Oil & Gas Science and Technology - Rev. IFP Energies nouvelles **73**, 78 (2018)

https://doi.org/10.2516/ogst/2018067

Available online at:
ogst.ifpenergiesnouvelles.fr

**Numerical methods and HPC**
A. Anciaux-Sedrakian and Q. H. Tran (Guest editors)

REGULAR ARTICLE OPEN ACCESS

# Energy stable numerical methods for porous media flow type problems

Clément Cancès*

INRIA, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, 59000 Lille, France

Received: 28 February 2018 / Accepted: 19 September 2018

**Abstract.** Many problems arising in the context of multiphase porous media flows that take the form of degenerate parabolic equations have a dissipative structure, so that the energy of an isolated system is decreasing along time. In this paper, we discuss two approaches to tune a rather large family of numerical method in order to ensure a control on the energy at the discrete level as well. The first methodology is based on upwinding of the mobilities and leads to schemes that are unconditionally positivity preserving but only first order accurate in space. We present a second methodology which is based on the construction of local positive dissipation tensors. This allows to recover a second order accuracy w.r.t. space, but the preservation of the positivity is conditioned to some additional assumption on the nonlinearities. Both methods are based on an underlying numerical method for a linear anisotropic diffusion equation. We do not suppose that this building block is monotone.

## 1 Introduction

### 1.1 Two-phase porous media flows

Incompressible two-phase porous media flows are often modeled by the following set of equations. In the absence of source terms, the volume of the phase $\alpha \in \{\mathrm{n}, \mathrm{w}\}$ (n and w stand for non-wetting and wetting respectively) is locally conserved along time as a consequence of

$$\phi \partial_t s_\alpha + \nabla \cdot \boldsymbol{v}_\alpha = 0, \tag{1}$$

where $\phi$ denotes the porosity of the rock (supposed to be constant w.r.t. time), where $s_\alpha$ denotes the saturation of the phase $\alpha$ and $\boldsymbol{v}_\alpha$ denotes the filtration speed of the phase $\alpha$. It is classically assumed that the phase filtration speeds obey the generalized Darcy law

$$\boldsymbol{v}_\alpha = -\frac{k_{\mathrm{r},\alpha}(s_\alpha)}{\mu_\alpha} \mathbb{K}(\nabla p_\alpha - \rho_\alpha \mathbf{g}). \tag{2}$$

The intrinsic permeability tensor $\mathbb{K}$ of the porous medium is a definite positive and symmetric tensor field whereas the relative permeability $k_{\mathrm{r},\alpha}$ of the phase $\alpha$ is an increasing function of the saturation satisfying $k_{\mathrm{r},\alpha}(0) = 0$ (we neglect the residual saturations). The variations of the viscosities $\mu_\alpha$ and of the densities $\rho_\alpha$ are neglected, and $\mathbf{g} = \nabla(\mathbf{g} \cdot \boldsymbol{x})$ denotes the gravity vector. The phase pressures $\boldsymbol{p} = (p_\mathrm{n}, p_\mathrm{w})$ are the main unknowns of the system together with the saturations $\boldsymbol{s} = (s_\mathrm{n}, s_\mathrm{w})$. Two algebraic relations are imposed to close the problem. The first one is a constraint coming from the fact that the whole pore volume is saturated by the fluid:

$$s_\mathrm{n} + s_\mathrm{w} = 1. \tag{3}$$

The second one links the capillary pressure to the non-wetting phase saturation in a monotone way:

$$p_\mathrm{n} - p_\mathrm{w} \in \pi(s_\mathrm{n}) \tag{4}$$

where $\pi$ is a maximal monotone graph from $[0,1]$ to $\mathbb{R}$ that may also depend on the space variable in the case where the geological environment is made of several different rocks (see for instance [1–5]).

Once complemented by no-flux conditions on the boundary of the porous domain $\Omega$, the model (1)–(4) has a very particular variational structure. As depicted in [6] (see also [7–9]), this problem can be reinterpreted as the generalized gradient flow [10] of the energy

$$\mathcal{E}(\boldsymbol{s}) = \int_\Omega E(\boldsymbol{s}) \phi \mathrm{d}\mathbf{x}, \tag{5}$$

with

$$E(\boldsymbol{s}) = \Pi(s_\mathrm{n}) - \sum_{\alpha \in \{\mathrm{n},\mathrm{w}\}} s_\alpha \rho_\alpha \mathbf{g} \cdot \boldsymbol{x} + \chi(\boldsymbol{s}).$$

This energy is made of the capillary energy, of the gravitational potential energy, and of a contribution related to the constraint (3):

$$\chi(\boldsymbol{s}) = \begin{cases} 0 & \text{if } s_\mathrm{n} + s_\mathrm{w} = 1, \\ +\infty & \text{otherwise.} \end{cases}$$

