly positive reals and let $\overline{\alpha} \geq \underline{\alpha} > 0$ such that $\{\alpha_K, K \in \mathcal{M}\} \subset [\underline{\alpha}, \overline{\alpha}]$.
We notice that, for any $u \in X_{\mathcal{D}}$ which is the interpolation of a linear function, the linear form $R_{K,\sigma}(u)$ vanishes, since $u_\sigma = u_K + \nabla_K u \cdot (x_\sigma - x_K)$. Hence an idea is to look for the solution of a modified scheme, given by

$$u = \text{argmin}_{v\in X^{(\varepsilon)}_{\mathcal{D},0}} J_{\mathcal{D},\alpha}(v), \tag{19}$$

where, for a given $\varepsilon > 0$, we define

$$G_K^{(\varepsilon)}(v) = \frac{1}{2}\sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma}\overline{n}_{K,\sigma} \cdot \overline{\overline{\Lambda}}_K \overline{n}_{K,\sigma} R_{K,\sigma}(v)^2 - m_K\varepsilon,\ \forall K \in \mathcal{M}, \tag{20}$$

and

$$X^{(\varepsilon)}_{\mathcal{D},0} = \{v \in X_{\mathcal{D},0},\ G_K^{(\varepsilon)}(v) \leq 0, \forall K \in \mathcal{M}\}. \tag{21}$$

Let us remark that, for any function $\phi \in C_c^2(\Omega)$ (that are functions in $C^2(\Omega)$ with compact support), Lemma 2.4 shows that the interpolation $P_{\mathcal{D}}\phi$ satisfies $P_{\mathcal{D}}\phi \in X^{(\varepsilon)}_{\mathcal{D},0}$ provided $h_{\mathcal{D}}$ be sufficiently small. We now have the following lemma.

Lemma 3.1 (Characterization of the solution of the constrained problem)
Let us assume Hypothesis (2). Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 and let $\theta \geq \theta_{\mathcal{D}}$ be given, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ α be a family of strictly positive reals and let $\underline{\alpha} \in (0, \min\{\alpha_K, K \in \mathcal{M}\})$. Let $\varepsilon > 0$ be given. Then there exists one and only one solution u to (19), which moreover satisfies: there exists a family of non negative reals $\beta = (\beta_K)_{K\in\mathcal{M}}$ such that $(u, \beta) \in X_{\mathcal{D},0}^{(\varepsilon)} \times (\mathbb{R}_+)^{\mathcal{M}}$ is a saddle point on $X_{\mathcal{D},0} \times (\mathbb{R}_+)^{\mathcal{M}}$ of the function $\mathcal{L}$ defined by

$$\mathcal{L}(v, \beta) = J_{\mathcal{D},\alpha}(v) + \sum_{K\in\mathcal{M}} \beta_K G_K^{(\varepsilon)}(v), \tag{22}$$

and the so-called Kuhn-Tucker relations

$$\beta_K G_K^{(\varepsilon)}(u) = 0, \ \forall K \in \mathcal{M}, \tag{23}$$

hold. Finally, the following relation holds:

$$\langle u, v\rangle_{\mathcal{D},\alpha+\beta} = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx, \ \forall v \in X_{\mathcal{D},0}. \tag{24}$$

Proof. The existence and the uniqueness of u of (19) is an immediate consequence of Lemma 2.2 and of the fact that $X_{\mathcal{D},0}^{(\varepsilon)}$ is a closed non empty convex set. The existence of the multipliers $\beta \in (\mathbb{R}_+)^{\mathcal{M}}$ such that (23) holds is a consequence of Theorem 5.1 given in the appendix. The saddle point property implies that

$$u = \text{argmin}_{v\in X_{\mathcal{D},0}} \left(J_{\mathcal{D},\alpha}(v) + \sum_{K\in\mathcal{M}} \beta_K G_K^{(\varepsilon)}(v) \right).$$

The variational formulation of the above relation exactly is (24). □
Thanks to the characterization given by Lemma 3.1, we can state the following estimate on the solution of the constrained problem.

Lemma 3.2 (Estimate on the solution of the constrained scheme)
Let us assume Hypothesis (2). Let $\theta > 0$, $\underline{\alpha} > 0$ be given. Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 such that $\theta \geq \theta_{\mathcal{D}}$, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ be a family of reals such that $\min\{\alpha_K, K \in \mathcal{M}\} \geq \underline{\alpha}$. Let $\varepsilon > 0$ be given. Then, for all $\beta \in (\mathbb{R}_+)^{\mathcal{M}}$, there exists one and only one $u_{\mathcal{D}}$ solution to (24), which satisfies

$$\|u_{\mathcal{D}}\|_{1,\mathcal{D}} \leq \frac{\|f\|_{L^2(\Omega)} C_1}{\alpha_0}, \tag{25}$$

where α_0 only depends on d, $\underline{\alpha}$, $\underline{\lambda}$ and on θ (see Lemma 2.2) and C_1 only depends on d, Ω and on θ (see Lemma 2.1). Moreover, in the case where (u, β) is a saddle point of the function defined by (22), then

$$\sum_{K\in\mathcal{M}} \beta_K m_K \leq \|f\|_{L^2(\Omega)}^2 \frac{C_1^2}{2\alpha_0\varepsilon}. \tag{26}$$

Proof. Letting $v = u$ in (24) provides

$$\langle u, u\rangle_{\mathcal{D},\alpha+\beta} = \int_\Omega f(x) P_{\mathcal{M}} u(x) dx. \tag{27}$$

Since $\langle u, u\rangle_{\mathcal{D},\alpha} \leq \langle u, u\rangle_{\mathcal{D},\alpha+\beta}$, we get, from Lemmas 2.2 and 2.1,

$$\alpha_0 [u, u]_{\mathcal{D}} \leq \|f\|_{L^2(\Omega)} C_1 ([u, u]_{\mathcal{D}})^{1/2},$$

where α_0 only depends on d, $\underline{\alpha}$, $\underline{\lambda}$ and on θ (in particular, it does not depend on β). This gives (25). We now assume that (u, β) is a saddle point of the function defined by (22). We then get, from (27),

$$\sum_{K\in\mathcal{M}} \beta_K \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} \overline{n}_{K,\sigma} \cdot \overline{\overline{\Lambda}}_K \overline{n}_{K,\sigma} R_{K,\sigma}(u)^2 \leq \int_\Omega f(x) P_{\mathcal{M}} u(x) dx,$$

which provides

$$\sum_{K\in\mathcal{M}} \beta_K (G_K^{(\varepsilon)}(u) + m_K\varepsilon) \leq \|f\|_{L^2(\Omega)}^2 \frac{C_1^2}{2\alpha_0}.$$

Thanks to (23) which holds in this case, we get

$$\sum_{K\in\mathcal{M}} \beta_K m_K \leq \|f\|^2_{L^2(\Omega)} \frac{C_1^2}{2\alpha_0\varepsilon}.$$

□
We can now state the following theorem.

