HAL
open science

Numerical analysis of a robust free energy diminishing Finite Volume scheme for parabolic equations with gradient structure

Clément Cancès, Cindy Guichard

To cite this version:

Clément Cancès, Cindy Guichard. Numerical analysis of a robust free energy diminishing Finite Volume scheme for parabolic equations with gradient structure. Foundations of Computational Mathematics, 2017, 17 (6), pp.1525-1584. 10.1007/s10208-016-9328-6 . hal-01119735v4

HAL Id: hal-01119735

https://hal.science/hal-01119735v4

Submitted on 7 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NUMERICAL ANALYSIS OF A ROBUST FREE ENERGY DIMINISHING FINITE VOLUME SCHEME FOR PARABOLIC EQUATIONS WITH GRADIENT STRUCTURE

CLÉMENT CANCÈS AND CINDY GUICHARD

Abstract. We present a numerical method for approximating the solutions of degenerate parabolic equations with a formal gradient flow structure. The numerical method we propose preserves at the discrete level the formal gradient flow structure, allowing the use of some nonlinear test functions in the analysis. The existence of a solution to and the convergence of the scheme are proved under very general assumptions on the continuous problem (nonlinearities, anisotropy, heterogeneity) and on the mesh. Moreover, we provide numerical evidences of the efficiency and of the robustness of our approach.

Communicated by Eitan Tadmor

AMS classification subjects: 65M12, 35K65, 65M08

Keywords: degenerate parabolic equation, finite volumes, nonlinear stability, general grids, convergence analysis

1. Introduction

Many problems coming from physics (like e.g. porous media flows modeling [11, 10, 35]) or biology (like e.g. chemotaxis modeling [67]) lead to degenerate parabolic equations or systems. Many of these models can be interpreted as gradient flows in appropriate geometries. For instance, such variational structures were depicted for porous media flows [86, 70, 32], chemotaxis processes in biology [19], superconductivity [5, 4], or semiconductor devices modeling [84, 68] (this list is far from being complete).

Designing accurate numerical schemes for approximating their solutions is therefore a major issue. In the case of porous media flow models — used e.g. in oil-engineering, water resources management or nuclear waste repository management — the problems may moreover be highly anisotropic and heterogeneous. As an additional difficulty, the meshes are often prescribed by geological data, yielding non-conformal grids made of elements of various shapes. This situation can also be encountered in mesh adaptation procedures. Hence, the robustness of the method w.r.t. anisotropy and to the grid is an important quality criterion for a numerical method in view of practical applications.

In this contribution, we focus on the numerical approximation of a single nonlinear Fokker-Planck equation. Since it contains crucial difficulties arising in the applications, namely degeneracy and possibly strong anisotropy, the discretization

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

of this nonlinear Fokker-Planck equation appears to be a keystone for the approximation of more complex problems.

1.1. **Presentation of the continuous problem.** Let Ω be a polyhedral connected open bounded subset of $\mathbb{R}^d$ ($d = 2$ or 3), and let $t_{\mathrm{f}} > 0$ be a finite time horizon. In this contribution, we focus on the discretization of the model problem

$$\begin{cases} \partial_t u - \boldsymbol{\nabla}\cdot(\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}(p(u)+V)) = 0 & \text{in } Q_{t_{\mathrm{f}}} := \Omega\times(0,t_{\mathrm{f}}),\\ (\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}(p(u)+V))\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega\times(0,t_{\mathrm{f}}),\\ u_{|t=0} = u_0 & \text{in } \Omega, \end{cases} \tag{1}$$

which appears to be a keystone before discretizing more complex problems. We do the following assumptions on the data of the continuous problem (1).

(**A**1) The function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous function such that $\eta(0) = 0$, $\eta(u) > 0$ if $u \neq 0$ and η is non-decreasing on $\mathbb{R}_+$. The function η is continuously extended on the whole $\mathbb{R}$ into an even function. It is called the *mobility* function in reference to the porous media flow context.

(**A**2) The so-called *(entropy) pressure* function $p \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ is absolutely continuous and increasing on $(0,+\infty)$ (i.e., $0 < p' \in L^1_{\mathrm{loc}}((0,+\infty))$), and satisfies $\lim_{u\to+\infty} p(u) = +\infty$. In the case where $p(0) = \lim_{u\searrow 0} p(u)$ is finite, the function p is extended into an increasing absolutely continuous function $p : \mathbb{R} \to \mathbb{R}$ defined by

$$p(u) = 2p(0) - p(-u), \qquad \forall u \leq 0. \tag{2}$$

We denote by

$$\mathcal{I}_p = \begin{cases} \mathbb{R}_+^* & \text{if } p(0) = -\infty,\\ \mathbb{R} & \text{if } p(0) > -\infty. \end{cases}$$

and by $\overline{\mathcal{I}}_p$ its closure in $\mathbb{R}$. We additionally require that the function $\sqrt{\eta}p'$ belongs to $L^1_{\mathrm{loc}}(\mathbb{R}_+)$ (and is in particular integrable near 0) and that $\lim_{u\searrow 0}\sqrt{\eta(u)}p(u) = 0$.

(**A**3) The tensor field $\boldsymbol{\Lambda} : \Omega \to (L^\infty(\mathbb{R}))^{d\times d}$ is such that $\boldsymbol{\Lambda}(\boldsymbol{x})$ is symmetric for almost every $\boldsymbol{x} \in \Omega$. Moreover, we assume that there exist $\lambda_\star > 0$ and $\lambda^\star \in [\lambda_\star, +\infty)$ such that

$$\lambda_\star|\mathbf{v}|^2 \leq \boldsymbol{\Lambda}(\boldsymbol{x})\mathbf{v}\cdot\mathbf{v} \leq \lambda^\star|\mathbf{v}|^2, \qquad \forall \mathbf{v} \in \mathbb{R}^d, \text{ for a.e. } \boldsymbol{x} \in \Omega. \tag{3}$$

$\boldsymbol{\Lambda}$ is called the *intrinsic permeability* tensor field still in reference to the porous media flow context.

(**A**4) The initial data u_0 is assumed to belong to $L^1(\Omega)$. Moreover, defining the convex function $\Gamma : \overline{\mathcal{I}}_p \to \mathbb{R}_+$ (called *entropy function* in the following) by

$$\Gamma(u) = \int_1^u (p(a) - p(1))\mathrm{d}a, \qquad \forall u \in \mathcal{I}_p, \tag{4}$$

we assume that the following positivity and finite entropy conditions are fulfilled:

$$u_0 \geq 0 \text{ a.e. in } \Omega, \qquad \int_\Omega u_0 \mathrm{d}\boldsymbol{x} > 0, \qquad \int_\Omega \Gamma(u_0)\mathrm{d}\boldsymbol{x} < +\infty. \tag{5}$$

(**A**5) The *exterior potential* $V : \Omega \to \mathbb{R}$ is Lipschitz continuous.

Throughout this paper, we adopt the convention

$$\Gamma(u) = +\infty \quad \text{if } p(0) = -\infty \text{ and } u < 0. \tag{6}$$

In order to give a proper mathematical sense to the solution of (1), we need to introduce the function $\xi : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$\xi(u) = \int_0^u \sqrt{\eta(a)} p'(a) \mathrm{d}a, \qquad \forall u \geq 0. \tag{7}$$

Note that ξ is well defined since we assumed that $\sqrt{\eta} p'$ belongs to $L^1_{\mathrm{loc}}(\mathbb{R}_+)$. Moreover, in the case where $p(0)$ is finite, then the formula (7) can be extended to the whole $\mathbb{R}$, leading to an odd function. We additionally assume that the following relations between ξ, η and Γ hold:

(**A**6) There exists $C > 0$ such that

$$0 \leq \xi(u) \leq C\left(1 + \Gamma(u)\right), \qquad \forall u \in [0, +\infty). \tag{8}$$

Moreover, we assume that

$$\frac{\Gamma(u)}{\eta(u)} \to +\infty \quad \text{as } u \to +\infty, \tag{9}$$

and that the function

$$\sqrt{\eta \circ \xi^{-1}} \text{ is uniformly continuous on the range of } \xi. \tag{10}$$

Definition 1 (weak solution)**.** *A measurable function u is said to be a weak solution to problem (1) if*

- **i.** *the functions u and $\eta(u)$ belong to $L^\infty((0, t_\mathrm{f}); L^1(\Omega))$;*
- **ii.** *the function $\xi(u)$ belongs to $L^2((0, t_\mathrm{f}); H^1(\Omega))$;*
- **iii.** *for all function $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_\mathrm{f}); \mathbb{R})$, one has*

$$\begin{aligned} &\iint_{Q_{t_\mathrm{f}}} u \partial_t \psi \, \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega u_0 \psi(\cdot, 0) \mathrm{d}\boldsymbol{x} \\ &\qquad - \iint_{Q_{t_\mathrm{f}}} \eta(u) \boldsymbol{\Lambda} \boldsymbol{\nabla} V \cdot \boldsymbol{\nabla} \psi \mathrm{d}\boldsymbol{x}\mathrm{d}t - \iint_{Q_{t_\mathrm{f}}} \boldsymbol{\Lambda} \boldsymbol{\nabla} \xi(u) \cdot \sqrt{\eta(u)} \boldsymbol{\nabla} \psi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0. \end{aligned} \tag{11}$$

Following the seminal work of [2], there exists at least one weak solution u to the problem (1). Denoting by

$$\phi(u) = \int_0^u \eta(a) p'(a) \mathrm{d}a, \qquad \forall u \in \mathcal{I}_p,$$

the uniqueness of the solution (and even a L^1-contraction principle) is ensured as soon as $\eta \circ \phi^{-1} \in C^{0,1/2}$ (cf. [85], see also [7] for a slightly weaker condition in the case of a smooth domain Ω). Moreover, u belongs to $C([0, t_\mathrm{f}]; L^1(\Omega))$ (cf. [31]) and $u(\cdot, t) \geq 0$ for all $t \in [0, t_\mathrm{f}]$ thanks to classical monotonicity arguments.

Remark 1.1. *Assumptions (**A**1)–(**A**6) formulated above deserve some comments.*

- *First of all, let us stress that Assumptions (**A**1)– (**A**2) and (**A**6) are satisfied if $\eta(u) = u$ and $p(u) = \log(u)$ as in the seminal paper of Jordan, Kinderlehrer, and Otto [65]. One can also deal with power like pressure functions $p(u) = u^{m-1}$, but only for $m > 1$. Our study does not cover the case of the fast-diffusion equation $m < 1$ with linear mobility function (see e.g. [86]) because of the technical assumption (**A**6).*

- *The most classical choice for the mobility function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is $\eta(u) = u$. In this case, the convection is linear. In this situation, the formal gradient flow structure highlighted in §1.2 can be made rigorous following the program proposed in [65, 86, 1, 3, 77] and many others. The gradient flow structure can also be made rigorous in the case where η is concave (cf. [43]) and in the non-degenerate case $\eta(u) \geq \alpha > 0$.*
- *One assumes in (**A**1) that η is nondecreasing on $\mathbb{R}_+$. This assumption is natural in all the applications we have in mind. However, it is not mandatory in the proof and can be easily relaxed: it would have been sufficient to assume that there exists $\gamma > 0$ such that*

$$\frac{\eta(a) + \eta(b)}{2} = \gamma \max_{s \in [a,b]} \eta(s), \qquad \forall [a,b] \subset \mathbb{R}.$$

  This relation is clearly satisfied with $\gamma = 1/2$ when η is nondecreasing.
- *In Assumption (**A**2), the condition $\lim_{u\to\infty} p(u) = +\infty$ ensures that*

$$\lim_{u\to\infty} \frac{\Gamma(u)}{u} = +\infty,$$

  *where Γ was defined in (**A**4). Given a sequence $(u_n)_n \subset L^1(\Omega)$ with bounded entropy, i.e., such that $\int_\Omega \Gamma(u_n)\mathrm{d}\boldsymbol{x}$ is bounded, then $(u_n)_n$ is uniformly equi-integrable thanks to the de La Vallée Poussin's theorem [42]. Therefore, a sequence $(u_n)_n$ with bounded entropy relatively compact for the weak topology of $L^1(\Omega)$.*
- *Since the unique weak solution to the problem remains non-negative, the extension of η and p on the whole $\mathbb{R}$ could seem to be useless. However, in the case where $p(0)$ is finite, the non-negativity of the solution may not be preserved by the numerical method we propose. The extension of the functions η and p on $\mathbb{R}_-$ is then necessary.*
- *Only the regularity of the potential V is prescribed by (**A**5). Confining potentials like e.g. $V(\boldsymbol{x}) = \frac{|\boldsymbol{x}-\boldsymbol{x}_\star|^2}{2}$ for some $\boldsymbol{x}_\star \in \Omega$, or gravitational potential $V(\boldsymbol{x}) = -\boldsymbol{g}\cdot\boldsymbol{x}$, where $\boldsymbol{g}$ is the (downward) gravity vector can be considered.*

1.2. **Formal gradient flow structure of the continuous problem.** Let us highlight the (formal) gradient flow structure of the system (1). Following the path of [86, §1.3] (see also [84, 87]), the calculations carried out in this section are formal. They can be made rigorous under the non-degeneracy assumption $\eta(u) \geq \alpha > 0$ for all $u \geq 0$.

Define the affine space

$$\mathfrak{M} = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega u_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right\}$$

of the admissible states, called *state space*.

In order to define a Riemannian geometry on $\mathfrak{M}$, we need to introduce the tangent space $T_u\mathfrak{M}$, given by

$$T_u\mathfrak{M} = \left\{ w : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0 \right\}.$$

We also need to define the metric tensor $\mathfrak{g}_u : T_u\mathfrak{M} \times T_u\mathfrak{M} \to \mathbb{R}$, which consists in a scalar product on $T_u\mathcal{M}$ (depending on the state u)

$$\mathfrak{g}_u(w_1, w_2) = \int_\Omega \phi_1 w_2 \,\mathrm{d}\boldsymbol{x} = \int_\Omega w_1 \phi_2 \mathrm{d}\boldsymbol{x} = \int_\Omega \eta(u) \boldsymbol{\nabla}\phi_1 \cdot \boldsymbol{\Lambda}\boldsymbol{\nabla}\phi_2 \,\mathrm{d}\boldsymbol{x}, \tag{12}$$

for all $w_1, w_2 \in T_u\mathfrak{M}$, where ϕ_i are defined *via* the elliptic problem

$$\begin{cases} -\boldsymbol{\nabla}\cdot(\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}\phi_i) = w_i \text{ in } \Omega, \\ \eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}\phi_i \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega, \\ \displaystyle\int_\Omega \phi_i \,\mathrm{d}\boldsymbol{x} = 0. \end{cases} \tag{13}$$

Note that $T_u\mathfrak{M}$ does not depend on u (at least in the non-degenerate case), but the metric tensor $\mathfrak{g}_u(\cdot,\cdot)$ does. So we are not in a Hilbertian framework.

Define the *free energy* functional (cf. [64])

$$\mathfrak{E} : \begin{cases} \mathfrak{M} \to \mathbb{R} \cup \{+\infty\} \\ u \mapsto \mathfrak{E}(u) = \displaystyle\int_\Omega (\Gamma(u(\boldsymbol{x})) + u(\boldsymbol{x})V(\boldsymbol{x})) \,\mathrm{d}\boldsymbol{x}, \end{cases} \tag{14}$$

and the *hydrostatic pressure* function

$$\mathfrak{h} : \begin{cases} \mathcal{I}_p \times \Omega \to \mathbb{R} \\ (u, \boldsymbol{x}) \mapsto \mathfrak{h}(u, \boldsymbol{x}) = p(u) + V(\boldsymbol{x}) = D_u\mathfrak{E}(u). \end{cases}$$

Then given $w \in T_u\mathfrak{M}$, one has

$$D_u\mathfrak{E}(u) \cdot w = \int_\Omega \mathfrak{h}(u(\boldsymbol{x}), \boldsymbol{x}) w(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int_\Omega \eta(u) \boldsymbol{\nabla}\mathfrak{h}(u,\cdot) \cdot \boldsymbol{\Lambda}\boldsymbol{\nabla}\phi \,\mathrm{d}\boldsymbol{x}, \tag{15}$$

where ϕ is deduced from w using the elliptic problem (13). Moreover, thanks to (12), one has

$$\mathfrak{g}_u(\partial_t u, w) = \int_\Omega \partial_t u \, \phi \,\mathrm{d}\boldsymbol{x}, \qquad \forall w \in T_u\mathfrak{M}. \tag{16}$$

In view of (15)–(16), the problem (1) is equivalent to

$$\mathfrak{g}_u(\partial_t u, w) = -D_u\mathfrak{E}(u) \cdot w = -\mathfrak{g}_u(\boldsymbol{\nabla}_u\mathfrak{E}(u), w), \qquad \forall w \in T_u\mathfrak{M}, \tag{17}$$

where the cotangent vector $D_u\mathfrak{E}(u) \in (T_u\mathfrak{M})^*$ has been identified to the tangent vector $\boldsymbol{\nabla}_u\mathfrak{E}(u) \in T_u\mathfrak{M}$ thanks to Riesz theorem applied on $T_u\mathfrak{M}$ with the scalar product $\mathfrak{g}_u$. This relation can be rewritten as

$$\partial_t u = -\boldsymbol{\nabla}_u\mathfrak{E}(u) = \boldsymbol{\nabla}\cdot(\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}\mathfrak{h}(u,\cdot)) \quad \text{in } T_u\mathfrak{M}, \tag{18}$$

justifying the gradient flow denomination.

Choosing $w = \partial_t u$ in (17) and using (18), we get that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u) = D_u\mathfrak{E}(u) \cdot \partial_t u = -D_u\mathfrak{E}(u) \cdot \boldsymbol{\nabla}_u\mathfrak{E}(u) = \int_\Omega \eta(u) \boldsymbol{\nabla}\mathfrak{h}(u,\cdot) \cdot \boldsymbol{\Lambda}\boldsymbol{\nabla}\mathfrak{h}(u,\cdot) \mathrm{d}\boldsymbol{x}.$$

An integration w.r.t. time yields the classical energy/dissipation relation: $\forall t \in [0, t_\mathrm{f}]$,

$$\begin{aligned} &\mathfrak{E}(u(\cdot,t)) - \mathfrak{E}(u_0) \\ &\qquad + \int_0^t \int_\Omega \eta(u(\boldsymbol{x},\tau))\boldsymbol{\Lambda}(\boldsymbol{x})\boldsymbol{\nabla}\mathfrak{h}(u(\boldsymbol{x},\tau),\boldsymbol{x}) \cdot \boldsymbol{\nabla}\mathfrak{h}(u(\boldsymbol{x},\tau),\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}\tau = 0. \end{aligned} \tag{19}$$

The fact that a physical problem has a gradient flow structure provides some informations concerning its evolution. The physical system aims at decreasing its free energy as fast as possible. As highlighted by (19), the whole energy decay corresponds to the dissipation. As a byproduct, the free energy is a Liapunov functional and the total dissipation (integrated w.r.t. time) is bounded by the free energy associated to the initial data. The variational structure was exploited for instance in [86, 21, 22, 93] to study the long-time asymptotic of the system.

1.3. **Goal and positioning of the paper.** The goal of this paper is to propose and analyse a numerical scheme that mimics at the discrete level the gradient flow structure highlighted in §1.2. Since the point of view adopted in our presentation concerning the gradient flow structure is formal, the rigorous numerical analysis of the scheme will rather rely on the well established theory of weak solutions in the sense of Definition 1. But as a byproduct of the formal gradient flow structure, the discrete free energy will decrease along time, yielding the non-linear stability of the scheme.

There are some existing numerical methods based on Eulerian coordinates (as the one proposed in this paper). This is for instance the case of monotone discretizations, that can be reinterpreted as Markov chains [81, 45], for which one can even prove a rigorous gradient flow structure. Classical ways to construct monotone discretizations in the isotropic setting $\mathbf{\Lambda}(\boldsymbol{x}) = \lambda(\boldsymbol{x})\mathbf{I}_d$ are to use finite volumes schemes with two-points flux approximation (TPFA, see e.g. [51, 50]) or finite elements on Delaunay's meshes (see e.g. [37]). An advanced second order in space finite volume method was proposed in [18] and discontinuous Galerkin schemes in [80, 79]. However, these approaches — as well as the finite difference scheme proposed in [78] — require strong assumptions on the mesh. Moreover, the extension to the anisotropic framework of TPFA finite volume scheme fails for consistency reasons (cf. [50]), while finite elements are no longer monotone on a prescribed mesh for general anisotropy tensors $\mathbf{\Lambda}$. In [61, 56], it appears that all the linear finite volume schemes (i.e., schemes leading to linear systems when linear equations are approximated) able to handle general grids and anisotropic tensors may lose at the discrete level the monotonicity of the continuous problem.

The monotonicity at the discrete level can be restored thanks to nonlinear corrections [27, 71, 30, 72]. Another approach consists in designing directly monotone nonlinear schemes (see, e.g., [66, 92, 75, 76, 44, 90]). However, the monotonicity of the method is not sufficient to ensure the decay of non-quadratic energies. Moreover, the available convergence proofs [44, 30] require some numerical assumptions involving the numerical solution itself. Finally, let us mention here the recent contribution [59] where linear monotone schemes are constructed on cartesian grids for possibly anisotropic tensors $\mathbf{\Lambda}$.

In the case where $\mathbf{\Lambda} = \mathbf{I}_d$ and $\eta(u) = u$, the formal gradient flow structure can be made rigorous in the metric space

$$\mathcal{P}(\Omega) := \left\{ u \in L^2(\Omega; \mathbb{R}_+) \;\middle|\; \int_\Omega u \mathrm{d}\boldsymbol{x} = \int_\Omega u_0 \mathrm{d}\boldsymbol{x} \right\}$$

endowed with the Wasserstein metric with quadratic cost function. Several approach were proposed in the last years for solving the JKO minimization scheme (cf. [65, 3]). This requires the computation of Wasserstein distances. If $d = 1$, switching to Lagrangian coordinates is a natural choice that has been exploited for

example in [69, 20, 82, 83]. The case $d \geq 2$ is more intricate. Methods based on a so-called *entropic regularization* [13, 88] of the transport plan appear to be costly, but very tractable. Another approach consists in solving the so-called Monge-Ampère equation in order to compute the optimal transport plan [15]. Let us finally mention the application [14] to the Wasserstein gradient flows of the *CFD relaxation approach* of Benamou and Brenier [12] to solve the Monge-Kantorovich problem.

Motivated by applications in the context of complex porous media flows where irregular grids are often prescribed by the geology, the scheme we propose was designed to be able to handle highly anisotropic and heterogeneous diffusion tensor $\mathbf{\Lambda}$ and very general grids (non-conformal grids, cells of various shapes). It relies on the recently developed *Vertex Approximate Gradient* (VAG) method [52, 54, 53, 26], but alternative versions can be inspired from most of the existing symmetric coercive methods for approximating the solutions of linear elliptic equations [61, 56]. Moreover, we want our scheme to mimic at the discrete level the gradient flow structure highlighted in §1.2. This ensures in particular the decay of the discrete counterpart of the free energy, and thus the nonlinear stability of the scheme.

A nonlinearly stable control volume finite elements (CVFE) scheme was proposed in our previous contribution [34] (see also [33]). The nonlinear CVFE scheme [34] is based on a suitable upwinding of the mobility. It is only first order accurate in space while linear CVFE schemes are second order accurate in space. Moreover, it appears in the numerical simulations presented in [34] that this nonlinear CVFE scheme lacks robustness with respect to anisotropy: its convergence is slow in particularly unfavorable situations.

The main goal of this paper is to propose a scheme that preserves some important features of the one introduced in [34] (possible use of some prescribed nonlinear test function, decay of the physically motivated energy, convergence proof for discretization parameters tending to 0), without jeopardizing the accuracy of the scheme compared to the more classical approach based on formulations with Kirchhoff transforms (see for instance [57, 89, 58, 9]). Convincing numerical results are provided in §5 as an evidence of the efficiency of our approach. Two theorems are stated in §2.4 (and proved in §3 and §4) in order to ensure the following properties:

(1) Theorem 2.3. At a fixed mesh, the scheme, that consists in a nonlinear system, admits (at least) one solution. This allows in particular to speak about the discrete solution provided by the scheme. Moreover, we take advantage of the gradient structure of the scheme for deriving some nonlinear stability estimates.
(2) Theorem 2.4. Letting the discretization parameters tend to 0 (while controlling some regularity factors related to the discretization), the discrete solution converges in some appropriate sense towards the unique weak solution to the problem (1) in the sense of Definition 1.

Remark 1.2. *We only consider potential convection in the paper. A more general convection with speed* $\mathbf{v}$ *can be split into a potential part and a divergence free part:*

$$\mathbf{v} = -\nabla V + \nabla \times \mathbf{w}.$$

We suggest for instance to use classical (entropy stable) finite volume schemes (see for instance [74]*) for the divergence free part combined with our method for the potential part.*

2. Definition of the scheme and main results

As already mentioned, the scheme we propose is based on the so-called VAG scheme [52]. In §2.1, we state our assumptions on the spatial mesh and the time discretization of $(0, t_{\mathrm{f}})$. Then in §2.2, we define the nonlinear scheme we will study in this paper. The gradient flow structure of the discretized problem is highlighted in §2.3, where a variational interpretation is given to the scheme. Finally, in §2.4, we state the existence of discrete solutions to the scheme and their convergence towards the unique weak solution as the discretization parameters tend to 0.