* Corresponding author: clement.cances@inria.fr

The capillary potential $\Pi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a convex function that is finite in [0,1], infinite outside of [0,1], and satisfies

$$\pi(s) = \partial\Pi(s) \quad \text{for } s \in [0,1], \tag{6}$$

where

$$\partial\Pi(s) = \{p \in \mathbb{R} \,|\, \Pi(\check{s}) - \Pi(s) - p(\check{s} - s) \geq 0 \quad \text{for all } \check{s} \in D(\Pi)\}$$

is the subdifferential of $\Pi$ at $s$. The functional $E : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ is convex and its subdifferential $\partial E(\boldsymbol{s})$ is made of the couples $\mathbf{h} = (h_{\rm n}, h_{\rm w}) = (p_{\rm n} - \rho_{\rm n}\mathbf{g}\cdot\boldsymbol{x}, p_{\rm w} - \rho_{\rm w}\mathbf{g}\cdot\boldsymbol{x})$ such that the relation (4) holds.

We will not go deep into details on the description of the gradient flow structure since it is not the purpose of this paper. We refer to [7, 11] for a more complete discussion on the Wasserstein gradient flow interpretation of porous media flows and to [7, 12] for a general presentation on gradient flows in metric spaces. Let us only stress important points that will motivate our discussion on numerical schemes. First, the gradient flow structure implies that $\mathcal{E}(\boldsymbol{s})$ is a decreasing function of time and that the dynamics aims at maximizing this decay. This yields an energy/dissipation of the form

$$\mathcal{E}(\boldsymbol{s})(t) + \int_0^t \int_\Omega \sum_{\alpha\in\{\rm n,w\}} \frac{k_{{\rm r},\alpha}(s_\alpha)}{\mu_\alpha} \mathbb{K}\boldsymbol{\nabla} h_\alpha \cdot \boldsymbol{\nabla} h_\alpha {\rm d}\mathbf{x}{\rm d}\tau = \mathcal{E}(\boldsymbol{s}^{\rm ini}) \tag{7}$$

for all $t \geq 0$. The energy/dissipation relation (7) can be obtained by multiplying formally the equation (1) by $h_\alpha$ and by summing over $\alpha \in \{\rm n, w\}$. As a consequence, stable steady states $\boldsymbol{s}^\infty$ are local minimizers of $\mathcal{E}$ for which each phase is hydrostatic on its support, *i.e.*,

$$s_\alpha^\infty = 0 \qquad \text{or} \qquad \boldsymbol{\nabla} p_\alpha^\infty = \rho_\alpha \mathbf{g}. \tag{8}$$

In order to compute in a accurate way the long-time behavior of the system (this is very important in the context of basin modeling), the numerical scheme has to be designed to make the discrete counterpart of the energy $\mathcal{E}(\boldsymbol{s})$ decay along time and, as much as possible, to be exact on the equilibrium (8) (see [13]). In the hydrostatic zones, the equilibrium consists in a balance between the diffusion and the convection. Thus we will avoid operator splitting and discretize the convection and the diffusion simultaneously to recover this equilibrium.

### 1.2 A simplified model problem

Our purpose can already be illustrated on the single scalar equation

$$\partial_t s - \boldsymbol{\nabla}\cdot(\eta(s)\mathbb{K}(\boldsymbol{\nabla} p + \Psi)) = 0, \quad \text{with } p \in \pi(s) = \partial\Pi(s), \tag{9}$$

$\Pi$ being a convex and coercive internal energy functional as previously and $\Psi$ being a smooth external (possibly gravitational) potential. The mobility function $\eta$ is nondecreasing on $D(\Pi) \cap \mathbb{R}_+$ and satisfies $\eta(0) = 0$ and $\eta(s) > 0$ if $s > 0$. Here, we used the notation $D(\Pi) = \{s \in \mathbb{R} \,|\, \Pi(s) < \infty\}$ for the domain of $\Pi$ and we assume that $0 \in D(\Pi)$. If $\Pi$ is defined on the whole $\mathbb{R}_+$, *i.e.*, $D(\Pi) \cap \mathbb{R}_+ = \mathbb{R}_+$, then we assume moreover that $\Pi$ is superlinear:

$$\lim_{s\to+\infty} \frac{\Pi(s)}{s} = +\infty.$$

The solutions to (9) corresponding to nonnegative initial data remain nonnegative along time. This property might be destroyed by the numerical approximation, so that we need to extend the definitions of the functions $\Pi$, $\pi$, and $\eta$ for negative values of $s$ when this is possible. In the case where $\partial\Pi(0)$ contains a finite value, *i.e.*, if

