Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations

Clément Cancès, Cindy Guichard

CONVERGENCE OF A NONLINEAR ENTROPY DIMINISHING CONTROL VOLUME FINITE ELEMENT SCHEME FOR SOLVING ANISOTROPIC DEGENERATE PARABOLIC EQUATIONS

CLÉMENT CANCÈS AND CINDY GUICHARD

ABSTRACT. In this paper, we propose and analyze a Control Volume Finite Elements (CVFE) scheme for solving possibly degenerated parabolic equations. This scheme does not require the introduction of the so-called Kirchhoff transform in its definition. We prove that the discrete solution obtained *via* the scheme remains in the physical range, and that the natural entropy of the problem decreases with time. The convergence of the method is proved as the discretization steps tend to 0. Finally, numerical examples illustrate the efficiency of the method.

1. INTRODUCTION

1.1. **Presentation of the problem.** Degenerate parabolic equations appear in the modeling many real life applications, and in particular in the modeling of multiphase or unsaturated flows in porous media [6]. Since it appears to be a keystone for studying complex porous media flows, we focus in the paper on the numerical approximation of an anisotropic *Generalized Porous Medium Equation* (in the sense of [42]). Moreover, despite our study can be extended to the 3-dimensional case, we restrict our study to the 2-dimensional case.

In what follows, we denote by Ω a connected polygonal open bounded subset of $\mathbb{R}^2$, by $t_{\rm f} > 0$ a finite time horizon, and by $Q_{t_{\rm f}} = \Omega \times (0, t_{\rm f})$. We are interested in approximating a convenient function p (p stands for *pressure* in analogy to porous media flows [6, 37]) such that

$$\begin{cases} \partial_t \beta(p) - \nabla \cdot (\eta(p) \Lambda \nabla p) = 0 & \text{in } Q_{t_{\rm f}}, \\ \beta(p)_{|_{t=0}} = \beta_0 & \text{in } \Omega, \\ \eta(p) \Lambda \nabla p \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0, t_{\rm f}). \end{cases} \tag{1}$$

In (1), the function β is a nondecreasing continuous function. Moreover, we assume that there exist $p_\star$ and $p^\star$ in $\overline{\mathbb{R}}$ with $p_\star < p^\star$ such that β is strictly increasing on $[p_\star, p^\star]$. Additionally, we assume that $\beta^{-1} : [\beta(p_\star), \beta(p^\star)] \to \overline{\mathbb{R}}$ belongs to $L^1(\beta(p_\star), \beta(p^\star))$. In particular, this enforces that

$$-\infty < \beta(p_\star) < \beta(p^\star) < +\infty.$$

The mobility function η is assumed to be bounded, continuous, and to fulfill

$$\eta(p) > 0 \text{ for a.e. } p \in (p_\star, p^\star) \text{ and } \eta(p) = 0 \text{ otherwise.} \tag{2}$$

Moreover, we assume all along this paper that

$$\int_{\mathbb{R}} \sqrt{\eta(a)} \mathrm{d}a < \infty, \qquad \lim_{p \to \pm\infty} p\eta(p) = 0. \tag{3}$$

This work was supported by the French National Research Agency ANR (project GeoPor, grant ANR-13-JS01-0007-01).

The permeability tensor Λ belongs to $(L^\infty(\Omega))^{2\times 2}$, and it is supposed to be symmetric and uniformly elliptic on Ω, i.e. there exist $\underline{\lambda}, \overline{\lambda} \in \mathbb{R}$ such that

$$\underline{\lambda}|\mathbf{v}|^2 \leq \Lambda(\mathbf{x})\mathbf{v}\cdot\mathbf{v} \leq \overline{\lambda}|\mathbf{v}^2|, \qquad \forall \mathbf{v} \in \mathbb{R}^2, \text{ for a.e. } \mathbf{x} \in \Omega. \tag{4}$$

The initial data β_0 belong to $L^\infty(\Omega; [\beta(p_\star), \beta(p^\star)])$. This assumption (together with the assumptions on β) ensures that there exists a measurable function $p_0 : \Omega \to [p_\star, p^\star]$ such that $\beta_0 = \beta(p_0)$.

We define the function $\Gamma : \mathbb{R} \to \mathbb{R}$ by

$$\Gamma(p) = \int_0^p a\beta'(a)\mathrm{d}a. \tag{5}$$

The the function $\Gamma \circ \beta^{-1}$ is convex on $[\beta(p_\star), \beta(p^\star)]$, and it follows from the definition (5) of Γ that

$$p\partial_t\beta(p) = \partial_t\Gamma(p), \qquad \text{for all } p \in \mathbb{R}. \tag{6}$$

Now, remark that since Ω is bounded, since $\beta_0 \in L^\infty(\Omega; [\beta(p_\star), \beta(p^\star)])$, and since β^{-1} belongs to $L^1((\beta(p_\star), \beta(p_\star)); \mathbb{R})$, then the finite entropy condition on the initial data

$$\int_\Omega \Gamma(p_0(\mathbf{x}))\mathrm{d}\mathbf{x} < \infty. \tag{7}$$

holds (see e.g. [14]).

Multiplying (formally) the first equation of (1) by p and integrating on Ω yields the following entropy dissipation property:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \Gamma(p(\mathbf{x},t))\mathrm{d}\mathbf{x} = -\int_\Omega \eta(p(\mathbf{x},t))\Lambda(\mathbf{x})\nabla p(\mathbf{x},t)\cdot\nabla p(\mathbf{x},t)\mathrm{d}\mathbf{x} \leq 0, \qquad \forall t > 0. \tag{8}$$

In particular, integrating w.r.t. $t \in (0, t_\mathrm{f})$ and using that $\Gamma(p) \geq 0$ and (7), one gets that

$$\iint_{Q_{t_\mathrm{f}}} \eta(p)\Lambda\nabla p\cdot\nabla p\,\mathrm{d}\mathbf{x}\mathrm{d}t < \infty. \tag{9}$$

The natural set the pressure p has to belong to is then

$$E = \{p : Q_{t_\mathrm{f}} \to \mathbb{R} \mid \text{ the estimate (9) holds}\}.$$

One of the main difficulty in the study of degenerate parabolic equations of type (1) comes from the fact that the natural topology for the pressure p degenerates when $\eta(p)$ vanishes. Indeed, given a solution p of (1), define by $\omega_\star$ (resp. $\omega^\star$) the subset of Q_{t_f} such that $p(\mathbf{x},t) \leq p_\star$ in $\omega_\star$ (resp $p(\mathbf{x},t) \geq p^\star$ in $\omega^\star$). Then, because of the degeneracy of the topology (9) any function $\tilde{p} : Q_{t_\mathrm{f}} \in \mathbb{R}$ such that

$$\tilde{p}(\mathbf{x},t) = \begin{cases} p(\mathbf{x},t) & \text{if } (\mathbf{x},t) \in Q_{t_\mathrm{f}} \setminus (\omega_\star \cup \omega^\star), \\ \tilde{p}(\mathbf{x},t) \leq p_\star & \text{if } (\mathbf{x},t) \in \omega_\star, \\ \tilde{p}(\mathbf{x},t) \geq p^\star & \text{if } (\mathbf{x},t) \in \omega^\star, \end{cases} \tag{10}$$

is a solution of (1). This lead for example to the notion of extended pressure proposed in [14, 9] (see also [12, 13]). Therefore, in order to select one solution, we impose that $p_\star \leq p \leq p^\star$.

Remark 1.1. *In the case where $\beta^{-1} \notin L^1(\beta(p_\star), \beta(p^\star))$, the analysis carried out throughout this paper still holds under the assumption (7) that the initial condition has a finite entropy.*

1.2. **Kirchhoff transform or not Kirchhoff transform.** As pointed out previously, the degeneracy of the system leads to a severe lack of control on the variations of p. In order to circumvent this difficulty from a mathematical point of view, the so-called *Kirchhoff transform* $\varphi(p)$ can be introduced.

More precisely, defining the nondecreasing Lipschitz continuous function $\phi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(p) = \int_0^p \eta(s)\mathrm{d}s, \qquad \forall p \in \mathbb{R},$$

then for regular enough functions p, one has

$$\eta(p)\nabla p = \nabla\varphi(p), \tag{11}$$

so that (1) can be rewritten

$$\begin{cases} \partial_t \beta(p) - \nabla \cdot (\Lambda \nabla \varphi(p)) = 0 & \text{in } Q_{t_\mathrm{f}}, \\ \beta(p)_{|_{t=0}} = \beta_0 & \text{in } \Omega, \\ \Lambda \nabla \varphi(p) \cdot \mathbf{n} = 0 & \text{in } \partial\Omega \times (0, t_\mathrm{f}). \end{cases} \tag{12}$$

Let us now introduce another closely related function. Define the Lipschitz continuous nondecreasing function $\xi : \mathbb{R} \to \mathbb{R}$ by

$$\xi(p) = \int_0^p \sqrt{\eta(s)}\mathrm{d}s, \qquad \forall p \in \mathbb{R}, \tag{13}$$

then the function ξ belongs to $L^\infty(\mathbb{R})$ thanks to (3). The natural entropy estimate (9) rewrites

$$\iint_{Q_{t_\mathrm{f}}} \Lambda \nabla \xi(p) \cdot \nabla \xi(p)\, \mathrm{d}\mathbf{x}\mathrm{d}t < \infty. \tag{14}$$

Due to the assumption (4), the relation (14) enforces that $\xi(p) \in L^2((0, t_\mathrm{f}); H^1(\Omega))$.

Both φ and ξ remain constant on $(-\infty, p_\star)$ and $(p^\star, +\infty)$. Hence, whatever the choice of a function $\tilde{p}$ as in (10), one has $\varphi(\tilde{p}) = \varphi(p)$ and $\xi(\tilde{p}) = \xi(p)$. Since φ and ξ are increasing on $[p_\star, p^\star]$, one can define their inverses $\varphi^{-1}, \xi^{-1} : [\xi(p_\star), \xi(p^\star)] \to [p_\star, p^\star]$.

This leads to the following notion of weak solution.

Definition 1.1 (weak solution)**.** *A measurable function $p : Q_{t_\mathrm{f}} \to [p_\star, p^\star]$ is said to be a weak solution of (1) if*

- $\xi(p) \in L^2((0, t_\mathrm{f}); H^1(\Omega))$*;*
- *for all $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_\mathrm{f}))$, one has*

$$\iint_{Q_{t_\mathrm{f}}} \beta(p)\partial_t\psi \mathrm{d}\mathbf{x}\mathrm{d}t + \int_\Omega \beta_0\psi(\cdot, 0)\mathrm{d}\mathbf{x} - \iint_{Q_{t_\mathrm{f}}} \Lambda\nabla\varphi(p) \cdot \nabla\psi \mathrm{d}\mathbf{x}\mathrm{d}t = 0. \tag{15}$$

Let us now state that the problem is well-posed in some convenient sense. The existence of a weak solution has been established in [2], while a complete uniqueness proof was given in [36].

Theorem 1.2 ([2, 36])**.** *Under the assumptions stated in §1.1, there exists a unique weak solution to the problem (1) in the sense of Definition 1.1.*

Despite they are of great help from a mathematical point of view, both $\varphi(p)$ and $\xi(p)$ have no clear physical meaning in opposition with p. This was underlined for example in the seminal work of Otto [37] in the case of the porous medium equation. As a consequence, they are of no use in the case where the physics becomes more complicated, like for example in presence of hysteresis (see e.g. [40]). Moreover, the control of the decay of the entropy $t \mapsto \int_\Omega \Gamma(p(\mathbf{x}, t))\mathrm{d}\mathbf{x}$ has been shown to play a crucial role for the long time behavior of the discrete solution in [18, 17].

Therefore, from a numerical point of view, one aims to discretize the problem in its form (1) rather than in its form (12), avoiding the introduction of the Kirchhoff transforms φ and ξ in the scheme. Nevertheless, the introduction of these function will allow to perform the numerical analysis.

Several contributions have been proposed for solving problems related to (1), either thanks to Finite Volumes or to Finite Elements. In the isotropic case, i.e. $\Lambda = \lambda \mathrm{I_d}$, let us mention the work of Eymard *et al.* [27] (see also [28, 25, 26]) for the convergence of a two-point flux approximation under a restrictive assumption on the mesh (the so-called *orthogonality condition*). In the anisotropic case, the convergence of a nonconforming finite element schemes is proposed in [30, 31], while the convergence of mixed/hybrid finite volume discretization is proven in [3]. The convergence of very general schemes based on the Kirchhoff transform towards the solution of a Stefan problem is addressed in [22]. In [27, 25, 26, 30, 31, 3], the discretization always uses the Kirchhoff φ.

Let us now focus on schemes that do not require the introduction of the Kirchhoff transform in their definition. Let us first mention the scheme proposed in [7]. Its convergence is addressed in [8] under the assumption that Λ is isotropic and that the mesh the so-called orthogonality condition. In the more complex case of multiphase flows in porous media, schemes avoiding the introduction of the Kirchhoff transform have been proposed in [34, 29, 38]. While no convergence proof is provided in [34], the isotropy of the porous medium and a restrictive orthogonality condition on the mesh are required in [29, 38].

1.3. **Outline of the paper.** The goal of this paper is to establish the convergence of a scheme inspired from one proposed in [34] in the anisotropic case, allowing further extensions to more complex physics and more general schemes in future works.

In §2, we define the scheme and we state the main result of our paper in Theorem 2.1. In §2.1.1, we introduce the primal triangular mesh and the corresponding dual barycentric mesh that are needed to define a CVFE scheme (see e.g. [5]).This leads to the introduction in §2.1.2 and §2.1.3 of two different reconstructions of the discrete solution, one being the usual $\mathbb{P}_1$ reconstruction, the other being piecewise constant. Difficulties in considering nonlinear test functions in linear numerical methods are underlined in §2.2. The nonlinear CVFE scheme is defined in §2.3, and the main result of the paper, i.e. the convergence of such a scheme, is stated in Theorem 2.1 of §2.4. In §3, we derive *a priori* estimates on the discrete solution, allowing to state the existence of such a discrete solution in §3.4. §4 is devoted to the convergence proof of the discrete solution when the size of the time and space discretization tends to 0. This is based first on the obtention of compactness on the family of the discrete solutions in §4.1, and then on the identification of the limit as the size of the time and space discretization tends to 0 in §4.2. Numerical results are finally proposed in §5. In particular, we give an evidence that using nonlinear schemes can be a very powerful tool for avoiding numerical locks even for solving linear equations.