2.1. Discretization of $Q_{t_{\mathrm{f}}}$ and discrete functional spaces.

2.1.1. *Spatial discretization and discrete reconstruction operators.* Following [53, 26], we consider generalized polyhedral discretizations of Ω. Let $\mathcal{M}$ be the set of the cells, that are disjoint polyhedral open subsets of Ω such that $\bigcup_{\kappa\in\mathcal{M}} \overline{\kappa} = \overline{\Omega}$. Each cell $\kappa \in \mathcal{M}$ is assumed to be star-shaped with respect to its so-called *center*, denoted by $\boldsymbol{x}_\kappa$. We denote by $\mathcal{F}$ the set of the faces of the mesh, which are not assumed to be planar if $d = 3$ (whence the term "generalized polyhedral"). We denote by $\mathcal{V}$ the set of the vertices of the mesh. We denote by $\boldsymbol{x}_{\mathrm{s}} \in \overline{\Omega}$ the location of the vertex $\mathrm{s} \in \mathcal{V}$. The sets $\mathcal{V}_\kappa$, $\mathcal{F}_\kappa$ and $\mathcal{V}_\sigma$ denote respectively the vertices and faces of a cell κ, and the vertices of a face σ. For any face $\sigma \in \mathcal{F}_\kappa$, one has $\mathcal{V}_\sigma \subset \mathcal{V}_\kappa$. Let $\mathcal{M}_{\mathrm{s}}$ denote the set of the cells sharing the vertex s. The set of edges of the mesh (defined only if $d = 3$) is denoted by $\mathcal{E}$ and $\mathcal{E}_\sigma$ denotes the set of edges of the face $\sigma \in \mathcal{F}$, while $\mathcal{E}_\kappa$ denotes the set of the edges of the cell κ. The set $\mathcal{V}_{\mathrm{e}}$ denotes the pair of vertices at the extremities of the edge $\mathrm{e} \in \mathcal{E}$. In the 3-dimensional case, it is assumed that for each face $\sigma \in \mathcal{F}$, there exists a so-called "center" of the face $\boldsymbol{x}_\sigma$ such that

$$\boldsymbol{x}_\sigma = \sum_{\mathrm{s}\in\mathcal{V}_\sigma} \beta_{\sigma,\mathrm{s}} \boldsymbol{x}_{\mathrm{s}}, \qquad \text{with} \qquad \sum_{\mathrm{s}\in\mathcal{V}_\sigma} \beta_{\sigma,\mathrm{s}} = 1, \tag{20}$$

and $\beta_{\sigma,\mathrm{s}} \geq 0$ for all $\mathrm{s} \in \mathcal{V}_\sigma$. The face σ is then assumed to match with the union of the triangles $T_{\sigma,\mathrm{e}}$ defined as by the face center $\boldsymbol{x}_\sigma$ and each of its edge $\mathrm{e} \in \mathcal{E}_\sigma$. A two-dimensional example of mesh $\mathcal{M}$ is drawn on figure 1.

The previous discretization is denoted by $\mathcal{D}$, and we define the discrete space

$$W_{\mathcal{D}} = \left\{ \boldsymbol{v} = (v_\kappa, v_{\mathrm{s}})_{\kappa\in\mathcal{M},\mathrm{s}\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{M}+\#\mathcal{V}} \right\}.$$

In the 3-dimensional case, we introduce for all $\sigma \in \mathcal{F}$ the operator $I_\sigma : W_{\mathcal{D}} \to \mathbb{R}$ defined by

$$I_\sigma(\boldsymbol{v}) = \sum_{\mathrm{s}\in\mathcal{V}_\sigma} \beta_{\sigma,\mathrm{s}} v_{\mathrm{s}}, \qquad \forall \boldsymbol{v} \in W_{\mathcal{D}},$$

yielding a second order interpolation at $\boldsymbol{x}_\sigma$ thanks to the definition (20) of $\boldsymbol{x}_\sigma$.

We introduce the simplicial submesh $\mathcal{T}$ (a two-dimensional illustration is provided on Figure 2) defined by

- $\mathcal{T} = \{T_{\kappa,\sigma}, \kappa \in \mathcal{M}, \sigma \in \mathcal{F}_\kappa\}$ in the two-dimensional case, where $T_{\kappa,\sigma}$ denotes the triangle whose vertices are $\boldsymbol{x}_\kappa$ and $\boldsymbol{x}_{\mathrm{s}}$ for $\mathrm{s} \in \mathcal{V}_\sigma$;
- $\mathcal{T} = \{T_{\kappa,\sigma,\mathrm{e}}, \kappa \in \mathcal{M}, \sigma \in \mathcal{F}_{\mathrm{s}}, \mathrm{e} \in \mathcal{E}_\sigma\}$ in the three-dimensional case, where $T_{\kappa,\sigma,\mathrm{e}}$ denotes the tetrahedron whose vertices are $\boldsymbol{x}_\kappa$, $\boldsymbol{x}_\sigma$ and $\boldsymbol{x}_{\mathrm{s}}$ for $\mathrm{s} \in \mathcal{V}_{\mathrm{e}}$.

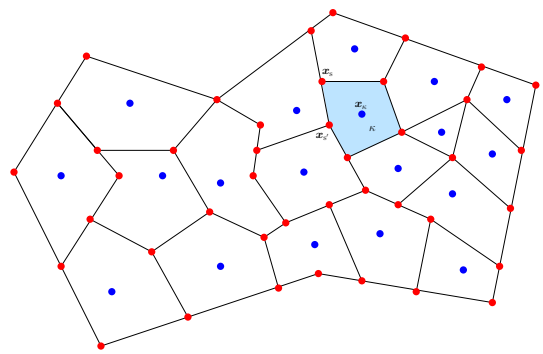

FIGURE 1. The primal mesh $\mathcal{M}$ can be made of cells with various and general shapes. Degrees of freedom are located at the so-called cell center (blue dots) and at the so-called vertices (red dots).

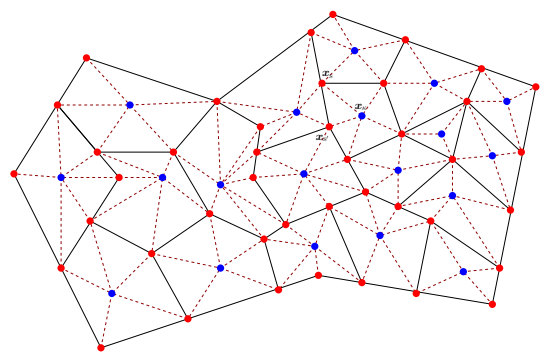

FIGURE 2. The simplicial submesh $\mathcal{T}$ is derived from the primal mesh $\mathcal{M}$ by decomposing the primal cells $\kappa \in \mathcal{M}$ into triangles if $d = 2$ or tetrahedra if $d = 3$. This construction is possible since κ was assumed to be star-shaped with respect to a ball centered in $\boldsymbol{x}_\kappa$.

We define the regularity $\theta_{\mathcal{T}}$ of the simplicial mesh $\mathcal{T}$ by

$$\theta_{\mathcal{T}} = \max_{T \in \mathcal{T}} \frac{h_T}{\rho_T}, \tag{21}$$

where h_T and ρ_T respectively denote the diameter of T and the insphere diameter of T. We denote by

$$h_{\mathcal{T}} = \max_{T \in \mathcal{T}} h_T \tag{22}$$

the maximum diameter of the simplicial mesh. We also define the quantities ℓ_κ and ℓ_{s} quantifying the number of vertices of the cell κ and the number of neighboring cells for the vertex s respectively:

$$\ell_\kappa = \#\mathcal{V}_\kappa, \quad \ell_{\mathrm{s}} = \#\mathcal{M}_{\mathrm{s}}, \quad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}. \tag{23}$$

This allows to introduce the quantity

$$\ell_{\mathcal{D}} = \max\left(\max_{\beta\in\mathcal{M}\cup\mathcal{V}} \ell_\beta, \max_{\kappa\in\mathcal{M}}\{\#\mathcal{F}_\kappa\}\right), \tag{24}$$

controlling the regularity of the general discretization $\mathcal{D}$ of Ω.

Denoting by $H_{\mathcal{T}} \subset W^{1,\infty}(\Omega)$ the usual $\mathbb{P}_1$-finite element space on the simplicial mesh $\mathcal{T}$, we define the reconstruction operator $\pi_{\mathcal{T}} : W_{\mathcal{D}} \to H_{\mathcal{T}}$ by setting, for all $\boldsymbol{v} \in W_{\mathcal{D}}$ and all $(\mathrm{s}, \kappa, \sigma) \in \mathcal{V} \times \mathcal{M} \times \mathcal{F}$,

$$\pi_{\mathcal{T}}\boldsymbol{v}(\boldsymbol{x}_{\mathrm{s}}) = v_{\mathrm{s}}, \quad \pi_{\mathcal{T}}\boldsymbol{v}(\boldsymbol{x}_\kappa) = v_\kappa, \quad \text{and} \quad \pi_{\mathcal{T}}\boldsymbol{v}(\boldsymbol{x}_\sigma) = I_\sigma(\boldsymbol{v}). \tag{25}$$

This allows to define the operator $\boldsymbol{\nabla}_{\mathcal{T}} : W_{\mathcal{D}} \to (L^\infty(\Omega))^d$ by

$$\boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{v} = \boldsymbol{\nabla}\pi_{\mathcal{T}}\boldsymbol{v}, \qquad \forall \boldsymbol{v} \in W_{\mathcal{D}}. \tag{26}$$

We aim now to reconstruct piecewise constant functions. To this end, we introduce a so-called *mass lumping* mesh $\mathcal{D}$ depending on additional parameters that appear to play an important role in practical applications [54]. Let $\kappa \in \mathcal{M}$, then introduce some weights $(\alpha_{\kappa,\mathrm{s}})_{\mathrm{s}\in\mathcal{V}_\kappa}$ such that

$$\alpha_{\kappa,\mathrm{s}} \geq 0, \quad \text{and} \quad \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \alpha_{\kappa,\mathrm{s}} \leq 1, \quad \forall \mathrm{s} \in \mathcal{V}_\kappa. \tag{27}$$

Denoting by $\mathrm{meas}(\kappa) = \int_\kappa \mathrm{d}\boldsymbol{x}$ the volume of κ, then we define the quantities

$$\begin{cases} m_{\kappa,\mathrm{s}} = \alpha_{\kappa,\mathrm{s}}\mathrm{meas}(\kappa), & \forall \kappa \in \mathcal{M}, \forall \mathrm{s} \in \mathcal{V}_\kappa, \\ m_{\mathrm{s}} = \sum_{\kappa\in\mathcal{M}_{\mathrm{s}}} m_{\kappa,\mathrm{s}}, & \forall \mathrm{s} \in \mathcal{V}, \\ m_\kappa = \mathrm{meas}(\kappa) - \sum_{\mathrm{s}\in\mathcal{V}_\kappa} m_{\kappa,\mathrm{s}}, & \forall \kappa \in \mathcal{M}, \end{cases} \tag{28}$$

so that one has

$$\sum_{\beta\in\mathcal{M}\cup\mathcal{V}} m_\beta = \mathrm{meas}(\Omega).$$

For all $\kappa \in \mathcal{M}$, we denote by ω_κ and $\omega_{\kappa,\mathrm{s}}$ some disjointed open subsets of κ, such that

$$\overline{\omega}_\kappa \cup \left(\bigcup_{\mathrm{s}\in\mathcal{V}_\kappa} \overline{\omega}_{\kappa,\mathrm{s}}\right) = \overline{\kappa},$$

and such that

$$\mathrm{meas}(\omega_\kappa) = m_\kappa \quad \text{and} \quad \mathrm{meas}(\omega_{\kappa,\mathrm{s}}) = m_{\kappa,\mathrm{s}}, \qquad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}_\kappa.$$

Note that such a decomposition always exists thanks to (27)–(28). Then we denote by

$$\omega_{\mathrm{s}} = \bigcup_{\kappa\in\mathcal{M}_{\mathrm{s}}} \omega_{\kappa,\mathrm{s}}, \qquad \forall \mathrm{s} \in \mathcal{V}.$$

The mass-lumping mesh $\mathcal{D}$ is made of the cells ω_κ and ω_{s} for $\kappa \in \mathcal{M}$ and $\mathrm{s} \in \mathcal{V}$. An illustration is presented in Figure 3. In [54], a study focused on the repartition of the porous volume between nodes and centers is proposed, in the case of a coupled problem (transport of a species in the flow of a fluid in a porous medium). The influence of this repartition mainly concerns the transport part, and not the fluid flow. In the framework of the present paper, we numerically observe that this repartition has not a strong influence in the cases studied. However, since the function η can vanish, the limit choice $m_\kappa = 0$ (resp. $m_{\mathrm{s}} = 0$) can lead to singular cases, and therefore must be prevented.

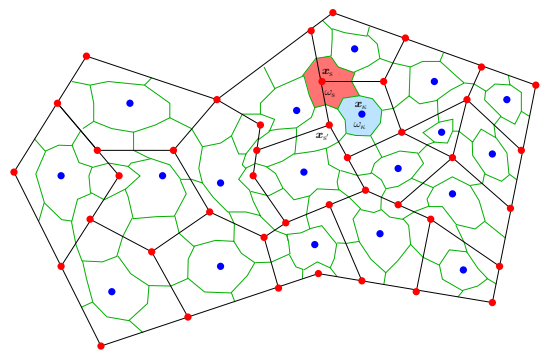

FIGURE 3. The mass-lumping mesh $\mathcal{D}$ contains one cell ω_κ and ω_{s} by degree of fredom. The mass repartition depends on the factors $\alpha_{\kappa,\mathrm{s}}$ introduced at (27).

In what follows, we denote by

$$\zeta_{\mathcal{D}} = \min_{\beta\in\mathcal{M}\cup\mathcal{V}} \frac{m_\beta}{\int_\Omega \pi_{\mathcal{T}}\mathbf{e}_\beta \mathrm{d}\boldsymbol{x}}, \tag{29}$$

where $\mathbf{e}_\beta$, $\beta \in \mathcal{M}\cup\mathcal{V}$ is the unique element of $W_{\mathcal{D}}$ such that

$$\pi_{\mathcal{T}}\mathbf{e}_\beta(\boldsymbol{x}_\gamma) = \delta_\beta^\gamma, \qquad \forall \gamma \in \mathcal{M}\cup\mathcal{V},$$

and δ_β^γ is the kronecker symbol.

We can now define the piecewise constant reconstruction operators $\pi_{\mathcal{D}} : W_{\mathcal{D}} \to L^\infty \cap \mathrm{BV}(\Omega)$ and $\pi_{\mathcal{M}} : W_{\mathcal{D}} \to L^\infty \cap \mathrm{BV}(\Omega)$ by

$$\pi_{\mathcal{D}}(\boldsymbol{v})(\boldsymbol{x}) = \sum_{\kappa\in\mathcal{M}} v_\kappa \mathbf{1}_{\omega_\kappa}(\boldsymbol{x}) + \sum_{\mathrm{s}\in\mathcal{V}} v_{\mathrm{s}} \mathbf{1}_{\omega_{\mathrm{s}}}(\boldsymbol{x}), \qquad \forall \boldsymbol{x}\in\Omega, \quad \forall \boldsymbol{v}\in W_{\mathcal{D}}, \tag{30}$$

and

$$\pi_{\mathcal{M}}(\boldsymbol{v})(\boldsymbol{x}) = \sum_{\kappa\in\mathcal{M}} v_\kappa \mathbf{1}_\kappa(\boldsymbol{x}) \qquad \forall \boldsymbol{x}\in\Omega, \quad \forall \boldsymbol{v}\in W_{\mathcal{D}}. \tag{31}$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a possibly nonlinear function, then denote by

$$f(\boldsymbol{v}) = \left(f(v_\kappa), f(v_{\mathrm{s}})\right)_{\kappa\in\mathcal{M},\mathrm{s}\in\mathcal{V}}, \qquad \forall \boldsymbol{v} = \left(v_\kappa, v_{\mathrm{s}}\right)_{\kappa\in\mathcal{M},\mathrm{s}\in\mathcal{V}} \in W_{\mathcal{D}}.$$

Notice that in general,

$$\pi_{\mathcal{T}}(f(\boldsymbol{v})) \neq f(\pi_{\mathcal{T}}(\boldsymbol{v})) \quad \text{and} \quad \boldsymbol{\nabla}_{\mathcal{T}}(f(\boldsymbol{v})) \neq \boldsymbol{\nabla} f(\pi_{\mathcal{T}}(\boldsymbol{v})),$$

but that

$$\pi_{\mathcal{D}}(f(\boldsymbol{v})) = f(\pi_{\mathcal{D}}(\boldsymbol{v})) \quad \text{and} \quad \pi_{\mathcal{M}}(f(\boldsymbol{v})) = f(\pi_{\mathcal{M}}(\boldsymbol{v})), \qquad \forall \boldsymbol{v}\in W_{\mathcal{D}}. \tag{32}$$

2.1.2. *Time-and-space discretizations and discrete functions.* Let $N \geq 1$ and let $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = t_{\mathrm{f}}$ be some subdivision of $[0, t_{\mathrm{f}}]$. We denote by $\Delta t_n = t_n - t_{n-1}$ for all $n \in \{1, \ldots N\}$, by $\boldsymbol{\Delta} t = (\Delta t_1, \ldots, \Delta t_N)^T \in \mathbb{R}^N$, and by

$$\overline{\Delta t} = \max_{1\leq n\leq N} \Delta t_n. \tag{33}$$

The time and space discrete space is then defined by

$$W_{\mathcal{D},\boldsymbol{\Delta} t} = \left\{ \boldsymbol{v} = \left(v_\kappa^n, v_{\mathrm{s}}^n\right)_{\kappa\in\mathcal{M},\mathrm{s}\in\mathcal{V},1\leq n\leq N} \in \mathbb{R}^{N(\#\mathcal{M}+\#\mathcal{V})} \right\}.$$

For $\boldsymbol{v} \in W_{\mathcal{D},\boldsymbol{\Delta} t}$ and $n \in \{1, \dots, N\}$, we denote by

$$\boldsymbol{v}^n = (v_\kappa^n, v_{\mathrm{s}}^n)_{\kappa \in \mathcal{M}, \mathrm{s} \in \mathcal{V}} \in W_{\mathcal{D}}.$$

We deduce from the space reconstruction operators $\pi_{\mathcal{D}}$, $\pi_{\mathcal{M}}$ and $\pi_{\mathcal{T}}$ some time and space reconstructions operators $\pi_{\mathcal{D},\boldsymbol{\Delta} t}, \pi_{\mathcal{M},\boldsymbol{\Delta} t}, \pi_{\mathcal{T},\boldsymbol{\Delta} t} : W_{\mathcal{D},\boldsymbol{\Delta} t} \to L^\infty(Q_{t_{\mathrm{f}}})$ mapping the elements of $W_{\mathcal{D},\boldsymbol{\Delta} t}$ into constant w.r.t. time functions defined by

$$\pi_{\mathcal{D},\boldsymbol{\Delta} t} \boldsymbol{v}(\cdot, t) = \pi_{\mathcal{D}}(\boldsymbol{v}^n), \quad \pi_{\mathcal{M},\boldsymbol{\Delta} t} \boldsymbol{v}(\cdot, t) = \pi_{\mathcal{M}}(\boldsymbol{v}^n) \quad \text{and} \quad \pi_{\mathcal{T},\boldsymbol{\Delta} t} \boldsymbol{v}(\cdot, t) = \pi_{\mathcal{T}}(\boldsymbol{v}^n)$$

if $t \in (t_{n-1}, t_n]$. The gradient reconstruction operator $\boldsymbol{\nabla}_{\mathcal{T},\boldsymbol{\Delta} t} : W_{\mathcal{D},\boldsymbol{\Delta} t} \to (L^\infty(Q_{t_{\mathrm{f}}}))^d$ is then defined by

$$\boldsymbol{\nabla}_{\mathcal{T},\boldsymbol{\Delta} t} \boldsymbol{v} = \boldsymbol{\nabla} \pi_{\mathcal{T},\boldsymbol{\Delta} t} \boldsymbol{v}, \qquad \forall \boldsymbol{v} \in W_{\mathcal{D},\boldsymbol{\Delta} t}.$$

2.2. **The nonlinear scheme for degenerate parabolic equations.** For $\kappa \in \mathcal{M}$, we denote by $\mathbf{A}_\kappa = \left(a_{\mathrm{s},\mathrm{s}'}^\kappa\right)_{\mathrm{s},\mathrm{s}' \in \mathcal{V}_\kappa} \in \mathbb{R}^{\ell_\kappa \times \ell_\kappa}$ the symmetric positive definite matrix whose coefficients are defined by

$$a_{\mathrm{s},\mathrm{s}'}^\kappa = \int_\kappa \boldsymbol{\Lambda}(\boldsymbol{x}) \boldsymbol{\nabla}_{\mathcal{T}} \mathbf{e}_{\mathrm{s}}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\mathcal{T}} \mathbf{e}_{\mathrm{s}'}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = a_{\mathrm{s}',\mathrm{s}}^\kappa. \tag{34}$$

It results from the relation

$$\pi_{\mathcal{T}} \mathbf{e}_\kappa(\boldsymbol{x}) + \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \pi_{\mathcal{T}} \mathbf{e}_{\mathrm{s}}(\boldsymbol{x}) = 1, \qquad \forall \boldsymbol{x} \in \kappa, \ \forall \kappa \in \mathcal{M},$$

that, for all $\boldsymbol{u}, \boldsymbol{v} \in W_{\mathcal{D}}$ and all $\kappa \in \mathcal{M}$, one has

$$\int_\kappa \boldsymbol{\Lambda}(\boldsymbol{x}) \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \sum_{\mathrm{s}' \in \mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa (u_\kappa - u_{\mathrm{s}})(v_\kappa - v_{\mathrm{s}'}). \tag{35}$$

For $\kappa \in \mathcal{M}$, we denote by $\boldsymbol{\delta}_\kappa : W_{\mathcal{D}} \to \mathbb{R}^{\ell_\kappa}$ the linear operator defined by

$$(\boldsymbol{\delta}_\kappa \boldsymbol{v})_{\mathrm{s}} = v_\kappa - v_{\mathrm{s}}, \qquad \forall \mathrm{s} \in \mathcal{V}_\kappa, \ \forall \boldsymbol{v} \in W_{\mathcal{D}}.$$

With this notation, we obtain that (35) rewrites as

$$\int_\kappa \boldsymbol{\Lambda}(\boldsymbol{x}) \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \boldsymbol{\delta}_\kappa \boldsymbol{v} \cdot \mathbf{A}_\kappa \boldsymbol{\delta}_\kappa \boldsymbol{u}, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in W_{\mathcal{D}}, \ \forall \kappa \in \mathcal{M}.$$

In order to deal with the nonlinearities of the problem, we introduce the sets $W_{\mathcal{D}}^{\mathrm{ad}} \subset W_{\mathcal{D}}$ and $W_{\mathcal{D},\boldsymbol{\Delta} t}^{\mathrm{ad}} \subset W_{\mathcal{D},\boldsymbol{\Delta} t}$ of the admissible states defined by

$$\boldsymbol{v} \in W_{\mathcal{D}}^{\mathrm{ad}} \quad \text{iff} \qquad v_\nu \in \mathcal{I}_p, \quad \forall \nu \in \mathcal{M} \cup \mathcal{V},$$

and

$$\boldsymbol{v} \in W_{\mathcal{D},\boldsymbol{\Delta} t}^{\mathrm{ad}} \quad \text{iff} \qquad \boldsymbol{v}^n \in W_{\mathcal{D}}^{\mathrm{ad}}, \quad \forall n \in \{1, \dots, N\},$$

while we denote by $W_{\mathcal{D}}^{\mathrm{en}} \subset W_{\mathcal{D}}$ the set of finite entropy vectors:

$$\boldsymbol{v} \in W_{\mathcal{D}}^{\mathrm{en}} \quad \text{iff} \quad \mathfrak{E}_{\mathcal{D}}(\boldsymbol{v}) := \int_\Omega \left(\Gamma(\pi_{\mathcal{D}} \boldsymbol{v}) + \pi_{\mathcal{D}}(\boldsymbol{v}) \pi_{\mathcal{D}}(\mathbf{V})\right) \mathrm{d}\boldsymbol{x} < \infty, \tag{36}$$

where $\mathbf{V} = (V_\kappa, V_{\mathrm{s}})_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}$ is defined by

$$V_\kappa = V(\boldsymbol{x}_\kappa) \quad V_{\mathrm{s}} = V(\boldsymbol{x}_{\mathrm{s}}) \qquad \forall \kappa \in \mathcal{M}, \forall \mathrm{s} \in \mathcal{V}. \tag{37}$$

It is easy to check that, thanks to assumptions ($\mathbf{A}2$) and to the definition (4) of the convex function Γ, one has $W_{\mathcal{D}}^{\mathrm{ad}} \subset W_{\mathcal{D}}^{\mathrm{en}}$.

Given $\boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{ad}}$ and $\mathbf{V} \in W_{\mathcal{D}}$, we define the discrete hydrostatic pressure $\boldsymbol{h}(\boldsymbol{u}) = (\mathfrak{h}_\kappa(u_\kappa), \mathfrak{h}_{\mathrm{s}}(u_s))_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}$ by

$$\mathfrak{h}_\kappa(u_\kappa) = p(u_\kappa) + V_\kappa, \qquad \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}) = p(u_{\mathrm{s}}) + V_{\mathrm{s}}, \qquad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}. \tag{38}$$

The initial data u_0 is discretized into an element $\boldsymbol{u}^0 \in W_{\mathcal{D}}^{\mathrm{en}}$ by

$$u_\beta^0 = \frac{1}{m_\beta} \int_{\omega_\beta} u_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad \forall \beta \in \mathcal{M} \cup \mathcal{V}, \tag{39}$$

so that, thanks to (5),

$$\int_\Omega \pi_{\mathcal{D}}(\boldsymbol{u}^0) \mathrm{d}\boldsymbol{x} = \int_\Omega u_0 \mathrm{d}\boldsymbol{x} > 0. \tag{40}$$

Let us state a first lemma that ensure that the discretized initial data has a finite discrete entropy.

Lemma 2.1. *Let $u_0 \in L^1(\Omega)$ be such that (**A**4) holds, V be such that (**A**5) holds, $\boldsymbol{u}^0$ be defined by (39), and $\mathbf{V}$ defined by (37). Then there exists C depending only on $\|u_0\|_{L^1(\Omega)}$ and $\|\boldsymbol{\nabla} V\|_{L^\infty(\Omega)^d}$ such that*

$$\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^0) \le \mathfrak{E}(u_0) + C h_{\mathcal{T}} \le \mathfrak{E}(u_0) + C \mathrm{diam}(\Omega), \tag{41}$$

where the entropy functional $\mathfrak{E}$ is defined by (14) and its discrete counterpart $\mathfrak{E}_{\mathcal{D}}$ is defined by (36). In particular, $\boldsymbol{u}^0$ belongs to $W_{\mathcal{D}}^{\mathrm{en}}$.