Theorem 3.1 (Convergence of the constrained scheme)
Let us assume Hypothesis (2). Let $\theta > 0$, $\overline{\alpha} \geq \underline{\alpha} > 0$ be given. Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 such that $\theta \geq \theta_\mathcal{D}$, where $\theta_\mathcal{D}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ be a family of reals such that $\{\alpha_K, K \in \mathcal{M}\} \subset [\underline{\alpha}, \overline{\alpha}]$. Let $\varepsilon_\mathcal{D} > 0$ be given, and $u_\mathcal{D}$ be the unique solution of (19) for $\varepsilon = \varepsilon_\mathcal{D}$. Then $u_\mathcal{D}$ tends in $L^2(\Omega)$ to u, the unique solution of (3), as $h_\mathcal{D}$ and $h_\mathcal{D}/\sqrt{\varepsilon_\mathcal{D}}$ tend to 0. Moreover, $\nabla_\mathcal{D} u_\mathcal{D}$ tends in $L^2(\Omega)^d$ to ∇u.

Proof. Lemmas 3.2 and 2.3 allow to extract from any sequence of solutions defined by a sequence of discretizations, a sub-sequence which converges in $L^2(\Omega)$ to some $u \in H^1_0(\Omega)$. Let $\phi \in C^\infty_c(\Omega)$ be given. Let $\mathcal{D}$ be a discretization belonging to this extracted sub-sequence, and let us take $v = P_\mathcal{D}\phi$ in (24). We get

$$\langle u_\mathcal{D}, P_\mathcal{D}\phi\rangle_{\mathcal{D},\alpha} + T_1(u_\mathcal{D}, P_\mathcal{D}\phi) = \int_\Omega f(x) P_\mathcal{M} P_\mathcal{D}\phi(x)dx,$$

where T_1 is defined by

$$T_1(w,v) = \sum_{K\in\mathcal{M}} \beta_K \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} \overline{n}_{K,\sigma} \cdot \overline{\overline{\Lambda}}_K \overline{n}_{K,\sigma} R_{K,\sigma}(w) R_{K,\sigma}(v),\ \forall w, v \in X_\mathcal{D}. \tag{28}$$

Let us apply the Cauchy-Schwarz inequality. We get

$$T_1(u_\mathcal{D}, P_\mathcal{D}\phi)^2 \leq T_1(u_\mathcal{D}, u_\mathcal{D}) T_1(P_\mathcal{D}\phi, P_\mathcal{D}\phi).$$

Since, applying Lemma 2.4, we get that there exists C_4, only depending on d, θ and ϕ, such that

$$|R_{K,\sigma}(P_\mathcal{D}\phi)| \leq C_4 h_\mathcal{D},$$

we can write

$$T_1(P_\mathcal{D}\phi, P_\mathcal{D}\phi) \leq C_4^2 h_\mathcal{D}^2 \overline{\lambda} d \sum_{K\in\mathcal{M}} \beta_K m_K.$$

The above relations lead to the existence of C_5, only depending on θ, ϕ, $\overline{\overline{\Lambda}}$, d, Ω and $\underline{\alpha}$ such that

$$T_1(u_\mathcal{D}, P_\mathcal{D}\phi)^2 \leq C_5 \frac{h_\mathcal{D}^2}{\varepsilon}.$$

Hence, under the condition that $h_\mathcal{D}/\sqrt{\varepsilon_\mathcal{D}}$ tends to 0, we get that $T_1(u_\mathcal{D}, P_\mathcal{D}\phi)$ tends to 0 as well. This completes the proof of the convergence of the scheme, since the convergence of $\langle u_\mathcal{D}, P_\mathcal{D}\phi\rangle_{\mathcal{D},\alpha}$ is provided by Lemma 2.5, and the proof of convergence of $\int_\Omega f(x) P_\mathcal{M} P_\mathcal{D}\phi(x)dx$ to $\int_\Omega f(x)\phi(x)dx$ is straightforward. This implies that $\langle u_\mathcal{D}, u_\mathcal{D}\rangle_{\mathcal{D},\alpha}$ converges to $\int_\Omega \overline{\overline{\Lambda}}(x)\nabla u(x).\nabla u(x)dx$ and that $T_1(u_\mathcal{D}, u_\mathcal{D})$ tends to 0. From the preceding convergence property of $\langle u_\mathcal{D}, u_\mathcal{D}\rangle_{\mathcal{D},\alpha}$, we deduce as in [8] that $\nabla_\mathcal{D} u_\mathcal{D}$ converges to ∇u in $L^2(\Omega)^d$.
□
Let us underline that the above convergence result is proven in the general framework of diffusion matrix $\overline{\overline{\Lambda}}$ which are only assumed to be bounded, without further assumptions. This implies that the only regularity assumed on the continuous solution of (3) is $u \in H^1_0(\Omega)$. On the contrary, the following error estimate result is only stated in the particular case $\overline{\overline{\Lambda}} = \overline{\overline{\mathrm{Id}}}$, and $u \in C^2(\overline{\Omega})$. Although these assumptions are quite restrictive (for example, we could handle the cases where $\overline{\overline{\Lambda}}$ is piecewise C^1 and u is piecewise H^2), they prevent from entering into too technical details but allow for getting an indication of the origin of the error (in many similar cases, a priori error estimates for finite volume methods are known to be not sharp and pessimistic compared to the results obtained numerically).

Theorem 3.2 (Error estimate) *Let us assume Hypothesis (2). We also assume that $\overline{\overline{\Lambda}} = \overline{\overline{\mathrm{Id}}}$, and that the solution u of (3) satisfies $u \in C^2(\overline{\Omega})$.*

Let $\theta > 0$, $\overline{\alpha} \geq \underline{\alpha} > 0$ be given. Let $\mathcal{D}$ be a discretization of Ω in the sense of Definition 2.1 such that $\theta \geq \theta_{\mathcal{D}}$, where $\theta_{\mathcal{D}}$ is defined by (4). Let $\alpha = (\alpha_K)_{K\in\mathcal{M}}$ be a family of reals such that $\{\alpha_K, K \in \mathcal{M}\} \subset [\underline{\alpha}, \overline{\alpha}]$. Let $\varepsilon_{\mathcal{D}} > 0$ be given, and $u_{\mathcal{D}}$ be the unique solution of (19) for $\varepsilon = \varepsilon_{\mathcal{D}}$. Then there exists C_6 depending only on d, Ω, θ, $\underline{\alpha}$, $\overline{\alpha}$ and u such that:

$$\|u_{\mathcal{D}} - P_{\mathcal{D}}(u)\|_{1,\mathcal{D}} \leq C_6 \left(\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}_{\mathcal{D}}} + h_{\mathcal{D}}^2\right)^{1/2}, \tag{29}$$

there exists C_7 depending only on d, Ω, θ, $\underline{\alpha}$, $\overline{\alpha}$ and u such that:

$$\|P_{\mathcal{M}} u_{\mathcal{D}} - u\|_{L^2(\Omega)} \leq C_7 \left(\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}_{\mathcal{D}}} + h_{\mathcal{D}}^2\right)^{1/2}, \tag{30}$$

and there exists C_8 depending only on d, Ω, θ, $\underline{\alpha}$, $\overline{\alpha}$ and u such that:

$$\|\nabla_D u_{\mathcal{D}} - \nabla u\|_{L^2(\Omega)^d} \leq C_8 \left(\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}_{\mathcal{D}}} + h_{\mathcal{D}}^2\right)^{1/2}. \tag{31}$$

Remark 3.1 *The above error estimate is in accordance with the convergence theorem 3.1, which requires that $\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}_{\mathcal{D}}}$ tends to zero.*

Proof. Let $v \in X_{\mathcal{D},0}$. We integrate (1) (which resumes to $-\Delta u = f$) in $K \in \mathcal{M}$, we multiply by v_K and we sum the result on $K \in \mathcal{M}$. We get

$$-\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} (v_K - v_\sigma) \int_\sigma \nabla u(x) \cdot \overline{n}_{K,\sigma} ds(x) = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx.$$