2. The numerical scheme and main result

2.1. Discretization of Q_{t_f}.

2.1.1. *Space discretizations of* Ω. The CVFE method requires the introduction of two different space discretizations of Ω, namely the *primal triangular discretization* and the *dual barycentric discretization*.

We denote by $\mathcal{T}$ a triangular discretization of Ω made of a finite number $\#\mathcal{T}$ of triangles. We assume that $\bigcup_{T\in\mathcal{T}} \overline{T} = \overline{\Omega}$ (recall that Ω is assumed to be polygonal), and that $T \cap T' = \emptyset$ if $T \neq T'$. We denote by $\mathcal{V}$ the set of the vertices of the discretization $\mathcal{T}$, located at positions $(\mathbf{x}_K)_{K\in\mathcal{V}}$. The set $\mathcal{E}$ of the edges of $\mathcal{T}$ is made of straight segments σ joining two vertices of $\mathcal{V}$. Given $T, T' \in \mathcal{T}$, we assume that $\overline{T} \cap \overline{T}'$ is either empty, or it is reduced to $\mathbf{x}_K$ for some $K \in \mathcal{V}$, or it consists in an edge σ belonging $\mathcal{E}$. For $T \in \mathcal{T}$, we denote by $\mathcal{E}_T$ the subset of $\mathcal{E}$ made of the edges σ such that $\bigcup_{\sigma\in\mathcal{E}_T} \sigma = \partial T$. We assume that $\mathcal{E} = \bigcup_{T\in\mathcal{T}} \mathcal{E}_T$. Given two vertices $K, L \in \mathcal{V}$ of a triangle T, then the edge joining $\mathbf{x}_K$ and $\mathbf{x}_L$ is denoted by σ_{KL}.

For $T \in \mathcal{T}$, we denote by $\mathbf{x}_T$ the center of gravity of T, by h_T the diameter of the triangle T, and by ρ_T the diameter of the incircle of T. Then, we define the mesh diameter h and the mesh regularity $\theta_{\mathcal{T}}$ by

$$h = \max_{T\in\mathcal{T}} h_T, \qquad \theta_{\mathcal{T}} = \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}.$$

For $K \in \mathcal{V}$, one denotes by $\mathcal{T}_K$ the subset of $\mathcal{T}$ made the triangles admitting K as a vertex, by $\mathcal{E}_K$ the set of edges having the vertex K at an extremity, and by $\mathcal{V}_K$ the subset of $\mathcal{V}$ such that, if $L \in \mathcal{V}_K$, then $[\mathbf{x}_K, \mathbf{x}_L]$ is an edge of $\mathcal{E}_K$.

Once the primal triangular mesh has been built, we can define its dual barycentric mesh $\mathcal{M}$ as follows. To each $K \in \mathcal{V}$ corresponds a cell ω_K whose vertices are the isobarycenters $\mathbf{x}_T$ of the triangles $T \in \mathcal{T}_K$ and the isobarycenters $\mathbf{x}_\sigma$ of the edges $\sigma \in \mathcal{E}_K$. Note that $\bigcup_{K\in\mathcal{V}} \overline{\omega}_K = \overline{\Omega}$. The 2-dimensional Lebesgue measure of ω_K is denoted by m_K.

The construction of the primal triangular and dual barycentric meshes is illustrated on Figure 1.

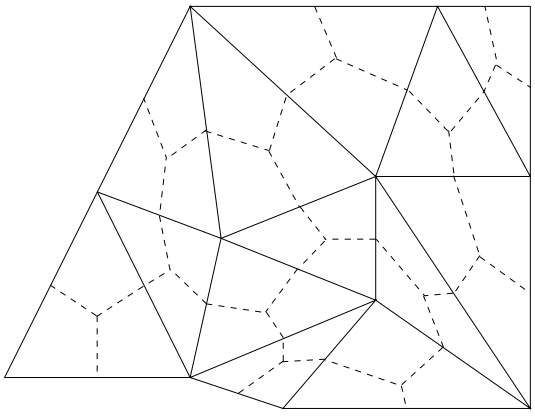


FIGURE 1. The triangular mesh $\mathcal{T}$ (solid line) and its corresponding dual barycentric dual mesh $\mathcal{M}$ (dashed line).

2.1.2. *The discrete spaces $V_{\mathcal{T}}$ and $X_{\mathcal{M}}$.* From the two meshes built in §2.1.1, we construct two discrete functional spaces. The first one, namely $V_{\mathcal{T}}$, is the usual conforming $\mathbb{P}_1$-finite elements space corresponding to the triangular mesh $\mathcal{T}$, containing piecewise affine elements:

$$V_{\mathcal{T}} = \{f \in H^1(\Omega) \mid f_{|_T} \in \mathbb{P}_1(\mathbb{R}), \quad \forall T \in \mathcal{T}\}.$$

On the other hand, we define the space $X_{\mathcal{M}}$ of the piecewise constant functions on the dual mesh $\mathcal{M}$ by

$$X_{\mathcal{M}} = \{f : \Omega \to \overline{\mathbb{R}} \text{ measurable} \mid f_{|_{\omega_K}} \in \overline{\mathbb{R}} \text{ is constant}, \quad \forall K \in \mathcal{V}\}.$$

Given a vector $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, there exists a unique $u_\mathcal{T} \in V_\mathcal{T}$ and a unique $u_\mathcal{M} \in X_\mathcal{M}$ such that

$$u_\mathcal{T}(\mathbf{x}_K) = u_\mathcal{M}(\mathbf{x}_K) = u_K, \qquad \forall K \in \mathcal{V}. \tag{16}$$

As a consequence, there are one-to-one mappings between $\mathbb{R}^{\#\mathcal{V}}$, $V_\mathcal{T}$ and $X_\mathcal{M}$.

In what follows, we denote by $(e_K)_{K\in\mathcal{V}}$ the canonical basis of $V_\mathcal{T}$, characterized by

$$e_K(\mathbf{x}_K) = 1, \quad e_K(\mathbf{x}_L) = 0 \text{ if } L \neq K, \quad \forall K \in \mathcal{V}.$$

The following geometrical property holds:

$$\int_\Omega e_K(\mathbf{x})\mathrm{d}\mathbf{x} = m_K, \qquad \forall K \in \mathcal{V}.$$

2.1.3. *Space-time discretization of Q_{t_f}.* In order to simplify the notations, we restrict our study to the case of a uniform time discretization of $(0, t_\mathrm{f})$, but all the results presented in this paper can be extended to the case of a general time discretization without any difficulty. Moreover, we assume, still for the sake of simplicity, that the spacial meshes do not change with the time step.

Let N be a nonnegative integer, then we define $\Delta t = t_\mathrm{f}/(N+1)$, and $t_n = n\Delta t$, so that $t_0 = 0$, and $t_{N+1} = t_\mathrm{f}$. Then we define the space and time discrete spaces $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ as the set of piecewise constant functions in time with values in $V_\mathcal{T}$ and $X_\mathcal{M}$ respectively:

$$V_{\mathcal{T},\Delta t} = \left\{ f \in L^2((0,t_\mathrm{f}); H^1(\Omega)) \mid f(x,t) = f(x,t^{n+1}) \in V_\mathcal{T}, \quad \forall t \in (t_n, t_{n+1}] \right\},$$
$$X_{\mathcal{M},\Delta t} = \left\{ f : Q_{t_\mathrm{f}} \to \overline{\mathbb{R}} \text{ measurable} \mid f(x,t) = f(x,t^{n+1}) \in X_\mathcal{M}, \quad \forall t \in (t_n, t_{n+1}] \right\},$$

where

Given $\left(u_K^{n+1}\right)_{n\in\{0,\dots,N\},K\in\mathcal{V}} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$, we denote by $u_{\mathcal{T},\Delta t}$ and $u_{\mathcal{M},\Delta t}$ the unique elements of $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ respectively such that

$$u_{\mathcal{T},\Delta t}(\mathbf{x}_K, t) = u_\mathcal{M}(\mathbf{x}_K, t) = u_K^{n+1}, \qquad \forall K \in \mathcal{V}, \forall t \in (t_n, t_{n+1}]. \tag{17}$$

For all $n \in \{0,\dots,N\}$, we also denote introduce the functions

$$u_\mathcal{T}^{n+1} = u_{\mathcal{T},\Delta t}(\cdot, t_{n+1}), \qquad u_\mathcal{M}^{n+1} = u_{\mathcal{M},\Delta t}(\cdot, t_{n+1}).$$

2.2. **Finite elements and nonlinear test functions.** For $(K,L) \in \mathcal{V}^2$, we define the coefficient

$$a_{KL} = -\int_\Omega \Lambda(\mathbf{x})\nabla e_K(\mathbf{x}) \cdot \nabla e_L(\mathbf{x})\mathrm{d}\mathbf{x} = a_{LK}. \tag{18}$$

Since $\sum_{K\in\mathcal{V}} \nabla e_K = 0$, one obtains that

$$a_{KK} = \sum_{L\neq K} a_{KL} > 0. \tag{19}$$

As a consequence of (18)–(19), given $u_\mathcal{T}$ and $v_\mathcal{T}$ two elements of $V_\mathcal{T}$, one has

$$\int_\Omega \Lambda\nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T}\mathrm{d}\mathbf{x} = \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(u_K - u_L)(v_K - v_L). \tag{20}$$

Let ϕ be a locally Lipschitz continuous nondecreasing function, then denoting by ζ a nondecreasing function such that $(\zeta')^2 = \phi'$, the Cauchy-Schwarz inequality yields that for all a, b, one has $(a-b)(\phi(a)-\phi(b)) \geq (\zeta(a)-\zeta(b))^2$. As a consequence, denoting by $\phi_\mathcal{T}$ (resp. $\zeta_\mathcal{T}$) the function of $V_\mathcal{T}$ with nodal values $(\phi(u_K))_{K\in\mathcal{V}}$ (resp. $(\zeta(u_K))_{K\in\mathcal{V}}$), then if

$$a_{KL} \geq 0 \text{ for all } \sigma_{KL} \in \mathcal{E}, \tag{21}$$

one has

$$\int_\Omega \Lambda\nabla u_\mathcal{T}\cdot\nabla\phi_\mathcal{T}\mathrm{d}\mathbf{x} = \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(u_K-u_L)(\phi(u_K)-\phi(u_L)) \geq \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(\zeta(u_K)-\zeta(u_L))^2 = \int_\Omega \Lambda\nabla\zeta_\mathcal{T}\cdot\nabla\zeta_\mathcal{T}\mathrm{d}\mathbf{x}\geq 0. \tag{22}$$

In particular, the fact that (22) holds for any nondecreasing function ϕ yields the monotonicity of the discrete counterpart of the anisotropic diffusion operator $(u\mapsto -\nabla\cdot(\Lambda\nabla u))^{-1}$.

In the case where Λ is isotropic, and if all the triangles of $\mathcal{T}$ have only acute angles, the condition (21) holds, as well as property (22). This was used for example in [16] for proving the convergence of the finite element approximation towards the renormalized solution of an elliptic equation, or in [39] for proving the convergence of a finite-volume finite element approximation for compressible multiphase flows in porous media.

A property similar to (22) holds for the two-point flux finite volume approximation (see e.g. [24]). This was, among other, used in [26] for proving the convergence of a monotone scheme towards the entropy solution of a strongly degenerate parabolic problem (see [15]), and in [20] for the approximation of the solution to non-coercive parabolic equations.

Nevertheless, for general triangulations $\mathcal{T}$ of Ω, and/or in the case of anisotropic tensors Λ, it is well known that some coefficients a_{KL} may become negative. Therefore, the relation (22) (and thus the monotonicity of the discrete diffusion operator) is lost.

2.3. **The nonlinear CVFE scheme.** The discretization $\beta^0_\mathcal{M}\in X_\mathcal{M}$ of the initial data is defined by

$$\beta^0_\mathcal{M}(\mathbf{x}) = \beta^0_K = \frac{1}{m_K}\int_{\omega_K}\beta_0(\mathbf{y})\mathrm{d}\mathbf{y},\quad \forall\mathbf{x}\in\omega_K, \forall K\in\mathcal{V}. \tag{23}$$

As a consequence of Jensen's inequality and of (7), one has

$$\int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M},0})\mathrm{d}\mathbf{x}\leq\int_\Omega\Gamma\circ\beta^{-1}(\beta_0)\mathrm{d}\mathbf{x}<\infty. \tag{24}$$

Let us now define what we call the *nonlinear control volume finite element scheme*. For all $K\in\mathcal{V}$, and for all $n\in\{0,\dots,N\}$,

$$\frac{\beta(p_K^{n+1})-\beta_K^n}{\Delta t}m_K + \sum_{\sigma_{KL}\in\mathcal{E}_K}\eta_{KL}^{n+1}a_{KL}\left(p_K^{n+1}-p_L^{n+1}\right)=0, \tag{25}$$

where $\beta_K^0=\frac{1}{m_K}\int_{\omega_K}\beta_0(\mathbf{x})\mathrm{d}\mathbf{x}$ and $\beta_K^n=\beta(p_K^n)$ if $n\geq 1$, and where, denoting by

$$I_{KL}^{n+1} = \begin{cases}[p_K^{n+1},p_L^{n+1}] & \text{if } p_K^{n+1}\leq p_L^{n+1},\\ [p_L^{n+1},p_K^{n+1}] & \text{otherwise,}\end{cases}$$

we have set

$$\eta_{KL}^{n+1} = \begin{cases}\max_{p\in I_{KL}^{n+1}}\eta(p) & \text{if } a_{KL}\geq 0,\\ \min_{p\in I_{KL}^{n+1}}\eta(p) & \text{if } a_{KL}\leq 0.\end{cases} \tag{26}$$

This scheme, whose construction is based on finite elements *via* (18), can be interpreted as a finite volume scheme. Indeed, denoting by

$$F_{KL}^{n+1} = a_{KL}\eta_{KL}^{n+1}(p_K-p_L),$$

the scheme (25) can be rewritten under the locally conservative form on the dual cells ω_K:

$$\begin{cases} F_{KL}^{n+1} + F_{LK}^{n+1} = 0, & \text{for all } \sigma_{KL} \in \mathcal{E}, \\ \dfrac{\beta(p_K^{n+1}) - \beta_K^n}{\Delta t} m_K + \displaystyle\sum_{\sigma_{KL} \in \mathcal{E}_K} F_{KL}^{n+1} = 0, & \text{for all } K \in \mathcal{V}. \end{cases} \tag{27}$$

As a straightforward consequence, we can claim that the scheme is globally conservative:

$$\sum_{K \in \mathcal{V}} \beta(p_K^{n+1}) m_K = \sum_{K \in \mathcal{V}} \beta_K^n m_K, \qquad \forall n \in \{0, \dots, N\}. \tag{28}$$

2.4. **Main result.** Consider a sequence $(\mathcal{T}_m)_{m \geq 1}$ of triangulations of Ω such that

$$h_m = \max_{T \in \mathcal{T}_m} \operatorname{diam}(T) \to 0 \quad \text{as } m \to \infty. \tag{29}$$

We assume that there exists $\theta > 0$ such that

$$\theta_{\mathcal{T}_m} \leq \theta, \qquad \forall m \geq 1. \tag{30}$$

As a by-product, a sequence of dual meshes $(\mathcal{M}_m)_{m \geq 1}$ is given.