Proof. We deduce from Jensen inequality that

$$\Gamma(u_\beta^0) \le \frac{1}{m_\beta} \int_{\omega_\beta} \Gamma(u_0) \mathrm{d}\boldsymbol{x},$$

whence, thanks to (**A**4) and to the definition (36) of the discrete entropy functional $\mathfrak{E}_{\mathcal{D}}$, one has

$$\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^0) \le \mathfrak{E}(u_0) + \int_\Omega u_0(\pi_{\mathcal{D}} \mathbf{V} - V) \mathrm{d}\boldsymbol{x} + \int_\Omega (\pi_{\mathcal{D}} \boldsymbol{u}^0 - u_0) \pi_{\mathcal{D}} \mathbf{V} \mathrm{d}\boldsymbol{x}.$$

The last term in the above inequality is equal to zero thanks to (40). The Lipschitz regularity of V yields

$$\|\pi_{\mathcal{D}} \mathbf{V} - V\|_{L^\infty(\Omega)} \le \|\boldsymbol{\nabla} V\|_{L^\infty(\Omega)^d} h_{\mathcal{T}},$$

so that

$$\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^0) \le \mathfrak{E}(u_0) + \|u_0\|_{L^1(\Omega)} \|\boldsymbol{\nabla} V\|_{L^\infty(\Omega)^d} h_{\mathcal{T}}.$$

□

With all this setting, we can present the scheme we will analyze in this contribution. For $\boldsymbol{u} \in W_{\mathcal{D},\boldsymbol{\Delta t}}^{\mathrm{ad}}$, we introduce the notation

$$\eta_{\kappa,\mathrm{s}}^n = \frac{\eta(u_\kappa^n) + \eta(u_{\mathrm{s}}^n)}{2}, \qquad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}_\kappa, \forall n \in \{1, \dots, N\}. \tag{42}$$

Given $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}^{\mathrm{en}}$, the vector $\boldsymbol{u}^n \in W_{\mathcal{D}}^{\mathrm{ad}}$ is obtained by solving the following nonlinear system:

$$m_\kappa \frac{u_\kappa^n - u_\kappa^{n-1}}{\Delta t_n} + \sum_{\mathrm{s} \in \mathcal{V}_\kappa} F_{\kappa,\mathrm{s}}(\boldsymbol{u}^n) = 0, \qquad \forall \kappa \in \mathcal{M}, \tag{43a}$$

$$m_{\mathrm{s}} \frac{u_{\mathrm{s}}^n - u_{\mathrm{s}}^{n-1}}{\Delta t_n} + \sum_{\kappa \in \mathcal{M}_{\mathrm{s}}} F_{\mathrm{s},\kappa}(\boldsymbol{u}^n) = 0, \qquad \forall \mathrm{s} \in \mathcal{V}, \tag{43b}$$

$$F_{\kappa,\mathrm{s}}(\boldsymbol{u}^n) = \sqrt{\eta_{\kappa,\mathrm{s}}^n} \sum_{\mathrm{s}' \in \mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa \sqrt{\eta_{\kappa,\mathrm{s}'}^n} \left(\mathfrak{h}_\kappa(u_\kappa^n) - \mathfrak{h}_{\mathrm{s}'}(u_{\mathrm{s}'}^n) \right), \quad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}_\kappa, \tag{43c}$$

$$F_{\kappa,\mathrm{s}}(\boldsymbol{u}^n) + F_{\mathrm{s},\kappa}(\boldsymbol{u}^n) = 0, \qquad \forall \kappa \in \mathcal{M},\ \forall \mathrm{s} \in \mathcal{V}_\kappa. \tag{43d}$$

The scheme (43) can be interpreted as a finite volume scheme, the conservation being established on the cells ω_κ and ω_{s} for $\kappa \in \mathcal{M}$ and $\mathrm{s} \in \mathcal{V}$. As a direct consequence of the conservativity of the scheme, one has

$$\int_\Omega \pi_{\mathcal{D}}(\boldsymbol{u}^n) \mathrm{d}\boldsymbol{x} = \int_\Omega \pi_{\mathcal{D}}(\boldsymbol{u}^{n-1}) \mathrm{d}\boldsymbol{x}. \tag{44}$$

However, contrarily to usual finite volume schemes, the fluxes $F_{\kappa,\mathrm{s}}^n$ are not issued from the computation of $\int_\sigma \eta(u)\Lambda \boldsymbol{\nabla}(p(u)+V)\cdot \boldsymbol{n}_\sigma$ on a specific boundary σ between identified control volumes. They result from the variational formulation of the scheme and are viewed as fluxes between control volumes located at nodes and centers.

Defining, for all $\kappa \in \mathcal{M}$ and $\boldsymbol{u} = (u_\kappa, u_{\mathrm{s}})_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}$, the diagonal matrix $\mathbf{M}_\kappa(\boldsymbol{u}) \in \mathbb{R}^{\ell_\kappa \times \ell_\kappa}$ by

$$\left(\mathbf{M}_\kappa(\boldsymbol{u})\right)_{\mathrm{s},\mathrm{s}'} = \begin{cases} \sqrt{\frac{\eta(u_\kappa)+\eta(u_{\mathrm{s}})}{2}} & \text{if } \mathrm{s} = \mathrm{s}', \\ 0 & \text{otherwise,} \end{cases}$$

the systems (43) is equivalent to the following compact formulation: $\forall \boldsymbol{v} \in W_{\mathcal{D}}$,

$$\int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^n \pi_{\mathcal{D}} \boldsymbol{v} \mathrm{d}\boldsymbol{x} + \Delta t_n \sum_{\kappa \in \mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n) \cdot \mathbf{B}_\kappa(\boldsymbol{u}^n) \boldsymbol{\delta}_\kappa \boldsymbol{v} = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^{n-1} \pi_{\mathcal{D}} \boldsymbol{v} \mathrm{d}\boldsymbol{x}, \tag{45}$$

where

$$\mathbf{B}_\kappa(\boldsymbol{u}) := \mathbf{M}_\kappa(\boldsymbol{u}) \mathbf{A}_\kappa \mathbf{M}_\kappa(\boldsymbol{u}), \qquad \forall \kappa \in \mathcal{M},\ \forall \boldsymbol{u} \in W_{\mathcal{D}}, \tag{46}$$

is a symmetric semi-positive matrix since $\mathbf{A}_\kappa$ and $\mathbf{M}_\kappa(\boldsymbol{u})$ have this property.

2.3. **Gradient flow interpretation for the scheme.** The goal of this section is to transpose the formal variational structure pointed out in §1.2 to the discrete setting. A natural discretization of the manifold $\mathfrak{M}$ consists in

$$\mathfrak{M}_{\mathcal{D}} = \left\{ \boldsymbol{u} \in W_{\mathcal{D}} \;\middle|\; \int_\Omega \pi_{\mathcal{D}} \boldsymbol{u} \mathrm{d}\boldsymbol{x} = \int_\Omega u_0 \mathrm{d}\boldsymbol{x} \right\}, \tag{47}$$

leading to

$$T_{\boldsymbol{u}} \mathfrak{M}_{\mathcal{D}} = \left\{ \boldsymbol{v} \in W_{\mathcal{D}} \;\middle|\; \int_\Omega \pi_{\mathcal{D}} \boldsymbol{v} \mathrm{d}\boldsymbol{x} = 0 \right\}. \tag{48}$$

In order to define the discrete counterpart $\mathfrak{g}_{\mathcal{D},\boldsymbol{u}}$ of the metric tensor $\mathfrak{g}_u$ defined by (12)–(13), one needs a discrete counterpart of

- the classical $L^2(\Omega)$ scalar product: we will use

$$(\boldsymbol{w}_1, \boldsymbol{w}_2) \mapsto \int_\Omega \pi_{\mathcal{D}} \boldsymbol{w}_1 \, \pi_{\mathcal{D}} \boldsymbol{w}_2 \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in W_{\mathcal{D}};$$

- the weighted $H^1(\Omega)$ "scalar product" with weight $\eta(u)$: we use

$$(\boldsymbol{w}_1, \boldsymbol{w}_2) \mapsto \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{w}_1 \cdot \mathbf{B}_\kappa(\boldsymbol{u}) \boldsymbol{\delta}_\kappa \boldsymbol{w}_2, \qquad \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in W_{\mathcal{D}}.$$

This allows to define the discrete metric tensor $\mathfrak{g}_{\mathcal{D},\boldsymbol{u}}$ by: $\forall \boldsymbol{u} \in W_{\mathcal{D}}$, $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in T_{\boldsymbol{u}}\mathfrak{M}_{\mathcal{D}}$,

(49)
$$\mathfrak{g}_{\mathcal{D},\boldsymbol{u}}(\boldsymbol{w}_1, \boldsymbol{w}_2) = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{w}_1 \pi_{\mathcal{D}} \boldsymbol{\phi}_2 \mathrm{d}\boldsymbol{x} = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{\phi}_1 \pi_{\mathcal{D}} \boldsymbol{v}_2 \mathrm{d}\boldsymbol{x} = \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{\phi}_1 \cdot \mathbf{B}_\kappa(\boldsymbol{u}) \boldsymbol{\delta}_\kappa \boldsymbol{\phi}_2,$$

where $\boldsymbol{\phi}_i \in W_{\mathcal{D}}$ solves the discrete counterpart of (13), that is

$$\sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{\phi}_i \cdot \mathbf{B}_\kappa(\boldsymbol{u}) \boldsymbol{\delta}_\kappa \boldsymbol{\psi} = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{w}_i \, \pi_{\mathcal{D}} \boldsymbol{\psi} \, \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{\psi} \in W_{\mathcal{D}}. \tag{50}$$

In this setting, we can define the semi-discrete in space gradient flow by

$$\begin{aligned} \mathfrak{g}_{\mathcal{D},\boldsymbol{u}}(\partial_t \boldsymbol{u}, \boldsymbol{w}) &= \int_\Omega \pi_{\mathcal{D}}(\partial_t \boldsymbol{u}) \, \pi_{\mathcal{D}} \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \\ &= \mathfrak{g}_{\mathcal{D},\boldsymbol{u}}(-\boldsymbol{\nabla}_u \mathfrak{E}(\boldsymbol{u}), \boldsymbol{w}) = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{h}(\boldsymbol{u}) \, \pi_{\mathcal{D}} \boldsymbol{w} \, \mathrm{d}\boldsymbol{x} = \sum_\kappa \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}) \cdot \mathbf{B}_\kappa(\boldsymbol{u}) \boldsymbol{\delta}_\kappa \boldsymbol{v}, \end{aligned} \tag{51}$$

where $\boldsymbol{v}$ solve the discrete elliptic problem

$$\sum_{\kappa\in\mathfrak{M}} \boldsymbol{\delta}_\kappa \boldsymbol{v} \cdot \mathbf{B}_\kappa(\boldsymbol{u}) \boldsymbol{\delta}_\kappa \boldsymbol{\psi} = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{w} \, \pi_{\mathcal{D}} \boldsymbol{\psi} \, \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{\psi} \in W_{\mathcal{D}}.$$

In order to recover (45) from (51), one applies the backward Euler scheme.

Remark 2.2. *In their seminal paper [65], Jordan, Kinderlehrer and Otto proposed to approximate the solution of gradient flows thanks to the minimizing movement scheme*

$$\boldsymbol{u}^n \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathfrak{M}_{\mathcal{D}}} \Big\{ \frac{\mathfrak{d}(\boldsymbol{u}, \boldsymbol{u}^{n-1})}{2\Delta t} + \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) \Big\}, \tag{52}$$

where $\mathfrak{d}$ denotes the distance on $\mathfrak{M}_{\mathcal{D}}$ induced by the metric tensor field $\mathfrak{g}_{\mathcal{D}}$. Several practical and theoretical difficulties arise when one aims at using (52). First of all, the Riemannian structure is formal, even in the continuous case. It is unclear if one can define rigorously a distance $\mathfrak{d}$ if $\eta(0) = 0$ even if η is concave (cf. [43, 81]). But even if $\mathfrak{d}$ is a distance, yielding a metric structure for $\mathfrak{M}_{\mathcal{D}}$, computing this distance is a complex problem we avoid by using an backward Euler scheme rather than (52).

2.4. **Main results.** The first result we want to point out concerns the scheme for a fixed mesh. The following theorem states that the scheme (43) admits at least one solution, and justifies the *free energy diminishing* denomination for the scheme.

Theorem 2.3. *Let $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}^{\mathrm{en}}$, then there exists (at least) one vector $\boldsymbol{u}^n \in W_{\mathcal{D}}^{\mathrm{ad}}$ solution to the system (43), and the following dissipation property holds:*

$$\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^n) + \Delta t_n \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n) \cdot \mathbf{B}_\kappa(\boldsymbol{u}^n) \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n) \le \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^{n-1}), \tag{53}$$

where $\mathfrak{E}_{\mathcal{D}}$ is defined by (36) and $\boldsymbol{h}(\boldsymbol{u}^n) = (\mathfrak{h}_\kappa(u_\kappa^n), \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}^n))_{\kappa,\mathrm{s}}$ is defined by (38).

Since $\boldsymbol{u}^0 \in W_{\mathcal{D}}^{\mathrm{en}}$ and since $W_{\mathcal{D}}^{\mathrm{ad}} \subset W_{\mathcal{D}}^{\mathrm{en}}$, Theorem 2.3 allows to define the *iterated solution* $\boldsymbol{u} = (\boldsymbol{u}^n)_{1\leq n\leq N} \in W_{\mathcal{D},\boldsymbol{\Delta} t}^{\mathrm{ad}}$ to the scheme (43).

The proof of Theorem 2.3 is contained in §3, together with some supplementary material that allows to carry out the convergence analysis when the discretization steps tend to 0. More precisely, we consider a sequence $(\mathcal{D}_m)_{m\geq 1} = (\mathcal{M}_m, \mathcal{T}_m)_{m\geq 1}$ of discretizations of Ω as introduced in §2.1.1, such that

$$h_{\mathcal{T}_m} = \max_{T\in\mathcal{T}_m} h_T \underset{m\to+\infty}{\longrightarrow} 0, \tag{54a}$$

and such that there exists $\theta^\star > 0$ and $\ell^\star > 0$ satisfying

$$\sup_{m\geq 1} \theta_{\mathcal{T}_m} \leq \theta^\star, \qquad \sup_{m\geq 1} \ell_{\mathcal{D}_m} \leq \ell^\star, \tag{54b}$$

where $\theta_{\mathcal{T}_m}$ and $\ell_{\mathcal{D}_m}$ are defined by (21) and (24) respectively.

Even though it can be avoided in some specific situations, we also do the following assumption, allowing to circumvent some technical difficulties:

$$\inf_{m\geq 1} \zeta_{\mathcal{D}_m} = \zeta^\star > 0. \tag{54c}$$

This means that there is a minimum ratio of volume allocated to the cell centers and to the nodes in the mass lumping procedure.

Concerning the time-discretizations, we consider a sequence $(\boldsymbol{\Delta} t_m)_{m\geq 1}$ of discretizations of $(0, t_{\mathrm{f}})$ as prescribed in §2.1.2 :

$$\boldsymbol{\Delta} t_m = (\Delta t_{1,m}, \ldots, \Delta t_{N_m,m}), \qquad \forall m \geq 1.$$

We assume that the time discretization step tends to 0, i.e.,

$$\overline{\Delta t}_m = \max_{1\leq n\leq N_m} \Delta t_{n,m} \underset{m\to+\infty}{\longrightarrow} 0. \tag{54d}$$

Theorem 2.4. *Let $(\mathcal{D}_m, \boldsymbol{\Delta} t_m)_m$ be a sequence of discretizations of $Q_{t_{\mathrm{f}}}$ satisfying Assumptions (54), and let $(\boldsymbol{u}_m)_{m\geq 1}$, with $\boldsymbol{u}_m \in W_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}^{\mathrm{ad}}$, be a corresponding sequence of iterated discrete solutions, then*

$$\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m \underset{m\to+\infty}{\longrightarrow} u \quad \textit{strongly in } L^1(Q_{t_{\mathrm{f}}}),$$

where u is the unique weak solution to (1) in the sense of Definition 1.

Proving the Theorem 2.4 is the purpose of §4. The practical implementation of the scheme (43) is discussed in §5, where we also give evidences of the efficiency of the scheme.

3. Proof of Theorem 2.3 and additional estimates

In order to ease the reading of the paper, several technical lemmas have been postponed to Appendix.

3.1. One-step *A priori* estimates.

Lemma 3.1. *Let $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}^{\mathrm{en}}$, and let $\boldsymbol{u}^n \in W_{\mathcal{D}}^{\mathrm{ad}}$ be a solution to the scheme (43), then (53) holds.*

Proof. Substituting $\boldsymbol{v}$ by $\boldsymbol{h}(\boldsymbol{u}^n) = (\mathfrak{h}_\kappa(u_\kappa^n), \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}^n))_{\kappa,\mathrm{s}}$ defined by (38) in (45) yields

$$\int_\Omega \left(\pi_{\mathcal{D}}\boldsymbol{u}^n - \pi_{\mathcal{D}}\boldsymbol{u}^{n-1}\right)\pi_{\mathcal{D}}\boldsymbol{h}(\boldsymbol{u}^n)\mathrm{d}\boldsymbol{x} + \Delta t_n \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n)\cdot \mathbf{B}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n) = 0. \tag{55}$$

It follows from the convexity of Γ that

$$\Gamma(a) - \Gamma(b) \le (a-b)\left(p(a) - p(1)\right), \qquad \forall\, a, b \in \mathbb{R} \ \text{ s.t. } \ \Gamma(a), \Gamma(b) < +\infty.$$

Hence, using (44), one has

$$\begin{aligned}
&\int_\Omega \left(\pi_{\mathcal{D}}\boldsymbol{u}^n - \pi_{\mathcal{D}}\boldsymbol{u}^{n-1}\right)\pi_{\mathcal{D}}\boldsymbol{h}(\boldsymbol{u}^n)\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_\Omega \left(\pi_{\mathcal{D}}\boldsymbol{u}^n - \pi_{\mathcal{D}}\boldsymbol{u}^{n-1}\right)\left(p(\pi_{\mathcal{D}}\boldsymbol{u}^n) + \pi_{\mathcal{D}}(\mathbf{V})\right)\mathrm{d}\boldsymbol{x} \\
&\qquad \ge \int_\Omega \left(\Gamma(\pi_{\mathcal{D}}\boldsymbol{u}^n) - \Gamma(\pi_{\mathcal{D}}\boldsymbol{u}^{n-1}) + \pi_{\mathcal{D}}(\boldsymbol{u}^n - \boldsymbol{u}^{n-1})\pi_{\mathcal{D}}\mathbf{V}\right)\mathrm{d}\boldsymbol{x} \\
&\qquad\qquad = \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^n) - \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^{n-1}).
\end{aligned}$$

Using this inequality in (55) provides (53). □

Lemma 3.2. *For all $\epsilon > 0$, there exists $C_\epsilon \in \mathbb{R}$ depending on ϵ and p such that*

$$|u| \le \epsilon\Gamma(u) + C_\epsilon, \qquad \forall u \in W_{\mathcal{D}}^{\mathrm{en}}.$$

Proof. Fix $\epsilon > 0$, then in view of Assumption (**A**2), the intermediate value theorem ensures the existence of $u_\epsilon \ge 1$ such that $p(u_\epsilon) = p(1) + 1/\epsilon$. Then for all $u \in \overline{\mathcal{I}}_p$, one has

$$\Gamma(u) = \int_1^u (p(a) - p(1))\mathrm{d}a = \Gamma(u_\epsilon) + \int_{u_\epsilon}^u (p(a) - p(1))\mathrm{d}a.$$

The function p being increasing, we deduce that

$$\Gamma(u) \ge \Gamma(u_\epsilon) + (p(u_\epsilon) - p(1))|u - u_\epsilon| \ge \Gamma(u_\epsilon) + \frac{1}{\epsilon}\left(|u| - |u_\epsilon|\right), \qquad \forall u \in \overline{\mathcal{I}}_p.$$

Lemma 3.2 follows with $C_\epsilon = |u_\epsilon| - \epsilon\Gamma(u_\epsilon)$. □

Lemma 3.3. *For all $\epsilon > 0$, there exists $C_\epsilon \in \mathbb{R}$ depending on ϵ, η and Γ such that*

$$\eta(u) \le \epsilon\Gamma(u) + C_\epsilon, \qquad \forall u \in \overline{\mathcal{I}}_p.$$

Proof. The function $u \mapsto \frac{\eta(u)}{\Gamma(u)}$ tends to 0 as $|u| \to \infty$ thanks to Assumption (**A**6). Let $\epsilon > 0$, then there exists $r_\epsilon > 0$ such that

$$|u| > r_\epsilon \quad \Longrightarrow \quad 0 \le \eta(u) \le \epsilon\Gamma(u).$$

The function η being continuous and nonnegative according to Assumption (**A**1), we know that

$$0 \le C_\epsilon := \max_{u\in[-r_\epsilon, r_\epsilon]} \eta(u) < +\infty.$$

The result of Lemma 3.3 follows. □

Lemma 3.4. *There exist C_1 and C_2 depending only on p, V and Ω such that*

$$\frac{1}{2}\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) + C_1 \le \int_\Omega \Gamma(\pi_{\mathcal{D}}\boldsymbol{u})\mathrm{d}\boldsymbol{x} \le 2\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) + C_2, \qquad \forall \boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{en}}. \tag{56}$$

In particular, the discrete entropy functional $\mathfrak{E}_{\mathcal{D}}$ is bounded from below uniformly w.r.t. the discretization $\mathcal{D}$.

Proof. Recall that the discrete entropy functional $\mathfrak{E}_{\mathcal{D}}$ is defined by

$$\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) = \int_\Omega \left(\Gamma(\pi_{\mathcal{D}}\boldsymbol{u}) + \pi_{\mathcal{D}}\mathbf{V}\,\pi_{\mathcal{D}}\boldsymbol{u}\right)\mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{en}}.$$

Hence, one has

$$\int_\Omega \Gamma(\pi_{\mathcal{D}}\boldsymbol{u})\mathrm{d}\boldsymbol{x} \leq \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) + \|\pi_{\mathcal{D}}\mathbf{V}\|_{L^\infty(\Omega)}\|\pi_{\mathcal{D}}\boldsymbol{u}\|_{L^1(\Omega)}, \qquad \forall \boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{en}}. \tag{57}$$

Let $\epsilon > 0$ a parameter to be fixed later on. Thanks to Lemma 3.2, there exists a quantity C_ϵ depending only on p and ϵ such that

$$|u| \leq \epsilon\Gamma(u) + C_\epsilon, \qquad \forall u \in \overline{\mathcal{I}}_p,$$

ensuring that

$$\|\pi_{\mathcal{D}}\boldsymbol{u}\|_{L^1(\Omega)} \leq \epsilon \int_\Omega \Gamma(\pi_{\mathcal{D}}\boldsymbol{u})\mathrm{d}\boldsymbol{x} + C_\epsilon \mathrm{meas}(\Omega), \qquad \forall \boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{en}}. \tag{58}$$

On the other hand, Assumption (**A**5) together with the definition (37) of $\mathbf{V} = (V_\kappa, V_{\mathrm{s}})_{\kappa,\mathrm{s}}$ ensure that

$$\|\pi_{\mathcal{D}}\mathbf{V}\|_{L^\infty(\Omega)} \leq \|V\|_\infty.$$

Setting $\epsilon = \frac{1}{2\|V\|_\infty}$ in (58) and injecting the resulting estimate in (57) ends the proof of the second inequality of (56). The proof of the first inequality of (56) being similar, it is left to the reader. □

Lemma 3.5. *There exists C depending only on $\mathbf{\Lambda}$, Ω, $\theta_{\mathcal{T}}$, $\zeta_{\mathcal{D}}$, $\ell_{\mathcal{D}}$, η, p and V such that, for all $\boldsymbol{v} = (v_\kappa, v_{\mathrm{s}})_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}^{\mathrm{ad}}$, one has*

$$\begin{aligned}\sum_{\kappa\in\mathcal{M}}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\eta_{\kappa,\mathrm{s}}(\boldsymbol{v})\left(p(v_\kappa)-p(v_{\mathrm{s}})\right)^2 \\ \leq C\left(1+\mathfrak{E}_{\mathcal{D}}(\boldsymbol{v})+\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\boldsymbol{h}(\boldsymbol{v})\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\boldsymbol{h}(\boldsymbol{v})\right),\end{aligned} \tag{59}$$

where we have set $\eta_{\kappa,\mathrm{s}}(\boldsymbol{v}) = \frac{\eta(v_\kappa)+\eta(v_{\mathrm{s}})}{2}$ for all $\kappa \in \mathcal{M}$ and all $\mathrm{s} \in \mathcal{V}_\kappa$.