This expression implies that

$$-\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} (v_K - v_\sigma) m_\sigma \nabla_K P_{\mathcal{D}} u \cdot \overline{n}_{K,\sigma} = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx + T_2(v),$$

defining $T_2(v)$ by

$$T_2(v) = -\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} (v_K - v_\sigma) m_\sigma \left(\nabla_K P_{\mathcal{D}} u - \frac{1}{m_\sigma} \int_\sigma \nabla u(x) ds(x)\right) \cdot \overline{n}_{K,\sigma}.$$

Hence we get

$$\sum_{K\in\mathcal{M}} m_K \nabla_K P_{\mathcal{D}} u \cdot \nabla_K v = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx + T_2(v),$$

which leads to

$$\langle P_{\mathcal{D}} u, v\rangle_{\mathcal{D},\alpha} = \int_\Omega f(x) P_{\mathcal{M}} v(x) dx + T_2(v) + T_3(v),$$

defining $T_3(v)$ by

$$T_3(v) = \sum_{K\in\mathcal{M}} \alpha_K \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} R_{K,\sigma}(P_{\mathcal{D}} u) R_{K,\sigma}(v).$$

We can then write, subtracting (24),

$$\langle P_{\mathcal{D}} u - u_{\mathcal{D}}, v\rangle_{\mathcal{D},\alpha} = T_2(v) + T_3(v) + T_1(u_{\mathcal{D}}, v). \tag{32}$$

We get, thanks to Lemma 2.4 and to the Cauchy-Schwarz inequality,

$$T_2(v) \leq C_9 h_{\mathcal{D}} \|v\|_{1,\mathcal{D}},$$

where C_9 only depends on u, θ, Ω and d. Similarly, thanks again to Lemma 2.4, we have

$$|R_{K,\sigma}(P_{\mathcal{D}} u)| \leq C_{10} h_{\mathcal{D}},$$

where C_{10} only depends on u, θ, Ω and d. Thanks to the Cauchy-Schwarz inequality, we then get that

$$T_3(v)^2 \leq C_{10}^2 h_{\mathcal{D}}^2 d\, m_\Omega \overline{\alpha}^2 \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma} R_{K,\sigma}(v)^2.$$

Remarking that, thanks to the Young inequality, we have

$$R_{K,\sigma}(v)^2 \leq 2\left(\frac{(v_\sigma - v_K)^2}{d_{K,\sigma}^2} + |\nabla_K v|^2 \frac{|x_\sigma - x_K|^2}{d_{K,\sigma}^2}\right)$$

using (4) and (9), we conclude that

$$T_3(v) \leq C_{11} h_{\mathcal{D}} \|v\|_{1,\mathcal{D}},$$

where C_{11} only depends on u, θ, Ω, $\overline{\alpha}$ and d. We then take $v = P_{\mathcal{D}}u - u_{\mathcal{D}}$. The last term in the right hand side of (32) can be handled as follows:

$$T_1(u_{\mathcal{D}}, P_{\mathcal{D}}u - u_{\mathcal{D}}) \leq T_1(u_{\mathcal{D}}, P_{\mathcal{D}}u) \leq C_{12}\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}},$$

where C_{12} is computed following the same steps as in the proof of Theorem 3.1. On the other hand, we have

$$\alpha_0 \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}}^2 \leq \langle P_{\mathcal{D}}u - u_{\mathcal{D}}, P_{\mathcal{D}}u - u_{\mathcal{D}}\rangle_{\mathcal{D},\alpha},$$

where α_0 only depends on θ, $\underline{\alpha}$ and d. Gathering these results, we get

$$\alpha_0 \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}}^2 \leq (C_9 + C_{11}) h_{\mathcal{D}} \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}} + C_{12}\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}}.$$

Thanks to the Young inequality, we get

$$(C_9 + C_{11}) h_{\mathcal{D}} \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}} \leq \frac{1}{2}\alpha_0 \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}}^2 + C_{13} h_{\mathcal{D}}^2,$$

where C_{13} only depends on α_0, C_9 and C_{11}. We finally get

$$\frac{1}{2}\alpha_0 \|P_{\mathcal{D}}u - u_{\mathcal{D}}\|_{1,\mathcal{D}}^2 \leq C_{13} h_{\mathcal{D}}^2 + C_{12}\frac{h_{\mathcal{D}}}{\sqrt{\varepsilon}},$$

which concludes the proof of (29). The proof of (30) is a straightforward consequence of Lemma 2.1, and that of (31) is an easy consequence of Lemma 2.4 and of (9).
□

4 Implementation and numerical results

In order to find an approximation of the solution u to Problem (19), let us consider the use of Uzawa's algorithm, consisting, for a given real $\rho > 0$ and an initial family $\beta^{(0)} \in (\mathbb{R}_+)^{\mathcal{M}}$ (we take it null in the examples below), in the definition of the sequence $(u^{(n)}, \beta^{(n)})_{n\in\mathbb{N}}$ such that

$$\begin{array}{l} u^{(n)} = \mathrm{argmin}_{u\in X_{\mathcal{D},0}} \mathcal{L}(u, \beta^{(n)}), \\ \beta_K^{(n+1)} = \max(\beta_K^{(n)} + \rho G_K^{(\varepsilon)}(u^{(n)}), 0),\ \forall K \in \mathcal{M},\ \forall n \in \mathbb{N}. \end{array} \tag{33}$$

Indeed, this algorithm is such that, at each iteration, the minimization problem $u^{(n)} = \mathrm{argmin}_{u\in X_{\mathcal{D},0}} \mathcal{L}(u, \beta^{(n)})$ to be solved is given by (24), hence the structure of the code for solving the non-constrained problem is not modified (it suffices to replace α by $\alpha + \beta^{(n)}$). Thanks to Estimate (25), we get that the hypotheses of Theorem 5.2 given in the appendix, stating the convergence of Uzawa's algorithm, are verified. Nevertheless, from a practical point of view, since the constraints are arbitrarily chosen by fixing the value of ε, we are not interested in finding an accurate approximation of the saddle point (which can be too much expensive). We therefore stop the iteration procedure once the family $u^{(n)}$ is such that $G_K^{(\varepsilon)}(u^{(n)}) \leq 0$ for all $K \in \mathcal{M}$. Moreover, we do not allow the Lagrangian multipliers to decrease within the iteration procedure, which is ensured by the following slight modification of (33)

$$\begin{array}{l} u^{(n)} = \mathrm{argmin}_{u\in X_{\mathcal{D},0}} \mathcal{L}(u, \beta^{(n)}), \\ \beta_K^{(n+1)} = \max(\beta_K^{(n)} + \rho G_K^{(\varepsilon)}(u^{(n)}), \beta_K^{(n)}),\ \forall K \in \mathcal{M},\ \forall n \in \mathbb{N}. \end{array} \tag{34}$$

Remark 4.1 *Although the above procedure does not ensure the convergence to the solution of Problem (19), we have numerically observed that the solution was only weakly modified by using* (34) *instead of* (33)*.*

As we show below, this procedure allows to reach significant improvement in accuracy and stability, compared to non-constrained scheme, within a small number of iterations. Note that values for the numerical parameters must be selected for running the algorithm, and that, in particular, the selection of an optimal value for ρ appears to be a complicate problem. Although theorem 5.2 suggests an order of magnitude for ρ, namely $\frac{\alpha}{2M}$ using the notations of the statement, which could indicate an order $1/\varepsilon^2$ for ρ since M should behave as the square of the gradient of the constraints, we have preferred in the following results to study the sensitivity of the results with respect to the values of the parameters.