Let $(N_m)_m$ be an increasing sequence of integers, then we define the corresponding sequence of time steps $(\Delta t_m)_m$ tending to 0 as m tends to ∞.

The purpose of this paper is to prove the following theorem.

Theorem 2.1. *Given a sequence of spatial discretization $(\mathcal{T}_m)_m$ and time discretization $(\Delta t_m)_m$, then there exists a corresponding sequence $(p_{\mathcal{M}_m, \Delta t_m})_m$ of solutions to the scheme (25) such that $p_\star \leq p_{\mathcal{M}_m, \Delta t_m} \leq p^\star$ a.e. in Q_{t_f}.*

Moreover, if (29) and (30) hold, then

$$p_{\mathcal{M}_m, \Delta t_m} \to p \qquad \text{a.e. in } Q_{t_\mathrm{f}} \text{ as } m \to \infty,$$

where p is the unique weak solution to the problem (1) in the sense of Definition 1.1.

In view of Theorem 1.2, it is sufficient to prove appropriate compactness properties for the sequence $(p_{\mathcal{M}_m, \Delta t_m})_m$, and to prove that, up to a subsequence, it converges towards a weak solution. Indeed, the uniqueness of the weak solution ensures automatically the convergence of the whole sequence.

3. *A priori* estimates and existence of a discrete solution

3.1. **Some preliminary material.** The first Lemma we state is of great importance, since it will replace partially the property (22) that is lost if some coefficients a_{KL} are negative.

Lemma 3.1. *Let $\left(p_K^{n+1}\right)_{K,n} \in \overline{\mathbb{R}}^{(N+1)\#\mathcal{V}}$, then denoting by $\xi_{\mathcal{T},\Delta t}$ the unique function of $V_{\mathcal{T},\Delta t}$ with nodal values $\left(\xi(p_K^{n+1})\right)_{K,n} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$, one has*

$$\begin{aligned} &\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 \\ &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 = \iint_{Q_{t_\mathrm{f}}} \Lambda \nabla \xi_{\mathcal{T},\Delta t} \cdot \nabla \xi_{\mathcal{T},\Delta t} \mathrm{d}\mathbf{x}\mathrm{d}t. \end{aligned} \tag{31}$$

Proof. For all $\sigma_{KL} \in \mathcal{E}$, it follows from the definition (13) of ξ and the definition (26) that for all $n \geq 0$,

$$a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \geq a_{KL}\left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2.$$

The result of Lemma 3.1 follows. □

Let $T \in \mathcal{T}$, and let $(K, L) \in \mathcal{V}^2$, then we denote by

$$\alpha_{KL}^T = -\int_T \Lambda \nabla e_K \cdot \nabla e_L \mathrm{d}\mathbf{x} = \alpha_{LK}^T. \tag{32}$$

In particular, if $\alpha_{KL}^T \neq 0$, then both K and L belong to $\mathcal{V}_T$, and

$$\sum_{T \in \mathcal{T}} \alpha_{KL}^T = a_{KL}, \qquad \forall \sigma_{KL} \in \mathcal{E}.$$

Remark 3.1. *Despite* $\left(p_K^{n+1}\right)_{K,n} \in \overline{\mathbb{R}}^{(N+1)\#\mathcal{V}}$ *can contain infinite values, the sums appearing in* (31) *are well defined. Indeed, thanks to* (3)*, the function* ξ *is uniformly bounded, ensuring that the right-hand side has a sense. Concerning the left-hand side, if* $a_{KL} > 0$*, then the corresponding term in the sum can be equal to* $+\infty$ *(it is anyway nonnegative). If* $a_{KL} < 0$*, then either* p_K^{n+1} *and* p_L^{n+1} *belong to the open interval* $(p_\star, p^\star)$*, yielding a finite contribution, or* $\eta_{KL}^{n+1} = 0$*, yielding a null contribution thanks to* (3)*.*

Lemma 3.2. *Let* $\Psi_{\mathcal{T}} = \sum_{K \in \mathcal{V}} \psi_K e_K \in V_{\mathcal{T}}$*, then there exists* C_0 *depending only on* Λ, θ *such that*

$$\sum_{\sigma_{KL} \in \mathcal{E}} \sum_{T \in \mathcal{T}} |\alpha_{KL}^T|(\psi_K - \psi_L)^2 \leq C_0 \int_\Omega \Lambda \nabla \Psi_{\mathcal{T}} \cdot \nabla \Psi_{\mathcal{T}} \mathrm{d}\mathbf{x}.$$

Proof. In the proof below, unless specified, C denotes a generic quantity depending only on Λ and $\theta_{\mathcal{T}}$. First, notice that

$$\|\nabla \Psi_{\mathcal{T}}\|_{L^2(\Omega)}^2 \leq \frac{1}{\underline{\lambda}} \int_\Omega \Lambda \nabla \Psi_{\mathcal{T}} \cdot \nabla \Psi_{\mathcal{T}} \mathrm{d}\mathbf{x},$$

so that it only remains to prove that

$$\sum_{\sigma_{KL} \in \mathcal{E}} \sum_{T \in \mathcal{T}} |\alpha_{KL}^T|(\psi_K - \psi_L)^2 \leq C\|\nabla \Psi_{\mathcal{T}}\|_{L^2(\Omega)}^2.$$

It follows from Cauchy-Schwarz inequality that

$$|\alpha_{KL}^T| \leq \overline{\lambda}\|\nabla e_K\|_{L^2(T)}\|\nabla e_L\|_{L^2(T)}.$$

Using the classical inequality (see e.g. [21, 10]), one has

$$\|\nabla e_K\|_{L^2(T)} \leq c\theta_{\mathcal{T}} \frac{|T|}{(h_T)^2}, \qquad \forall K \in \mathcal{V}, \forall T \in \mathcal{T},$$

where c is an absolute constant, so that

$$|\alpha_{KL}^T| \leq C, \qquad \forall T \in \mathcal{T}, \quad \forall \sigma_{KL} \in \mathcal{E}_T. \tag{33}$$

This implies that

$$\sum_{\sigma_{KL} \in \mathcal{E}_T} |\alpha_{KL}^T|(\psi_K - \psi_L)^2 \leq C\frac{|T|}{(h_T)^2} \sum_{\sigma_{KL} \in \mathcal{E}_T} (\psi_K - \psi_L)^2.$$

Now, it follows from the analysis carried out for example in [10] that for all $T \in \mathcal{T}$, K, L, M being its vertices,

$$\frac{|T|}{(h_T)^2}\left((\psi_K - \psi_L)^2 + (\psi_K - \psi_M)^2 + (\psi_L - \psi_M)^2\right) \leq C\|\nabla \Psi_{\mathcal{T}}\|_{L^2(T)}^2,$$

so that, since $\sigma_{KL} \in \mathcal{E}$ is shared by at most two triangles, one has

$$\sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} |\alpha_{KL}^T|(\psi_K - \psi_L)^2 \leq C \sum_{T\in\mathcal{T}} \|\nabla\Psi_\mathcal{T}\|_{L^2(T)}^2 = C\|\nabla\Psi_\mathcal{T}\|_{L^2(\Omega)}^2,$$

concluding the proof of Lemma 3.2. □

Lemma 3.3. *There exists C_1 depending only on Λ and $\theta_\mathcal{T}$ such that*

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 \leq C_1 \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2. \tag{34}$$

Proof. Denoting by $\mathcal{E}^- = \{\sigma_{KL} \in \mathcal{E} \mid a_{KL} < 0\}$, one has

$$\begin{aligned}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 &= \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 \\ &\quad + 2\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}^-} |a_{KL}|\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2.\end{aligned} \tag{35}$$

It follows from the definition (26) of η_{KL}^{n+1} that

$$\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 \leq \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2, \quad \forall \sigma_{KL} \in \mathcal{E}^-.$$

Plugging it into (35) yields

$$\begin{aligned}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 &\leq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1} \left(p_K^{n+1} - p_L^{n+1}\right)^2 \\ &\quad + 2\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}| \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2.\end{aligned} \tag{36}$$

The technical Lemma 3.2 ensures the existence of $C_0 > 0$ depending only on Λ and $\theta_\mathcal{T}$ such that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}| \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 \leq C_0 \iint_{Q_{t_\mathrm{f}}} \Lambda\nabla\xi_{\mathcal{T},\Delta t} \cdot \nabla\xi_{\mathcal{T},\Delta t} \mathrm{d}\mathbf{x}\mathrm{d}t.$$

We deduce from Lemma 3.1 that estimate (34) holds with $C_1 = 1 + 2C_0(\Lambda, \theta_\mathcal{T})$. □

Now, we introduce the notion of *transmissive path*.

Definition 3.4. *A* transmissive path w *joining $K_\mathrm{i} \in \mathcal{V}$ to $K_\mathrm{f} \in \mathcal{V}$ consists in a list of vertices $(K_q)_{0\leq q\leq M}$ such that $K_\mathrm{i} = K_0$, $K_\mathrm{f} = K_M$, with $K_q \neq K_\ell$ if $q \neq \ell$, and such that $\sigma_{K_qK_{q+1}} \in \mathcal{E}$ with $a_{K_qK_{q+1}} > 0$ for all $q \in \{0, \dots, M-1\}$. We denote by $\mathcal{W}(K_\mathrm{i}, K_\mathrm{f})$ the set of the transmissive path joining $K_\mathrm{i} \in \mathcal{V}$ to $K_\mathrm{f} \in \mathcal{V}$.*

Lemma 3.5. *For all $(K_\mathrm{i}, K_\mathrm{f}) \in \mathcal{V}^2$ there exists a transmissive path* w $\in \mathcal{W}(K_\mathrm{i}, K_\mathrm{f})$.

Proof. Let $K_{\rm i} \in \mathcal{V}$, then define $\overline{\mathcal{V}}_{K_{\rm i}}$ the subset of $\mathcal{V}$ made of the vertices connected to $K_{\rm i}$ *via* a transmissive path. Assume that $\overline{\mathcal{V}}_{K_{\rm i}} \subsetneq \mathcal{V}$.

We introduce the function $u_{\mathcal{T}} \in V_{\mathcal{T}}$ such that

$$u_K = \begin{cases} 1 & \text{if } K \in \overline{\mathcal{V}}_{K_{\rm i}} \\ 0 & \text{otherwise.} \end{cases}$$

The lack of transmissive path between the elements of $\overline{\mathcal{V}}_{K_{\rm i}}$ and the elements of $\mathcal{V} \setminus \overline{\mathcal{V}}_{K_{\rm i}}$ leads to

$$\sum_{\sigma_{KL} \in \mathcal{E}} (a_{KL})^+ (u_K - u_L)^2 = 0.$$

On the other hand, since $\mathcal{V} \setminus \overline{\mathcal{V}}_{K_{\rm i}} \neq \emptyset$, the function $u_{\mathcal{T}}$ is not constant. Therefore, since Ω is assumed to be connected,

$$\sum_{\sigma_{KL} \in \mathcal{E}} (a_{KL})^+ (u_K - u_L)^2 \geq \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} (u_K - u_L)^2 = \int_\Omega \Lambda \nabla u_{\mathcal{T}} \cdot \nabla u_{\mathcal{T}} \mathrm{d}\mathbf{x} > 0.$$

providing a contradiction. The fact that the path is necessarily of finite length stems from the finite number of possible combinations for designing a path. □

3.2. **A uniform L^∞-estimate on $\beta(p_{\mathcal{M},\Delta t})$.** In what follows, $\left(p_K^{n+1}\right)_{K,n}$ denotes a solution to the scheme (25) (whose existence will be established latter).