Proof. Let $\boldsymbol{v} \in W_{\mathcal{D}}^{\mathrm{ad}} \subset W_{\mathcal{D}}^{\mathrm{en}}$, then it follows from the definition (38) of $\boldsymbol{h}(\boldsymbol{v}) = (\mathfrak{h}_\kappa(v_\kappa), \mathfrak{h}_{\mathrm{s}}(v_s))_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}$ that

$$\boldsymbol{\delta}_\kappa p(\boldsymbol{v})\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa p(\boldsymbol{v}) \leq 2\boldsymbol{\delta}_\kappa\boldsymbol{h}(\boldsymbol{v})\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\boldsymbol{h}(\boldsymbol{v}) + 2\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V}, \qquad \forall\kappa\in\mathcal{M}.$$

It follows from Lemma A.2 that there exists C depending only on $\mathbf{\Lambda}$, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\eta_{\kappa,\mathrm{s}}(\boldsymbol{v})\left(p(v_\kappa)-p(v_{\mathrm{s}})\right)^2 \leq C\boldsymbol{\delta}_\kappa p(\boldsymbol{v})\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa p(\boldsymbol{v}), \qquad \forall\kappa\in\mathcal{M}.$$

Therefore, it only remains to prove that

$$\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V} \leq \mathfrak{E}_{\mathcal{D}}(\boldsymbol{v}) + C, \tag{60}$$

for some C depending only on the prescribed data. Using Lemma A.3, we get that

$$\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V} \leq \sum_{\kappa\in\mathcal{M}}\max_{\mathrm{s}\in\mathcal{V}_\kappa}\eta_{\kappa,\mathrm{s}}(\boldsymbol{v})\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\left(V_\kappa - V_{\mathrm{s}}\right)^2. \tag{61}$$

It results from Lemma A.2 that for all $\kappa \in \mathcal{M}$,

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(V_\kappa - V_\mathrm{s})^2 \le C\int_\kappa \boldsymbol{\nabla}_\mathcal{T}\mathbf{V}\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}_\mathcal{T}\mathbf{V}\mathrm{d}\boldsymbol{x} \le C|\kappa|\lambda^\star\|\boldsymbol{\nabla}V\|^2_\infty. \tag{62}$$

Denote by $\overline{\eta}(\boldsymbol{v}) = (\overline{\eta}_\kappa(\boldsymbol{v}), \overline{\eta}_\mathrm{s}(\boldsymbol{v}))_{\kappa,\mathrm{s}} \in W_\mathcal{D}$ the vector defined by

$$\overline{\eta}_\kappa(\boldsymbol{v}) = \max\left(\eta(v_k); \max_{\mathrm{s}'\in\mathcal{V}_\kappa}\eta(v'_\mathrm{s})\right), \qquad \overline{\eta}_\mathrm{s}(\boldsymbol{v}) = 0, \qquad \forall\kappa\in\mathcal{M},\ \forall\mathrm{s}\in\mathcal{V},$$

and remark that

$$\max_{\mathrm{s}\in\mathcal{V}_\kappa}\eta_{\kappa,\mathrm{s}}(\boldsymbol{v}) \le \overline{\eta}_\kappa(\boldsymbol{v}), \qquad \forall\kappa\in\mathcal{M}.$$

Hence, we deduce using (62) in (61) that

$$\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V} \le C\int_\Omega \pi_\mathcal{M}\overline{\eta}(\boldsymbol{v})\mathrm{d}\boldsymbol{x},$$

for some C depending only on $\theta_\mathcal{T}$, $\boldsymbol{\Lambda}$, $\ell_\mathcal{D}$ and $\|\boldsymbol{\nabla}V\|_\infty$, the operator $\pi_\mathcal{M}$ being defined by (31). Let us now use Lemma A.8 to obtain that

$$\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V} \le \widetilde{C}\int_\Omega \pi_\mathcal{D}\eta(\boldsymbol{v})\mathrm{d}\boldsymbol{x} \tag{63}$$

for some $\widetilde{C}$ depending only on the prescribed data, namely $\theta_\mathcal{T}$, $\boldsymbol{\Lambda}$, $\ell_\mathcal{D}$, $\zeta_\mathcal{D}$ and $\|\boldsymbol{\nabla}V\|_\infty$. Using Lemma 3.3, we know that for all $\epsilon > 0$, there exists C_ϵ depending only on ϵ, η, Γ and $\mathrm{meas}(\Omega)$ such that

$$\int_\Omega \pi_\mathcal{D}\eta(\boldsymbol{v})\mathrm{d}\boldsymbol{x} \le \epsilon\int_\Omega \Gamma(\pi_\mathcal{D}\boldsymbol{v})\mathrm{d}\boldsymbol{x} + C_\epsilon.$$

Combining this result with Lemma 3.4 and (63), we deduce that for all $\epsilon > 0$, there exists C_ϵ depending only on ϵ, $\boldsymbol{\Lambda}$, Ω, $\theta_\mathcal{T}$, $\zeta_\mathcal{D}$, $\ell_\mathcal{D}$, η, p and V such that

$$\sum_{\kappa\in\mathcal{M}}\boldsymbol{\delta}_\kappa\mathbf{V}\cdot\mathbf{B}_\kappa(\boldsymbol{v})\boldsymbol{\delta}_\kappa\mathbf{V} \le \epsilon\widetilde{C}\mathfrak{E}_\mathcal{D}(\boldsymbol{v}) + C_\epsilon, \qquad \forall\boldsymbol{v}\in W^{\mathrm{en}}_\mathcal{D}.$$

We obtain (60) by choosing $\epsilon = \frac{1}{\widetilde{C}}$. This ends the proof of Lemma 3.5. □

Lemma 3.6. *Let $\boldsymbol{u}^{n-1} \in W^{\mathrm{en}}_\mathcal{D}$, and let $\boldsymbol{u}^n \in W^{\mathrm{ad}}_\mathcal{D}$ be a solution to the scheme (43). There exist C_1 and C_2 depending on Δt_n, $\boldsymbol{\Lambda}$, Ω, $\theta_\mathcal{T}$, $\zeta_\mathcal{D}$, $\ell_\mathcal{D}$, η, p and V such that*

$$\sum_{\kappa\in\mathcal{M}}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\eta^n_{\kappa,\mathrm{s}}\left(p(u^n_\kappa) - p(u^n_\mathrm{s})\right)^2$$
$$\le C_1\left(1 + \mathfrak{E}_\mathcal{D}(\boldsymbol{u}^{n-1})\right) \le C_2\left(1 + \int_\Omega \Gamma(\pi_\mathcal{D}\boldsymbol{u}^{n-1})\mathrm{d}\boldsymbol{x}\right).$$

Proof. Since $\boldsymbol{u}^n$ is a solution of the scheme (43), the nonlinear discrete stability estimate (53) holds. Therefore, taking (53) into account in (59) yields

$$\sum_{\kappa\in\mathcal{M}}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\eta^n_{\kappa,\mathrm{s}}\left(p(u^n_\kappa) - p(u^n_\mathrm{s})\right)^2 \le C_1\left(1 + \mathfrak{E}_\mathcal{D}(\boldsymbol{u}^{n-1})\right)$$

for some C_1 depending on the prescribed data. Then it only remains to use Lemma 3.4 to conclude the proof of Lemma 3.6. □

3.2. **Existence of a discrete solution.** The scheme (43) can be rewritten in the form of a nonlinear system

$$\mathcal{F}(\boldsymbol{u}^n) = \mathbf{0}_{\mathbb{R}^{\#\mathcal{M}+\#\mathcal{V}}}.$$

In the case where $p(0) = -\infty$, the function $\mathcal{F}$ is continuous on $W_{\mathcal{D}}^{\mathrm{ad}}$, but not uniformly continuous. The existence proof for a discrete solution we propose relies on a topological degree argument (see e.g. [73, 41]), whence we need to restrict the set of the possible $\boldsymbol{u}^n$ for recovering the uniform continuity by avoiding the singularity near 0. This is the purpose of the following lemma, which is an adaptation of [34, Lemma 3.10].

Lemma 3.7. *Let $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}^{\mathrm{en}}$ be such that $\int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^{n-1} \mathrm{d}\boldsymbol{x} > 0$ and let $\boldsymbol{u}^n$ be a solution to the scheme (43). Assume that $p(0) = -\infty$, then there exists $\epsilon_{\mathcal{D},\Delta t_n} > 0$ depending on Δt_n, $\mathcal{D}$, $\boldsymbol{\Lambda}$, Ω, η, p, V, and $\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^{n-1})$ such that*

$$u_\nu^n \geq \epsilon_{\mathcal{D},\Delta t_n}, \qquad \forall \nu \in \mathcal{M} \cup \mathcal{V}.$$

Proof. First of all, remark that proving Lemma 3.7 is equivalent to proving that there exists $C_{\mathcal{D},\Delta t_n} > 0$ such that

$$p(u_\nu^n) \geq -C_{\mathcal{D},\Delta t_n}, \qquad \forall \nu \in \mathcal{M} \cup \mathcal{V}. \tag{64}$$

Because of the conservation of mass (44), we have

$$\int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^n \mathrm{d}\boldsymbol{x} = \int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^{n-1} \mathrm{d}\boldsymbol{x} > 0.$$

Therefore, we can claim that there exists $\nu_{\mathrm{i}} \in \mathcal{M} \cup \mathcal{V}$ such that

$$u_{\nu_{\mathrm{i}}}^n \geq \frac{1}{\mathrm{meas}(\Omega)} \int_\Omega \pi_{\mathcal{D}} \boldsymbol{u}^{n-1} \mathrm{d}\boldsymbol{x} > 0. \tag{65}$$

Let $(\nu_{\mathrm{f}}) \in \mathcal{M} \cup \mathcal{V}$ be arbitrary, and $(\nu_q)_{q=0,\dots,\ell}$ be a path from ν_{i} to ν_{f}, i.e.

- $\nu_0 = \nu_{\mathrm{i}}$, $\nu_\ell = \nu_{\mathrm{f}}$, and $\nu_p \neq \nu_q$ if $p \neq q$;
- for all $q \in \{0, \dots, \ell - 1\}$, one has:

$$\nu_q \in \mathcal{M} \implies \nu_{q+1} \in \mathcal{V}_{\nu_q}, \quad \text{and} \quad \nu_q \in \mathcal{V} \implies \nu_{q+1} \in \mathcal{M}_{\nu_q}.$$

Let $q \in \{0, \dots, \ell - 1\}$

It follows from Lemma 3.6 that there exists $C_{\mathcal{D},\Delta t_n}$ depending on $\mathcal{D}, \Delta t_n$ $\boldsymbol{\Lambda}$, Ω, η, p, V, and $\mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^{n-1})$ such that

$$\sum_{\kappa \in \mathcal{M}} \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \eta_{\kappa,\mathrm{s}}^n (p(u_\kappa^n) - p(u_{\mathrm{s}}^n))^2 \leq C_{\mathcal{D},\Delta t_n}.$$

This ensures in particular that

$$\sum_{q=0}^{\ell-1} \eta_{\nu_q,\nu_{q+1}}^n (p(u_{\nu_q}^n) - p(u_{\nu_{q+1}}^n))^2 \leq C_{\mathcal{D},\Delta t_n}, \tag{66}$$

where we have set $\eta_{\nu_q,\nu_{q+1}}^n = \eta_{\kappa,\mathrm{s}}^n$ if $\{\nu_q, \nu_{q+1}\} = \{\kappa, \mathrm{s}\}$.

We can now prove (64) thanks to an induction along the path. Assume that $u_{\nu_q}^n > \epsilon_{\mathcal{D},\Delta t_n}$ for some $\epsilon_{\mathcal{D},\Delta t_n} > 0$, whence $\eta_{\nu_q,\nu_{q+1}}^n \geq \frac{\eta(u_{\nu_q}^n)}{2} \geq \epsilon'_{\mathcal{D},\Delta t_n} > 0$. Then it follows from (66) that

$$p(u_{\nu_{q+1}}^n) \geq p(u_{\nu_q}^n) - \sqrt{\frac{C_{\mathcal{D},\Delta t_n}}{\epsilon'_{\mathcal{D},\Delta t_n}}} \geq -C'_{\mathcal{D},\Delta t_n} \quad \implies \quad u_{\nu_{q+1}}^n \geq \epsilon''_{\mathcal{D},\Delta t_n} > 0.$$

We conclude as in [34, Lemma 3.10] thanks to the finite number of possible paths. □

Thanks to Lemma 3.7, one can apply the same strategy as in [34] for proving the existence of a solution to the scheme (43).

Proposition 3.8. *Let $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}^{\mathrm{en}}$, then there exists (at least) one vector $\boldsymbol{u}^n \in W_{\mathcal{D}}^{\mathrm{ad}}$ solution to the system (43).*

Proof. As in [34, Proposition 3.11], the proof relies on a topological degree argument (cf. [73, 41]) applied twice. More precisely, start from the parametrized nonlinear problem that consists in looking for $\boldsymbol{u}^{n,\gamma}$ solution to: $\forall \boldsymbol{v} \in W_{\mathcal{D}}$,

$$\begin{aligned}(67)\quad \int_\Omega \pi_{\mathcal{D}} \left(\frac{\boldsymbol{u}^{n,\gamma} - \boldsymbol{u}^{n-1}}{\Delta t_n} \right) \pi_{\mathcal{D}} \boldsymbol{v} \mathrm{d}\boldsymbol{x} \\ + \gamma \sum_{\kappa \in \mathcal{M}} \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \sum_{\mathrm{s}' \in \mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa \left(\mathfrak{h}_\kappa(u_\kappa^{n,\gamma}) - \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}^{n,\gamma}) \right) (v_\kappa - v_{\mathrm{s}}') \\ + (1-\gamma) \sum_{\kappa \in \mathcal{M}} \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \left(\sum_{\mathrm{s}' \in \mathcal{V}_\kappa} |a_{\mathrm{s},\mathrm{s}'}^\kappa| \right) \left(\mathfrak{h}_\kappa(u_\kappa^{n,\gamma}) - \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}^{n,\gamma}) \right) (v_\kappa - v_{\mathrm{s}}) = 0.\end{aligned}$$

For $\gamma = 0$, the corresponding scheme is monotone, hence the nonlinear system (67) admits a unique solution $\boldsymbol{u}^{n,0} \in W_{\mathcal{D}}^{\mathrm{ad}}$, and its corresponding topological degree is equal to 1 (see for instance [49] for the application of this argument to the case of a pure hyperbolic equation). In the case where $\lim_{u \searrow 0} p(u) = -\infty$, one can prove as in Lemma 3.7 that any solution $\boldsymbol{u}^{n,\gamma} \in W_{\mathcal{D}}^{\mathrm{ad}}$ to (67) satisfies

$$u_\beta^{n,\gamma} \geq \epsilon_{\mathcal{D},\Delta t_n}, \quad \forall \beta \in \mathcal{M} \cup \mathcal{V}$$

for some $\epsilon_{\mathcal{D},\Delta t_n} > 0$ not depending on γ. The convex subset on which one looks for a solution $\boldsymbol{u}^{n,\gamma}$ can be restricted to the subset $W_{\mathcal{D}}^{\mathrm{ad}}$ defined by

$$K := \left\{ \boldsymbol{u} \in W_{\mathcal{D}}^{\mathrm{ad}} \;\middle|\; u_\beta \geq \frac{\epsilon_{\mathcal{D},\Delta t_n}}{2} \text{ and } \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}) \leq \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^{n-1}) + 1 \right\},$$

the corresponding topological degree being still equal to 1. Note that the bound $u_\beta \geq \frac{\epsilon_{\mathcal{D},\Delta t_n}}{2}$ must be removed if $p(0)$ is finite. This ensures the existence of at least one solution to the nonlinear system (67) when γ is equal to 1.

Starting from the system (67) with $\gamma = 1$, one defines a second homotopy parametrized by $\mu \in [0,1]$ to get (45). More precisely (the superscript $\gamma = 1$ has been removed for clarity), we set

$$\eta^\mu : u \mapsto 1 + \mu(\eta(u) - 1), \qquad \forall \mu \in [0,1],$$

so that η^0 is constant equal to 1 and $\eta^1 \equiv \eta$. Define $\boldsymbol{u}^{n,\mu} \in W_{\mathcal{D}}^{\mathrm{ad}}$ as a solution to

$$\begin{aligned}(68)\quad \int_\Omega \pi_{\mathcal{D}} \left(\frac{\boldsymbol{u}^{n,\mu} - \boldsymbol{u}^{n-1}}{\Delta t_n} \right) \pi_{\mathcal{D}} \boldsymbol{v} \mathrm{d}\boldsymbol{x} \\ + \sum_{\kappa \in \mathcal{M}} \sum_{\mathrm{s} \in \mathcal{V}_\kappa} \sqrt{\eta_{\kappa,\mathrm{s}}^{n,\mu}} \sum_{\mathrm{s}' \in \mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa \sqrt{\eta_{\kappa,\mathrm{s}'}^{n,\mu}} \left(\mathfrak{h}_\kappa(u_\kappa^{n,\mu}) - \mathfrak{h}_{\mathrm{s}}(u_{\mathrm{s}}^{n,\mu}) \right) (v_\kappa - v_{\mathrm{s}}') = 0,\end{aligned}$$

where $\boldsymbol{v} \in W_{\mathcal{D}}$ is arbitrary, and where

$$\eta_{\kappa,\mathrm{s}'}^{n,\mu} = \frac{\eta^\mu(u_\kappa^n) + \eta^\mu(u_{\mathrm{s}}^n)}{2}, \qquad \forall \kappa \in \mathcal{M}, \forall \mathrm{s} \in \mathcal{V}_\kappa.$$

A priori estimates similar to those derived previously in the paper ensure that the solutions to (68) cannot belong to ∂K whatever the value of $\mu \in [0,1]$. The existence of a solution to the scheme (43) follows. □

3.3. **Multistep *a priori* estimates.** As a byproduct of the existence of a discrete solution $\boldsymbol{u}^n$ for all $n \in \{1,\dots,N\}$, we can now derive *a priori* estimates on functions reconstructed thanks to the discrete solution $\boldsymbol{u} \in W_{\mathcal{D},\boldsymbol{\Delta} t}$.

The first estimate we get is obtained by summing Ineq. (53) w.r.t. n, and by using the positivity of the dissipation. This provides

$$\max_{n\in\{1,\dots,N\}} \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^n) \le \mathfrak{E}_{\mathcal{D}}(\boldsymbol{u}^0) \le \mathfrak{E}(u_0) + C \le C, \tag{69}$$

where C only depends on V, u_0, p and Ω thanks to Lemma 2.1. Since the discrete entropy functional $\mathfrak{E}_{\mathcal{D}}$ is bounded from below by a quantity depending only on p, V and Ω (cf. Lemma 3.4), we deduce also from the summation of (53) w.r.t. n that there exists C depending only on p, V, Ω, and u_0 (but not on $\mathcal{D}$) such that

$$\sum_{n=1}^{N} \Delta t_n \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n)\cdot \mathbf{B}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa \boldsymbol{h}(\boldsymbol{u}^n) \le C. \tag{70}$$

Mimicking the proof of Lemma 3.5, this yields

$$\sum_{n=1}^{N} \Delta t_n \sum_{\kappa\in\mathcal{M}} \boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n)\cdot \mathbf{B}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n) \le C \tag{71}$$

for some quantity C depending on $\boldsymbol{\Lambda}$, Ω, $\theta_{\mathcal{T}}$, $\zeta_{\mathcal{D}}$, $\ell_{\mathcal{D}}$, η, p, V and t_{f}.

The following lemma is a direct consequence of Estimate (69) and Lemma 3.4. Its detailed proof is left to the reader.

Lemma 3.9. *There exists C depending only on Ω, p, V, u_0 and Ω (but not on the discretization) such that*

$$\|\Gamma(\pi_{\mathcal{D},\boldsymbol{\Delta} t}\boldsymbol{u})\|_{L^\infty((0,t_{\mathrm{f}});L^1(\Omega))} \le C.$$

The introduction of the Kirchhoff transform was avoided in our scheme. Its extension to complex problems (like e.g., systems, problems with hysteresis) in therefore easier. However, the (semi-)Kirchhoff transform ξ defined by (7) is useful for carrying the analysis out. The purpose of the following lemma is to provide a discrete $L^2((0,T);H^1(\Omega))$ estimate on $\xi(\boldsymbol{u})$.

Lemma 3.10. *There exists $C > 0$depending only on $\boldsymbol{\Lambda}$, η, V, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$, Ω, t_{f}, u_0 such that*

$$\iint_{Q_{t_{\mathrm{f}}}} \boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T},\boldsymbol{\Delta} t}\xi(\boldsymbol{u})\cdot\boldsymbol{\nabla}_{\mathcal{T},\boldsymbol{\Delta} t}\xi(\boldsymbol{u})\mathrm{d}\boldsymbol{x} \le C.$$

Proof. Since the function η was assumed to be nondecreasing, see Assumption ($\mathbf{A}1$). We know that for all interval $[a,b] \subset \mathcal{I}_p$, $\max_{c\in[a,b]}\eta(c) = \max\{\eta(a),\eta(b)\}$, hence, denoting by $\mathcal{I}^n_{\kappa,\mathrm{s}}$ the interval with extremities u^n_κ and u^n_{s}, we obtain that

$$\eta^n_{\kappa,\mathrm{s}} \ge \frac{1}{2}\max_{c\in\mathcal{I}^n_{\kappa,\mathrm{s}}}\eta(c), \qquad \forall\kappa\in\mathcal{M},\ \forall \mathrm{s}\in\mathcal{V}_\kappa. \tag{72}$$

The definition (7) of the function ξ implies that

$$\left(\xi(u^n_{\mathrm{s}}) - \xi(u^n_\kappa)\right)^2 \le \left(\max_{c\in\mathcal{I}^n_{\kappa,\mathrm{s}}}\eta(c)\right)\left(p(u^n_{\mathrm{s}}) - p(u^n_\kappa)\right)^2,$$

whence we obtain that for all $\kappa \in \mathcal{M}$,

$$\begin{aligned}|\mathbf{M}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n)|^2 &= \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \eta^n_{\kappa,\mathrm{s}} \left(p(u^n_{\mathrm{s}}) - p(u^n_\kappa)\right)^2 \\ &\geq \frac{1}{2}\sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\xi(u^n_{\mathrm{s}}) - \xi(u^n_\kappa)\right)^2 = \frac{1}{2}|\boldsymbol{\delta}_\kappa \xi(\boldsymbol{u}^n)|^2.\end{aligned}$$

Using that

$$\boldsymbol{v}\cdot\mathbf{A}_\kappa\boldsymbol{v} \geq \boldsymbol{w}\cdot\mathbf{A}_\kappa\boldsymbol{w}, \qquad \forall\, \boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^{\ell_\kappa} \ \text{ s.t. } \ |\boldsymbol{v}|^2 \geq \mathrm{Cond}_2(\mathbf{A}_\kappa)|\boldsymbol{w}|^2,$$

and that $\mathbf{B}_\kappa(\boldsymbol{u}) = \mathbf{M}_\kappa(\boldsymbol{u})\mathbf{A}_\kappa\mathbf{M}_\kappa(\boldsymbol{u})$, we get that for all $\kappa\in\mathcal{M}$,

$$\begin{aligned}\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n)\cdot\mathbf{B}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n) &\geq \frac{1}{2\mathrm{Cond}_2(\mathbf{A}_\kappa)}\boldsymbol{\delta}_\kappa\xi(\boldsymbol{u}^n)\cdot\mathbf{A}_\kappa\boldsymbol{\delta}_\kappa\xi(\boldsymbol{u}^n) \\ &= \frac{1}{2\mathrm{Cond}_2(\mathbf{A}_\kappa)}\int_\kappa \boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T}}\xi(\boldsymbol{u}^n)\cdot\boldsymbol{\nabla}_{\mathcal{T}}\xi(\boldsymbol{u}^n)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Thanks to Lemma A.1 stated in appendix, we know that $C>0$ depending only on $\boldsymbol{\Lambda}, \theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that $\mathrm{Cond}_2(\mathbf{A}_\kappa)\leq C$, for all $\kappa\in\mathcal{M}$, so that: $\forall\kappa\in\mathcal{M}$, $\forall n\in\{1,\dots,N\}$,

$$\int_\kappa \boldsymbol{\nabla}_{\mathcal{T}}\xi(\boldsymbol{u}^n)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T}}\xi(\boldsymbol{u}^n)\mathrm{d}\boldsymbol{x} \leq C\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n)\cdot\mathbf{B}_\kappa(\boldsymbol{u}^n)\boldsymbol{\delta}_\kappa p(\boldsymbol{u}^n). \tag{73}$$

In order to conclude the proof, it only remains to multiply (73) by Δt_n, to sum over $\kappa\in\mathcal{M}$ and $n\in\{1,\dots,N\}$, and finally to use (71). □

Combining Estimate (71) and Lemma A.2 yields the following lemma, whose complete proof is left to the reader.

Lemma 3.11. *There exists C depending only on $\boldsymbol{\Lambda}$, Ω, $\theta_{\mathcal{T}}$, $\zeta_{\mathcal{D}}$, $\ell_{\mathcal{D}}$, η, p, V and t_{f} such that*

$$\sum_{n=1}^N \Delta t_n \sum_{\kappa\in\mathcal{M}}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)\eta^n_{\kappa,\mathrm{s}}\left(p(u^n_\kappa)-p(u^n_{\mathrm{s}})\right)^2 \leq C.$$

4. Proof of Theorem 2.4

In what follows, we consider a sequence $(\mathcal{D}_m, \boldsymbol{\Delta}t_m)_{m\geq1}$ of discretizations of $Q_{t_{\mathrm{f}}}$ such that (54) holds. In order to prove the convergence of the reconstructed discrete solution $\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\boldsymbol{u}_m$ towards the weak solution of (1) as m tends to ∞, we adopt the classical strategy that consists in showing first that the family $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\boldsymbol{u}_m)_{m\geq1}$ is precompact in $L^1(Q_{t_{\mathrm{f}}})$ (this is the purpose of §4.1), then to identify in §4.2 the limit as a weak solution of (1) in the sense of Definition 1.

As a direct consequence of Theorem 2.3, one knows that the scheme admits a solution $\boldsymbol{u}_m = \left(u^n_{\kappa,m}, u^n_{\mathrm{s},m}\right)$ that, thanks to the regularity assumptions (54b)–(54c) on the discretization and thanks to Lemmas 3.9, 3.10 and 3.11, satisfies the following uniform estimates w.r.t. m:

$$\|\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\Gamma(\boldsymbol{u}_m)\|_{L^\infty((0,t_{\mathrm{f}});L^1(\Omega))} \leq C, \tag{74}$$

$$\iint_{Q_{t_{\mathrm{f}}}} \boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m)\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C, \tag{75}$$

$$\sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}\in\mathcal{V}_\kappa} |a^\kappa_{\mathrm{s},\mathrm{s}'}| \right) \eta^n_{\kappa,\mathrm{s}} \left(p(u^n_{\kappa,m}) - p(u^n_{\mathrm{s},m})\right)^2 \leq C, \tag{76}$$

where C may depend on the data of the continuous problem, and on the discretization regularity factors $\theta^\star$, $\ell^\star$ and $\zeta^\star$ but not on m.

4.1. Compactness properties of the discrete solutions.

Lemma 4.1. *Let $(\mathcal{D}_m, \boldsymbol{\Delta} t_m)$ be a sequence of discretizations of Q_{t_f} satisfying Assumptions (54), there exists C depending only on $\boldsymbol{\Lambda}$, $\theta^\star$, $\ell^\star$, Ω, t_f, p and u_0 such that, for all $m \geq 1$, one has*

$$\|\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^2((0,t_\mathrm{f});H^1(\Omega))} \leq C \quad \text{and} \quad \|\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^2(Q_{t_\mathrm{f}})} \leq C.$$

Proof. It follows from the Estimate (75) and from Lemma A.5 that for all $m \geq 1$,

$$\begin{aligned} &\|\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m) - \pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^2((0,t_\mathrm{f});L^1(\Omega))} \\ &\qquad \leq \mathrm{meas}(\Omega)^{1/2} \|\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m) - \pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^2(Q_{t_\mathrm{f}})} \leq C \end{aligned} \tag{77}$$

for some C depending only on $\boldsymbol{\Lambda}$, Ω, t_f, $\ell^\star$, $\theta^\star$, p, and u_0 (but not on m).