4.1 Anisotropic and heterogeneous case with an analytical solution

This test is inspired from [2; 12], and induces numerical locking for some schemes. It has been selected as test case in the Benchmark on discretization schemes for anisotropic diffusion problems on general grids [11]. In the domain $\Omega = (0;1)^2$, we consider a case where the diffusion operator $\overline{\overline{\Lambda}}$ is heterogeneous and anisotropic, and given by

$$\overline{\overline{\Lambda}}(x_1, x_2) = \frac{1}{(x_1^2 + x_2^2)} \begin{pmatrix} 10^{-3}x_1^2 + x_2^2 & (10^{-3} - 1)x_1x_2 \\ (10^{-3} - 1)x_1x_2 & x_1^2 + 10^{-3}x_2^2 \end{pmatrix}.$$

The analytical solution and the right-hand-side of (3) are defined in this test case by

$$u(x_1, x_2) = \sin \pi x_1 \sin \pi x_2,\ f = -\mathrm{div}(\overline{\overline{\Lambda}}\nabla u).$$

We now compare the solution provided by different grids and different selections for the numerical parameters. We present in Figure 1 the two irregular grids used for this test, also performed with regular square grids (not shown). The interest

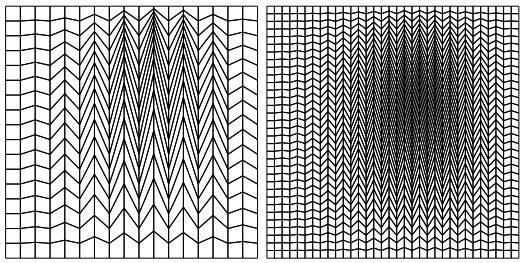

Figure 1: Left: Coarse irregular grid (mesh 1). Right: Refined irregular grid (mesh 2).

of these irregular grids is that they include a few features of irregular grids generated by the geological observations in the underground engineering framework. In order to get a comparison point, we have compared the results obtained with $\alpha = 1$ and $\beta = 0$ (non-constrained scheme) using mesh 1, with the results obtained with $\alpha = 10^{-3}$ and β given by Uzawa's algorithm (constrained scheme) using mesh 1 and mesh 2. Note that the choice $\alpha = 1$ and $\beta = 0$ is the one which leads to two-point fluxes on rectangular meshes and on some regular triangular meshes in the case of isotropic problems (see [9]). We thus provide in Figure 2 the difference between the three solutions and the analytical solution, computed along the line with equation $x_2 = 0.5$. We see that the respect of the constraints decreases strongly the error, which decreases again using the finest mesh. This is confirmed by the L^2 errors of the solution and of its gradient, shown in Table 1, using a variety of grids and numerical parameters.
We again used "(nc)" for non-constrained scheme, with $\alpha = 1$, and "(c)" for constrained scheme. We again remark that the results obtained using the constrained scheme are much more accurate than the results without constraint. Moreover we observe that, as could be expected, the values of the multipliers which are required are lower and lower as the mesh size decrease (recall that, for the size of the mesh tending to 0, the constraints are necessarily satisfied by regular solutions).
We also remark that the results are not highly sensitive to the value of ε, which is not the case for the parameter ρ. In all these cases, 3 iterations where sufficient to get a point satisfying all the constraints. We observe that, for the highest values of ρ, we get more precise values of the gradient. This shows that the obtained final point verifies stronger constraints, leading to a more regular gradient. By choosing a sufficiently high value for ρ as the grid size decreases, we observe that the order of convergence is not far from 2, showing that the result proven in this paper is not sharp (note that this is the case for the great majority of the error estimates provided in the finite volume framework). We finally remark in Figure 3 that the highest values of β are obtained on the regular grid and the coarse irregular grid (mesh 1) around the point $(1, 1)$, whereas they are obtained at the location of the highest perturbations of the refined irregular grid mesh 2.

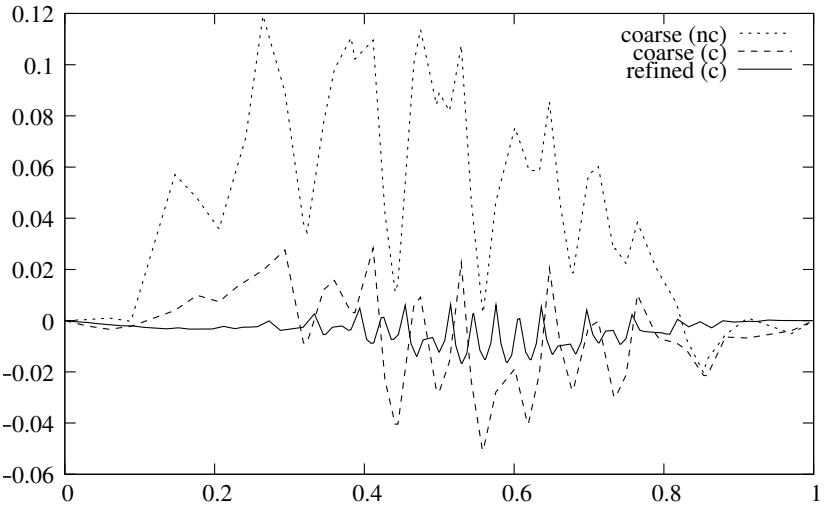

Figure 2: Profiles with irregular grids along the line $x_2 = 0.5$: Difference with the analytical solution computed using the non-constrained scheme with the coarse irregular grid, the constrained scheme with the coarse irregular grid and the constrained scheme with the refined irregular.

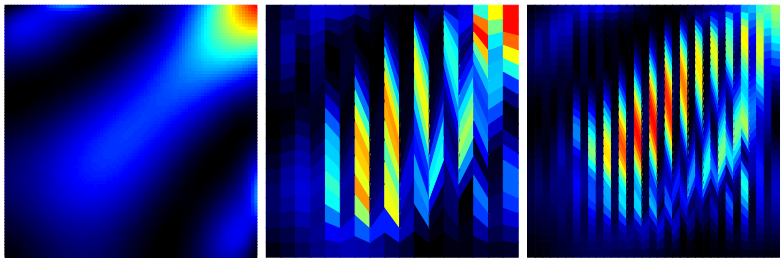

Figure 3: Distribution of β on 80×80 grid (left), mesh 1 (middle), mesh 2 (right).

Remark 4.2 *In all colored figures representing the field of a scalar value, the red color stands for the highest value, the dark blue color for the lowest value.*

The conclusion of this first test case is that the constrained scheme leads to accurate tuning of the Lagrange multiplier, ensuring the best precision for the solution and its gradient.