Proposition 3.6. *For all $K \in \mathcal{V}$ and all $n \in \{0, \dots, N+1\}$, one has*

$$-\infty < \beta(p_\star) \leq \beta_K^n \leq \beta(p^\star) < +\infty. \tag{37}$$

Proof. The property (37) clearly holds for $n = 0$ thanks to the assumption on β_0 and to (23). Assume now (37) holds at time step n, and assume that

$$\beta_{K_m}^{n+1} = \max_{L \in \mathcal{V}} \beta(p_L^{n+1}) > \beta(p^\star),$$

implying that $p_{K_m}^{n+1} > p^\star$. In view of the definition (26) of $\eta_{K_m L}^{n+1}$, and of the fact that $\eta(p) = 0$ if $p \geq p^\star$, one has that

$$\eta_{K_m L}^{n+1} = 0 \quad \text{if} \quad a_{K_m L} \leq 0.$$

Therefore, the scheme (25) at the vertex K_m rewrites

$$0 \leq \frac{\Delta t}{m_K} \sum_{L \in \mathcal{V}_K} (a_{K_m L})^+ \eta_{K_m L}^{n+1} (p_{K_m}^{n+1} - p_L^{n+1}) = \beta_{K_m}^n - \beta_{K_m}^{n+1} < 0,$$

yielding a contradiction, ensuring that

$$\beta_K^{n+1} \leq \beta_{K_m}^{n+1} \leq \beta(p^\star), \qquad \forall K \in \mathcal{V}.$$

Proving that $\beta_K^{n+1} \geq \beta(p_\star)$ for all $K \in \mathcal{V}$ is similar. □

Remark 3.2. *The Proposition 3.6 holds for any choice of nondecreasing function β. This means that the behavior of β outside of $[p_\star, p^\star]$ does not affect the solution to the scheme (25). We then choose arbitrarily to extend β by*

$$\beta(p) = \begin{cases} \beta(p_\star) + p - p_\star & \text{if } p < p_\star, \\ \beta(p^\star) + p - p^\star & \text{if } p > p_\star. \end{cases} \tag{38}$$

Recall here that $p_\star$ (resp. $p^\star$) can be equal to $-\infty$ (resp. $+\infty$). As a consequence, Proposition 3.6 yields

$$p_\star \leq p_K^{n+1} \leq p^\star, \qquad \forall K \in \mathcal{V}, \forall n \in \{0, \dots, N\}. \tag{39}$$

Note that this estimate only provides some information for finite $p_\star$ or $p^\star$.

Remark 3.3. *Defining the functions* $(M_K)_{K\in\mathcal{V}}$ *and* $(R_K)_{K\in\mathcal{V}}$ *by*

$$M_K\left(p_K^{n+1}, p_K^n, \left(p_L^{n+1}\right)_{L\neq K}\right) = \frac{\beta(p_K^{n+1}) - \beta(p_K^n)}{\Delta t} m_K + \sum_{L\in\mathcal{V}_K} (a_{KL})^+ \eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1}),$$

$$R_K\left(p_K^{n+1}, \left(p_L^{n+1}\right)_{L\neq K}\right) = \sum_{L\in\mathcal{V}_K} (a_{KL})^- \eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1}),$$

then M_K *is nondecreasing w.r.t.* p_K^{n+1} *and non-increasing w.r.t. its other arguments, while* $R_K\left(p_K^{n+1}, \left(p_L^{n+1}\right)_{L\neq K}\right) = 0$ *as soon as* $p_K^{n+1} \notin (p_\star, p^\star)$*. Hence, the scheme* (25) *can be decomposed into the sum of a monotone part (corresponding to* M_K*) and a residual part (corresponding to* R_K*) which vanishes in the degenerate regions.*

3.3. **A "nonlinear" entropy estimate.** The goal of this section is to establish a discrete counterpart to the entropy estimates (8) and (14). In what follows, $\left(p_K^{n+1}\right)_{K,n}$ denotes a solution to the scheme (25) (whose existence will be established latter).

Proposition 3.7. *For all* $n \in \{0, \dots, N\}$*, one has*

$$\int_\Omega \Gamma(p_{\mathcal{M}}^{n+1})\mathrm{d}\mathbf{x} + \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \leq \int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M}}^n)\mathrm{d}\mathbf{x}. \tag{40}$$

Proof. Multiplying the scheme (25) by $\Delta t p_K^{n+1}$ and summing over $K \in \mathcal{V}$ provides after reorganizing the sums

$$\mathcal{A} + \mathcal{B} = 0,$$

where

$$\mathcal{A} = \sum_{K\in\mathcal{V}} \left(\beta(p_K^{n+1}) - \beta_K^n\right) p_K^{n+1} m_K,$$
$$\mathcal{B} = \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2,$$

By convexity of $\Gamma_{\mathcal{M}}(\beta^{-1})$, one deduces that $\mathcal{A}$ can be underestimated as follows:

$$\mathcal{A} \geq \int_\Omega \Gamma(p_{\mathcal{M}}^{n+1})\mathrm{d}\mathbf{x} - \int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M}}^n)\mathrm{d}\mathbf{x}.$$

concluding the proof of Proposition 3.7. □

Remark 3.4. *In the discrete counterpart* (40) *of* (9)*, the equality is replaced by an inequality. But as well as in the continuous setting, the function*

$$t \mapsto \int_\Omega \Gamma(p_{\mathcal{M},\Delta t})(\mathbf{x}, t)\mathrm{d}\mathbf{x}$$

is nonincreasing, justifying the denomination Entropy diminishing scheme*.*

Corollary 3.8. *There exists* C *depending only on* Ω *and* β *such that*

$$\iint_{Q_{t_\mathrm{f}}} \Lambda\nabla\xi_{\mathcal{T},\Delta t}\cdot\nabla\xi_{\mathcal{T},\Delta t}\mathrm{d}\mathbf{x}\mathrm{d}t \leq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \leq C. \tag{41}$$

Proof. Summing (40) over $n \in \{0, \dots, N\}$ provides

$$\int_\Omega \Gamma(p_{\mathcal{M}}^{N+1})\mathrm{d}\mathbf{x} + \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \le \int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M}}^0)\mathrm{d}\mathbf{x}.$$

Since $\Gamma(p) \ge 0$ for all p, we get that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \le \int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M}}^0)\mathrm{d}\mathbf{x}.$$

It remains to check that for all $b \in (\beta(p_\star), \beta(p^\star))$,

$$\Gamma\circ\beta^{-1}(b) = \int_0^{\beta^{-1}(b)} a\beta'(a)\mathrm{d}a = \int_{\beta(0)}^{b} \beta^{-1}(a)\mathrm{d}a,$$

ensuring that

$$\int_\Omega \Gamma\circ\beta^{-1}(\beta_{\mathcal{M}}^0)\mathrm{d}\mathbf{x} \le |\Omega|\|\beta^{-1}\|_{L^1(\beta(p_\star),\beta(p^\star))}.$$

The first inequality in (41) follows from Lemma 3.1. □

The following statement is a straightforward consequence of Corollary 3.8 and of Lemma 3.3.

Corollary 3.9. *There exists C depending only on Ω, β, Λ and $\theta_{\mathcal{T}}$ such that*

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}\left(p_K^{n+1} - p_L^{n+1}\right)^2 \le C.$$

3.4. **Existence of a discrete solution.** In order to prove the existence of a solution $\left(p_K^{n+1}\right)_K$ to the scheme (25), we need an additional mesh-depending estimate on the solution. In what follows, m_Ω denotes the Lebesgue measure of Ω, and $\overline{\beta}_0 = \frac{1}{m_\Omega}\int \beta_0 \mathrm{d}\mathbf{x}$.

Lemma 3.10. *Assume that $\overline{\beta}_0 < \beta(p^\star)$ (resp. $\overline{\beta}_0 > \beta(p_\star)$), then there exists $C^\star < \infty$ (resp. $C_\star > -\infty$) depending only on $\Delta t, \mathcal{T}, \Lambda, \Omega, \beta, \overline{\beta}_0$ such that*

$$p_K^{n+1} \le C^\star \quad (\text{resp. } p_K^{n+1} \ge C_\star), \qquad \forall K \in \mathcal{V}, \quad \forall n \in \{0, \dots, N\}.$$

Proof. Let us prove that $p_K^{n+1} \le C^\star$, the proof of $p_K^{n+1} \ge C_\star$ being similar. Assume first that $p^\star < \infty$, then we can choose $C^\star = p^\star$ thanks to (39), so that we can focus on the case $p^\star = +\infty$.

Since, in view of the definition (23) of β_K^0 and of the global conservation property (28), one has that

$$\sum_{K\in\mathcal{V}} (\beta(p_K^{n+1}) - \overline{\beta}_0)m_K = 0,$$

ensuring the existence of at least one vertex K_{i} such that $\beta(p_{K_{\mathrm{i}}}^{n+1}) \le \overline{\beta}_0$. In particular,

$$p_{K_{\mathrm{i}}}^{n+1} \le \beta^{-1}(\overline{\beta}_0) < \infty. \tag{42}$$

Let $K_{\mathrm{f}} \in \mathcal{V} \setminus \{K_{\mathrm{i}}\}$, then thanks to Lemma 3.5, there exists a transmissive path $\mathrm{w} \in \mathcal{W}(K_{\mathrm{i}}, K_{\mathrm{f}}) = (K_q)_{0\le q\le M}$ of finite length in the sense of Definition 3.4. Let us show that for all $p_{K_q}^{n+1} < \infty$ for all $q \in \{0, \dots, M\}$.

First, we have checked in (42) that $p_{K_0}^{n+1} < \infty$. Assume now that $p_{K_q}^{n+1} < \infty$, then it follows from Corollary 3.9 that

$$\Delta t a_{K_qK_{q+1}}\eta_{K_qK_{q+1}}^{n+1}\left(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}\right)^2 \le C.$$

It follows from the definition (26) of $\eta^{n+1}_{K_qK_{q+1}}$ that

$$\eta^{n+1}_{K_qK_{q+1}} \geq \eta(p^{n+1}_{K_q}) > 0.$$

Since $a_{K_qK_{q+1}} > 0$, we obtain that

$$\left(p^{n+1}_{K_q} - p^{n+1}_{K_{q+1}}\right)^2 \leq \frac{C}{\Delta t a_{K_qK_{q+1}} \eta(p^{n+1}_{K_q})} < \infty,$$

ensuring that $p^{n+1}_{K_{q+1}} < \infty$.

We have proved the existence of a finite quantity $(C_{K_\mathrm{i},K_\mathrm{f},\mathrm{w}})_{K_\mathrm{f}\in\mathcal{V}}$ (depending on $\Delta t, \mathcal{T}, \Lambda, \Omega, \beta, \overline{\beta}_0$) such that

$$\beta\left(p^{n+1}_{K_\mathrm{i}}\right) \leq \overline{\beta}_0 \quad \Longrightarrow \quad p^{n+1}_{K_\mathrm{f}} \leq C_{K_\mathrm{i},K_\mathrm{f},\mathrm{w}}.$$

As a consequence,

$$p^{n+1}_K \leq \max_{K_\mathrm{i}\in\mathcal{V}} \max_{K_\mathrm{f}\in\mathcal{V}} \min_{\mathrm{w}\in\mathcal{W}(K_\mathrm{i},K_\mathrm{f})} C_{K_\mathrm{i},K_\mathrm{f},\mathrm{w}} < \infty, \qquad \forall K \in \mathcal{V}, \forall n \in \{0,\dots,N\},$$

concluding the proof of Lemma 3.10. □

Proposition 3.11. *Given $(\beta^n_K)_{K\in\mathcal{V}} \in [\beta(p_\star), \beta(p^\star)]^{\#\mathcal{V}}$ such that $\sum_{K\in\mathcal{V}} \beta^n_K m_K = \overline{\beta}_0 m_\Omega$, there exists (at least) one solution $\left(p^{n+1}_K\right)_{K\in\mathcal{V}} \in [p_\star, p^\star]^{\#\mathcal{V}}$ of the scheme (25), that moreover satisfies $\sum_{K\in\mathcal{V}} \beta^{n+1}_K m_K = \overline{\beta}_0 m_\Omega$.*

Proof. Assume first that $\overline{\beta}_0 = \beta(p_\star)$ (resp. $\overline{\beta}_0 = \beta(p^\star)$), then the constant function $p^{n+1}_K = p_\star$ (resp. $p^{n+1}_K = p^\star$) for all $K \in \mathcal{V}$ is a solution.

We assume now that $\overline{\beta}_0 \in (\beta(p_\star), \beta(p^\star))$. In order to prove the existence of a weak solution, we follow the methodology proposed in [25], using a topological degree argument [35, 19]. We detail the proof, since some unusual difficulties occur because of the discretization of the degenerate problem.

Let $\mu \in [0,1]$, and define $\left(p^{n+1}_{K,\mu}\right)_K$ as the solution of the scheme: $\forall K \in \mathcal{V}$,

$$\frac{\beta(p^{n+1}_{K,\mu}) - \beta^n_K}{\Delta t} m_K + \mu \sum_{L\in\mathcal{V}_K} a_{KL} \eta^{n+1}_{KL,\mu} \left(p^{n+1}_{K,\mu} - p^{n+1}_{L,\mu}\right) + (1-\mu)\|\eta\|_\infty \sum_{L\in\mathcal{V}_K} |a_{KL}| \left(p^{n+1}_{K,\mu} - p^{n+1}_{L,\mu}\right) = 0,$$

where $\eta^{n+1}_{KL,\mu}$ is defined by (26) with $p^{n+1}_{K,\mu}$ and $p^{n+1}_{L,\mu}$ instead of p^{n+1}_K and p^{n+1}_L.

Reproducing carefully the analysis carried out in §3.3, one can show that there exists C depending only on Ω, β, Λ and $\theta_\mathcal{T}$ such that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}| \eta^{n+1}_{KL,\mu} \left(p^{n+1}_{K,\mu} - p^{n+1}_{L,\mu}\right)^2 \leq C, \qquad \forall \mu \in [0,1].$$

In view of Lemma 3.10, we can claim that any solution $\left(p^{n+1}_{K,\mu}\right)_K$ satisfies

$$p_\star \leq \underline{p} \leq p^{n+1}_{K,\mu} \leq \overline{p} \leq p^\star, \qquad \forall K \in \mathcal{V}. \tag{43}$$

for some finite $\underline{p}, \overline{p}$ depending on the discretization and on the data.