Moreover, it follows from Assumption (8) that

$$\|\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^\infty((0,t_\mathrm{f});L^1(\Omega))} \leq C \left(1 + \|\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\Gamma(\boldsymbol{u}_m)\|_{L^\infty((0,t_\mathrm{f});L^1(\Omega))}\right) \leq C$$

thanks to the estimate (74). Combining this inequality with (77) provides that

$$\|\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^1(Q_{t_\mathrm{f}})} \leq C,$$

whence the sequence $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m))_{m\geq 1}$ is bounded in $L^2((0,t_\mathrm{f});W^{1,1}(\Omega))$ thanks to (75). A classical bootstrap argument using Sobolev inequalities allows to claim that it is bounded in $L^2((0,t_\mathrm{f});H^1(\Omega))$, thus in particular in $L^2(Q_{t_\mathrm{f}})$. One concludes that $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m))_{m\geq 1}$ is also bounded in $L^2(Q_{t_\mathrm{f}})$ thanks to (77). □

Remark 4.2. *An alternative way to prove the key-point of Lemma 4.1, namely*

$$\|\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)\|_{L^2((0,t_\mathrm{f});H^1(\Omega))} \leq C,$$

would consist in using [63, Lemma A.1], *that shows that*

$$v \in L^1(\Omega) \text{ and } \boldsymbol{\nabla}\xi(v) \in L^2(\Omega)^d \quad \Longrightarrow \quad \xi(v) \in H^1(\Omega).$$

As a consequence of Lemma 4.1, we know that the sequence $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m))_{m\geq 1}$ is relatively compact for the $L^2((0,t_\mathrm{f});H^1(\Omega))$-weak topology. Moreover, the space $H^1(\Omega)$ being locally compact in $L^2(\Omega)$, a uniform information on the time translates of $\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m)$ will provide the relative compactness of $(\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m}\xi(\boldsymbol{u}_m))_{m\geq 1}$ in the $L^2(Q_{t_\mathrm{f}})$-strong topology (see e.g. [91]). Such a uniform time-translate estimate can be obtained by using directly the numerical scheme (see e.g. [51, 34]). One can also make use of black-boxes like e.g. [6, 8]. Note that the result of [60] does not apply here because of the degeneracy of the problem. We do not provide the proof of next proposition here, since a suitable black-box will be contained in the forthcoming contribution [8]. A more classical but calculative possibility would consist in mimicking the proof of [34, Lemmas 4.3 and 4.5].

Proposition 4.3. *Let $(\mathcal{D}_m, \boldsymbol{\Delta} t_m)$ be a sequence of discretizations of Q_{t_f} satisfying Assumptions (54), and let $(\boldsymbol{u}_m)_{m\geq 1}$ be the corresponding sequence of solutions to the scheme (43). Then there exists a measurable function $u : Q_{t_\mathrm{f}} \to \mathbb{R}$ with $\xi(u) \in L^2((0,t_\mathrm{f}); H^1(\Omega))$ such that, up to a subsequence, one has*

$$\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m \underset{m\to\infty}{\longrightarrow} u \quad \text{a.e. in } Q_{t_\mathrm{f}}. \tag{78}$$

Corollary 4.4. *Keeping the assumption and notations of Proposition 4.3, one has*

$$\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m \underset{m\to\infty}{\longrightarrow} u \quad \text{strongly in } L^1(Q_{t_\mathrm{f}}).$$

Proof. As a result of Proposition 4.3, the almost everywhere convergence property (78) holds. On the other hand, it follows from Assumption (**A**2), more precisely from the fact that $\lim_{u\to\infty} p(u) = +\infty$ that the function Γ defined by (4) is superlinear, i.e.,

$$\lim_{u\to+\infty} \frac{\Gamma(u)}{u} = +\infty.$$

Therefore, Estimate (74) implies that $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m)_{m\geq 1}$ is uniformly equi-integrable. Hence we can apply Vitali's convergence theorem to conclude the proof of Corollary 4.4. □

Lemma 4.5. *Under the assumptions of Proposition 4.3, one has*

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} \xi(u) \quad \text{weakly in } L^2((0,t_\mathrm{f}); H^1(\Omega)),$$

where u is the solution exhibited in Proposition 4.3.

Proof. Thanks to Lemma 4.1, the sequence $(\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m))_{m\geq 1}$ is uniformly bounded in $L^2(Q_{t_\mathrm{f}})^d$. Therefore, there exists $\Xi \in L^2((0,t_\mathrm{f}); H^1(\Omega))$ such that

$$\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} \Xi \quad \text{weakly in } L^2((0,t_\mathrm{f}); H^1(\Omega)).$$

But in view of Proposition 4.3 and of the continuity of ξ, we know that

$$\xi\left(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m\right) = \pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} \xi(u) \quad \text{a.e. in } Q_{t_\mathrm{f}}.$$

Since $\pi_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m)$ and $\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \xi(\boldsymbol{u}_m)$ have the same limit (cf. Lemma A.5), we get that $\Xi = \xi(u)$. □

Lemma 4.6. *Let u be the limit value of $\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m$ obtained in Proposition 4.3, then*

$$\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \eta(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} \eta(u) \quad \text{strongly in } L^1(Q_{t_\mathrm{f}}), \tag{79}$$

and

$$\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \eta(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} \eta(u) \quad \text{strongly in } L^1(Q_{t_\mathrm{f}}). \tag{80}$$

Proof. Let us first establish (79). Thanks to the entropy estimate (74), we know that the sequence $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \Gamma(\boldsymbol{u}_m))_{m\geq 1}$ is uniformly bounded in $L^\infty((0,t_\mathrm{f}); L^1(\Omega))$, thus in $L^1(Q_{t_\mathrm{f}})$. Then Assumption (9) allows to use the de la Vallée-Poussin theorem to claim that $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \eta(\boldsymbol{u}_m))_{m\geq 1}$ is uniformly equi-integrable on Q_{t_f}. Moreover, the continuity of η and Proposition 4.3 provide that

$$(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \eta(\boldsymbol{u}_m))_{m\geq 1} \underset{m\to\infty}{\longrightarrow} \eta(u) \quad \text{a.e. in } Q_{t_\mathrm{f}}.$$

Therefore, we obtain (79) by applying Vitali's theorem.

Let us now prove (80) by proving that $\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)$ and $\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)$ (that is uniformly equi-integrable for the same reasons as $\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)$ is) have the same limit $\eta(u)$ as m tends to ∞. It follows from a combination of (75) with Lemma A.5 that, still up to an unlabeled subsequence,

$$\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m) - \pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m) \underset{m\to\infty}{\longrightarrow} 0 \quad \text{a.e. in } Q_{t_\mathrm{f}}. \tag{81}$$

Since the function $\sqrt{\eta\circ\xi^{-1}}$ is assumed to be uniformly continuous (cf. (10)), it admits a non-decreasing modulus of continuity $\varpi \in C(\mathbb{R}_+;\mathbb{R}_+)$ with $\varpi(0)=0$ such that, for all $v,\widehat{v}$ in the range of ξ,

$$\left|\sqrt{\eta\circ\xi^{-1}}(v) - \sqrt{\eta\circ\xi^{-1}}(\widehat{v})\right| \le \varpi(|v-\widehat{v}|), \tag{82}$$

so that

$$\left|\sqrt{\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)} - \sqrt{\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)}\right| \le \varpi(|\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m) - \pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m)|).$$

Therefore, it follows from (81) that

$$\sqrt{\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)} - \sqrt{\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)} \underset{m\to\infty}{\longrightarrow} 0 \quad \text{a.e. in } Q_{t_\mathrm{f}}.$$

Thus $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m))_{m\ge1}$ and $(\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m))_{m\ge1}$ share the same limit. □

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

Proposition 4.7. *Let u be a limit value of the sequence $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}\boldsymbol{u}_m)_{m\ge1}$ exhibited in Proposition 4.3, then u is the unique weak solution to the problem (1) in the sense of Definition 1.*

Proof. In order to check that u is a weak solution, it only remains to check that the weak formulation (11) holds. Let $\psi \in C_c^\infty(\overline{\Omega}\times[0,t_\mathrm{f}))$, then, for all $m\ge1$, for all $\beta\in\mathcal{M}_m\cup\mathcal{V}_m$ and all $n\in\{0,\dots,N_m\}$, we denote by $\boldsymbol{\Delta}t_m = (\Delta t_{1,m},\dots,\Delta t_{N_m,m})$, by $t_{n,m} = \sum_{i=1}^n \Delta t_{i,m}$, by $\psi_\beta^n = \psi(\boldsymbol{x}_\beta, t_{n,m})$, by $\boldsymbol{\psi}_m^n = \left(\psi_\beta^n\right)_{\beta\in\mathcal{M}_m\cup\mathcal{V}_m} \in W_{\mathcal{D}_m}$, and by $\boldsymbol{\psi}_m = (\boldsymbol{\psi}_m^n)_{0\le n\le N_m} \in W_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}$. Note that since $\psi(\cdot,t_\mathrm{f})=0$, one has $\boldsymbol{\psi}_m^{N_m} = \mathbf{0}$ for all $m\ge1$.

Setting $\boldsymbol{v} = \boldsymbol{\psi}_m^{n-1}$ in (45) and summing over $n\in\{1,\dots,N_m\}$ leads after a classical reorganization of the sums [51] to

$$A_m + B_m + C_m + D_m = 0, \tag{83}$$

where we have set

$$\begin{aligned}
A_m &= \sum_{n=1}^{N_m}\Delta t_{n,m}\int_\Omega \pi_{\mathcal{D}_m}\boldsymbol{u}_m^n(\boldsymbol{x})\pi_{\mathcal{D}_m}\left(\frac{\boldsymbol{\psi}^{n-1}-\boldsymbol{\psi}^n}{\Delta t_{n,m}}\right)(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\
B_m &= -\int_\Omega \pi_{\mathcal{D}_m}\boldsymbol{u}_m^0(\boldsymbol{x})\pi_{\mathcal{D}_m}\boldsymbol{\psi}^0(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\
C_m &= \sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\boldsymbol{\delta}_\kappa p(\boldsymbol{u}_m^n)\cdot\mathbf{B}_\kappa(\boldsymbol{u}_m^n)\boldsymbol{\delta}_\kappa\boldsymbol{\psi}_m^{n-1},\\
D_m &= \sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\boldsymbol{\delta}_\kappa\mathbf{V}_m\cdot\mathbf{B}_\kappa(\boldsymbol{u}_m^n)\boldsymbol{\delta}_\kappa\boldsymbol{\psi}_m^{n-1},
\end{aligned}$$

and $\mathbf{V}_m = (V(\boldsymbol{x}_\kappa), V(\boldsymbol{x}_\mathrm{s}))_{\kappa\in\mathcal{M}_m,\mathrm{s}\in\mathcal{V}_m}$.

The regularity of ψ yields

$$\sum_{n=1}^{N_m} \pi_{\mathcal{D}_m}\left(\frac{\boldsymbol{\psi}^{n-1}-\boldsymbol{\psi}^n}{\Delta t_{n,m}}\right)\mathbf{1}_{[t_{n-1,m},t_{n,m})} \underset{m\to\infty}{\longrightarrow} -\partial_t\psi \quad \text{uniformly on } Q_{t_\mathrm{f}}$$

where $t_{n,m}=\sum_{i=1}^n \Delta t_{i,m}$, so that, using Corollary 4.4, one gets

$$\lim_{m\to\infty} A_m = -\iint_{Q_{t_\mathrm{f}}} u\partial_t\psi \,\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{84}$$

The function $\pi_{\mathcal{D}_m}\boldsymbol{u}_m^0(\boldsymbol{x})$ tends strongly in $L^1(\Omega)$ towards u_0 and $\pi_{\mathcal{D}_m}\boldsymbol{\psi}^0$ converges uniformly towards $\psi(\cdot,0)$ as m tends to $+\infty$, leading to

$$\lim_{m\to\infty} B_m = -\int_\Omega u_0(\boldsymbol{x})\psi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}. \tag{85}$$

We split the term C_m into three parts

$$C_m = C_{1,m}+C_{2,m}+C_{3,m}, \qquad m\geq 1, \tag{86}$$

where, setting $\widehat{\boldsymbol{\psi}}_m^0=\boldsymbol{\psi}_m^0$, $\widehat{\boldsymbol{\psi}}_m^n=\boldsymbol{\psi}_m^{n-1}\in W_{\mathcal{D}_m}$ for all $n\in\{1,\dots,N_m\}$, and $\widehat{\boldsymbol{\psi}}_m=\left(\widehat{\boldsymbol{\psi}}_m^n\right)_{0\leq n\leq N_m}\in W_{\mathcal{D}_m,\boldsymbol{\Delta}t_m}$, one has

$$C_{1,m}=\iint_{Q_{t_\mathrm{f}}} \pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\sqrt{\eta(\boldsymbol{u}_m)}\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\widehat{\boldsymbol{\psi}}_m\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$C_{2,m}=\sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}\sqrt{\eta_{\kappa,\mathrm{s}}^n}\left(p(u_\kappa^n)-p(u_\mathrm{s}^n)\right)a_{\mathrm{s},\mathrm{s}'}^\kappa \times\left(\sqrt{\eta_{\kappa,\mathrm{s}'}^n}-\sqrt{\eta(u_\kappa^n)}\right)\left(\psi_\kappa^{n-1}-\psi_{\mathrm{s}'}^{n-1}\right),$$

$$C_{3,m}=\sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\sqrt{\eta(u_\kappa^n)}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sqrt{\eta_{\kappa,\mathrm{s}}^n}(p(u_\kappa^n)-p(u_\mathrm{s}^n))-(\xi(u_\kappa^n)-\xi(u_\mathrm{s}^n))\right) \times\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}a_{\mathrm{s},\mathrm{s}'}^\kappa\left(\psi_\kappa^{n-1}-\psi_{\mathrm{s}'}^{n-1}\right).$$

Thanks to Lemma 4.6, we know that

$$\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\sqrt{\eta(\boldsymbol{u}_m)}\underset{m\to\infty}{\longrightarrow}\sqrt{\eta(u)} \quad \text{strongly in } L^2(Q_{t_\mathrm{f}}).$$

Hence, it follows from the weak convergence in $L^2(Q_{t_\mathrm{f}})$ of $\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\xi(\boldsymbol{u}_m)$ towards $\boldsymbol{\nabla}\xi(u)$ (cf. Lemma 4.5) and from the uniform convergence of $\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\widehat{\boldsymbol{\psi}}_m$ towards $\boldsymbol{\nabla}\psi$ as m tends to $+\infty$ (see for instance [40, Theorem 16.1]) that

$$\lim_{m\to\infty} C_{1,m}=\iint_{Q_{t_\mathrm{f}}}\sqrt{\eta(u)}\boldsymbol{\nabla}\xi(u)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}\psi\,\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{87}$$

Let us focus now of $C_{2,m}$. Using the inequality $ab\leq\epsilon a^2+\frac{1}{4\epsilon}b^2$, one gets that

$$C_{2,m}\leq\epsilon C_{2,m}'+\frac{C_{2,m}''}{4\epsilon}, \qquad \forall\epsilon>0, \tag{88}$$

where

$$C'_{2,m} = \sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} \left|a^\kappa_{\mathrm{s},\mathrm{s}'}\right| \right) \eta^n_{\kappa,\mathrm{s}} \left(p(u^n_\kappa) - p(u^n_\mathrm{s})\right)^2,$$

$$C''_{2,m} = \sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} \left|a^\kappa_{\mathrm{s},\mathrm{s}'}\right| \right) \left(\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u^n_\kappa)}\right)^2 \left(\psi^{n-1}_\kappa - \psi^{n-1}_\mathrm{s}\right)^2.$$

We deduce from Estimate (76) that

$$C'_{2,m} \le C, \qquad \forall m \ge 1, \tag{89}$$

for some C depending only on u_0, p, Ω, $\mathbf{\Lambda}$, $\theta^\star$ and $\ell^\star$.

Define $\boldsymbol{\mu}_m = (\mu^n_\kappa, \mu^n_\mathrm{s})_{\kappa,\mathrm{s},n} \in W_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}$ by

$$\begin{cases} \mu^n_\kappa = \max\limits_{\mathrm{s}\in\mathcal{V}_\kappa} \sup\limits_{u\in\mathcal{I}^n_{\kappa,\mathrm{s}}} \left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u)}\right|, & \forall \kappa \in \mathcal{M}_m, \\ \mu^n_\mathrm{s} = 0, & \forall \mathrm{s} \in \mathcal{V}_m, \end{cases} \qquad \forall n \in \{0,\dots,N_m\}. \tag{90}$$

The definition of μ^n_κ implies directly that

$$\left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u^n_\kappa)}\right| \le \mu^n_\kappa, \qquad \forall \kappa \in \mathcal{M}_m,\ \forall \mathrm{s} \in \mathcal{V}_\kappa.$$

Therefore, we get that

$$C''_{2,m} \le \sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} (\mu^n_\kappa)^2 \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} \left|a^\kappa_{\mathrm{s},\mathrm{s}'}\right| \right) \left(\psi^{n-1}_\kappa - \psi^{n-1}_\mathrm{s}\right)^2.$$

Then thanks to Lemma A.2, there exists C depending only on $\mathbf{\Lambda}$, $\theta^\star$ and $\ell^\star$ such that

$$C''_{2,m} \le C \iint_{Q_{t_\mathrm{f}}} \left(\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \boldsymbol{\mu}_m\right)^2 \ \boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \widehat{\boldsymbol{\psi}}_m \cdot \mathbf{\Lambda} \boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \widehat{\boldsymbol{\psi}}_m \mathrm{d}\boldsymbol{x}\mathrm{d}t, \qquad \forall m \ge 1.$$

Since $\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \boldsymbol{\mu}_m$ converges to 0 strongly in $L^2(Q_{t_\mathrm{f}})$ as m tends to ∞ (this is the purpose of Lemma 4.8 hereafter), and since $\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta} t_m} \widehat{\boldsymbol{\psi}}_m$ remains bounded in $L^\infty(Q_{t_\mathrm{f}})$ uniformly w.r.t. $m \ge 1$, one gets that

$$\lim_{m\to\infty} C''_{2,m} = 0. \tag{91}$$

Therefore, it follows from (88)–(91) that $\limsup_{m\to\infty} C_{2,m} \le C\epsilon$ for arbitrary small values of $\epsilon > 0$, whence

$$\lim_{m\to\infty} C_{2,m} = 0. \tag{92}$$

As a preliminary before considering $C_{3,m}$, let us set, for all $\kappa \in \mathcal{M}_m$, all $\mathrm{s} \in \mathcal{V}_\kappa$, and all $n \in \{1,\dots,N_m\}$,

$$\widetilde{\eta}^n_{\kappa,\mathrm{s}} = \begin{cases} \left(\dfrac{\xi(u^n_\kappa) - \xi(u^n_\mathrm{s})}{p(u^n_\kappa) - p(u^n_\mathrm{s})}\right)^2 & \text{if } u^n_\kappa \ne u^n_\mathrm{s}, \\ \eta(u^n_\kappa) & \text{if } u^n_\kappa = u^n_\mathrm{s}. \end{cases}$$

Thanks to the mean value theorem, we can claim that, for all $\kappa \in \mathcal{M}_m$, all $\mathrm{s} \in \mathcal{V}_\kappa$ and all $n \in \{1,\dots,N_m\}$, there exists $\widetilde{u}^n_{\kappa,\mathrm{s}} \in \mathcal{I}^n_{\kappa,\mathrm{s}} = [\min(u^n_\kappa, u^n_\mathrm{s}), \max(u^n_\kappa, u^n_\mathrm{s})]$ such that $\widetilde{\eta}^n_{\kappa,\mathrm{s}} = \eta(\widetilde{u}^n_{\kappa,\mathrm{s}})$. In particular, this ensures that

$$\left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\widetilde{\eta}^n_{\kappa,\mathrm{s}}}\right| \le \mu^n_\kappa, \qquad \forall \kappa \in \mathcal{M}_m,\ \forall \mathrm{s} \in \mathcal{V}_\kappa$$

where μ_κ was defined by (90). Using moreover that $\eta(u)_\kappa^n \leq 2\eta_{\kappa,\mathrm{s}}^n$, one gets that

$$C_{3,m} \leq 2\sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \mu_\kappa^n \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \sqrt{\eta_{\kappa,\mathrm{s}}^n}(p(u_\kappa^n)-p(u_\mathrm{s}^n)) \sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa \left(\psi_\kappa^{n-1}-\psi_{\mathrm{s}'}^{n-1}\right).$$

Cauchy-Schwarz inequality and Estimate (76) yield

$$C_{3,m} \leq C\left(\sum_{\kappa\in\mathcal{M}_m} (\mu_\kappa^n)^2 \sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} |a_{\mathrm{s},\mathrm{s}'}^\kappa|\right)\left(\psi_\kappa^{n-1}-\psi_\mathrm{s}^{n-1}\right)^2\right)^{1/2}.$$

Lemma A.2, and the regularity of ψ provide that

$$C_{3,m} \leq C\left\|\pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\boldsymbol{\mu}_m\right\|_{L^2(Q_{t_\mathrm{f}})}.$$

Applying Lemma 4.8, we get

$$\lim_{m\to\infty} C_{3,m} = 0. \tag{93}$$

Putting (86) together with (87), (92), and (93), one gets that

$$\lim_{m\to\infty} C_m = \iint_{Q_{t_\mathrm{f}}} \sqrt{\eta(u)}\boldsymbol{\nabla}\xi(u)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}\psi\,\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{94}$$

Now, we focus on the term D_m that can be decomposed into

$$D_m = D_{1,m} + D_{2,m} + D_{3,m}, \qquad \forall m \geq 1, \tag{95}$$

where we have set

$$D_{1,m} = \iint_{Q_{t_\mathrm{f}}} \pi_{\mathcal{M}_m,\boldsymbol{\Delta}t_m}\eta(\boldsymbol{u}_m)\boldsymbol{\nabla}_{\mathcal{T}_m}\mathbf{V}_m\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\widehat{\boldsymbol{\psi}}_m\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$\begin{aligned} D_{2,m} =& \frac{1}{2}\sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa\left(\sqrt{\eta_{\kappa,\mathrm{s}}^n}-\sqrt{\eta(u_\kappa^n)}\right)(V_\kappa-V_\mathrm{s}) \\ &\times\left(\sqrt{\eta_{\kappa,\mathrm{s}'}^n}+\sqrt{\eta(u_\kappa^n)}\right)\left(\psi_\kappa^{n-1}-\psi_{\mathrm{s}'}^{n-1}\right), \\ D_{3,m} =& \frac{1}{2}\sum_{n=1}^{N_m}\Delta t_{n,m}\sum_{\kappa\in\mathcal{M}_m}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa\left(\sqrt{\eta_{\kappa,\mathrm{s}}^n}+\sqrt{\eta(u_\kappa^n)}\right)(V_\kappa-V_\mathrm{s}) \\ &\times\left(\sqrt{\eta_{\kappa,\mathrm{s}'}^n}-\sqrt{\eta(u_\kappa^n)}\right)\left(\psi_\kappa^{n-1}-\psi_{\mathrm{s}'}^{n-1}\right). \end{aligned}$$

It follows from Lemma 4.6, from the uniform convergence of $\boldsymbol{\nabla}_{\mathcal{T}_m}\mathbf{V}_m$ towards $\boldsymbol{\nabla}V$ and of $\boldsymbol{\nabla}_{\mathcal{T}_m,\boldsymbol{\Delta}t_m}\widehat{\boldsymbol{\psi}}_m$ towards $\boldsymbol{\nabla}\psi$ as m tends to $+\infty$ that

$$\lim_{m\to\infty} D_{1,m} = \int_{Q_{t_\mathrm{f}}} \eta(u)\boldsymbol{\nabla}V\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}\psi\,\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{96}$$

Let $\epsilon > 0$, using again the inequality $|ab| \leq \epsilon a^2 + \frac{b^2}{4\epsilon}$, we obtain that

$$|D_{2,m}| \leq \epsilon D_{2,m}' + \frac{1}{16\epsilon}D_{2,m}'', \qquad \forall m \geq 1, \tag{97}$$

where we have set

$$D'_{2,m} = \sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} |a^\kappa_{\mathrm{s},\mathrm{s}'}| \right) \left(\sqrt{\eta^n_{\kappa,\mathrm{s}}} + \sqrt{\eta(u^n_\kappa)} \right)^2 \left(\psi^{n-1}_\kappa - \psi^{n-1}_\mathrm{s} \right)^2,$$

$$D''_{2,m} = \sum_{n=1}^{N_m} \Delta t_{n,m} \sum_{\kappa\in\mathcal{M}_m} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} |a^\kappa_{\mathrm{s},\mathrm{s}'}| \right) \left(\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u^n_\kappa)} \right)^2 \left(V_\kappa - V_\mathrm{s} \right)^2.$$

Define $\overline{\boldsymbol{\eta}}_m = (\overline{\eta}^n_\kappa, \overline{\eta}^n_\mathrm{s})_{\kappa,\mathrm{s}} \in W_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}$ by

$$\overline{\eta}^n_\kappa = \max\left(\eta(u^n_\kappa), \max_{\mathrm{s}\in\mathcal{V}_\kappa} \eta(u^n_\mathrm{s}) \right), \qquad \overline{\eta}^n_\mathrm{s} = 0, \qquad \forall \kappa \in \mathcal{M}_m,\ \forall \mathrm{s} \in \mathcal{V}_m,$$

then one has

$$\left(\sqrt{\eta^n_{\kappa,\mathrm{s}}} + \sqrt{\eta(u^n_\kappa)} \right)^2 \leq 4\overline{\eta}^n_\kappa, \qquad \forall \kappa \in \mathcal{M}_m,\ \forall \mathrm{s} \in \mathcal{V}_\kappa,$$

whence

$$D'_{2,m} \leq C \iint_{Q_{t_\mathrm{f}}} \pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \overline{\boldsymbol{\eta}}_m\, \boldsymbol{\Lambda} \boldsymbol{\nabla} \mathcal{T}_{m,\boldsymbol{\Delta} t_m} \boldsymbol{\psi}_m \cdot \boldsymbol{\nabla} \mathcal{T}_{m,\boldsymbol{\Delta} t_m} \boldsymbol{\psi}_m \mathrm{d}\boldsymbol{x}\mathrm{d}t, \qquad \forall m \geq 1$$

thanks to Lemma A.2. Using Lemma A.8, we know that there exists C depending on the data of the continuous problem and of the regularity factors $\theta^\star$, $\ell^\star$ and $\zeta^\star$ such that

$$\|\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \overline{\boldsymbol{\eta}}_m\|_{L^1(Q_{t_\mathrm{f}})} \leq C, \qquad \forall m \geq 1,$$

while the regularity of ψ ensures that

$$\|\boldsymbol{\nabla} \mathcal{T}_{m,\boldsymbol{\Delta} t_m} \boldsymbol{\psi}_m\|_{L^\infty(Q_{t_\mathrm{f}})} \leq C, \qquad \forall m \geq 1.$$