4.2 Anisotropic case without source terms

We consider a test similar to the test 3 of the Benchmark on discretization schemes for anisotropic diffusion problems on general grids [11]. In this test, the main directions of an anisotropic diffusion matrix are tilted with respect to the boundary conditions and the mesh. In the domain $\Omega = (0;1)^2$, we consider a non-homogeneous Dirichlet problem, without right-hand-side. Let us consider Problem (1) where the diffusion $\overline{\overline{\Lambda}}$ is homogeneous, and given by:

$$\overline{\overline{\Lambda}} = R_\theta \begin{pmatrix} 1 & 0 \\ 0 & 10^{-1} \end{pmatrix} R_\theta^{-1},$$

where R_θ is the matrix of a rotation of angle $\theta = 40$ degrees and $f = 0$. The non-homogeneous boundaries conditions are continuous and piecewise linear on $\partial\Omega$, and such that

$$u(x) = \begin{cases} 1 & \text{on} \quad ((0.;0.2) \times 0. \quad \cup \quad 0. \times (0.;0.2)) \\ 0 & \text{on} \quad ((0.8;1.) \times 1. \quad \cup \quad 1. \times (0.8;1.)) \\ \frac{1}{2} & \text{on} \quad ((0.3;1.) \times 0. \quad \cup \quad 0. \times (0.3;1.)) \\ \frac{1}{2} & \text{on} \quad ((0.;0.7) \times 1. \quad \cup \quad 1. \times (0.;0.7)) \end{cases} \tag{35}$$

We compare the solution, again obtained with a reference 200×200 mesh, $\alpha = 1$ and $\beta = 0$, with the results obtained using a regular 20×20 mesh and using the irregular grid mesh 2 (see Figure 1). For the computations performed with the

grid	ε	Nb. iterations	ρ	$\beta_{\min}$	$\beta_{\max}$	L^2 error	L^2 error on gradient
mesh 1 (nc)	–	1	–	0	0	$3.0\ 10^{-2}$	1.35
mesh 1 (c)	10^{-7}	3	10^4	$8.\ 10^3$	$6.\ 10^6$	$2.5\ 10^{-2}$	$2.73\ 10^{-1}$
mesh 2 (nc)	–	1	–	0	0	$7.\ 10^{-3}$	$7.4\ 10^{-1}$
mesh 2 (c)	10^{-7}	3	10^4	$1.9\ 10^1$	$7.\ 10^5$	$7.\ 10^{-3}$	$5.9\ 10^{-2}$
10×10 (nc)	–	1	–	0	0	$2.3\ 10^{-2}$	$5.9\ 10^{-1}$
10×10 (c)	10^{-7}	3	10^4	$4.\ 10^3$	$4.\ 10^7$	$7.5\ 10^{-3}$	$2.75\ 10^{-2}$
20×20 (nc)	–	1	–	0	0	$6.1\ 10^{-3}$	$1.8\ 10^{-1}$
20×20 (c)	10^{-7}	3	10^4	$1.6\ 10^3$	$3.\ 10^6$	$1.9\ 10^{-3}$	$7.3\ 10^{-3}$
40×40 (nc)	–	1	–	0	0	$1.6\ 10^{-3}$	$5.7\ 10^{-2}$
40×40 (c)	10^{-7}	3	10^4	8.	$2.\ 10^6$	$4.\ 10^{-4}$	$2.26\ 10^{-3}$
80×80 (c)	10^{-7}	3	10^4	$7.9\ 10^{-2}$	$1.4\ 10^4$	$1.18\ 10^{-4}$	$2.09\ 10^{-3}$
80×80 (c)	10^{-8}	3	10^4	784.	$1.4\ 10^4$	$1.18\ 10^{-4}$	$2.08\ 10^{-3}$
80×80 (c)	10^{-8}	3	10^5	784.	$1.4\ 10^5$	$1.18\ 10^{-4}$	$7.7\ 10^{-4}$
80×80 (c)	10^{-8}	3	10^6	784.	$1.4\ 10^6$	$1.18\ 10^{-4}$	$4.7\ 10^{-4}$
80×80 (c)	10^{-8}	3	10^7	784.	$1.4\ 10^7$	$1.18\ 10^{-4}$	$4.6\ 10^{-4}$
80×80 (c)	10^{-8}	3	10^8	784.	$1.4\ 10^8$	$1.18\ 10^{-4}$	$4.6\ 10^{-4}$

Table 1: Comparison of the errors of the solution and its gradient, using different grids and values of numerical parameters.

constrained scheme, the iterations start from the value $\alpha = 10^{-9}$ (value also used for the computations without constraint). We see in Figure 4 that the approximate solutions are in good agreement, as far as one can compare piecewise constant functions of different meshes. In any case, this qualitative comparison is not sufficient to give precise indications about the precision and the regularity of the solutions. These indications are provided by the profiles of the solutions along given

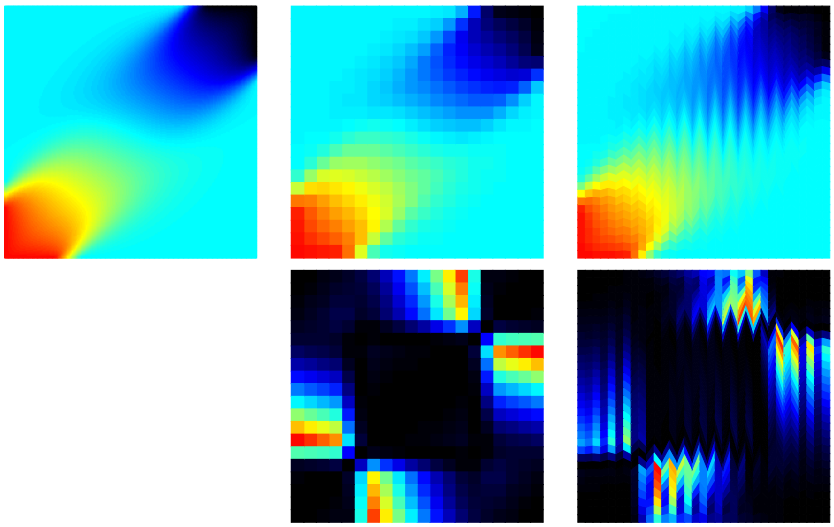

Figure 4: Left top: solution with fine grid. Middle top: constrained solution with regular coarse grid. Right top: constrained solution with irregular grid. Middle bottom: values of β with regular coarse grid. Right bottom: values of β with irregular grid.

lines. Indeed, we present in Figures 5 and 6 the solution along the lines of equations $x_2 = 0.2$ and $x_2 = 0.4$. We first notice that the unconstrained solution with $\alpha = 10^{-9}$ is dramatically far from the reference solution, and that it presents numerous oscillations in the case of the irregular grid. On the contrary, the constrained solution happens to be close to the reference one, and is regular even in the case of the irregular grid.
In order to get indications of the capability of the fluxes provided by the constrained scheme to be used for the coupled transport of any scalar or vector quantity, we have computed the streamlines given by the approximate solution. The method is classical in 2D: we compute a scalar potential, initialized by 0 at a corner of the domain, and increased along each edge of the mesh by the oriented value of the flux across the edge. Hence we get a value of the potential at all the vertices of the mesh, the isovalues of which are the streamlines along the velocity $-\overline{\overline{\Lambda}}\nabla u$. We observe that these

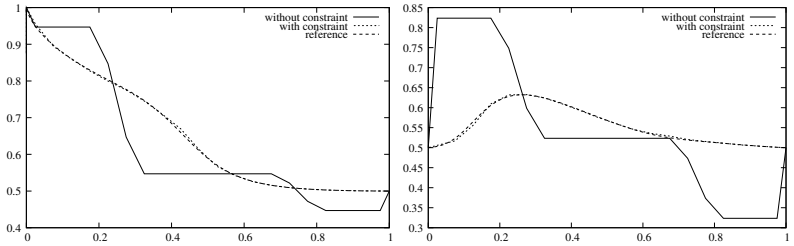

Figure 5: Profiles with coarse regular and fine regular grids along the line $x_2 = 0.2$ (left) and along the line $x_2 = 0.4$ (right).