Define the compact subset $\mathcal{K} = [\underline{p} - 1, \overline{p} + 1]^{\#\mathcal{V}}$ of $\mathbb{R}^{\#\mathcal{V}}$, and define the function $\mathcal{H}\left((p_K)_K, \mu\right) : \mathbb{R}^{\#\mathcal{V}} \times [0,1] \to \mathbb{R}^{\#\mathcal{V}}$ by: $\forall K \in \mathcal{V}$,

$$\mathcal{H}_K\left((p_K)_K, \mu\right) = \frac{\beta(p_{K,\mu}^{n+1}) - \beta_K^n}{\Delta t} m_K + \mu \sum_{L \in \mathcal{V}_K} a_{KL} \eta_{KL,\mu}^{n+1} \left(p_{K,\mu}^{n+1} - p_{L,\mu}^{n+1}\right) + (1-\mu)\|\eta\|_\infty \sum_{L \in \mathcal{V}_K} |a_{KL}| \left(p_{K,\mu}^{n+1} - p_{L,\mu}^{n+1}\right).$$

The function $\mathcal{H}$ is uniformly continuous on $\mathcal{K} \times [0,1]$ (recall that β is artificially extended by (38)), and it follows from (43) that for all $\mu \in [0,1]$, the nonlinear system

$$\mathcal{H}\left((p_K)_K, \mu\right) = 0 \tag{44}$$

cannot admit any solution on $\partial\mathcal{K}$. Therefore, the corresponding topological degree $\delta(\mathcal{H}, \mathcal{K})(\mu)$ is constant w.r.t. μ. For $\mu = 0$, the topological degree is equal to 1 thanks to the monotonicity of the system (this can be proved for example by adapting the existence proof of a discrete solution to the monotone implicit scheme for an hyperbolic scalar conservation law studied in [23]). Hence, the system (44) admits at least one solution for $\mu = 1$, ensuring the existence of a solution to the scheme (25). □

4. Convergence towards a weak solution

4.1. **Compactness result for the family of the discrete solutions.** We define the (strictly) increasing function $\Upsilon : [p_\star, p^\star] \to \mathbb{R}$ by

$$\Upsilon(p) = \int_0^p \min\left(\beta'(a), \sqrt{\eta(a)}\right) \mathrm{d}a. \tag{45}$$

For all $K \in \mathcal{V}_m$ and for all $n \geq 1$, we denote by $\Upsilon_K^n = \Upsilon(p_K^n)$, and by $\Upsilon_{\mathcal{T}_m, \Delta t_m}$ the corresponding piecewise affine in space and constant in time reconstruction in $V_{\mathcal{T}_m, \Delta t_m}$, and by $\Upsilon_{\mathcal{M}_m, \Delta t_m}$ the piecewise constant reconstruction in $X_{\mathcal{M}, \Delta t}$.

4.1.1. *Space translates estimates.* For all $m \geq 0$, the function $\Upsilon_{\mathcal{T}_m, \Delta t_m}$ (resp. $\Upsilon_{\mathcal{M}_m, \Delta t_m}$) belongs to $V_{\mathcal{T}, \Delta t}$ (resp. $X_{\mathcal{M}, \Delta t}$), and is extended by 0 outside of Q_{t_f}.

Lemma 4.1. *There exists C depending only on $T, \Omega, \eta, \beta, u_0, \Lambda$ and θ (but not on m) such that*

$$\int_0^{t_\mathrm{f}} \int_{\mathbb{R}^2} |\Upsilon_{\mathcal{T}_m, \Delta t_m}(\mathbf{x} + \mathbf{w}, t) - \Upsilon_{\mathcal{T}_m, \Delta t_m}(\mathbf{x}, t)| \, \mathrm{d}\mathbf{x}\mathrm{d}t \leq C|\mathbf{w}|. \tag{46}$$

Proof. The function $\Upsilon \circ \xi^{-1}$ is 1-Lipschitz continuous, so that, thanks to Corollary 3.8, one has

$$\|\nabla \Upsilon_{\mathcal{T}_m, \Delta t_m}\|_{(L^2(Q_{t_\mathrm{f}}))^2} \leq C.$$

The set Q_{t_f} being of finite measure, Hölder inequality yields

$$\|\nabla \Upsilon_{\mathcal{T}_m, \Delta t_m}\|_{(L^1(Q_{t_\mathrm{f}}))^2} \leq C. \tag{47}$$

Since Υ is uniformly bounded (because ξ is), its extension to the whole $\mathbb{R}^3$, still denoted by $\Upsilon_{\mathcal{T}_m, \Delta t_m}$, belongs to $L^\infty \cap \mathrm{BV}(\mathbb{R}^3)$ and satisfies

$$\mathrm{TV}\left(\Upsilon_{\mathcal{T}_m, \Delta t_m}\right) \leq \|\nabla \Upsilon_{\mathcal{T}_m, \Delta t_m}\|_{(L^1(Q_{t_\mathrm{f}}))^2} + \|\Upsilon\|_\infty \left(T m_{\partial\Omega} + 2 m_\Omega\right) < \infty, \tag{48}$$

where $m_{\partial\Omega}$ denotes the length of $\partial\Omega$. The inequality (46) is a classical consequence of (48) (see e.g. [11, 4]). □

Lemma 4.2. *There exists $C_{\Upsilon,\mathrm{s}}$ depending only on $T, \Omega, \eta, \beta, u_0, \Lambda$ and θ (but neither on m nor on $\mathbf{w}$) such that*

$$\int_0^{t_\mathrm{f}} \int_{\mathbb{R}^2} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x}+\mathbf{w},t) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x}\mathrm{d}t \leq C_{\Upsilon,\mathrm{s}} \left(|\mathbf{w}| + h\right). \tag{49}$$

Proof. The triangle inequality implies that

$$\int_0^{t_\mathrm{f}} \int_{\mathbb{R}^2} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x}+\mathbf{w},t) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x} \leq A + B,$$

where

$$\begin{aligned} A =& \int_0^{t_\mathrm{f}} \int_{\mathbb{R}^2} |\Upsilon_{\mathcal{T}_m,\Delta t_m}(\mathbf{x}+\mathbf{w},t) - \Upsilon_{\mathcal{T}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x}\mathrm{d}t, \\ B =& 2\int_0^{t_\mathrm{f}} \int_{Q_{t_\mathrm{f}}} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t) - \Upsilon_{\mathcal{T}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x}\mathrm{d}t. \end{aligned}$$

The term A is overestimated thanks to Lemma 4.1, while $B \leq Ch$ thanks to (47) and to a straightforward generalization of Lemma A.2 given in appendix. □

4.1.2. *Time translates estimate.* Because of the possible degeneracy of the problem (1), the discrete version [32] of the Aubin-Simon Lemma [41] is of no use in our case. Therefore, following Eymard *et al.* [25, 29], we adapt the path proposed by Alt & Luckhaus [2] to obtain some estimates on the times translates for the family $(\Upsilon(p)_{\mathcal{M}_m,\Delta t_m})_m$. For all $\tau \in (0, t_\mathrm{f})$, we denote by $Q_{t_\mathrm{f}-\tau} = \Omega \times (0, t_\mathrm{f} - \tau)$.

Lemma 4.3. *There exists $C_{\Upsilon,\mathrm{t}}$ depending only on $t_\mathrm{f}, \Omega, \beta, \beta_0, \eta, \Lambda$ and θ (but neither on m nor on τ) such that, for all $\tau \in (0, t_\mathrm{f})$, one has*

$$\iint_{Q_{t_\mathrm{f}-\tau}} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x}\mathrm{d}t \leq C_{\Upsilon,\mathrm{t}} \left(\sqrt{\tau} + \sqrt{\Delta t_m}\right). \tag{50}$$

Proof. For $t \in (0, t_\mathrm{f} - \tau)$, we denote by $A_m(t)$ the quantity defined by

$$A_m(t) = \int_\Omega |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|^2 \,\mathrm{d}\mathbf{x}.$$

Cauchy-Schwarz inequality ensures that

$$\iint_{Q_{t_\mathrm{f}-\tau}} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)| \,\mathrm{d}\mathbf{x}\mathrm{d}t \leq C\sqrt{\int_0^{t_\mathrm{f}-\tau} A_m(t)\mathrm{d}t}. \tag{51}$$

For $t \in (0, T]$, we denote by $\nu(t)$ the unique positive integer such that $t_{\nu(t)} < t \leq t_{\nu(t)+1}$, so that, for almost all $t \in (0, t_\mathrm{f} - \tau)$,

$$A_m(t) = \sum_{K \in \mathcal{V}_{m,\mathrm{int}}} \left(\Upsilon_K^{\nu(t+\tau)+1} - \Upsilon_K^{\nu(t)+1}\right)^2 m_K.$$

The definition (45) of Υ implies that

$$\begin{aligned} \left(\Upsilon_K^{\nu(t+\tau)+1} - \Upsilon_K^{\nu(t)+1}\right)^2 \leq& \left(\beta(p_K^{\nu(t+\tau)+1}) - \beta(p_K^{\nu(t)+1})\right) \\ &\times \left(\xi(p_K^{\nu(t+\tau)+1}) - \xi(p_K^{\nu(t)+1})\right) \\ =& \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \left(\beta(p_K^{n+1}) - \beta(p_K^n)\right) \\ &\times \left(\xi(p_K^{\nu(t+\tau)+1}) - \xi(p_K^{\nu(t)+1})\right). \end{aligned}$$

Therefore, using the scheme (25) and a classical reorganization of the sum, provides that

$$A_m(t) \le \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} \Big[a_{KL}\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1}) \\ \times \Big(\big(\xi(p_K^{\nu(t)+1}) - \xi(p_L^{\nu(t)+1})\big) - \big(\xi(p_K^{\nu(t+\tau)+1}) - \xi(p_L^{\nu(t+\tau)+1})\big)\Big)\Big].$$

Then, using Young's inequality provides that

$$A_m(t) \le A_{1,m}(t) + A_{2,m}(t) + A_{3,m}(t), \tag{52}$$

where we have set

$$A_{1,m}(t) = \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} |a_{KL}|\,\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2,$$

$$A_{2,m}(t) = \frac{\|\eta\|_\infty}{2} \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} |a_{KL}| \left(\xi(p_K^{\nu(t+\tau)+1}) - \xi(p_L^{\nu(t+\tau)+1})\right)^2$$

$$A_{3,m}(t) = \frac{\|\eta\|_\infty}{2} \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} |a_{KL}| \left(\xi(p_K^{\nu(t)+1}) - \xi(p_L^{\nu(t)+1})\right)^2$$

Using the technical result proposed in [29, Proposition 9.3], the Lemma 3.3, and the Proposition 3.7, we can claim that there exists C such that

$$\int_0^{t_{\rm f}-\tau} A_{\kappa,m}(t)\mathrm{d}t \le C(\tau+\Delta t_m), \qquad \forall \kappa\in\{1,2,3\}. \tag{53}$$

Hence, the function $s\mapsto\sqrt{s}$ being concave, estimates (51), (52), and (53) allow us to conclude that (50) holds. □

Lemma 4.4. *There exists C depending only on $t_{\rm f}, \Omega, \beta, \beta_0, \eta, \Lambda$ and θ (but not on m) such that, for all $\tau\in(0,t_{\rm f})$, one has*

$$\int_{\mathbb{R}}\int_{\mathbb{R}^2} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t \le C\left(\sqrt{\tau}+\sqrt{\Delta t_m}+\tau\right).$$

Proof. In view of the null extension of $\Upsilon_{\mathcal{M}_m,\Delta t_m}$ outside $Q_{t_{\rm f}}$, one has

$$\int_{\mathbb{R}}\int_{\mathbb{R}^2} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t \\ = \iint_{Q_{t_{\rm f}-\tau}} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t \\ + \int_{t_{\rm f}-\tau}^{t_{\rm f}}\int_\Omega |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t.$$

Thanks to Lemma 4.3, one has

$$\iint_{Q_{t_{\rm f}}} |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t+\tau) - \Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t \le C(\sqrt{\tau}+\sqrt{\Delta t_m}),$$

while, since Υ is bounded, one has

$$\int_{t_{\rm f}-\tau}^{t_{\rm f}}\int_\Omega |\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t \le \tau m_\Omega \|\Upsilon\|_\infty.$$

□

4.1.3. *Compactness for the approximate solutions.*

Lemma 4.5. *There exists a measurable function $p : Q_{t_\mathrm{f}} \to [p_\star, p^\star]$ such that*

$$p_{\mathcal{M}_m,\Delta t_m} \to p \quad \text{a.e. in } Q_{t_\mathrm{f}} \text{ as } m \to \infty. \tag{54}$$

Proof. As a consequence of Lemma 4.2 and Lemma 4.4, one can apply the Kolmogorov compactness criterion (see e.g. [25] or [29, Corollary 9.1]) to the sequence $(\Upsilon_{\mathcal{M}_m,\Delta t_m})_{m\geq 1}$, ensuring that, up to an unlabeled subsequence,

$$\Upsilon_{\mathcal{M}_m,\Delta t_m} \to \Upsilon^\star \quad \text{strongly in } L^1(Q_{t_\mathrm{f}}).$$

Since the function Υ is continuous and increasing on $[p_\star, p^\star]$ for all $\mathbf{x} \in \Omega$, this allows to define

$$p(\mathbf{x},t) = \Upsilon^{-1}\left(\Upsilon^\star(\mathbf{x},t)\right), \qquad \text{for a.e. } (\mathbf{x},t) \in Q_{t_\mathrm{f}}, \tag{55}$$

and since

$$p_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t) = \Upsilon^{-1}\left(\Upsilon_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t)\right), \quad \forall (\mathbf{x},t) \in \omega_K \times (0,t_\mathrm{f}), \forall K \in \mathcal{V},$$

the relation (54) holds. □

Lemma 4.6. *Up to an unlabeled subsequence, the sequence $(\xi(p)_{\mathcal{T}_m,\Delta t_m})_{m\geq 1}$ converges weakly in $L^2((0,t_\mathrm{f}); H^1(\Omega))$ towards $\xi(p)$, where p is defined by (55).*

Proof. It follows from Lemma 3.1 and Proposition 3.7 that $\nabla\xi(p)_{\mathcal{T}_m,\Delta t_m}$ is uniformly bounded in $\left(L^2(Q_{t_\mathrm{f}})\right)^2$. Moreover, the boundedness of ξ (due to (3)) ensures that $\xi(p)_{\mathcal{T}_m,\Delta t_m}$ is uniformly bounded in $L^2(Q_{t_\mathrm{f}})$, providing the boundedness of $(\xi(p)_{\mathcal{T}_m,\Delta t_m})_m$ in $L^2((0,t_\mathrm{f}); H^1(\Omega))$. The existence of $\xi^\star$ such that $\xi(p)_{\mathcal{T}_m,\Delta t_m} \to \xi^\star$ follows.