Therefore, there exists C depending only on the data of the continuous problem and the regularity factors $\theta^\star$, $\ell^\star$ and $\zeta^\star$ such that

$$D'_{2,m} \leq C, \qquad \forall m \geq 1. \tag{98}$$

The term $D''_{2,m}$ can be studied as $C''_{2,m}$ was, leading to

$$\lim_{m\to\infty} D''_{2,m} = 0, \tag{99}$$

whence, taking (98)–(99) into account in (97), one gets that

$$\lim_{m\to\infty} D_{2,m} = 0. \tag{100}$$

Reproducing the calculations carried out for dealing with $D_{2,m}$ allows to show that

$$\lim_{m\to\infty} D_{3,m} = 0. \tag{101}$$

Combining (95)–(96) and (100)–(101), we obtain that

$$\lim_{m\to\infty} D_m = \iint_{Q_{t_\mathrm{f}}} \eta(u) \boldsymbol{\Lambda} \boldsymbol{\nabla} V \cdot \boldsymbol{\nabla} \psi \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{102}$$

Finally, it follows from (83), (84)–(85), (94) and (102) that the limit u of the discrete reconstructions $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m} \boldsymbol{u}_m)_{m\geq 1}$ is a weak solution to the problem (1) in the sense of Definition 1. □

Lemma 4.8. *Let $\boldsymbol{\mu}_m = (\mu^n_\kappa, \mu^n_\mathrm{s})_{\kappa,\mathrm{s},n} \in W_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}$ be defined by (90), then*

$$\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m} \boldsymbol{\mu}_m \underset{m\to\infty}{\longrightarrow} 0 \quad \textit{strongly in } L^2(Q_{t_\mathrm{f}}),$$

Proof. Let $n \in \{1, \dots, N_m\}$, let $\kappa \in \mathcal{M}$, and let $\mathrm{s} \in \mathcal{V}_\kappa$, then, for all $u \in \mathcal{I}^n_{\kappa,\mathrm{s}}$, one has

$$\left(\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u)}\right)^2 \le 2\left(\eta^n_{\kappa,\mathrm{s}} + \eta(u)\right) \le 4\max\left(\eta(u^n_k), \max_{\mathrm{s}\in\mathcal{V}_\kappa} \eta(u^n_{\mathrm{s}})\right).$$

This provides that

$$(\mu^n_\kappa)^2 \le 4\max\left(\eta(u^n_k), \max_{\mathrm{s}\in\mathcal{V}_\kappa} \eta(u^n_{\mathrm{s}})\right), \qquad \forall \kappa \in \mathcal{M},\ \forall n \in \{1, \dots, N_m\}.$$

Using (54b)–(54c), which ensure that

$$m_\kappa \ge \frac{\zeta^\star}{d}\mathrm{meas}(\kappa) \quad \text{and} \quad m_{\mathrm{s}} \ge \frac{\zeta^\star}{d\ell^\star}\mathrm{meas}(\kappa),$$

we deduce that there exists C depending on d, $\ell^\star$ and $\zeta^\star$ such that

$$(\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\boldsymbol{\mu}_m)^2 \le C\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\eta(\boldsymbol{u}_m), \qquad \forall m \ge 1.$$

As a particular consequence of Lemma 4.6, we know that $(\pi_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}\eta(\boldsymbol{u}_m))_{m\ge1}$ is uniformly equi-integrable, whence

$$(\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\boldsymbol{\mu}_m)_{m\ge1} \text{ is uniformly } L^2\text{-equi-integrable.} \tag{103}$$

Let us introduce $\boldsymbol{w}_m = (w^n_\kappa, w^n_{\mathrm{s}})_{\kappa,\mathrm{s},n} \in W_{\mathcal{D}_m,\boldsymbol{\Delta} t_m}$ defined for all $\kappa \in \mathcal{M}_m$, all $\mathrm{s} \in \mathcal{V}_m$ and all $n \in \{1, \dots, N_m\}$ by

$$w^n_{\mathrm{s}} = 0, \quad w^n_\kappa = \max_{\mathrm{s}\in\mathcal{V}_\kappa} |\xi(u^n_\kappa) - \xi(u^n_{\mathrm{s}})|. \tag{104}$$

It follows from a straightforward generalization of Lemma A.9 and from estimate (75) that $\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\boldsymbol{w}_m$ converges strongly in $L^2(Q_{t_\mathrm{f}})$ towards 0. Therefore, up to an unlabeled subsequence, it converges almost everywhere. As a consequence,

$$\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\phi(\boldsymbol{w}_m) \underset{m\to\infty}{\longrightarrow} 0 \quad \text{a.e. in } Q_{t_\mathrm{f}} \tag{105}$$

for all continuous function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that $\phi(0) = 0$.

Let $n \in \{1, \dots, N_m\}$, $\kappa \in \mathcal{M}$ and $\mathrm{s} \in \mathcal{V}_\kappa$. Assume first that $0 \notin \mathcal{I}^n_{\kappa,\mathrm{s}}$, then then monotonicity of η on $\mathcal{I}^n_{\kappa,\mathrm{s}}$ yields, for all $u \in \mathcal{I}^n_{\kappa,\mathrm{s}}$,

$$\left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u)}\right| \le \left|\sqrt{\eta(u^n_\kappa)} - \sqrt{\eta(u^n_{\mathrm{s}})}\right| \le \varpi\left(|\xi(u^n_\kappa) - \xi(u^n_{\mathrm{s}})|\right) \le \varpi(w^n_\kappa),$$

where ϖ is a (non-decreasing) modulus of continuity of $\sqrt{\eta\circ\xi^{-1}}$ (cf. (82)). On the other hand, if $0 \in \mathcal{I}^n_{\kappa,\mathrm{s}}$, then for all $u \in \mathcal{I}^n_{\kappa,\mathrm{s}}$, one has

$$\left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u)}\right| \le \max\left(\sqrt{\eta(u^n_\kappa)}, \sqrt{\eta(u^n_{\mathrm{s}})}\right).$$

Since $\xi(u^n_\kappa)$ and $\xi(u^n_{\mathrm{s}})$ have opposite signs, one gets

$$\max\left(|\xi(u^n_\kappa)|, |\xi(u^n_{\mathrm{s}})|\right) \le |\xi(u^n_\kappa) - \xi(u^n_{\mathrm{s}})|,$$

hence

$$\left|\sqrt{\eta^n_{\kappa,\mathrm{s}}} - \sqrt{\eta(u)}\right| \le \max\left(\varpi(|\xi(u^n_\kappa)|), \varpi(|\xi(u^n_{\mathrm{s}})|)\right) \le \varpi\left(|\xi(u^n_\kappa) - \xi(u^n_{\mathrm{s}})|\right) \le \varpi(w^n_\kappa).$$

It follows from the definition (90) of $\boldsymbol{\mu}_m$ that

$$\mu^n_\kappa \le \varpi(w^n_\kappa), \quad \forall \kappa \in \mathcal{M}_m,\ \forall n \in \{1, \dots, N_m\},$$

hence we obtain that

$$0 \le \pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\boldsymbol{\mu}_m \le \pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\varpi(\boldsymbol{w}_m).$$

Thanks to (105), we obtain that

$$\pi_{\mathcal{M}_m,\boldsymbol{\Delta} t_m}\boldsymbol{\mu}_m \underset{m\to\infty}{\longrightarrow} 0 \quad \text{a.e. in } Q_{t_\mathrm{f}}. \tag{106}$$

In order to conclude, it only remains to remark that (103) and (106) allow us to use Vitali's convergence theorem. □

5. Numerical implementation and results

This section is devoted to the numerical resolution of the nonlinear system (43). First, we discuss in §5.1 the strategy that we used for solving the nonlinear system (43). Then we present in §5.2 2-dimensional cases with analytical solutions in order to illustrate the numerical convergence of the method. We also present a heterogeneous 2-dimensional test case in order to illustrate the robustness of the method.

5.1. **Newton method, Schur complement and time-step adaptation.** The nonlinear system (43) obtained at each time step is solved by a Newton-Raphson algorithm. Given $\boldsymbol{u}^{n-1} \in W_{\mathcal{D}}$, this leads to the computation of a sequence $\left(\boldsymbol{u}^{n,i}\right)_{i\geq 0} \subset W_{\mathcal{D}}$ such that $\boldsymbol{u}^n = \lim_{i\to\infty} \boldsymbol{u}^{n,i}$ is a solution to (43). The variation of the discrete unknowns between two Newton-Raphson algorithm iterations is denoted as follows,

$$\mathrm{d}\boldsymbol{u}^{n,i} \;=\; \left(\,\mathrm{d}u_{\mathrm{s}}^{n,i},\ \mathrm{d}u_{\kappa}^{n,i}\,\right)_{\mathrm{s}\in\mathcal{V},\kappa\in\mathcal{M}} = \boldsymbol{u}^{n,i+1} - \boldsymbol{u}^{n,i}, \qquad \forall i \geq 0.$$

Let us briefly detail the practical implementation of the iterative procedure allowing to deduce $\boldsymbol{u}^n$ from $\boldsymbol{u}^{n-1}$.

(1) In the case where $p(0)$ is finite, the initial guess for the Newton algorithm is, as usual, taken as $\boldsymbol{u}^{n,0} = \left(u_{\mathrm{s}}^{n-1}, u_{\kappa}^{n-1}\right)_{\kappa,\mathrm{s}}$ for all $\mathrm{s} \in \mathcal{V}, \kappa \in \mathcal{M}$. In the singular case $p(0) = -\infty$, it was proved in Lemma 3.7 that the solution $\boldsymbol{u}^n = (u_\kappa^n, u_\mathrm{s}^n)_{\kappa,\mathrm{s}}$ of (43) is such that $\min_{\beta\in\mathcal{M}\cup\mathcal{V}} u_\beta^n > 0$. Therefore, we can initialize the Newton algorithm by

$$\boldsymbol{u}^{n,0} = \left(\max\left(\epsilon, u_{\mathrm{s}}^{n-1}\right), \max\left(\epsilon, u_{\kappa}^{n-1}\right)\right)_{\kappa,\mathrm{s}}.$$

In the computations, we fixed $\epsilon = 10^{-10}$.

(2) The Newton-Raphson algorithm iterations are done until a convergence criterion on the $L^\infty(\Omega)$ norm of the variation of the discrete unknowns is reached or until the maximum number of iterations is reached. At each iteration, the Jacobian matrix resulting of (43) is computed and has the following block structure

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \mathrm{d}\boldsymbol{u}^{n,i} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

where the sub-matrices have the following sizes: $\mathbf{A} \in \mathbb{R}^{\#\mathcal{V}} \otimes \mathbb{R}^{\#\mathcal{V}}$, $\mathbf{B} \in \mathbb{R}^{\#\mathcal{V}} \otimes \mathbb{R}^{\#\mathcal{M}}$, $\mathbf{C} \in \mathbb{R}^{\#\mathcal{M}} \otimes \mathbb{R}^{\#\mathcal{V}}$, and $\mathbf{D} \in \mathbb{R}^{\#\mathcal{M}} \otimes \mathbb{R}^{\#\mathcal{M}}$. The sub-vectors at the right hand side have thus the following sizes: $\mathbf{b}_1 \in \mathbb{R}^{\#\mathcal{V}}$ and $\mathbf{b}_2 \in \mathbb{R}^{\#\mathcal{M}}$. The dependence of the sub-matrices and the sub-vectors w.r.t. n and i was not highlighted here for the ease of notations. A main characteristic of this block structure is that the block $\mathbf{D}$ is a non singular diagonal matrix, thus the Schur complement can be easily computed without fill-in to eliminate the variation of the cell unknowns. This allow to reduce the linear system to the variation of the vertices unknowns as is usual when using the VAG

scheme. The resulting linear system that we have to solve in order to obtain the variation of the vertices unknowns is given by,

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})(\mathrm{d}u_{\mathrm{s}}^{n,i})_{\mathrm{s}\in\mathcal{V}} = \mathbf{b}_1 - \mathbf{B}\mathbf{D}^{-1}\mathbf{b}_2, \tag{107}$$

and then the variation of the cell unknowns can be easily deduced by the matrix-vector product below,

$$(\mathrm{d}u_{\kappa}^{n,i})_{\kappa\in\mathcal{M}} = \mathbf{D}^{-1}(\mathbf{b}_2 - \mathbf{C}(\mathrm{d}u_{\mathrm{s}}^{n,i})_{\mathrm{s}\in\mathcal{V}}).$$

As for the initial step, we have to take into account the singular case at each Newton-Raphson iteration by,

$$\boldsymbol{u}^{n,i+1} \;=\; \max(\boldsymbol{u}^{n,i} + \mathrm{d}\boldsymbol{u}^{n,i}\ ,\ \epsilon).$$

(3) If the Newton-Raphson algorithm stops before the maximum number of iterations is reached, the next time iteration is proceeded by increasing the time step. Otherwise, the current time iteration is recomputed by reducing the time step. The time step is bounded by a maximum value denoted $\Delta t_{\max}$. A maximum number of convergence failures of the nonlinear methods is imposed in order to abort the simulation in case of a non-convergence.

5.2. **Definitions of the test-cases and numerical results.** We present here four 2-dimensional numerical cases where Ω is the unit square. The space domain is discretized by using meshes obtained from a benchmark on anisotropic diffusion problem [61]. In the following numerical experiments, the tensor is defined by

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

where l_x and l_y are chosen constant in Ω, and the exterior potential is defined by $V(\boldsymbol{x}) = -\mathbf{g}\cdot\boldsymbol{x}$ for all $\boldsymbol{x}\in\Omega$ where $\mathbf{g} = (g, 0)^t$ with $g \in \mathbb{R}_+$. The weights of the VAG scheme defined in (27) are defined by $\alpha_{\kappa,\mathrm{s}} = \frac{0.1}{\sharp\mathcal{V}_\kappa}$ for all $\kappa\in\mathcal{M}$, $\mathrm{s}\in\mathcal{V}_\kappa$. We refer to [54, 25] for a discussion on the mass distribution for heterogeneous problems. The linear solver applied to solve (107) is a home-made direct solver using Gaussian elimination with an optimal reordering.

In some of the test cases presented hereafter, Dirichlet boundary conditions are considered instead of no-flux boundary conditions. This allows to construct analytical solutions to the continuous problem and to perform a convergence study. Even though it has not been done in this paper, the convergence proof for the scheme can still be carried out when (sufficiently regular) Dirichlet boundary conditions are considered. However, the gradient flow structure is destroyed and the free energy might not be decreasing anymore in this case.

Errors are computed 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 provides the mesh size h, the initial and maximum time steps, the discrete errors, their associated convergence rate and the minimum value of the discrete solution. It also contains the total (integrated over time) number of Newton-Raphson iterations needed to compute the solution as a indicator of the cost of the numerical method.

5.2.1. *Test 1: Linear Fokker-Planck equation with no-flux boundary condition.* This first test case matches with the problem defined by (1) with the functions $\eta(u) = u$ on $\mathbb{R}_+$ and $p(u) = \log(u)$, and with the gravitational potential $V(x, y) = -gx$ where

the constant g is fixed to 1. Setting $\mathbf{g} = (g, 0)^T$, the problem (1) leads to the linear equation

$$\partial_t u - \boldsymbol{\nabla}\cdot(\boldsymbol{\Lambda}\ (\boldsymbol{\nabla} u - u\mathbf{g})\ ) = 0 \quad \text{in } Q_{t_\mathrm{f}}. \tag{108}$$

We compare the results obtained with the nonlinear scheme (43) with those obtained using the definition of the fluxes

$$\widetilde{F}_{\kappa,\mathrm{s}}(\boldsymbol{u}^n) = \sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a^\kappa_{\mathrm{s},\mathrm{s}'}(u^n_\kappa - u^n_{\mathrm{s}'}) + \frac{u^n_\kappa + u^n_\mathrm{s}}{2} \sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a^\kappa_{\mathrm{s},\mathrm{s}'}(V_\kappa - V_{\mathrm{s}'}), \qquad \forall\kappa\in\mathcal{M},\ \forall \mathrm{s}\in\mathcal{V}_\kappa \tag{109}$$

instead of (43c). The resulting scheme is called the *linear scheme*. The numerical convergence of both schemes has been compared on the following analytical solution (built from a 1-dimensional case):

$$\widetilde{u}(x,y,t) = \exp\left(-\alpha t + \frac{g}{2}x\right)\left(\pi\cos(\pi x) + \frac{g}{2}\sin(\pi x)\right) + \pi\exp\left(g(x-\frac{1}{2})\right), \qquad \forall((x,y),t)\in\Omega\times(0,t_\mathrm{f}), \tag{110}$$

where $\alpha = l_x\ (\pi^2 + \frac{g^2}{4})$. This function satisfies the homogeneous Neumann boundary condition and the property $\widetilde{u}(x,y,t) > 0$ for all $(x,y,t)\in Q_{t_\mathrm{f}}$.

In order to make a numerical convergence study, we have used a family of triangular meshes. These triangle meshes show no symmetry which could artificially increase the convergence rate. 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 4. Although the analytical solution is one-dimensional and the permeability tensor is diagonal, the discrete problem is really 2D because of the non-structured grids. The 2D aspect of the problem is amplified by the choice of a stronger diffusion in the transversal direction. For the tests on triangular grids, the final time t_f has been chosen to 0.25 and an anisotropic tensor has been consider: $l_x = 1$ and $l_y = 10$.

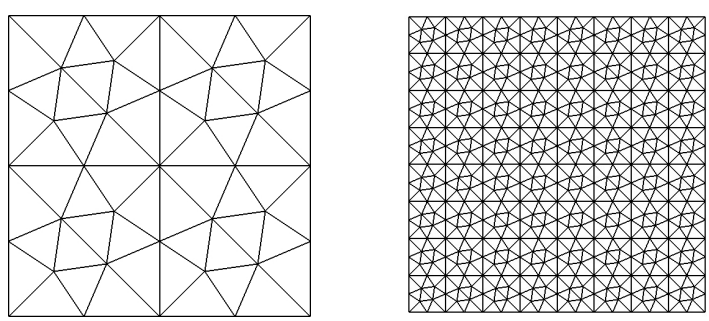

FIGURE 4. First and third mesh used in the numerical examples.

Let us first observe that the numerical order of convergence is close to 2 for both schemes. The nonlinear scheme is of course more expensive than the linear one but it preserves the positivity of the solution, unlike the linear scheme. This numerical behavior is a verification of the theoretical result mentioned in the Lemma 3.7. In the linear case, the number of Newton-Raphson iterations is equal to the number

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.250	37	0.001	0.01024	0.196E-01	-	0.754E-02	-	0.216E+00	-	0.022	204
0.125	129	0.00025	0.00256	0.512E-02	1.935	0.178E-02	2.084	0.600E-01	1.848	0.004	456
0.063	481	0.00006	0.00064	0.129E-02	1.986	0.430E-03	2.050	0.157E-01	1.931	0.001	1307
0.031	1857	0.00002	0.00016	0.324E-03	1.997	0.107E-03	2.007	0.473E-02	1.734	0.000	3935

TABLE 1. Triangles. Nonlinear scheme (43).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.250	37	0.001	0.01024	0.187E-01	-	0.708E-02	-	0.225E+00	-	-0.155	33
0.125	129	0.00025	0.00256	0.469E-02	1.993	0.165E-02	2.100	0.786E-01	1.515	-0.046	106
0.063	481	0.00006	0.00064	0.117E-02	1.999	0.406E-03	2.023	0.228E-01	1.784	-0.012	400
0.031	1857	0.00002	0.00016	0.293E-03	1.999	0.102E-03	1.999	0.611E-02	1.901	-0.003	1570

TABLE 2. Triangles. Linear scheme, fluxes defined by (109).

of time steps. On the finest mesh, the ratio of the number of Newton iterations between the nonlinear and the linear schemes is about 2.5. It seems to be acceptable in cases where preserving the positivity is mandatory.

Now, in order to exhibit the ability of the VAG scheme to deal with general meshes, the same test case has been applied on a so-called Kershaw grid (cf. Figure 5). Instead of an irrelevant numerical convergence study — it is difficult to define a refinement factor for this type of grids —, we aim to give an evidence that the scheme is free energy diminishing (thus positivity preserving) and that the long-time behavior of the continuous problem is preserved at the discrete level by the scheme. The final time t_{f} has been chosen to 250 and an anisotropic tensor has been consider: $l_x = 0.001$ and $l_y = 1$. The results are listed on the Table 3 and we can check again that the nonlinear scheme is positivity preserving despite the irregular grid.

	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	err_{L^1}	err_{L^∞}	$u_{\min}$	#Newton
nonlinear scheme	324	2.E-04	1	3.99E-02	0.404	1.42E-02	8.92E-04	1148
linear scheme	324	2.E-04	1	3.47E-02	0.377	2.01E-02	-1.49E-02	259

TABLE 3. Kershaw grid. Nonlinear and linear scheme, with an anisotropic tensor.

Denoting by $w = \pi \exp\left(g(x - \frac{1}{2})\right)$ the long-time asymptotic of $\widetilde{u}$ defined by (110), then the relative entropy of a function $u : \Omega \to \mathbb{R}_+$ w.r.t. w is defined by

$$E^w(u) = \int_\Omega \left(u \log\left(\frac{u}{w}\right) - u + w\right) \mathrm{d}\boldsymbol{x}. \tag{111}$$

It is simple to verify that

$$\frac{\mathrm{d}}{\mathrm{d}t} E^w(u) \frac{\mathrm{d}}{\mathrm{d}t} E(u) = 0.$$

Therefore the decay of the free energy is equivalent to the decay of the relative entropy. Note that E^w is undefined (or is set to $+\infty$) if $u < 0$ on a positive measure set. It is well known (see e.g. [36, 77]) that the relative entropy $E^w(\widetilde{u}(\cdot, t))$ converges exponentially fast towards 0 as t tends to $+\infty$. Exponential convergence results in the discrete setting were proved for instance in [39, 38, 17] in the case of a monotone discretization of dissipative equation (see also [16]). In order to check this asymptotic behavior at the discrete level, we introduce the discrete relative

entropy $E^w_{\mathcal{D}}(\boldsymbol{u})$ defined for all nonnegative $\boldsymbol{u} = (u_\kappa, u_{\mathrm{s}})_{\kappa,\mathrm{s}} \in W_{\mathcal{D}}$ (i.e., such that $u_\beta \geq 0$ for all $\beta \in \mathcal{M} \cup \mathcal{V}$) by

$$E^w_{\mathcal{D}}(\boldsymbol{u}) = \sum_{\beta \in \mathcal{M}\cup\mathcal{V}} m_\beta \left(u_\beta \log\left(\frac{u_\beta}{w(\boldsymbol{x}_\beta)}\right) - u_\beta + w(\boldsymbol{x}_\beta)\right). \tag{112}$$

The exponential convergence towards equilibrium is recovered as it appears clearly on Figure 5.

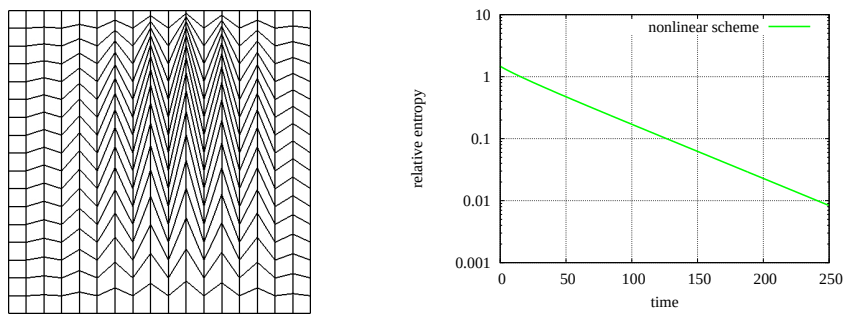

FIGURE 5. Left: Kershaw mesh. Right: Evolution of the relative entropy $t \mapsto E^w_{\mathcal{D}}(\boldsymbol{u}(\cdot,t))$ on a logarithmic scale in function of time.

5.2.2. *Test 2: Porous medium equation with Dirichlet boundary condition.* In this section, we apply our scheme to the case of the anisotropic porous medium equation

$$\partial_t u - \boldsymbol{\nabla}\cdot\left(\boldsymbol{\Lambda}\ \boldsymbol{\nabla}(u^2)\ \right) = 0 \quad \text{in } Q_{t_{\mathrm{f}}} \tag{113}$$

for different choices of functions η and p with $\int^u \eta(a)p'(a)\mathrm{d}a = u^2$, namely

- **(a)** $\eta(u) = 2u^2$ and $p(u) = \log(u)$,
- **(b)** $\eta(u) = 2u$ and $p(u) = u$,
- **(c)** $\eta(u) = 1$ and $p(u) = u^2$.

For the choice **(a)**, the function η is strictly convex. Therefore, the rigorous gradient flow structure of the problem corresponding to this choice of mobility function η is unclear [43]. The pressure function p is singular near 0, hence Lemma 3.7 implies that the corresponding scheme is positivity preserving.

The choice **(b)** with a linear mobility corresponds to the now classical setting highlighted in [86, 77].

Finally, the choice **(c)** corresponds to the usual approach for discretizing the porous medium equation. The corresponding scheme enters into the framework of [48], where its convergence is proved.

The problem is closed here with Dirichlet boundary conditions (destroying by the way the gradient flow structure but not the convergence of the scheme).

Comparison with a one-dimensional analytical solution. The numerical convergence of the three schemes has first been compared thanks to the following analytical solution (again built in 1-dimension),

$$\widehat{u}(x,y,t) = \max\left(2l_x t - x\ ,\ 0\right), \quad \forall((x,y),t) \in Q_{t_{\mathrm{f}}}. \tag{114}$$

Note that (114) is the unique weak solution corresponding to the initial condition $u_0(x,y) = \widehat{u}(x,y,0)$ and to the Dirichlet boundary condition $u_D(x,y,t) = \widehat{u}(x,y,t)$

on $\partial\Omega \times (0, t_{\mathrm{f}})$. Our numerical convergence study makes use of the family of triangular meshes already used for Test 1. Once again, the final time t_{f} is fixed to 0.25 and an anisotropic tensor is given by $l_x = 1$ and $l_y = 10$.