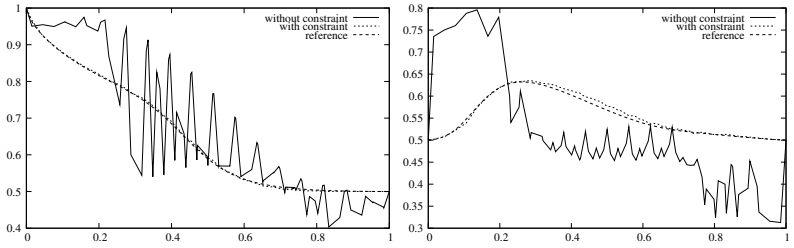

Figure 6: Profiles with irregular and fine regular grids along the line $x_2 = .2$ (left) and along the line $x_2 = .4$ (right).

streamlines, given in Figure 7, mainly follow the behavior of that provided by the reference solution (although there clearly remain effects of the distortion of the grid-blocks). Table 2 provides the number of iterations, with respect to

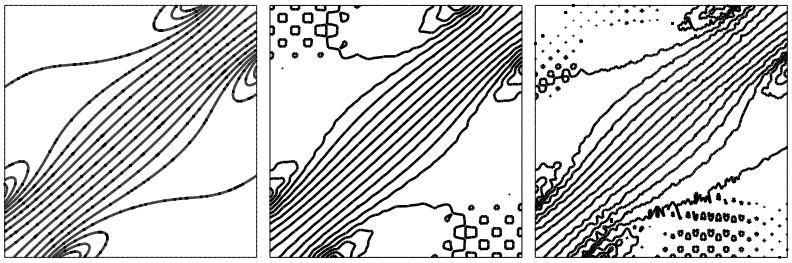

Figure 7: Stream lines: fine regular mesh (left), regular 20×20 grid (middle), irregular grid (right)

various values of the numerical parameters, for the regular 20×20 coarse grid. Parameter ρ is the length of the step along the gradient in Uzawa's algorithm (see (39)). Concerning the behavior of the method with respect to the numerical parameters, we observe similar results to those obtained in Section 4.1. Again, the strategy allowing the only increase of the Lagrangian multipliers within the iterative Uzawa's procedure provides $\beta_{\min}$ and $\beta_{\max}$-values strongly dependent of the parameter ρ but little sensitive to the values of ε, which determine the level for the approximate constraints. Table 3 provides the number of iterations, with respect to various values of the numerical parameters, for the irregular grid. We see that the same conclusions, applied to the case of the regular grid, hold for the irregular one. We remark that the number of iterations is higher in the case of the irregular grid, for the same choices of the numerical parameters. This seems to be compatible with the supplementary variations in the approximate solution due to the distortion of the grid blocks.

4.3 Two wells with anisotropy

We consider a test case which is very similar to test 9 described in [11], also provided in [1], which is focused on a problem with zero boundary flow, with two values of the unknowns imposed within two interior grid blocks of a given mesh. It is well-known that, under this form, this problem cannot be considered for a convergence study. We have therefore modified the data in order that this convergence to a reference solution could be numerically observed, providing the opportunity of exploring the behavior of the scheme on meshes with highly contrasted grid block sizes (such meshes are more and more used in oil engineering for modeling the source terms, since the classical approach by well indices fails in heterogeneous anisotropic or tilted wells cases). We consider the domain $\Omega = (0;1) \times (0;1) \setminus (\Omega_1 \cup \Omega_2)$, where $\Omega_1 = (0.308; 0.328) \times (0.49; 0.51)$ and $\Omega_2 = (0.672; 0.692) \times (0.49; 0.51)$, which represent two wells, with $u = 1$ on

ε	Nb. iterations	ρ	$\beta_{\min}$	$\beta_{\max}$
10^{-7}	3	10^6	2	150000
10^{-7}	10	10^5	0.2	15000
10^{-6}	10	10^5	0.1	15000
10^{-6}	22	10^4	0.01	1500

Table 2: Number of iterations for the regular 20×20 coarse grid.

ε	Nb. iterations	ρ	$\beta_{\min}$	$\beta_{\max}$
10^{-7}	11	10^6	4.5	300000
10^{-7}	18	10^5	0.5	30000
10^{-6}	16	10^5	10^{-9}	30000
10^{-6}	31	10^4	10^{-9}	3000

Table 3: Number of iterations for the irregular grid.

$\partial\Omega_1$ and $u = 0$ on $\partial\Omega_2$. The diffusion $\overline{\overline{\Lambda}}$ is assumed to be homogeneous, and given by:

$$\overline{\overline{\Lambda}} = R_\theta \begin{pmatrix} 1 & 0 \\ 0 & 10^{-3} \end{pmatrix} R_\theta^{-1},$$

where R_θ is the matrix of a rotation of angle $\theta = 67.5$ degrees and $f = 0$. We have computed this problem with two grids. The coarse one is depicted in the left part of Figure 8. The solution, obtained with the refined one (which is the coarse grid refined by a factor 10), is given in the right part of Figure 8. Note that the solution obtained on the refined grid

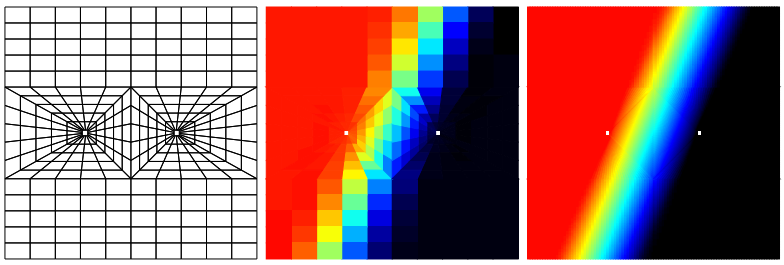

Figure 8: Left: The coarse grid for the two wells with anisotropy test case. Middle: the solution obtained with the coarse grid using the constrained scheme. Right: the solution obtained with the refined grid.

ensures $u(x) \in [0; 1]$, which is not the case for the coarse solution, whose minimum value is $-.04$ and maximum value is 1.04: in this case, the maximum principle is not satisfied by the coarse solution. We again get more precise indications of the regularity and the precision of the solutions by the study of the profiles of the solution along selected lines. We have therefore selected lines which cross the domains Ω_1 and Ω_2. We observe in Figure 9 that, again, the unconstrained solution is far from the reference one, but that the constrained solution is acceptably accurate and regular. We give in

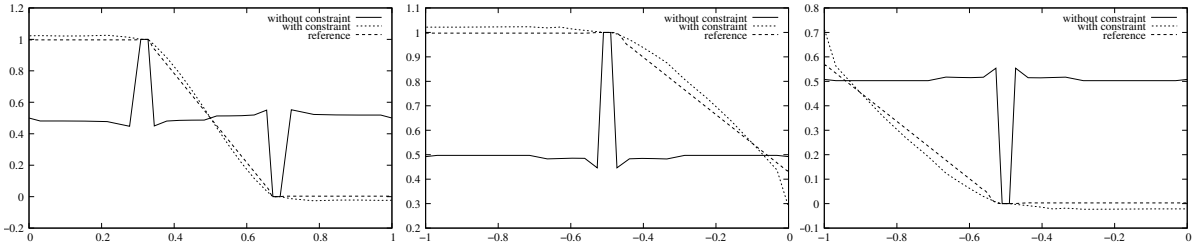

Figure 9: Left: Profiles along the line $x_2 = .5$. Middle: Profiles along the line $x_1 = .318$. Right: Profiles along the line $x_1 = .682$.