Let us identify $\xi^\star$ as $\xi(p)$. Due to Lemma A.2, we know that the sequences $(\xi(p)_{\mathcal{T}_m,\Delta t_m})_{m\geq 1}$ and $(\xi(p)_{\mathcal{M}_m,\Delta t_m})_{m\geq 1}$ have the same limit. Since $\xi(p)_{\mathcal{M}_m,\Delta t_m} = \xi\left(p_{\mathcal{M}_m,\Delta t_m}\right)$, then (54) ensures that $\lim_{m\to\infty} \xi(p)_{\mathcal{M}_m,\Delta t_m} = \xi(p)$. □

4.2. **Identification of the limit as the weak solution.**

Proposition 4.7. *Let p be a measurable function such that $p_{\mathcal{M}_m,\Delta t_m} \to p$ a.e. in Q_{t_f}, and such that $\xi(p)_{\mathcal{T}_m,\Delta t_m} \to \xi(p)$ weakly in $L^1((0,t_\mathrm{f}); H^1(\Omega))$. Then p is the unique weak solution to the problem (1) in the sense of Definition 1.1.*

Proof. Let $\psi \in C^\infty(\overline{\Omega} \times [0,t_\mathrm{f}))$, then, for all $K \in \mathcal{V}_m$ and all $n \in \{0,\dots,N_m\}$, we denote by $\psi_K^n = \psi(\mathbf{x}_K, t^n)$. Multiplying the scheme (25) by $\Delta t_m \psi_K^n$ and summing over $n \in \{0,\dots,N_m\}$ and $K \in \mathcal{V}_\mathrm{int}$ provides, after a classical reorganization of the sums (see e.g. [26]),

$$\mathcal{A}_m + \mathcal{B}_m + \mathcal{C}_m + \mathcal{D}_m = 0, \tag{56}$$

where

$$\begin{aligned}
\mathcal{A}_m &= \sum_{n=0}^{N_m} \sum_{K\in\mathcal{V}_m} \beta(p_K^{n+1})(\psi_K^n - \psi_K^{n+1}) m_K, \\
\mathcal{B}_m &= -\sum_{K\in\mathcal{V}_m} \beta_K^0 \psi_K^0 m_K, \\
\mathcal{C}_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} a_{KL} \begin{bmatrix} \eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1}) \\ -\sqrt{\eta_{KL}^{n+1}}(\xi(p_K^{n+1}) - \xi(p_L^{n+1})) \end{bmatrix} (\psi_K^n - \psi_L^n), \\
\mathcal{D}_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} a_{KL} \sqrt{\eta_{KL}^{n+1}}(\xi(p_K^{n+1}) - \xi(p_L^{n+1}))(\psi_K^n - \psi_L^n).
\end{aligned}$$

Define the function $\delta\psi_{\mathcal{M}_m,\Delta t_n}$ of $X_{\mathcal{M}_m,\Delta t_m}$ by

$$\delta\psi_{\mathcal{M}_m,\Delta t_m}(\mathbf{x},t) = \frac{\psi_K^{n+1}-\psi_K^n}{\Delta t_m} \quad \text{if } (\mathbf{x},t)\in\omega_K\times(t_n,t_{n+1}),$$

then, the term $\mathcal{A}_m$ can be rewritten

$$\mathcal{A}_m = -\iint_{Q_{t_\mathrm{f}}} \beta(p_{\mathcal{M}_m,\Delta t_m})\delta\psi_{\mathcal{M}_m,\Delta t_n}\mathrm{d}\mathbf{x}\mathrm{d}t.$$

Thanks to the regularity of ψ, the function $\delta\psi_{\mathcal{M}_m,\Delta t_n}$ converges uniformly towards $\partial_t\psi$ on Q_{t_f}, while, since β is bounded and since $p_{\mathcal{M}_m,\Delta t_m}$ converges almost everywhere in Q_{t_f} towards p,

$$\beta(p_{\mathcal{M}_m,\Delta t_m})\to\beta(p) \quad \text{in } L^q(Q_{t_\mathrm{f}}) \text{ as } m\to\infty$$

for all $q\in[1,\infty)$. Therefore,

$$\mathcal{A}_m \to -\iint_{Q_{t_\mathrm{f}}} \beta(p)\partial_t\psi\mathrm{d}\mathbf{x}\mathrm{d}t \quad \text{as } m\to\infty. \tag{57}$$

Moreover, it follows from classical arguments (see e.g. [25]) that

$$\mathcal{B}_m \to -\int_\Omega \beta_0\psi(\cdot,0)\mathrm{d}\mathbf{x} \quad \text{as } m\to\infty. \tag{58}$$

For all $\sigma_{KL}\in\mathcal{E}_m$ and all $n\geq 0$, we define

$$\overline{\eta}_{KL}^{n+1} = \begin{cases} \left(\frac{\xi(p_K^{n+1})-\xi(p_L^{n+1})}{p_K^{n+1}-p_L^{n+1}}\right)^2 & \text{if } p_K^{n+1}\neq p_L^{n+1},\\ \eta(p_K^{n+1}) & \text{if } p_K^{n+1}=p_L^{n+1}.\end{cases}$$

The term $\mathcal{C}_m$ then rewrites

$$\mathcal{C}_m = \sum_{n=0}^{N_m}\Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} a_{KL}\sqrt{\eta_{KL}^{n+1}}\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\overline{\eta}_{KL}^{n+1}}\right)(p_K^{n+1}-p_L^{n+1})(\psi_K^n-\psi_L^n).$$

Using Cauchy-Schwarz inequality provides that

$$|\mathcal{C}_m| \leq \left(\sum_{n=0}^{N_m}\Delta t_m\sum_{\sigma_{KL}\in\mathcal{E}_m}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2\right)^{1/2} \times \left(\underbrace{\sum_{n=0}^{N_m}\Delta t_m\sum_{\sigma_{KL}\in\mathcal{E}_m}|a_{KL}|\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\overline{\eta}_{KL}^{n+1}}\right)^2(\psi_K^n-\psi_L^n)^2}_{:=\mathcal{R}_m}\right)^{1/2}. \tag{59}$$

The first term in the right-hand side of (59) is bounded thanks to Lemma 3.3 and Proposition 3.7. Therefore, in order to prove that $\mathcal{C}_m$ tends to 0 as m tends to ∞, it remains to check that $\mathcal{R}_m$ does.

For $T\in\mathcal{T}_m$, we denote by

$$\overline{\xi}_T^{n+1} = \max_{\mathbf{x}\in T}\left(\xi(p)_{\mathcal{T}_m,\Delta t_n}(\mathbf{x},t_{n+1})\right), \qquad \underline{\xi}_T^{n+1} = \min_{\mathbf{x}\in T}\left(\xi(p)_{\mathcal{T}_m,\Delta t_n}(\mathbf{x},t_{n+1})\right),$$

and by

$$\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\mathbf{x},t) = \overline{\xi}_T^{n+1} \quad \text{if } (\mathbf{x},t)\in T\times(t_n,t_{n+1}),$$
$$\underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\mathbf{x},t) = \underline{\xi}_T^{n+1} \quad \text{if } (\mathbf{x},t)\in T\times(t_n,t_{n+1}).$$

For all $\sigma_{KL} \in \mathcal{E}_T$, one has

$$\left|\sqrt{\eta_{KL}^{n+1}} - \sqrt{\overline{\eta}_{KL}^{n+1}}\right| \leq \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right), \tag{60}$$

where μ is the continuity modulus of $\sqrt{\eta \circ \xi^{-1}}$ (which exists and is bounded since $\sqrt{\eta \circ \xi^{-1}}$ is continuous on the bounded interval $[\xi(p_\star), \xi(p^\star)]$). Using (60) in the definition (59) of $\mathcal{R}_m$ provides that

$$0 \leq \mathcal{R}_m \leq \mathcal{Q}_m, \tag{61}$$

where we have set

$$\mathcal{Q}_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T \in \mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL} \in \mathcal{E}_T} |\alpha_{KL}^T| (\psi_K^n - \psi_L^n)^2. \tag{62}$$

Following the proof of Lemma 3.2, there exists C depending only on Λ, θ and on ψ such that

$$\sum_{\sigma_{KL} \in \mathcal{E}_T} |\alpha_{KL}^T| (\psi_K^n - \psi_L^n)^2 \leq C m_T.$$

Therefore, (61) implies that

$$0 \leq \mathcal{R}_m \leq C \iint_{Q_{t_\mathrm{f}}} \mu\left(\overline{\xi}_{\mathcal{T}_m, \Delta t_m}(\mathbf{x}, t) - \underline{\xi}_{\mathcal{T}_m, \Delta t_m}(\mathbf{x}, t)\right) \mathrm{d}\mathbf{x}\mathrm{d}t, \tag{63}$$

where C depends on Λ, θ and ψ. Since μ is bounded, continuous, with $\mu(0) = 0$, it suffices to show that $\overline{\xi}_{\mathcal{T}_m, \Delta t_m} - \underline{\xi}_{\mathcal{T}_m, \Delta t_m}$ tends to 0 almost everywhere in Q_{t_f} as $m \to \infty$ (up to an unlabeled subsequence). It follows from a straightforward generalization of Lemma A.1 that

$$\begin{aligned} \iint_{Q_{t_\mathrm{f}}} \left|\overline{\xi}_{\mathcal{T}_m, \Delta t_m} - \underline{\xi}_{\mathcal{T}_m, \Delta t_m}\right| \mathrm{d}\mathbf{x}\mathrm{d}t \leq & C h_m \iint_{Q_{t_\mathrm{f}}} |\nabla \xi(p)_{\mathcal{T}_m, \Delta t_m}| \, \mathrm{d}\mathbf{x}\mathrm{d}t \\ \leq & C h_m \left(\iint_{Q_{t_\mathrm{f}}} |\nabla \xi(p)_{\mathcal{T}_m, \Delta t_m}|^2 \, \mathrm{d}\mathbf{x}\mathrm{d}t\right)^{1/2} \leq C h_m \end{aligned} \tag{64}$$

thanks to Lemma 3.1 and to Proposition 3.7. As a consequence of (59), (63) and (64), up to the extraction of an unlabeled subsequence, one has

$$\lim_{m \to \infty} \mathcal{C}_m = \lim_{m \to \infty} \mathcal{R}_m = \lim_{m \to \infty} \mathcal{Q}_m = 0. \tag{65}$$

Let us now focus on the last term $\mathcal{D}_m$. For all $T \in \mathcal{T}_m$, denoting by $\mathbf{x}_T$ its center of mass, we define the piecewise constant functions $\Xi_{\mathcal{T}_m, \Delta t_m}$ and $H_{\mathcal{T}_m, \Delta t_m}$ by

$$\Xi_{\mathcal{T}_m, \Delta t_m}(\mathbf{x}, t) = \xi(p)_{\mathcal{T}_m, \Delta t_m}(\mathbf{x}_T, t), \qquad \forall \mathbf{x} \in T, \forall t \in (t_n, t_{n+1}],$$

and $H_{\mathcal{T}_m, \Delta t_m} = \sqrt{\eta \circ \xi^{-1}(\Xi_{\mathcal{T}_m, \Delta t_m})}$. It is easy to check (one can for example use a slight adaptation of Lemma A.2 given in appendix) that

$$\Xi_{\mathcal{T}_m, \Delta t_m} \to \xi(p) \quad \text{ in } L^2(Q_{t_\mathrm{f}}) \text{ as } m \to \infty.$$

The function $\sqrt{\eta \circ \xi^{-1}}$ being continuous and bounded, one obtains that

$$H_{\mathcal{T}_m, \Delta t_m} \to \sqrt{\eta(p)} \quad \text{ in } L^2(Q_{t_\mathrm{f}}) \text{ as } m \to \infty. \tag{66}$$

We introduce the term

$$\mathcal{D}'_m = \iint_{Q_{t_\mathrm{f}}} H_{\mathcal{T}_m, \Delta t_m} \Lambda \nabla \xi(p)_{\mathcal{T}_m, \Delta t_m} \cdot \nabla \psi_{\mathcal{T}_m, \Delta t_m}(\cdot, t - \Delta t_m) \mathrm{d}\mathbf{x}\mathrm{d}t,$$

then since $\nabla\xi(p)_{\mathcal{T}_m,\Delta t_m}$ converges weakly in $L^2(Q_{t_f})$ towards $\nabla\xi(p)$, since $\nabla\psi_{\mathcal{T}_m,\Delta t_m}$ converges uniformly towards $\nabla\psi$, and thanks to (66), we obtain that

$$\lim_{m\to\infty}\mathcal{D}'_m=\iint_{Q_{t_f}}\sqrt{\eta(p)}\Lambda\nabla\xi(p)\cdot\nabla\psi\mathrm{d}\mathbf{x}\mathrm{d}t=\iint_{Q_{t_f}}\Lambda\nabla\varphi(p)\cdot\nabla\psi\mathrm{d}\mathbf{x}\mathrm{d}t. \tag{67}$$

Therefore, it only remains to verify that

$$|\mathcal{D}_m-\mathcal{D}'_m|\to 0\quad\text{ as } m\to\infty. \tag{68}$$

We denote by

$$\eta_T^{n+1}=\left(H_{\mathcal{T}_m,\Delta t_m}(\mathbf{x}_T,t_{n+1})\right)^2,\qquad\forall T\in\mathcal{T}_m,\forall n\in\{0,\dots,N_m\},$$

so that

$$\mathcal{D}_m-\mathcal{D}'_m=\sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\sum_{\sigma_{KL}\in\mathcal{E}_T}\left[\begin{array}{l}\alpha_{KL}^T\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\eta_T^{n+1}}\right)\\ \times\left(\xi(p_K^{n+1})-\xi(p_L^{n+1})\right)\left(\psi_K^n-\psi_L^n\right)\end{array}\right].$$