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.523E-02	-	0.997E-03	-	0.105E+00	-	0.000	479
0.153	129	0.00025	0.00256	0.205E-02	1.352	0.344E-03	1.535	0.522E-01	1.013	0.000	1143
0.077	481	0.00006	0.00064	0.898E-03	1.190	0.123E-03	1.490	0.259E-01	1.012	0.000	2218
0.038	1857	0.00002	0.00016	0.380E-03	1.240	0.417E-04	1.554	0.128E-01	1.012	0.000	5652

TABLE 4. Test 2: Choice **(a)** of mobility and pressure functions, convergence towards (114).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.769E-02	-	0.210E-02	-	0.645E-01	-	-0.032	138
0.153	129	0.00025	0.00256	0.263E-02	1.546	0.613E-03	1.775	0.326E-01	0.983	-0.017	383
0.077	481	0.00006	0.00064	0.897E-03	1.554	0.173E-03	1.823	0.164E-01	0.996	-0.009	1246
0.038	1857	0.00002	0.00016	0.306E-03	1.551	0.481E-04	1.849	0.821E-02	0.996	-0.005	4234

TABLE 5. Test 2: Choice **(b)** of mobility and pressure functions, convergence towards (114).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.116E-01	-	0.371E-02	-	0.764E-01	-	-0.065	148
0.153	129	0.00025	0.00256	0.423E-02	1.461	0.116E-02	1.672	0.388E-01	0.977	-0.039	436
0.077	481	0.00006	0.00064	0.149E-02	1.501	0.337E-03	1.788	0.233E-01	0.737	-0.021	1438
0.038	1857	0.00002	0.00016	0.524E-03	1.513	0.932E-04	1.856	0.129E-01	0.856	-0.010	4912

TABLE 6. Test 2: Choice **(c)** of mobility and pressure functions, convergence towards (114).

We observe in Tables 4–6 that second order convergence is destroyed for all the three schemes because of the lack of regularity of the exact solution. As expected, the discrete solution corresponding to the choice **(a)** remains positive while the discrete solutions to the schemes corresponding to the choices **(b)** and **(c)** suffer of undershoots. The choice **(b)** appears to be both cheaper and more accurate than the choice **(c)**, and the amplitude of the undershoots is smaller.

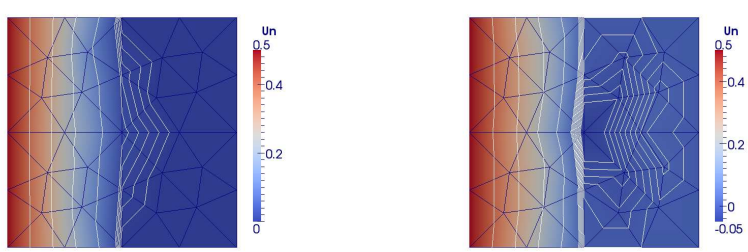

FIGURE 6. Test 2. Coarsest grid. Discrete unknown $(u_{\mathrm{s}})_{\mathrm{s}\in\mathcal{V}_\kappa}$ and its iso-values. Choice **(a)** (left) and **(c)** (right) for η and p.

Figure 6 illustrates the iso-values of the piecewise affine functions defined on the triangular mesh $\mathcal{M}$ reconstructed thanks to its nodal values $(u_{\mathrm{s}}^n)_{\mathrm{s}\in\mathcal{V}}$ for the coarsest triangle grid at the final time t_{f}. For the choice **(a)** of the mobility and the pressure (left), the iso-values are chosen from 0 to 0.025 by step of 0.05 and then from 0.1 to 0.5 by step of 0.1. For the choice **(c)** of the mobility and the pressure (right), the iso-values are taken from -0.025 to 0.025 by step of 0.05 and also from 0.1 to 0.5 by step of 0.1.

Comparison with a two-dimensional analytical solution. We test our approach on the two-dimensional analytical solution

$$\widehat{u}(x,y,t) = \frac{\alpha(x-0.5)^2 + \beta(y-0.5)^2}{1-t} \qquad \text{in } Q_{t_{\mathrm{f}}} \tag{115}$$

of the anisotropic porous medium equation (113), where t_{f} has been set to 0.25. The permeability tensor is still assumed to be diagonal with $l_x = 0.1$ and $l_y = 10$, and we have set $\alpha = \frac{1}{16 l_x}$ and $\beta = \frac{1}{16 l_y}$. The problem is closed with Dirichlet boundary conditions and the initial condition corresponding to (115). The results are gathered in Tables 7–9.

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.270E-02	-	0.114E-02	-	0.120E-01	-	0.000	89
0.153	129	0.00025	0.00256	0.942E-03	1.517	0.381E-03	1.578	0.473E-02	1.341	0.000	223
0.077	481	0.00006	0.00064	0.293E-03	1.688	0.119E-03	1.686	0.163E-02	1.534	0.000	805
0.038	1857	0.00002	0.00016	0.802E-04	1.868	0.334E-04	1.828	0.461E-03	1.826	0.000	3142

TABLE 7. Test 2: Choice **(a)** of mobility and pressure functions, convergence towards (115).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.645E-02	-	0.271E-02	-	0.271E-01	-	-0.027	99
0.153	129	0.00025	0.00256	0.202E-02	1.676	0.828E-03	1.709	0.992E-02	1.447	-0.008	237
0.077	481	0.00006	0.00064	0.604E-03	1.742	0.246E-03	1.753	0.341E-02	1.540	-0.002	801
0.038	1857	0.00002	0.00016	0.161E-03	1.905	0.667E-04	1.882	0.966E-03	1.821	0.000	3140

TABLE 8. Test 2: Choice **(b)** of mobility and pressure functions, convergence towards (115).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.102E-01	-	0.448E-02	-	0.575E-01	-	-0.046	128
0.153	129	0.00025	0.00256	0.321E-02	1.661	0.134E-02	1.739	0.194E-01	1.569	-0.011	250
0.077	481	0.00006	0.00064	0.933E-03	1.783	0.383E-03	1.811	0.553E-02	1.808	-0.003	810
0.038	1857	0.00002	0.00016	0.244E-03	1.933	0.101E-03	1.927	0.147E-02	1.914	-0.001	3140

TABLE 9. Test 2. Choice **(c)** of mobility and pressure functions, convergence towards (115).

As expected, the choice **(a)** leads to a positivity preserving scheme, contrarily to the choices **(b)** and **(c)**. Moreover, the scheme **(a)** is the most accurate and does not come with an additional cost.

5.2.3. *Test 3: Porous medium equation with drift.* In this third test case, we have set $\eta(u) = u$ on $\mathbb{R}_+$ and $p(u) = 2u$ and $g = 1$, leading to the degenerate problem

$$\partial_t u - \boldsymbol{\nabla}\cdot\big(\boldsymbol{\Lambda}\ (\boldsymbol{\nabla}(u^2) - u\mathbf{g})\ \big) = 0 \quad \text{in } Q_{t_\mathrm{f}}. \tag{116}$$

The problem is endowed with Dirichlet boundary conditions. The tensor $\boldsymbol{\Lambda}$ is chosen to be diagonal with $l_x = 1$ and $l_y = 100$. We compare the results obtained by (43) with those obtained using, instead of (43c), this particular definition of the fluxes

$$\widehat{F}_{\kappa,\mathrm{s}}(\boldsymbol{u}^n) = \sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a^\kappa_{\mathrm{s},\mathrm{s}'}((u^n_\kappa)^2 - (u^n_{\mathrm{s}'})^2) + \frac{u^n_\kappa + u^n_\mathrm{s}}{2}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} a^\kappa_{\mathrm{s},\mathrm{s}'}(V_\kappa - V_{\mathrm{s}'}), \quad \forall\kappa\in\mathcal{M},\ \forall \mathrm{s}\in\mathcal{V}_\kappa. \tag{117}$$

The resulting scheme is called the *quasilinear scheme*. The numerical convergence of both schemes has been compared on the sequence of triangular meshes already used in the previous tests, thanks to the following analytical solution (again built in 1-dimension),

$$\widehat{u}(x,y,t) = \max\left(\beta t - x\ ,\ 0\right), \quad \forall((x,y),t)\in\Omega\times(0,t_\mathrm{f}), \tag{118}$$

with $\beta = l_x(2+g)$. The profile (118) is the unique weak solution corresponding to the initial condition $u_0(x,y) = \widehat{u}(x,y,0)$ in Ω and the Dirichlet boundary condition $u_D(x,y,t) = \widehat{u}(x,y,t)$ on $\partial\Omega\times(0,t_\mathrm{f})$.

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.130E-01	-	0.423E-02	-	0.890E-01	-	-0.046	187
0.153	129	0.00025	0.00256	0.495E-02	1.398	0.133E-02	1.675	0.496E-01	0.843	-0.032	552
0.077	481	0.00006	0.00064	0.184E-02	1.428	0.397E-03	1.741	0.283E-01	0.808	-0.017	1609
0.038	1857	0.00002	0.00016	0.660E-03	1.479	0.116E-03	1.771	0.145E-01	0.970	-0.009	5586

TABLE 10. Test 3. Nonlinear scheme (43).

h	$\#\mathcal{V}$	Δt_{init}	$\Delta t_{\max}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	$u_{\min}$	#Newton
0.306	37	0.001	0.01024	0.154E-01	-	0.568E-02	-	0.939E-01	-	-0.068	193
0.153	129	0.00025	0.00256	0.671E-02	1.201	0.213E-02	1.416	0.613E-01	0.615	-0.048	642
0.077	481	0.00006	0.00064	0.271E-02	1.309	0.702E-03	1.600	0.326E-01	0.910	-0.027	2178
0.038	1857	0.00002	0.00016	0.104E-02	1.384	0.212E-03	1.725	0.170E-01	0.938	-0.015	7365

TABLE 11. Test 3. Quasilinear scheme, fluxes defined by (117).

Here again, the convergence orders of both scheme are similar, but strictly lower than 2 because of the lack of regularity of the exact solution. Both schemes violate the positivity of the solution in this case, but the amplitude of the undershoots is smaller for the nonlinear scheme. There is no contradiction here with Lemma 3.7 since p is not singular at $u = 0$. Our nonlinear scheme is slightly more accurate, produces undershoots with a smaller amplitude, and is cheaper than the quasilinear one.

5.2.4. *Test 4. A heterogeneous test case.* The last test aims to illustrate the ability of the scheme to deal with heterogeneous situations. Motivated by an application to complex flows in porous media (see for instance [35, 32]), we test the nonlinear VAG scheme in a slightly more complicated configuration where both the permeability tensor $\boldsymbol{\Lambda}$ and the pressure function p depend on $\boldsymbol{x}$ in a discontinuous way. More precisely, the domain $\Omega = (0,1)^2$ is made of two open subdomains Ω_1 (the *drain*)

and Ω_2 (the *barrier*) with $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ and $\Omega_1 \cap \Omega_2 = \emptyset$ (see Figure 7 for a representation of Ω_1 and Ω_2). The permeability tensor and the pressure function are defined by

$$\boldsymbol{\Lambda}(\boldsymbol{x}) = \begin{cases} \boldsymbol{\Lambda}_1 = \mathbf{I}_d & \text{if } \boldsymbol{x} \in \Omega_1, \\ \boldsymbol{\Lambda}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.01 \end{pmatrix} & \text{if } \boldsymbol{x} \in \Omega_2, \end{cases}$$

and

$$p(u, \boldsymbol{x}) = \begin{cases} p_1(u) = 3\log(u) & \text{if } \boldsymbol{x} \in \Omega_1, \\ p_2(u) = \log(u) & \text{if } \boldsymbol{x} \in \Omega_2. \end{cases}$$

The mobility function is linear and does not depend on $\boldsymbol{x}$, i.e., $\eta(u) = u$. For the sake of simplicity, we have set $V = 0$. At the interface J between Ω_1 and Ω_2, the flux and the pressure are assumed to be continuous, i.e., denoting by u_i the restriction of u to Ω_i and by $\boldsymbol{n}_i$ the normal to J outward w.r.t. Ω_i, we require

$$3\boldsymbol{\Lambda}_1 \boldsymbol{\nabla} u_1 \cdot \boldsymbol{n}_1 + \boldsymbol{\Lambda}_2 \boldsymbol{\nabla} u_2 \cdot \boldsymbol{n}_2 = 0, \quad \text{and} \quad p_1(u_1) = p_2(u_2) \quad \text{on } J \times (0, t_{\mathrm{f}}). \tag{119}$$

The problem is complemented with the boundary conditions

- $p(u, \boldsymbol{x}) = 0$ (hence $u = 1$) on the bottom boundary,
- $p(u, \boldsymbol{x}) = -4$ (hence $u \simeq 0.018$) on the top boundary,
- $\boldsymbol{\Lambda} \boldsymbol{\nabla} u \cdot \boldsymbol{n} = 0$ on the lateral boundaries.

The initial data is chosen at equilibrium, with $p(u_0, \boldsymbol{x}) = -4$ in the whole Ω. Existence and uniqueness for this problem follow from the analysis carried out in [28].

Since u is discontinuous across J (in opposition to the pressure $p(u, \cdot)$ following (119)), it is natural to choose p rather than u as the primary variable of the numerical scheme (cf. [62, 55], we refer to [23] for an alternate strategy that improves robustness) in order to avoid the complex treatment of the jump condition (119) at the interface performed for instance in [28, 29, 47, 24].

The mesh $\mathcal{M}$ is assumed to be compatible with the geometry of Ω, in the sense that $\kappa \in \mathcal{M}$ is either contained in Ω_1 or Ω_2, but $J \cap \kappa = \emptyset$ for all $\kappa \in \mathcal{M}$ (cf. Figure 7). Define the functions $u_\kappa : \mathbb{R} \to (0, \infty)$ as the inverse of $p(\cdot, \boldsymbol{x}_\kappa)$ for all $\kappa \in \mathcal{M}$. The subset of $\mathcal{V}$ made of vertices belonging to the top or bottom boundaries where Dirichlet boundary conditions hold is denoted by $\mathcal{V}_{\mathrm{ext}}$. We also make use of the notations $\mathcal{V}_{\mathrm{int}} = \mathcal{V} \setminus \mathcal{V}_{\mathrm{ext}}$ and $\mathcal{V}_{\kappa,\mathrm{int}} = \mathcal{V}_{\mathrm{int}} \cap \mathcal{V}_\kappa$ for $\kappa \in \mathcal{M}$. The scheme (43) expressed with p as a primary variable consists in finding $\boldsymbol{p} = (p_\kappa^n, p_{\mathrm{s}}^n)_{\kappa,\mathrm{s},n}$ in $W_{\mathcal{D},\boldsymbol{\Delta} t}$ such that for all $n \geq 1$,

$$\begin{cases} \dfrac{u_\kappa(p_\kappa^n) - u_\kappa(p_\kappa^{n-1})}{\Delta t_n} m_\kappa + \displaystyle\sum_{\mathrm{s} \in \mathcal{V}_\kappa} F_{\kappa,\mathrm{s}}^n = 0, & \forall \kappa \in \mathcal{M}, \\ \displaystyle\sum_{\kappa \in \mathcal{M}_{\mathrm{s}}} \dfrac{u_\kappa(p_{\mathrm{s}}^n) - u_\kappa(p_{\mathrm{s}}^{n-1})}{\Delta t_n} m_{\kappa,\mathrm{s}} + \sum_{\kappa \in \mathcal{M}_{\mathrm{s}}} F_{\mathrm{s},\kappa}^n = 0, & \forall \mathrm{s} \in \mathcal{V}_{\mathrm{int}}, \\ F_{\kappa,\mathrm{s}}^n + F_{\mathrm{s},\kappa}^n = 0, & \forall \kappa \in \mathcal{M}, \forall \mathrm{s} \in \mathcal{V}_{\kappa,\mathrm{int}}, \\ F_{\kappa,\mathrm{s}}^n = \sqrt{\eta_{\kappa,\mathrm{s}}^n} \displaystyle\sum_{\mathrm{s}' \in \mathcal{V}_\kappa} a_{\mathrm{s},\mathrm{s}'}^\kappa \sqrt{\eta_{\kappa,\mathrm{s}'}^n} (p_\kappa^n - p_{\mathrm{s}'}^n), & \forall \kappa \in \mathcal{M}, \ \forall \mathrm{s} \in \mathcal{V}_\kappa, \\ \eta_{\kappa,\mathrm{s}}^n = \dfrac{u_\kappa(p_\kappa^n) + u_\kappa(p_{\mathrm{s}}^n)}{2}, & \forall \kappa \in \mathcal{M}, \ \forall \mathrm{s} \in \mathcal{V}_\kappa. \end{cases}$$

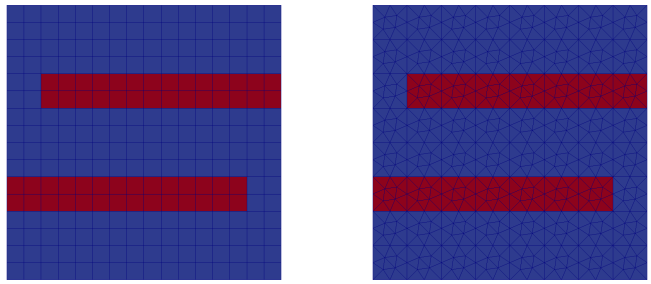

FIGURE 7. Test 4. Illustration of the two sub-domains: the *drain* Ω_1 in blue and the *barriers* Ω_2 in red. Left: cartesian grid. Right: unstructured triangular grid.

We observe on Figure 8 that the results on the triangular mesh (with 481 nodes) and on the cartesian mesh (with 289 nodes) are similar. Moreover, the numbers of Newton-Raphson iteration needed to compute both solutions are of the same order.

APPENDIX A. SOME LEMMAS RELATED TO THE VAG DISCRETIZATION

This appendix gathers lemmas on some properties of the VAG discretization that are independent of the continuous problem (and thus of the scheme). In what follows, $\mathcal{D} = (\mathcal{M}, \mathcal{T})$ denote a discretization of Ω as prescribed in §2.1.1, and $\pi_{\mathcal{T}}, \pi_{\mathcal{M}}, \pi_{\mathcal{D}}$ and $\boldsymbol{\nabla}_{\mathcal{T}}$ are the corresponding reconstruction operators.

Lemma A.1. *For $\kappa \in \mathcal{M}$, let $\mathbf{A}_\kappa = \left(a^\kappa_{\mathrm{s},\mathrm{s}'}\right)_{\mathrm{s},\mathrm{s}'\in\mathcal{V}_\kappa}$ be the matrix defined by (34), then there exists C depending only on $\boldsymbol{\Lambda}$, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ (but not on κ) such that* $\mathrm{Cond}_2(\mathbf{A}_\kappa) \le C$.

Proof. Following [26, Lemma 3.2], there exist $C_1, C_2 > 0$ depending only on $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that, for all $\boldsymbol{u} \in W_{\mathcal{D}}$ and all $\kappa \in \mathcal{M}$, one has

$$C_1 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} (u_\mathrm{s} - u_\kappa)^2 \le \|\boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}\|^2_{L^2(\kappa)} \le C_2 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} \sum_{\mathrm{s}\in\mathcal{V}_\kappa} (u_\mathrm{s} - u_\kappa)^2 ,$$

where h_κ denotes the diameter of the cell $\kappa \in \mathcal{M}$. As a consequence, one has

$$\lambda_\star C_1 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} |\boldsymbol{\delta}_\kappa \boldsymbol{u}|^2 \le \boldsymbol{\delta}_\kappa \boldsymbol{u} \cdot \mathbf{A}_\kappa \boldsymbol{\delta}_\kappa \boldsymbol{u} = \int_\kappa \boldsymbol{\Lambda} \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u} \cdot \boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u} \,\mathrm{d}\boldsymbol{x} \le \lambda^\star C_2 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} |\boldsymbol{\delta}_\kappa \boldsymbol{u}|^2.$$

Since the application $\boldsymbol{\delta}_\kappa : W_{\mathcal{D}} \to \mathbb{R}^{\ell_\kappa}$ is onto, we deduce that

$$\lambda_\star C_1 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} |\boldsymbol{v}|^2 \le \boldsymbol{v} \cdot \mathbf{A}_\kappa \boldsymbol{v} \le \lambda^\star C_2 \frac{\mathrm{meas}(\kappa)}{(h_\kappa)^2} |\boldsymbol{v}|^2, \qquad \forall \boldsymbol{v} \in \mathbb{R}^{\ell_\kappa},$$

and thus that $\mathrm{Cond}_2(\mathbf{A}_\kappa) \le \frac{\lambda^\star C_2}{\lambda_\star C_1}$. □

Lemma A.2. *There exists $C \ge 1$ depending only on $\boldsymbol{\Lambda}$, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that, for all $\kappa \in \mathcal{M}$ and all $\boldsymbol{v} = (v_\mathrm{s})_{\mathrm{s}\in\mathcal{V}_\kappa} \in \mathbb{R}^{\ell_\kappa}$, one has*

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa} \left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa} |a^\kappa_{\mathrm{s},\mathrm{s}'}| \right) (v_\mathrm{s})^2 \le C \, \boldsymbol{v} \cdot \mathbf{A}_\kappa \boldsymbol{v}.$$

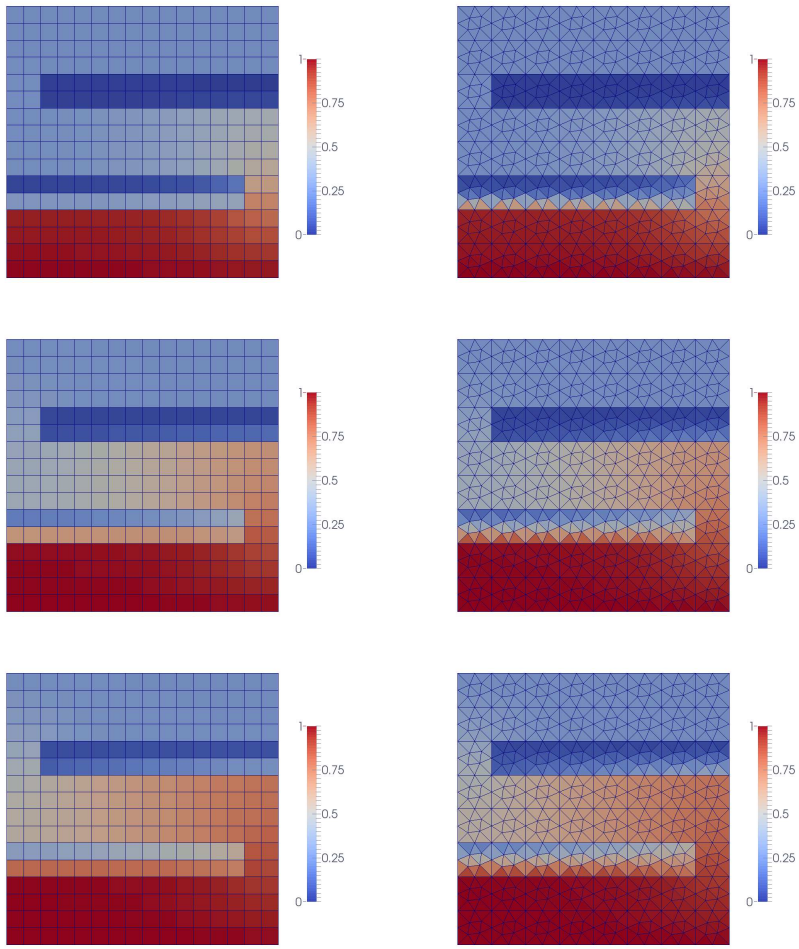

FIGURE 8. Approximation of $u(p, \boldsymbol{x})$ at times $t = 0.05$, $t = 0.2$, and $t = 1$ for the two different meshes.

Proof. Denoting by $\|\cdot\|_q$ the usual matrix q-norm, one has

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(v_\mathrm{s})^2 \leq \|\mathbf{A}_\kappa\|_1|\boldsymbol{v}|^2.$$

Since the dimension of the space $\mathbb{R}^{\ell_\kappa}$ is bounded by $\ell_\mathcal{D}$, there exists C_1 depending only on $\ell_\mathcal{D}$ such that $\|\mathbf{A}_\kappa\|_1 \leq C_1\|\mathbf{A}_\kappa\|_2$, so that

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(v_\mathrm{s})^2 \leq C_1\|\mathbf{A}_\kappa\|_2|\boldsymbol{v}|^2. \tag{120}$$

On the other hand, since $\mathbf{A}_\kappa$ is symmetric definite and positive, one has

$$\boldsymbol{v}\cdot\mathbf{A}_\kappa\boldsymbol{v}\geq\frac{\|\mathbf{A}_\kappa\|_2}{\mathrm{Cond}_2(\mathbf{A}_\kappa)}|\boldsymbol{v}|^2.$$

Using Lemma A.1, we obtain that there exists $C_2>0$ depending only on $\mathbf{\Lambda}$, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that

$$\boldsymbol{v}\cdot\mathbf{A}_\kappa\boldsymbol{v}\geq C_2\|\mathbf{A}_\kappa\|_2|\boldsymbol{v}|^2. \tag{121}$$

Putting (120) and (121) together, we conclude the proof of Lemma A.2 by choosing $C=\frac{C_1}{C_2}$. □

Lemma A.3. *Let $\kappa\in\mathcal{M}$ and $\mathbf{A}_\kappa=(a^\kappa_{\mathrm{s},\mathrm{s}'})_{\mathrm{s},\mathrm{s}'\in\mathcal{V}_\kappa}\in\mathbb{R}^{\ell_\kappa\times\ell_\kappa}$ be the matrix defined by (34). Let $\boldsymbol{\mu}_\kappa=(\mu_{\kappa,\mathrm{s}})_{\mathrm{s}\in\mathcal{V}_\kappa}\in\mathbb{R}^{\ell_\kappa}$ and $\boldsymbol{v}\in W_{\mathcal{D}}$, then*

$$\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}(v_\mathrm{s}-v_\kappa)\mu_{\kappa,\mathrm{s}}a^\kappa_{\mathrm{s},\mathrm{s}'}\mu_{\kappa,\mathrm{s}'}(v_{\mathrm{s}'}-v_\kappa)\leq\max_{\mathrm{s}\in\mathcal{V}_\kappa}(\mu_{\kappa,\mathrm{s}})^2\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(v_\mathrm{s}-v_\kappa)^2.$$

Proof. Using $ab\leq\frac{a^2}{2}+\frac{b^2}{2}$, we obtain that

$$\begin{aligned}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}&(v_\mathrm{s}-v_\kappa)\mu_{\kappa,\mathrm{s}}a^\kappa_{\mathrm{s},\mathrm{s}'}\mu_{\kappa,\mathrm{s}'}(v_{\mathrm{s}'}-v_\kappa)\\&\leq\max_{\mathrm{s}\in\mathcal{V}_\kappa}(\mu_{\kappa,\mathrm{s}})^2\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|v_\mathrm{s}-v_\kappa||a^\kappa_{\mathrm{s},\mathrm{s}'}||v_{\mathrm{s}'}-v_\kappa|\\&\leq\frac{\max_{\mathrm{s}\in\mathcal{V}_\kappa}(\mu_{\kappa,\mathrm{s}})^2}{2}\sum_{\mathrm{s}\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(v_\mathrm{s}-v_\kappa)^2\\&\qquad+\frac{\max_{\mathrm{s}\in\mathcal{V}_\kappa}(\mu_{\kappa,\mathrm{s}})^2}{2}\sum_{\mathrm{s}'\in\mathcal{V}_\kappa}\left(\sum_{\mathrm{s}\in\mathcal{V}_\kappa}|a^\kappa_{\mathrm{s},\mathrm{s}'}|\right)(v'_\mathrm{s}-v_\kappa)^2.\end{aligned}$$

One concludes the proof of Lemma A.3 by noticing that, since $\mathbf{A}_\kappa$ is symmetric, the two terms in the right-hand side of the above inequality are equal. □

Lemma A.4. *There exists C depending only on $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that*

$$\mathrm{meas}(T)\leq\mathrm{meas}(\kappa)\leq C\mathrm{meas}(T),\qquad\forall\kappa\in\mathcal{M},\ \forall T\in\mathcal{T}\text{ with }T\subset\kappa.$$

Proof. Let $\kappa\in\mathcal{M}$, then there exist $T_1,\dots,T_r$ simplexes, with $r=\ell_\kappa$ if $d=2$ and $r=2\#\mathcal{E}_\kappa$ if $d=3$, such that

$$\bigcup_{i=1}^{r_\kappa}\overline{T}_i=\overline{\kappa},\qquad T_i\cap T_j=\emptyset\text{ if }\ i\neq j.$$

The Euler-Descartes theorem ensures that $r\leq4(\ell_{\mathcal{D}}-1)$ if $d=3$.