Table 4 the relation between the number of Uzawa's iterations and the range of the Lagrange multipliers, for a variety of numerical parameters.

ε	Nb. iterations	ρ	$\beta_{\min}$	$\beta_{\max}$
10^{-5}	3	10^5	2	7000
10^{-7}	40	10^5	2	7000
10^{-8}	45	10^5	2	7000
10^{-8}	3	10^6	21	70000

Table 4: Number of iterations and range of the Lagrange multiplier.

We remark in this table that the values of the Lagrange multiplier belong to the range $[2; 7000]$ (the highest values taken around the wells) for the first three choices, and to the range $[21; 70000]$ for the last line, showing again that the convergence is not reached. As in Section 4.1, the final value for the multiplier, provided by this algorithm, strongly depends on the value of ρ, but again does not seem to depend on ε. We observed very close numerical solutions in all the cases presented in this table. We again remark that taking high value of ρ allows to satisfy the constraints within a very small number of iterations.

4.4 One well in a distorted quadrilateral domain

We finally consider a test case which is very similar to test 8 described in [11], again also provided in [1]. Since, again, our aim is to examine a convergence behavior, we modify it slightly, replacing the Dirac source term by a boundary condition $p = 1$ on the boundary of a very small polygonal domain Ω_1 with vertices given in the trigonometric order by the coordinates (x_1, x_2) equal to $(.481, .0156)$, $(.501, .0156)$, $(.499, .0176)$ and $(.479, .0176)$. Hence, the domain is given as $\Omega = \Omega_2 \setminus \Omega_1$, where Ω_2 is the polygonal domain with vertices given in the trigonometric order by $(0., 0.)$, $(1., 0.)$, $(-0.0192, .0333)$ and $(0.9808, .0333)$. Then the boundary condition $p = 0$ is prescribed on $\partial\Omega_2$. We show in Figure 10 the coarse mesh (the refined one is 10 times finer), and the values of the solution (comprised between -10^{-3} and .947 for the coarse one using the constrained scheme, between -10^{-24} and .994 for the fine one). We see that again, the maximum principle is not respected by the approximate solution. The profiles given in Figure 11 show the efficiency of

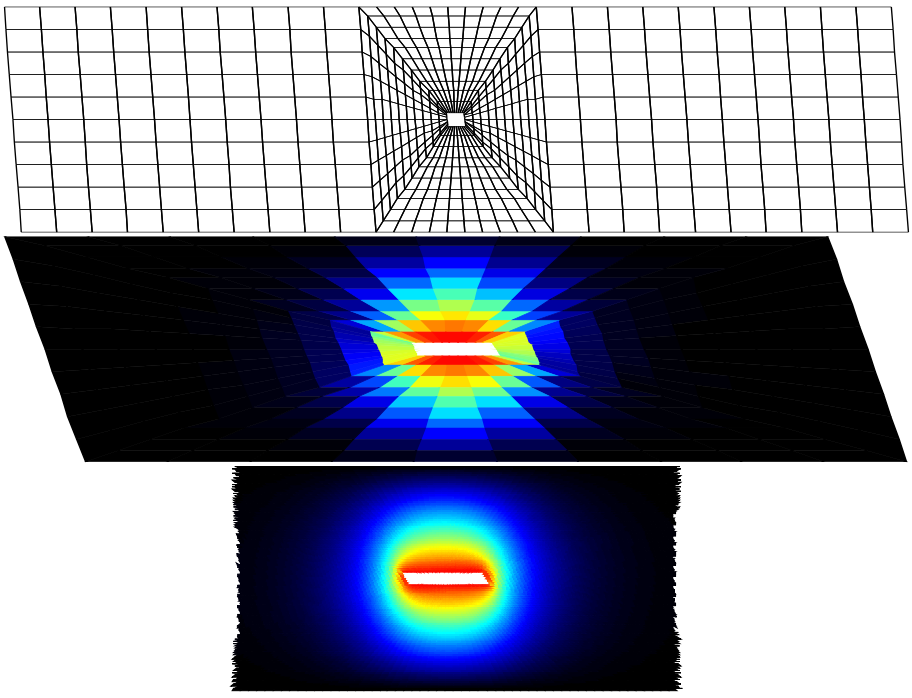

Figure 10: Top: The coarse grid for the one well in a distorted quadrilateral domain test case. Middle: zoom on the solution obtained with the coarse grid using the constrained scheme. Bottom: zoom on the solution obtained with the refined grid.

the constrained scheme (note that the unconstrained one is again far from the solution, and present dramatic oscillations). We show in Table 5 the number of iterations, together with the dependence of the final value of $\beta_{\max}$ on ρ.

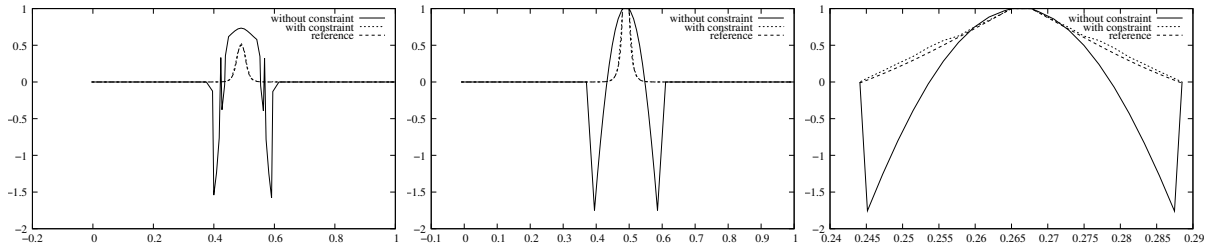

Figure 11: Left: Profiles along the line $x_2 = .00835$. Middle: Profiles along the line $x_2 = .0167$. Right: Profiles along the line $x_1 + \frac{x_2}{\sqrt{3}} = .5$.

ε	Nb. iterations	ρ	$\beta_{\min}$	$\beta_{\max}$
10^{-4}	6	10^5	10^{-9}	$3\ 10^6$
10^{-5}	61	10^5	10^{-9}	$3\ 10^6$
10^{-6}	95	10^5	10^{-9}	$3\ 10^6$
10^{-5}	3	10^6	10^{-9}	$3\ 10^7$

Table 5: Number of iterations and range of the Lagrange multiplier.

5 Conclusion

The method proposed in this paper aims at tuning the parameters of a numerical scheme, in this case the hybrid finite volume scheme, with identifying them as the Lagrange multipliers in a minimization problem under constraints. A series of numerical examples shows that this identification is successful, and that the constrained scheme satisfies the required criteria of stability and accuracy, even on distorted meshes and in the case of highly anisotropic problems. Note that further studies remain to be driven for the a priori assessment of the numerical parameters.

Nevertheless, this method does not cure in all cases one of the difficult up-to-date problems, which consists in computing solutions which do not violate the maximum principle. A research direction, opened by this paper, relies in formulations of this problem by minimization problems under constraints.

Appendix: the Lagrange multipliers (Kuhn-Tucker) theorem and Uzawa's algorithm

Let us first recall, for the sake of completeness, classical results which hold for regular minimization problems under constraints. Among many possible references, we refer to [7] for proofs of these results.