Since for all $n\in\{0,\dots,N_m\}$, for all $T\in\mathcal{T}_m$, and for all $\sigma_{KL}\in\mathcal{E}_T$, one has

$$\left|\sqrt{\eta_{KL}^{n+1}}-\sqrt{\eta_T^{n+1}}\right|\le\mu\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right),$$

where μ is the continuity modulus of $\sqrt{\eta\circ\xi^{-1}}$, one obtains that

$$\begin{aligned}|\mathcal{D}_m-\mathcal{D}'_m|\le\sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\Bigg[&\mu\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right)\\&\times\sum_{\sigma_{KL}\in\mathcal{E}_T}\alpha_{KL}^T\left(\xi(p_K^{n+1})-\xi(p_L^{n+1})\right)\left(\psi_K^{n+1}-\psi_L^{n+1}\right)\Bigg].\end{aligned}$$

We deduce from Cauchy-Schwarz inequality that

$$|\mathcal{D}_m-\mathcal{D}'_m|^2\le\mathcal{Q}_m\times\sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\sum_{\sigma_{KL}\in\mathcal{E}_T}\left|\alpha_{KL}^T\right|\left(\xi(p_K^{n+1})-\xi(p_L^{n+1})\right)^2, \tag{69}$$

where the term $\mathcal{Q}_m$ has been introduced in (62). Using Lemma 3.1, Lemma 3.2 and Proposition 3.7 provides that

$$\sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\sum_{\sigma_{KL}\in\mathcal{E}_T}\left|\alpha_{KL}^T\right|\left(\xi(p_K^{n+1})-\xi(p_L^{n+1})\right)^2\le C, \tag{70}$$

where C does not depend on m. We then deduce from (65), (69) and (70) that the relation (69) holds. Therefore, in view of (67), we obtain that

$$\lim_{m\to\infty}\mathcal{D}_m=\iint_{Q_{t_f}}\Lambda\nabla\varphi(p)\cdot\nabla\psi\mathrm{d}\mathbf{x}\mathrm{d}t. \tag{71}$$

Putting (57), (58), (65) and (71) together in (56) provides that p satisfies the weak formulation (15), then it is a weak solution. □

5. Numerical results

This section illustrates the numerical behavior of the scheme (25) through two test cases for which an analytical solution can be exhibited, thus enabling the numerical study of the convergence of the scheme. For both cases, the meshes used for the discretization of the domain $\Omega=(0,1)^2$ are issued from a 2D benchmark on anisotropic diffusion problem [33]. These triangle meshes show no symmetry which could artificially increase the convergence rate, and all angles of triangles are acute. This allows to compare situations where all coefficients a_{KL} defined by (18) are positive, with situations where some of them are negative by introducing

anisotropic permeability tensors. This family of meshes is built through the same pattern, which is reproduced at different scales: the first (coarsest) mesh and the third mesh are shown by Figure 2.

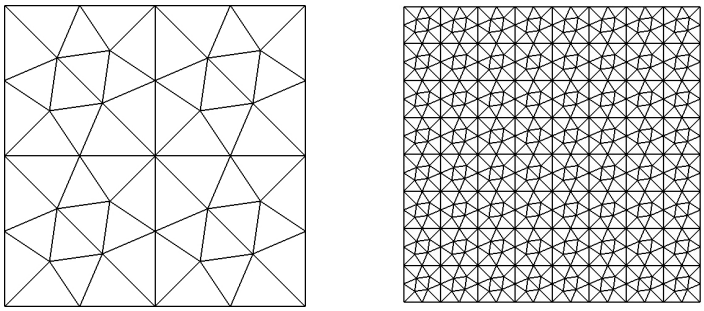


FIGURE 2. First and third mesh used in the numerical examples.

In the following numerical experiments, the permeability tensor is defined by

$$\Lambda = \begin{pmatrix} l_x & 0 \\ 0 & l_y \end{pmatrix}.$$

A first constant time step, denoted by Δt_1, is associated to the coarsest mesh and then between two successive meshes, the time step is divided by four since the mesh size is divided by two, so that the error due to the implicit Euler-time discretization remains negligible compared to that issued from the space discretization. The nonlinear systems obtained at each time step are solved by a Newton-Raphson algorithm. Notice that the scheme (25) has been implemented, not using the pressure p as main unknown, but using the saturation $u = \beta(p)$ as main unknown.

5.1. **First numerical example.** The first test case respects the framework of homogeneous Neumann boundary conditions of the model described in the theoretical part of the paper. The function β in (25) is defined through its reciprocal function β^{-1}

$$p = \beta^{-1}(u) = \log\left(\frac{u}{1-u}\right) \text{ hence } \beta(p) = \frac{e^p}{1+e^p}.$$

Then the function η of (25) is defined by

$$\eta(p) = u(1-u) = \beta(p)(1-\beta(p)) = \frac{e^p}{(1+e^p)^2}.$$

In this test case, note that (1) leads to

$$\partial_t u - \nabla \cdot (\Lambda \nabla u) = 0,$$

and therefore we compare the results with those obtained, using the linear scheme defined by $\beta_\ell^{-1}(u) = u$ and $\eta_\ell(p) = 1$ where the subscript ℓ is used here for the linear method. The numerical convergence of both cases has been compared on the following analytical solution,

$$\tilde{u}(\,(x,y)\,,t\,) = \frac{\cos(\pi x)\,\exp(-\pi^2 l_x t)\;+\;1}{2},$$

for $(x,y) \in \Omega$, $t \in (0,t_{\mathrm{f}})$, and where the final time t_{f} has been fixed to 0.07 and the first time step Δt_1 to 0.01024. Two different permeability tensors have been tested : the isotropic one $l_x = l_y = 1$ and an anisotropic one $l_x = 1$, $l_y = 10^3$. For all tests we have computed the errors in the classical discrete $L^2(Q_{t_{\mathrm{f}}})$, $L^1(Q_{t_{\mathrm{f}}})$ and $L^\infty(Q_{t_{\mathrm{f}}})$ norms. All the results are presented in the Tables below. Each table

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	$u_{\max}$
0.250	0.188E-02	-	0.387E-03	-	0.182E-01	-	0.041	0.959
0.125	0.478E-03	1.977	0.987E-04	1.970	0.473E-02	1.944	0.011	0.989
0.063	0.120E-03	1.995	0.250E-04	1.980	0.120E-02	1.976	0.003	0.997
0.031	0.300E-04	1.999	0.628E-05	1.994	0.305E-03	1.978	0.001	0.999
0.016	0.751E-05	2.000	0.157E-05	1.999	0.774E-04	1.979	0.000	1.000

TABLE 1. Test 1. Linear scheme, with an isotropic tensor.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	$u_{\max}$
0.250	0.129E-02	-	0.268E-03	-	0.133E-01	-	0.054	0.946
0.125	0.139E-02	-0.108	0.293E-03	-0.129	0.115E-01	0.211	0.016	0.984
0.063	0.889E-03	0.646	0.188E-03	0.639	0.693E-02	0.734	0.004	0.996
0.031	0.492E-03	0.855	0.104E-03	0.858	0.373E-02	0.895	0.001	0.999
0.016	0.257E-03	0.935	0.541E-04	0.938	0.193E-02	0.952	0.000	1.000

TABLE 2. Test 1. Nonlinear scheme, with an isotropic tensor.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	$u_{\max}$
0.250	0.980E-02	-	0.229E-02	-	0.677E-01	-	-0.020	1.020
0.125	0.235E-02	2.062	0.542E-03	2.076	0.201E-01	1.753	-0.008	1.008
0.063	0.579E-03	2.019	0.134E-03	2.019	0.531E-02	1.918	-0.002	1.002
0.031	0.144E-03	2.005	0.334E-04	2.005	0.136E-02	1.970	-0.001	1.001
0.016	0.360E-04	2.001	0.833E-05	2.001	0.342E-03	1.987	-0.000	1.000

TABLE 3. Test 1. Linear scheme, with an anisotropic tensor.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	$u_{\max}$
0.250	0.497E-01	-	0.115E-01	-	0.286E+00	-	0.300	0.700
0.125	0.486E-01	0.034	0.114E-01	0.017	0.294E+00	-0.038	0.180	0.820
0.063	0.438E-01	0.148	0.102E-01	0.155	0.269E+00	0.130	0.076	0.924
0.031	0.371E-01	0.240	0.856E-02	0.254	0.227E+00	0.240	0.025	0.975
0.016	0.295E-01	0.330	0.673E-02	0.349	0.181E+00	0.327	0.007	0.993

TABLE 4. Test 1. Nonlinear scheme, with an anisotropic tensor.

provides the mesh size h, the discrete errors and the associated convergence rate, and finally the minimum and maximum values of the discrete solution.

We observe on Tables 1 and 2 that the linear method shows higher convergence rates. However, the more the anisotropy ratio is important, the more the linear scheme exhibits over- and under-shoot (see in particular Table 3). On the contrary, following the theoretical properties proved in this paper, the nonlinear scheme satisfies the maximum principal whatever the anisotropy. Unfortunately, this property has a cost: the convergence is quite slow and accuracy decreases when the anisotropy ratio increases.

5.2. **Second numerical example.** The second test case deals with a degenerate parabolic equation with a Dirichlet boundary condition. The function β is defined by

$$\beta(p) = \begin{cases} 0 & \text{if } p \leq 0, \\ p & \text{if } p \in [0,1], \end{cases} \tag{72}$$

The mobility η is defined by

$$\eta(p) = \begin{cases} 0 & \text{if } p \leq 0 \text{ or } p \geq 2, \\ 2p & \text{if } p \in [0,1], \\ 2(2-p) & \text{if } p \in [1,2], \end{cases} \tag{73}$$

so that

$$\varphi(p) = \begin{cases} 0 & \text{if } p \leq 0, \\ p^2 & \text{if } p \in [0,1], \\ 2-(2-p)^2 & \text{if } p \in [1,2], \\ 2 & \text{if } p \geq 2. \end{cases} \tag{74}$$

In particular, any weak solution p of (1) that remains bounded between 0 and 1 is a solution to the anisotropic porous medium equation

$$\partial_t p - \nabla \cdot (\Lambda \nabla p^2) = 0.$$

We compare the results with the scheme obtained by taking the following functions $p = \beta_\ell^{-1}(u) = |u|u$ and $\eta_\ell(p) = 1$ where the subscript ℓ has been added for this formulation called the quasilinear one. Recall that u remains the numerical main unknown for the approximation of this degenerate parabolic equation. The numerical convergence of both schemes has been studied through the following analytical solution,

$$\tilde{u}(\ (x,y)\ ,t\ ) = \max(\ 2l_x t - x, 0\ ),$$

for $(x,y) \in \Omega, t \in (0,t_{\mathrm{f}})$, and where the final time t_{f} has been fixed to 0.5 seconds and the first time step is still given by $\Delta t_1 = 0.01024$ seconds. Two permeability tensors have been tested : the isotropic one $l_x = l_y = 1$ and an anisotropic $l_x = 1$, $l_y = 10^2$. For all tests we have again computed the errors in the classical discrete $L^2(Q_{t_{\mathrm{f}}})$, $L^1(Q_{t_{\mathrm{f}}})$ and $L^\infty(Q_{t_{\mathrm{f}}})$ norms. All the results are presented in the Tables below. Each table again provides the mesh size h, the discrete errors and the associated convergence rate, and finally the minimum and maximum values of the discrete solution.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$p_{\min}$	$p_{\max}$
0.250	0.671E-02	-	0.234E-02	-	0.431E-01	-	0.000	1.000
0.125	0.242E-02	1.472	0.746E-03	1.647	0.204E-01	1.081	0.000	1.000
0.063	0.828E-03	1.546	0.218E-03	1.776	0.949E-02	1.101	0.000	1.000
0.031	0.281E-03	1.560	0.611E-04	1.834	0.454E-02	1.066	0.000	1.000
0.016	0.959E-04	1.550	0.168E-04	1.863	0.223E-02	1.025	-0.000	1.000

TABLE 5. Test 2. Quasilinear diffusion scheme, with an isotropic tensor.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$p_{\min}$	$p_{\max}$
0.250	0.172E-01	-	0.673E-02	-	0.906E-01	-	0.000	1.000
0.125	0.104E-01	0.724	0.388E-02	0.793	0.650E-01	0.478	-0.000	1.000
0.063	0.604E-02	0.783	0.211E-02	0.880	0.424E-01	0.616	-0.000	1.000
0.031	0.339E-02	0.831	0.111E-02	0.924	0.263E-01	0.688	-0.000	1.000
0.016	0.185E-02	0.873	0.576E-03	0.951	0.159E-01	0.724	-0.000	1.000

TABLE 6. Test 2. Nonlinear scheme, with an isotropic tensor.

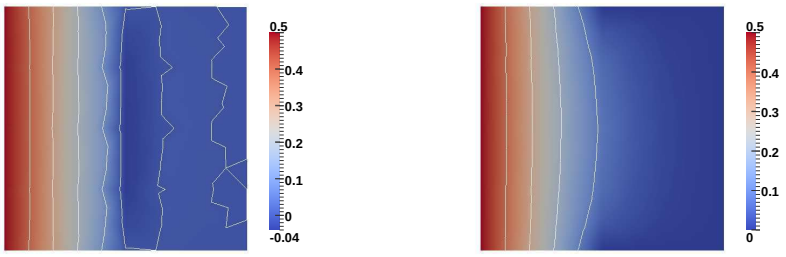
h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$p_{\min}$	$p_{\max}$
0.250	0.169E-01	-	0.734E-02	-	0.960E-01	-	-0.061	1.000
0.125	0.906E-02	0.897	0.356E-02	1.043	0.759E-01	0.338	-0.043	1.000
0.063	0.373E-02	1.282	0.128E-02	1.480	0.344E-01	1.142	-0.020	1.000
0.031	0.136E-02	1.450	0.395E-03	1.692	0.184E-01	0.907	-0.010	1.000
0.016	0.494E-03	1.466	0.116E-03	1.768	0.922E-02	0.994	-0.005	1.000

TABLE 7. Test 2. Quasilinear diffusion scheme with anisotropic tensor.

h	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$p_{\min}$	$p_{\max}$
0.250	0.226E-01	-	0.962E-02	-	0.110E+00	-	0.000	1.000
0.125	0.174E-01	0.376	0.720E-02	0.419	0.933E-01	0.243	0.000	1.000
0.063	0.132E-01	0.402	0.503E-02	0.517	0.753E-01	0.310	0.000	1.000
0.031	0.967E-02	0.444	0.334E-02	0.593	0.600E-01	0.329	0.000	1.000
0.016	0.691E-02	0.485	0.212E-02	0.654	0.472E-01	0.344	0.000	1.000

TABLE 8. Test 2. Nonlinear scheme with anisotropic tensor.