If T_i and T_j share a common edge, one gets that

$$\mathrm{meas}(T_i)\leq\theta^d\,\mathrm{meas}(T_j).$$

Let $i_0,i_1\in\{1,\dots,r_\kappa\}$ be arbitrary but different, we deduce from the previous inequality the following non-optimal estimate:

$$\mathrm{meas}(T_{i_0})\leq\theta^{4(\ell_{\mathcal{D}}-1)d}\,\mathrm{meas}(T_{i_1}).$$

Let $i_{\max}$ be such that $\mathrm{meas}(T_{i_{\max}})=\max_{1\leq i\leq r}\mathrm{meas}(T_i)$, then

$$\mathrm{meas}(\kappa)\leq r\,\mathrm{meas}(T_{i_{\max}})\leq4(\ell_{\mathcal{D}}-1)\theta^{4(\ell_{\mathcal{D}}-1)d}\mathrm{meas}(T_i),\qquad\forall i\in\{1,\dots,r\}.$$

□

We state now a slight generalization of [26, Lemma 3.4], where the same result is proven in the particular case $q = 2$. The straightforward adaptation of the proof given in [26] to the case $q \neq 2$ is left to the reader.

Lemma A.5. *There exists C depending only on $\ell_{\mathcal{D}}$ and $\theta_{\mathcal{T}}$ defined in (24) and (21) respectively such that, for all $\boldsymbol{v} \in W_{\mathcal{D}}$ and all $q \in [1, \infty]$, one has*

$$\|\pi_{\mathcal{D}}\boldsymbol{v} - \pi_{\mathcal{T}}\boldsymbol{v}\|_{L^q(\Omega)} + \|\pi_{\mathcal{D}}\boldsymbol{v} - \pi_{\mathcal{M}}\boldsymbol{v}\|_{L^q(\Omega)} \leq C h_{\mathcal{T}} \|\boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{v}\|_{L^q(\Omega)}.$$

Lemma A.6. *Let $\mathcal{D}$ be a discretization of Ω as introduced in §2.1.1 such that $\zeta_{\mathcal{D}} > 0$, then there exist $C_1 > 0$ depending only on q, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ and C_2 depending moreover on $\zeta_{\mathcal{D}}$ such that*

$$C_1\|\pi_{\mathcal{D}}\boldsymbol{v}\|_{L^q(\Omega)} \leq \|\pi_{\mathcal{T}}\boldsymbol{v}\|_{L^q(\Omega)} \leq C_2\|\pi_{\mathcal{D}}\boldsymbol{v}\|_{L^q(\Omega)}, \qquad \forall \boldsymbol{v} \in W_{\mathcal{D}}. \tag{122}$$

Proof. Let $\widehat{T}$ be a reference tetrahedron, and let $\widehat{v} : \widehat{T} \to \mathbb{R}$ be an affine function with nodal values v_i, $i \in \{1, \dots 4\}$, then for all $q > 0$, there exists C depending on q such that

$$\frac{1}{C}\sum_{i=1}^{4} |v_i|^q \leq \|\widehat{v}\|^q_{L^q(\widehat{T})} \leq C \sum_{i=1}^{4} |v_i|^q.$$

Therefore, using classical properties of the affine change of variable between simplexes, one gets the existence of C depending only on q, $\theta_{\mathcal{T}}$, and $\ell_{\mathcal{D}}$ such that, for all $\boldsymbol{v} \in W_{\mathcal{D}}$,

$$\begin{aligned} \frac{1}{C}\sum_{\kappa\in\mathcal{M}} \operatorname{meas}(\kappa) \left(|v_\kappa|^q + \sum_{\mathrm{s}\in\mathcal{V}_\kappa} |v_{\mathrm{s}}|^q \right) \\ \leq \|\pi_{\mathcal{T}}\boldsymbol{v}\|^q_{L^q(\Omega)} \leq C \sum_{\kappa\in\mathcal{M}} \operatorname{meas}(\kappa) \left(|v_\kappa|^q + \sum_{\mathrm{s}\in\mathcal{V}_\kappa} |v_{\mathrm{s}}|^q \right). \end{aligned} \tag{123}$$

On the other hand, one has

$$\|\pi_{\mathcal{D}}\boldsymbol{v}\|^q_{L^q(\Omega)} = \sum_{\kappa\in\mathcal{M}} m_\kappa |v_\kappa|^q + \sum_{\mathrm{s}\in\mathcal{V}} m_{\mathrm{s}} |v_{\mathrm{s}}|^q.$$

A classical geometrical property and (29) yield

$$m_\kappa \leq \operatorname{meas}(\kappa) = d \int_\Omega \pi_{\mathcal{T}}\mathbf{e}_\kappa(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq \frac{d}{\zeta_{\mathcal{D}}} m_\kappa \qquad \forall \kappa \in \mathcal{M}, \tag{124}$$

and similarly

$$m_{\mathrm{s}} \leq d \int_\Omega \pi_{\mathcal{T}}\mathbf{e}_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq \frac{d}{\zeta_{\mathcal{D}}} m_{\mathrm{s}}, \qquad \forall \mathrm{s} \in \mathcal{V}.$$

Notice now that the following geometrical identity holds:

$$d \int_\Omega \pi_{\mathcal{T}}\mathbf{e}_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\substack{T\in\mathcal{T} \\ \boldsymbol{x}_{\mathrm{s}}\in\partial T}} \operatorname{meas}(T), \qquad \forall \mathrm{s} \in \mathcal{V}.$$

Lemma A.4 yields the existence of $C > 0$ depending on $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that

$$\frac{1}{C}\sum_{\kappa\in\mathcal{M}_{\mathrm{s}}} \operatorname{meas}(\kappa) \leq d \int_\Omega \pi_{\mathcal{T}}\mathbf{e}_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq \sum_{\kappa\in\mathcal{M}_{\mathrm{s}}} \operatorname{meas}(\kappa), \qquad \forall \mathrm{s} \in \mathcal{V},$$

and the result of Lemma A.6 follows. □

Lemma A.7. *Let $\mathcal{D}$ be a discretization of Ω as introduced in §2.1.1 such that $\zeta_{\mathcal{D}} > 0$, then, for all $q \in [1, \infty]$, one has*

$$\|\pi_{\mathcal{M}} \boldsymbol{v}\|_{L^q(\Omega)} \leq \left(\frac{d}{\zeta_{\mathcal{D}}}\right)^{1/q} \|\pi_{\mathcal{D}} \boldsymbol{v}\|_{L^q(\Omega)}, \qquad \forall \boldsymbol{v} \in W_{\mathcal{D}}.$$

Proof. Let $\boldsymbol{v} = (v_\kappa, v_{\mathrm{s}})_{\kappa \in \mathcal{M}, \mathrm{s} \in \mathcal{V}} \in W_{\mathcal{D}}$, then it follows from (124) that

$$\begin{aligned}\|\pi_{\mathcal{M}} \boldsymbol{v}\|^q_{L^q(\Omega)} &= \sum_{\kappa \in \mathcal{M}} \mathrm{meas}(\kappa) \, |v_\kappa|^q \\ &\leq \left(\frac{d}{\zeta_{\mathcal{D}}}\right) \sum_{\kappa \in \mathcal{M}} m_\kappa \, |v_\kappa|^q \leq \left(\frac{d}{\zeta_{\mathcal{D}}}\right) \|\pi_{\mathcal{D}} \boldsymbol{v}\|^q_{L^q(\Omega)}.\end{aligned}$$

□

Lemma A.8. *Let $\boldsymbol{v} = (v_\kappa, v_{\mathrm{s}})_{\kappa, \mathrm{s}} \in W_{\mathcal{D}}$ be such that $v_\beta \geq 0$ for all $\beta \in \mathcal{M} \cup \mathcal{V}$, and define $\overline{\boldsymbol{v}} = (\overline{v}_\kappa, \overline{v}_{\mathrm{s}})_{\kappa, \mathrm{s}} \in W_{\mathcal{D}}$ by*

$$\overline{v}_{\mathrm{s}} = 0, \qquad \overline{v}_\kappa = \max\left(v_\kappa, \max_{\mathrm{s}' \in \mathcal{V}_\kappa} v_{\mathrm{s}'}\right), \qquad \forall \mathrm{s} \in \mathcal{V}, \forall \kappa \in \mathcal{M}.$$

Then there exists C depending only on $\theta_{\mathcal{T}}$, $\ell_{\mathcal{D}}$ and $\zeta_{\mathcal{D}}$ such that

$$\|\pi_{\mathcal{M}} \overline{\boldsymbol{v}}\|_{L^1(\Omega)} \leq C \, \|\pi_{\mathcal{D}} \boldsymbol{v}\|_{L^1(\Omega)}.$$

Proof. Let $\boldsymbol{v} \in W_{\mathcal{D}}$ be a vector with positives coordinates, and let $\overline{\boldsymbol{v}}$ be constructed as above. It follows from the construction of $\overline{\boldsymbol{v}}$ that

$$\overline{v}_\kappa \leq v_\kappa + \sum_{\mathrm{s} \in \mathcal{V}_\kappa} v_{\mathrm{s}}, \qquad \forall \kappa \in \mathcal{M},$$

whence, applying (123) with $q = 1$, one gets

$$\|\pi_{\mathcal{M}} \overline{\boldsymbol{v}}\|_{L^1(\Omega)} \leq C \|\pi_{\mathcal{T}} \boldsymbol{v}\|_{L^1(\Omega)}.$$

The result now directly follows from Lemma A.6. □

Lemma A.9. *Let $\boldsymbol{u} = (u_\kappa, u_{\mathrm{s}})_{\kappa \in \mathcal{M}, \mathrm{s} \in \mathcal{V}} \in W_{\mathcal{D}}$, then for all $\kappa \in \mathcal{M}$, we define $\overline{\boldsymbol{\delta}} \boldsymbol{u} = \left(\overline{\delta}_\kappa \boldsymbol{u}, \overline{\delta}_{\mathrm{s}} \boldsymbol{u}\right)_{\kappa \in \mathcal{M}, \mathrm{s} \in \mathcal{V}} \in W_{\mathcal{D}}$ by*

$$\overline{\delta}_{\mathrm{s}} \boldsymbol{u} = 0 \quad \textit{and} \quad \overline{\delta}_\kappa \boldsymbol{u} = \max_{\mathrm{s}' \in \mathcal{V}_\kappa} |u_\kappa - u_{\mathrm{s}}|, \qquad \forall \kappa \in \mathcal{M}, \ \forall \mathrm{s} \in \mathcal{V},$$

then, for all $q \in [1, \infty]$, there exists C depending only on q, $\theta_{\mathcal{T}}$, and $\ell_{\mathcal{D}}$ such that

$$\left\|\pi_{\mathcal{M}} \overline{\boldsymbol{\delta}} \boldsymbol{u}\right\|_{L^q(\Omega)} \leq C h_{\mathcal{T}} \|\boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}\|_{L^q(\Omega)}. \tag{125}$$

Proof. Let $\kappa \in \mathcal{M}$ and $\mathrm{s} \in \mathcal{V}_\kappa$, then there exists a simplicial sub-element $T \in \mathcal{T}$ of $\kappa \in \mathcal{M}$ such that $\boldsymbol{x}_\kappa$ and $\boldsymbol{x}_{\mathrm{s}}$ are vertices of T. Then it follows from classical finite element arguments (see e.g. [40, 46]) that

$$\mathrm{meas}(T)^{1/q} |u_\kappa - u_{\mathrm{s}}| \leq c \frac{(h_T)^2}{\rho_T} \|\boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}\|_{L^q(T)} \leq C h_{\mathcal{T}} \|\boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}\|_{L^q(\kappa)},$$

where c depends only on the dimension d and on q, while C depends additionally on $\theta_{\mathcal{T}}$. Thanks to Lemma A.4, we get the existence of C depending on d, q, $\theta_{\mathcal{T}}$ and $\ell_{\mathcal{D}}$ such that,

$$\mathrm{meas}(\kappa)^{1/q} |u_\kappa - u_{\mathrm{s}}| \leq C h_{\mathcal{T}} \|\boldsymbol{\nabla}_{\mathcal{T}} \boldsymbol{u}\|_{L^q(\kappa)}, \quad \forall \kappa \in \mathcal{M}, \ \forall \mathrm{s} \in \mathcal{V}_\kappa.$$

Summing over $\kappa \in \mathcal{M}$ provides that (125) holds. □

Acknowledgements. The authors are grateful to the anonymous referees for their valuable comments on the paper. They also warmly thank Flore Nabet and Thomas Rey for their precious feedback.

References

[1] M. Agueh. Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory. *Adv. Differential Equations*, 10(3):309–360, 2005.

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

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[4] L. Ambrosio, E. Mainini, and S. Serfaty. Gradient flow of the Chapman-Rubinstein-Schatzman model for signed vortices. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 28(2):217–246, 2011.

[5] L. Ambrosio and S. Serfaty. A gradient flow approach to an evolution problem arising in superconductivity. *Comm. Pure Appl. Math.*, 61(11):1495–1539, 2008.

[6] B. Andreianov. Time compactness tools for discretized evolution equations and applications to degenerate parabolic PDEs. In *Finite volumes for complex applications. VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 21–29. Springer, Heidelberg, 2011.

[7] B. Andreianov and F. Bouhsiss. Uniqueness for an elliptic-parabolic problem with Neumann boundary condition. *J. Evol. Equ.*, 4(2):273–295, 2004.

[8] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. HAL: hal-01142499, 2015.

[9] 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.

[10] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov. *Boundary value problems in mechanics of nonhomogeneous fluids*, volume 22 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1990. Translated from the Russian.

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

[12] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[13] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.

[14] J.-D. Benamou, G. Carlier, and M. Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. HAL: hal-01245184, 2015.

[15] J.-D. Benamou, G. Carlier, Q. Mérigot, and E. Oudet. Discretization of functionals involving the Monge-Ampère operator. *Numer. Math.*, online first:1–26, 2015.

[16] M. 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.

[17] M. Bessemoulin-Chatard and C. Chainais-Hillairet. Exponential decay of a finite volume scheme to the thermal equilibrium for drift-diffusion systems. HAL: hal-01250709, 2016.

[18] M. Bessemoulin-Chatard and F. Filbet. A finite volume scheme for nonlinear degenerate parabolic equations. *SIAM J. Sci. Comput.*, 34(5):B559–B583, 2012.

[19] A. Blanchet. A gradient flow approach to the Keller-Segel systems. RIMS Kokyuroku's lecture notes, vol. 1837, pp. 52–73, June 2013.

[20] A. Blanchet, V. Calvez, and J. A. Carrillo. Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM J. Numer. Anal.*, 46(2):691–721, 2008.

[21] F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012.

[22] F. Bolley, I. Gentil, and A. Guillin. Uniform convergence to equilibrium for granular media. *Arch. Ration. Mech. Anal.*, 208(2):429–445, 2013.

[23] K. Brenner and C. Cancès. Improving Newton's method performance by parametrization: the case of Richards equation. HAL: hal-01342386.

[24] 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.

[25] K. Brenner, Groza M., C. Guichard, and R. Masson. Vertex approximate gradient scheme for hybrid dimensional two-phase Darcy flows in fractured porous media. *ESAIM Math. Model. Numer. Anal.*, 49(2):303–330, 2015.

[26] 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.

[27] E. Burman and A. Ern. Discrete maximum principle for galerkin approximations of the laplace operator on arbitrary meshes. *C. R. Acad. Sci. Paris Sér. I Math.*, 338(8):641–646, 2004.

[28] C. Cancès. Nonlinear parabolic equations with spatial discontinuities. *NoDEA Nonlinear Differential Equations Appl.*, 15(4-5):427–456, 2008.

[29] C. Cancès. Finite volume scheme for two-phase flow in heterogeneous porous media involving capillary pressure discontinuities. *M2AN Math. Model. Numer. Anal.*, 43:973–1001, 2009.

[30] C. Cancès, M. Cathala, and C. Le Potier. Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations. *Numer. Math.*, 125(3):387–417, 2013.

[31] C. Cancès and T. Gallouët. On the time continuity of entropy solutions. *J. Evol. Equ.*, 11(1):43–55, 2011.

[32] C. Cancès, T. O. Gallouët, and L. Monsaingeon. The gradient flow structure of immiscible incompressible two-phase flows in porous media. *C. R. Acad. Sci. Paris Sér. I Math.*, 353:985–989, 2015.

[33] C. Cancès and C. Guichard. Entropy-diminishing CVFE scheme for solving anisotropic degenerate diffusion equations. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 187–196. Springer, Cham, 2014.

[34] C. Cancès and C. Guichard. Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.*, 85(298):549–580, 2016.

[35] 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.

[36] J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani, and A. Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatsh. Math.*, 133(1):1–82, 2001.

[37] 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.

[38] C. Chainais-Hillairet. Entropy method and asymptotic behaviours of finite volume schemes. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 17–35. Springer, Cham, 2014.

[39] C. Chainais-Hillairet, A. Jüngel, and S. Schuchnigg. Entropy-dissipative discretization of nonlinear diffusion equations and discrete Beckner inequalities. HAL : hal-00924282, 2014.

[40] P. G. Ciarlet. Basic error estimates for elliptic problems. Ciarlet, P. G. & Lions, J.-L. (ed.), in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 17–351, 1991.

[41] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.

[42] C. Dellacherie and P.-A. Meyer. *Probabilities and potential*, volume 29 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam-New York, 1978.

[43] J. Dolbeault, B. Nazaret, and G. Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.

[44] J. Droniou and Ch. Le Potier. Construction and convergence study of schemes preserving the elliptic local maximum principle. *SIAM J. Numer. Anal.*, 49(2):459–490, 2011.

[45] M. Erbar and J. Maas. Gradient flow structures for discrete porous medium equations. *Discrete Contin. Dyn. Syst.*, 34(4):1355–1374, 2014.

[46] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Series*. Springer, New York, 2004.

[47] A. Ern, I. Mozolevski, and L. Schuh. Discontinuous Galerkin approximation of two-phase flows in heterogeneous porous media with discontinuous capillary pressures. *Comput. Methods Appl. Mech. Engrg.*, 199(23-24):1491–1501, 2010.

[48] 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.

[49] 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.

[50] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. *Comput. Geosci.*, 18:285–296, 2014.

[51] 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.

[52] R. Eymard, C. Guichard, and R. Herbin. Benchmark 3D: the VAG scheme. In J. Fořt, J. Fürst, J. Halama, R. Herbin, and F. Hubert, editors, *Finite Volumes for Complex Applications VI Problems & Perspectives*, volume 4 of *Springer Proceedings in Mathematics*, pages 1013–1022. Springer Berlin Heidelberg, 2011.

[53] R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM Math. Model. Numer. Anal.*, 46(2):265–290, 2012.

[54] R. Eymard, C. Guichard, R. Herbin, and R. Masson. Vertex-centred discretization of multiphase compositional Darcy flows on general meshes. *Comput. Geosci.*, 16(4):987–1005, 2012.

[55] R. Eymard, C. Guichard, R. Herbin, and R. Masson. Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. *ZAMM - J. of App. Math. and Mech.*, 94(7-8):560–585, 2014.

[56] R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klöfkorn, and G. Manzini. 3d benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, Proceedings in Mathematics. Springer, 2011.

[57] 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.

[58] 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.

[59] J. Fehrenbach and J.-M. Mirebeau. Sparse non-negative stencils for anisotropic diffusion. *J. Math. Imaging Vision*, 49(1):123–147, 2014.

[60] 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.

[61] 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.

[62] H. Hoteit and A. Firoozabadi. Numerical modeling of two-phase flow in heterogeneous permeable media with different capillarity pressures. *Advances in Water Resources*, 31(1):56–73, 2008.

[63] N. Igbida. Hele-Shaw type problems with dynamical boundary conditions. *J. Math. Anal. Appl.*, 335(2):1061–1078, 2007.

[64] R. Jordan, D. Kinderlehrer, and F. Otto. Free energy and the fokker-planck equation. *Physica D: Nonlinear Phenomena*, 107(2):265–271, 1997.

[65] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[66] I. Kapyrin. A family of monotone methods for the numerical solution of three-dimensional diffusion problems on unstructured tetrahedral meshes. *Dokl. Math.*, 76:734–738, 2007.

[67] E. F. Keller and L. A. Segel. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2):225–234, 1971.

[68] D. Kinderlehrer, L. Monsaingeon, and X. Xu. A Wasserstein gradient flow approach to Poisson-Nernst-Planck equations. arXiv:1501.04437, to appear in ESAIM: COCV.

[69] D. Kinderlehrer and N. J. Walkington. Approximation of parabolic equations using the Wasserstein metric. *M2AN Math. Model. Numer. Anal.*, 33(4):837–852, 1999.

[70] P. Laurençot and B.-V. Matioc. A gradient flow approach to a thin film approximation of the muskat problem. *Calc. Var. Partial Differential Equations*, 47((1-2)):319–341, 2013.

[71] C. Le Potier. Correction non linéaire et principe du maximum pour la discrétisation d'opérateurs de diffusion avec des schémas volumes finis centrés sur les mailles. *C. R. Acad. Sci. Paris*, 348:691–695, 2010.

[72] C. Le Potier. Correction non linéaire d'ordre 2 et principe du maximum pour la discrétisation d'opérateurs de diffusion. *C. R. Math. Acad. Sci. Paris*, 352(11):947–952, 2014.

[73] J. Leray and J. Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup. (3)*, 51:45–78, 1934.

[74] Randall J LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[75] 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.

[76] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. A monotone finite volume method for advection-diffusion equations on unstructured polygon meshes. *J. Comput. Phys.*, 229(11):4017–4032, 2010.

[77] S. Lisini. Nonlinear diffusion equations with variable coefficients as gradient flows in Wasserstein spaces. *ESAIM Control Optim. Calc. Var.*, 15(3):712–740, 2009.

[78] H. Liu and Z. Wang. A free energy satisfying finite difference method for Poisson-Nernst-Planck equations. *J. Comput. Phys.*, 268:363–376, 2014.

[79] H. Liu and Z. Wang. An entropy satisfying discontinuous Galerkin method for nonlinear Fokker-Planck equations. arXiv:1601.02547, 2016.

[80] H. Liu and H. Yu. The entropy satisfying discontinuous Galerkin method for Fokker-Planck equations. *J. Sci. Comput.*, 62:803–830, 2015.

[81] J. Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, 261(8):2250–2292, 2011.

[82] D. Matthes and H. Osberger. Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM Math. Model. Numer. Anal.*, 48(3):697–726, 2014.

[83] D. Matthes and H. Osberger. A convergent Lagrangian discretization for a nonlinear fourth-order equation. *Found. Comput. Math.*, online first:1–54, 2015.

[84] A. Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

[85] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differential Equations*, 131:20–38, 1996.

[86] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[87] M. A. Peletier. Variational modelling: Energies, gradient flows, and large deviations. Lecture Notes, Würzburg. Available at http://www.win.tue.nl/∼mpeletie, Feb. 2014.

[88] G. Peyré. Entropic Approximation of Wasserstein Gradient Flows. *SIAM J. Imaging Sci.*, 8(4):2323–2351, 2015.

[89] F. A. Radu, I. S. Pop, and P. Knabner. Error estimates for a mixed finite element discretization of some degenerate parabolic equations. *Numer. Math.*, 109(2):285–311, 2008.

[90] Z. Sheng and G. Yuan. The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes. *J. Comput. Physics*, 230(7):2588–2604, 2011.

[91] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146:65–96, 1987.

[92] G. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. *J. Comput. Phys.*, 227(12):6288–6312, 2008.

[93] J. Zinsl and D. Matthes. Exponential convergence to equilibrium in a coupled gradient flow system modeling chemotaxis. *Anal. PDE*, 8(2):425–466, 2015.

Clément Cancès (clement.cances@inria.fr). Team RAPSODI, Inria Lille – Nord Europe, 40 av. Halley, F-59650 Villeneuve d'Ascq, France.

Cindy Guichard (guichard@ljll.math.upmc.fr). Sorbonne Universités, UPMC Univ. Paris 06, CNRS, INRIA, CEREMA, UMR 7598, Laboratoire Jacques-Louis Lions, Équipe ANGE, 4, place Jussieu 75005, Paris, France.