Theorem 5.1 (Lagrange multipliers)
Let V be a finite dimensional Euclidean space and let K be the convex closed subset of V, defined by

$$K = \{v \in V,\ G_i(v) \leq 0,\ \textit{for}\ 1 \leq i \leq p\}, \tag{36}$$

where $p \in \mathbb{N}^\star$ and for all $i = 1, \ldots, p$ the real function $G_i : V \to \mathbb{R}$ is convex and continuously differentiable. We assume that the set $\{v \in V,\ G_i(v) < 0, \textit{for all}\ 1 \leq i \leq p\}$ is non empty (sufficient condition for the qualification of the constraints). Let $J : V \to \mathbb{R}$ be a continuously differentiable strictly convex function such that $\lim_{|u|\to\infty} J(u) = +\infty$, and let $u^\star$ be the unique solution of the minimization problem

$$u^\star = \operatorname{argmin}_{u\in K} J(u). \tag{37}$$

Then there exists $\beta^\star = (\beta_i^\star)_{1\leq i\leq p} \in (\mathbb{R}_+)^p$ such that $(u^\star, \beta^\star)$ is a saddle point on $V \times (\mathbb{R}_+)^p$ of the function $\mathcal{L} : V \times \mathbb{R}^p \to \mathbb{R}$ defined by

$$\mathcal{L}(u, \beta) = J(u) + \sum_{i=1}^{p} \beta_i G_i(u).$$

This means that

$$\mathcal{L}(u^\star, \beta) \leq \mathcal{L}(u^\star, \beta^\star) \leq \mathcal{L}(u, \beta^\star),\ \forall (u, \beta) \in V \times (\mathbb{R}_+)^p.$$

Moreover, the so-called Kuhn and Tucker relations hold:

$$\left\{ \begin{array}{l} \nabla J(u^\star) + \sum_{i=1}^{p} \beta_i^\star \nabla G_i(u^\star) = 0, \\ \beta_i^\star G_i(u^\star) = 0, \ \forall i = 1, \ldots, p, \end{array} \right. \tag{38}$$

are satisfied. Reciprocally, if there exists $(u^\star, \beta^\star) \in K \times (\mathbb{R}_+)^p$ *such that relations (38) are satisfied, then* $u^\star = \mathrm{argmin}_{u \in K} J(u)$ *and* $(u^\star, \beta^\star)$ *is a saddle point on* $K \times (\mathbb{R}_+)^p$ *of the function* $\mathcal{L}$*.*

Let us now recall Uzawa's algorithm, to find an approximation of the solution $u^\star$ of (37). Let $\rho > 0$ be a given real and let $\beta^{(0)} \in (\mathbb{R}_+)^p$ be given. We define the sequence $(u^{(n)}, \beta^{(n)})_{n \in \mathbb{N}}$ by

$$\begin{array}{l} u^{(n)} = \mathrm{argmin}_{u \in V} \mathcal{L}(u, \beta^{(n)}), \\ \beta_i^{(n+1)} = \max(\beta_i^{(n)} + \rho G_i(u^{(n)}), 0), \ \forall i = 1, \ldots, p, \ \forall n \in \mathbb{N}. \end{array} \tag{39}$$

Theorem 5.2 (Convergence of Uzawa's algorithm) *Let* V *be a finite dimensional euclidean space and let* K *be the convex closed subset of* V*, defined by (36), where* $p \in \mathbb{N}^\star$ *and for all* $i = 1, \ldots, p$ *the real function* $G_i : V \to \mathbb{R}$ *is convex and continuously differentiable. We assume that the set* $\{v \in V, \ G_i(v) < 0,$ *for all* $1 \leq i \leq p\}$ *is non empty. Let* $J : V \to \mathbb{R}$ *be a continuously differentiable function such that there exists* $\alpha > 0$ *with*

$$(\nabla J(u) - \nabla J(v), u - v) \geq \alpha \|u - v\|^2, \ \forall u, v \in V, \tag{40}$$

(then J *is called "*α*-elliptic", which is sufficient to show that* J *is strictly convex and verifies* $\lim_{|u| \to \infty} J(u) = +\infty$*). Let us assume that there exists* $B \geq 0$ *such that, for all* $\beta \in (\mathbb{R}_+)^p$*,* $\| \mathrm{argmin}_{u \in V} \mathcal{L}(u, \beta)\| \leq B$*. Let* M *be defined by* $M = \max\{\sum_{i=1}^p \|\nabla G_i(u)\|^2, \ \|u\| \leq B\}$*. Then, for all* $\rho \in (0, \frac{\alpha}{2M})$ *and for all* $\beta^{(0)} \in (\mathbb{R}_+)^p$*, the sequence defined by (39) is such that* $(u^{(n)})_{n \in \mathbb{N}}$ *converges to the solution* $u^\star$ *of (37).*

References

[1] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, and J.M. Nordbotten. A compact multipoint flux approximation method with improved robustness. Numerical Methods for Partial Differential Equations, 24:1329–1360, 2008.

[2] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numer. Methods Partial Differential Equations, 23(1):145–195, 2007.

[3] E. Bertolazzi and G. Manzini. A second-order maximum principle preserving finite volume method for steady convection-diffusion problems. SIAM J. Numer. Anal., 43(5):2172–2199, 2005.

[4] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. SIAM J. Numer. Anal., 43(5):1872–1896, 2005.

[5] Franco Brezzi, Konstantin Lipnikov, and Valeria Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. Math. Models Methods Appl. Sci., 15(10):1533–1551, 2005.

[6] E. Burman and A. Ern. Discrete maximum principle for galerkin approximations of the laplace operator on arbitrary meshes. C. R. Acad. Sci. Paris, Ser. I, 338:641–646, 2004.

[7] J-C. Culioli. Introduction à l'optimisation, chapter Optimisation non-linéaire sous contraintes. Ellipses, 1994.

[8] R. Eymard, T. Gallouët, and R. Herbin. A new finite volume scheme for anisotropic diffusion problems on general grids: convergence analysis. C. R., Math., Acad. Sci. Paris, 344(6):403–406, 2007.

[9] R. Eymard, T. Gallouët, and R. Herbin. Discretisation of heterogeneous and anisotropic diffusion problems on general non-conforming meshes. SUSHI: a scheme using stabilisation and hybrid interfaces. Accepted for publication in IMAJNA, http://dx.doi.org/10.1093/imanum/drn084, 2009.

[10] R. Herbin. An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh. Numer. Methods Partial Differential Equations, 11(2):165–173, 1995.

[11] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In R. Eymard and J.-M. Hérard, editors, Finite Volumes for Complex Applications V, pages 659–692. Wiley, 2008.

[12] C. Le Potier. Schéma volumes finis pour des opérateurs de diffusion fortement anisotropes sur des maillages non structurés. C. R. Math. Acad. Sci. Paris, 340(12):921–926, 2005.

[13] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Yu. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. J. Comput. Phys., 227(1):492–512, 2007.

[14] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. J. Comput. Phys., 228(3):703–716, 2009.

[15] C. Le Potier. Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés. C.R.Acad.Sci.Paris Ser I, 341(12):787–792, 2005.

[16] C. Le Potier. Finite volume scheme satisfying maximum and minimum principles for anisotropic diffusion operators. In R. Eymard and J.-M. Hérard, editors, Finite Volumes for Complex Applications V, pages 103–118. Wiley, 2008.

[17] G. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. Journal of Computational Physics, 227:6288–6312, 2008.