We observe that, as expected, the convergence rates of the quasilinear implementation are better, but the numerical solution presents some undershoots. Nevertheless, in the anisotropic case, the L^1, L^2 and L^∞ errors are of the same order on the coarsest meshes, which are currently used in the industrial applications.



(A) Quasilinear diffusion scheme (B) Nonlinear scheme

FIGURE 3. Test 2 - 2nd mesh. Discrete unknown p and its isovalues, for each scheme, with an anisotropic tensor.

APPENDIX A. TECHNICAL LEMMAS

Let Ω be a polygonal subset of $\mathbb{R}^2$, let $\mathcal{T}$ be a conforming triangular mesh of Ω as described in §2.1.1. For all triangle $T \in \mathcal{T}$, we denote by $h_T = \mathrm{diam}(T)$, and by $h = \sup_{T\in\mathcal{T}} h_T$. We consider the set $\mathcal{V}$ of the vertices (called *nodes*), and $(\mathbf{x}_K)_{K\in\mathcal{V}}$ their coordinates in Ω, and the usual $\mathbb{P}_1$ finite element space $V_\mathcal{T}$ introduced in §2.1.2.

Let $u_\mathcal{T} \in V_\mathcal{T}$, then one defines the piecewise constant functions $\overline{u}_\mathcal{T}$ and $\underline{u}_\mathcal{T}$ mapping Ω to $\mathbb{R}$ by

$$\begin{aligned}\overline{u}_\mathcal{T}(\mathbf{x}) = &\overline{u}_T = \sup_{\mathbf{x}\in T} u_\mathcal{T}(\mathbf{x}) \quad \text{if } \mathbf{x} \in T \in \mathcal{T},\\ \underline{u}_\mathcal{T}(\mathbf{x}) = &\underline{u}_T = \inf_{\mathbf{x}\in T} u_\mathcal{T}(\mathbf{x}) \quad \text{if } \mathbf{x} \in T \in \mathcal{T}.\end{aligned}$$

Lemma A.1. *There exists an absolute constant $c > 0$ such that*

$$\int_\Omega |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})| \mathrm{d}\mathbf{x} \leq ch \int_\Omega |\nabla u_\mathcal{T}(x)| \, \mathrm{d}\mathbf{x}.$$

Proof. Let $T \in \mathcal{T}$ be a triangle whose nodes are located at $\mathbf{x}_K, \mathbf{x}_L$ and $\mathbf{x}_M$. We assume without loss of generality that

$$\overline{u}_T = u_K \geq u_L \geq u_M = \underline{u}_T. \tag{75}$$

We denote by

$$\mathbf{x}_T = \frac{1}{3}(\mathbf{x}_K + \mathbf{x}_L + \mathbf{x}_M) \quad \text{and} \quad \mathbf{x}_{KL} = \frac{1}{2}(\mathbf{x}_K + \mathbf{x}_L),$$

and by $\omega_{K,L}^T$ the triangle whose vertices are $\mathbf{x}_K, \mathbf{x}_{KL}$ and $\mathbf{x}_T$. Note that

$$m_{\omega_{K,L}^T} = \frac{m_T}{6}, \tag{76}$$

where $m_{\omega_{K,L}^T}$ (resp. m_T) denotes the Lebesgue measure of $\omega_{K,L}^T$ (resp. T).

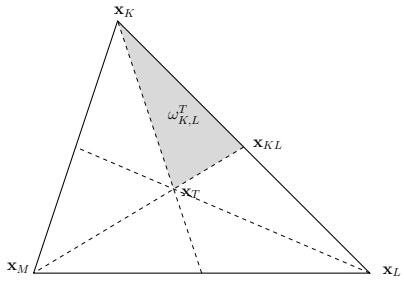


FIGURE 4. The subtriangle $\omega_{K,L}^T$.

Since $u_\mathcal{T} \in V_h$ is a linear function of $\mathbf{x}$ on T, then one has

$$u_T := \frac{1}{m_T} \int_T u_\mathcal{T}(\mathbf{x}) \mathrm{d}\mathbf{x} = u(\mathbf{x}_T) = \frac{1}{3}(u_K + u_L + u_M),$$

and, in view of (75), one has

$$u_\mathcal{T}(\mathbf{x}_{KL}) = \frac{1}{2}(u_K + u_L) \geq u_T.$$

As a consequence,

$$u_\mathcal{T}(\mathbf{x}) \geq u_T, \quad \forall \mathbf{x} \in \omega_{K,L}. \tag{77}$$

Hence, one has.

$$\begin{aligned}
\int_T |u_\mathcal{T}(\mathbf{x}) - u_T| \mathrm{d}\mathbf{x} &\geq \int_{\omega_{K,L}^T} |u_\mathcal{T}(\mathbf{x}) - u_T| \mathrm{d}\mathbf{x} = \int_{\omega_{K,L}^T} (u_\mathcal{T}(\mathbf{x}) - u_T) \mathrm{d}\mathbf{x} \\
&= \frac{m_{\omega_{K,L}^T}}{3} \left(u_K + \frac{u_K + u_L}{2} - 2u_T \right) = \frac{m_T}{18}(5u_K - u_L - 4u_M) \\
&\geq \frac{2m_T}{9}(u_K - u_M) = \frac{2}{9} \int_T |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})| \mathrm{d}\mathbf{x}. \qquad (78)
\end{aligned}$$

On the other hand, it follows from the Poincaré-Wirtinger inequality (cf. [1]) that

$$\int_T |u_\mathcal{T}(\mathbf{x}) - u_T| \mathrm{d}\mathbf{x} \leq \frac{h_T}{2} \int_T |\nabla u_\mathcal{T}(\mathbf{x})| \mathrm{d}\mathbf{x},$$

which, together with (78), yields

$$\int_\Omega |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})|\mathrm{d}\mathbf{x} = \sum_{T\in\mathcal{T}} |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})|\mathrm{d}\mathbf{x}$$
$$\leq \frac{9}{2}\sum_{T\in\mathcal{T}}\int_T |u_\mathcal{T}(\mathbf{x}) - u_T|\mathrm{d}\mathbf{x} \leq \frac{9h}{4}\int_\Omega |\nabla u_\mathcal{T}(\mathbf{x})|\mathrm{d}\mathbf{x},$$

concluding the proof of Lemma A.1. □

Lemma A.2. *Let $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, and let $u_\mathcal{T}$ and $u_\mathcal{M}$ be the corresponding piecewise linear and piecewise constant reconstructions defined in §2.1.2, then*

$$\int_\Omega |u_\mathcal{T}(\mathbf{x}) - u_\mathcal{M}(\mathbf{x})|\mathrm{d}\mathbf{x} \leq ch\|\nabla u_\mathcal{T}\|_{(L^1(\Omega))^2}, \tag{79}$$

where c is an absolute constant.

Proof. In order to prove (79), it suffices to check that, for a.e. $\mathbf{x}\in\Omega$,

$$|u_\mathcal{T}(\mathbf{x}) - u_\mathcal{M}(\mathbf{x})| \leq |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})|.$$

One concludes by using Lemma A.1. □

References

[1] G. Acosta and R. G. Durán. An optimal Poincaré inequality in L^1 for convex domains. *Proc. Amer. Math. Soc.*, 132(1):195–202 (electronic), 2004.

[2] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.

[3] O. Angelini, K. Brenner, and D. Hilhorst. A finite volume method on general meshes for a degenerate parabolic convection-reaction-diffusion equation. *Numer. Math.*, 123(2):219–257, 2013.

[4] H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Applications to PDEs and optimization.

[5] B. R. Baliga and S. V. Patankar. A control volume finite-element method for two-dimensional fluid flow and heat transfer. *Numerical Heat Transfer*, 6(3):245–261, 1983.

[6] J. Bear. *Dynamic of Fluids in Porous Media*. American Elsevier, New York, 1972.

[7] M. Bessemoulin-Chatard and F. Filbet. A finite volume scheme for nonlinear degenerate parabolic equations. *SIAM J. Scientific Computing*, 34(5):559–583, 2012.

[8] Marianne Bessemoulin-Chatard. *Développement et analyse de schémas volumes finis motivés par la présentation de comportements asymptotiques. Application à des modèles issus de la physique et de la biologie*. PhD thesis, Université Blaise Pascal - Clermont-Ferrand II, 2012.

[9] K. Brenner, C. Cancès, and D. Hilhorst. Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure. *Comput. Geosci.*, 17(3):573–597, 2013.

[10] K. Brenner and R. Masson. Convergence of a vertex centered discretization of two-phase darcy flows on general meshes. *Int. J. Finite Vol.*, 10:1–37, 2013.

[11] H. Brézis. *Analyse Fonctionnelle: Théorie et applications*. Masson, 1983.

[12] F. Buzzi, M. Lenzinger, and B. Schweizer. Interface conditions for degenerate two-phase flow equations in one space dimension. *Analysis*, 29:299–316, 2009.

[13] C. Cancès, T. Gallouët, and A. Porretta. Two-phase flows involving capillary barriers in heterogeneous porous media. *Interfaces Free Bound.*, 11(2):239–258, 2009.

[14] C. Cancès and M. Pierre. An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.*, 44(2):966–992, 2012.

[15] J. Carrillo. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.*, 147(4):269–361, 1999.

[16] J. Casado-Díaz, T. Chacón Rebollo, V. Girault, M. Gómez Mármol, and F. Murat. Finite elements approximation of second order linear elliptic equations in divergence form with right-hand side in L^1. *Numer. Math.*, 105(3):337–374, 2007.

[17] C. Chainais-Hillairet. Entropy method and asymptotic behaviours of finite volume schemes. In *FVCA7 conference proceedings*, 2014.

[18] C. Chainais-Hillairet, A. Jüngel, and Schuchnigg S. Entropy-dissipative discretization of nonlinear diffusion equations and discrete Beckner inequalities. HAL : hal-00924282.

[19] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.
[20] J. Droniou, T. Gallouët, and R. Herbin. A finite volume scheme for a noncoercive elliptic equation with measure data. *SIAM J. Numer. Anal.*, 41(6):1997–2031 (electronic), 2003.
[21] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Series*. Springer, New York, 2004.
[22] R. Eymard, P. Féron, T. Gallouët, C. Guichard, and R. Herbin. Gradient schemes for the stefan problem. *Int. J. Finite Vol.*, 13:1–37, 2013.
[23] R. Eymard, T. Gallouët, M. Ghilani, and R. Herbin. Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.*, 18(4):563–594, 1998.
[24] R. Eymard, T. Gallouet, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. HAL: hal-00801648, 2013.
[25] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.
[26] R. Eymard, T. Gallouët, R. Herbin, and A. Michel. Convergence of finite volume schemes for parabolic degenerate equations. *Numer. Math.*, 92:41–82, 2002.
[27] R. Eymard, T. Gallouët, D. Hilhorst, and Y. Naït Slimane. Finite volumes and nonlinear diffusion equations. *RAIRO Modél. Math. Anal. Numér.*, 32(6):747–761, 1998.
[28] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for richards equation. *Comput. Geosci.*, 3(3-4):259–294, 1999.
[29] R. Eymard, R. Herbin, and A. Michel. Mathematical study of a petroleum-engineering scheme. *M2AN Math. Model. Numer. Anal.*, 37(6):937–972, 2003.
[30] R. Eymard, D. Hilhorst, and M. Vohralík. A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numer. Math.*, 105(1):73–131, 2006.
[31] R. Eymard, D. Hilhorst, and M. Vohralík. A combined finite volume-finite element scheme for the discretization of strongly nonlinear convection-diffusion-reaction problems on nonmatching grids. *Numer. Methods Partial Differential Equations*, 26(3):612–646, 2010.
[32] T. Gallouët and J.-C. Latché. Compactness of discrete approximate solutions to parabolic PDEs—application to a turbulence model. *Commun. Pure Appl. Anal.*, 11(6):2371–2391, 2012.
[33] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In R. Eymard and J.-M. Herard, editors, *Finite Volumes for Complex Applications V*, pages 659–692. Wiley, 2008.
[34] R. Huber and R. Helmig. Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media. *Comput. Geosci.*, 4(2):141–164, 2000.
[35] J. Leray and J. Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup. (3)*, 51:45–78, 1934.
[36] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differential Equations*, 131:20–38, 1996.
[37] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.
[38] B. Saad and M. Saad. Study of full implicit petroleum engineering finite volume scheme for compressible two phase flow in porous media. *SIAM J. Numer. Anal.*, 51(1):716–741, 2013.
[39] B. Saad and M. Saad. Numerical analysis of a non equilibrium two-component two-compressible flow in porous media. *Discr. Cont. Dyn. Syst. S*, 7(2):317 – 346, 2014.
[40] B. Schweizer. The Richards equation with hysteresis and degenerate capillary pressure. *J. Differential Equations*, 252(10):5594–5612, 2012.
[41] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146:65–96, 1987.
[42] J. L. Vázquez. *The porous medium equation*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, Oxford, 2007. Mathematical theory.

Clément Cancès, Cindy Guichard
Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
cances@ann.jussieu.fr, guichard@ann.jussieu.fr