$$\pi^\circ(0) := \lim_{\varepsilon\to0^+} \frac{\Pi(\varepsilon) - \Pi(0)}{\varepsilon} > -\infty, \tag{10}$$

then $\Pi$ can be artificially extended on $(-\infty, 0) \cup D(\Pi)$ into a convex function (still denoted by $\Pi$) with a single valued subdifferential at 0, *i.e.*, $\partial\Pi(0) = \pi^\circ(0) = \pi(0)$. This can be done for instance by prescribing

$$\Pi(s) = \Pi(0) + s(\pi(0) + s) \tag{11}$$

hence

$$\pi(s) = \pi(0) + 2s$$

if $s < 0$. The function $\eta$ can for instance be extended by setting $\eta(s) = -s$ if $s < 0$.

The framework of (9) is already interesting since it includes Richards equation for which $s$ is the water saturation and $\Psi = -\rho\mathbf{g}\cdot\boldsymbol{x}$. The capillary energy function $\Pi$ is then defined on $\mathbb{R}_+$ as the antiderivative of the capillary pressure function. In usual models (see [14], pp. 343–345), $D(\Pi) = [0,1]$ and condition (10) does not hold. This also includes a linear Fokker-Planck equation (but written in a nonlinear form) when $\eta(s) = s$ and $\pi(s) = \log(s)$ or models for crowd motions [15]. We are interested in the computation of non-negative solutions corresponding to initial data $s^{\rm ini} \in L^1(\Omega; \mathbb{R}_+)$ with

$$\int_\Omega \Pi(s^{\rm ini}) {\rm d}\boldsymbol{x} < \infty. \tag{12}$$

There is still a generalized gradient flow structure corresponding to (9), the energy being given by

$$\mathcal{E}(s) = \int_\Omega (\Pi(s) + s\Psi) {\rm d}\boldsymbol{x}. \tag{13}$$

The counterpart to (7) is obtained by multiplying formally (9) by $p + \Psi$ and writes

$$\mathcal{E}(s)(t) + \int_0^t \int_\Omega \eta(s)\mathbb{K}\boldsymbol{\nabla}(p+\Psi)\cdot\boldsymbol{\nabla}(p+\Psi){\rm d}\boldsymbol{x}{\rm d}\tau = \mathcal{E}(s^{\rm ini}) < \infty. \tag{14}$$

Similarly to what was discussed in Section 1.1, one aims to build numerical schemes for (9) that are accurate in the long-time limit. The steady states for (9) are now given by

$$s^\infty = 0 \qquad \text{or} \qquad \nabla p^\infty = -\nabla \Psi. \tag{15}$$

One of the difficulties arising in the resolution of the problem (9) is that the variable $p$ and $s$ are related by the maximal monotone graph $\pi$ that may have vertical or horizontal parts. Similar difficulties occur for instance in the context of reactive flows in porous media, *cf.* [16]. For Richards equation, the graph $\pi$ has vertical parts as depicted on Figure 1. The solution $(s,p)$ to the problem (9) can be deduced from the knowledge of $p$, but not from $s$. But the choice of $p$ as a primary unknown in a naive numerical method can yield severe difficulties. Typically mass conservation can be lost, and severe troubles can be encountered in the convergence of the iterative procedure to compute the solution to the nonlinear system arising from the numerical scheme. These difficulties motivated the development of several strategies to optimize the convergence properties of the iterative procedures. An popular approach consist in making used of robust fixed point procedures with linear convergence speed rather than with Newton's method (see for instance [17–22]). There were also important efforts carried out to fix the difficulties of Newton's method [23–25]. Comparisons between the fixed point and the Newton's strategies are presented for instance in [26, 27] (see also [28]). In [29], the authors combine a Picard-type fixed point strategy with Newton's method (*i.e.*, they perform a few fixed points iterations before running Newton's algorithm). An alternative approach would consist in keeping both $s$ and $p$ as unknowns together with the additional equation $p \in \pi(s)$ that is often rephrased as a complementary constraint and then solving the problem with a non-smooth Newton method (see for instance [30–32]). Another classical solution consists in partitioning $\Omega$ at each time $t$ into a part $\Omega_s(t)$ where $s$ is chosen as a primary variable and a part $\Omega_p(t)$ where $p$ is chosen as a primary variable. Switching from one variable to another is then compulsory [33] in this case. This can be seen as a particular case of the approach proposed recently in [34, 35], which consists in parametrizing the graph $\pi$ in a suitable way. More precisely, the graph can be described by two functions $w \mapsto \overline{s}(w)$ and $w \mapsto \overline{p}(w)$ such that

$$p \in \pi(s) \iff \exists\, w \,\text{s.t.}\, s = \overline{s}(w) \quad \text{and} \quad p = \overline{p}(w), \tag{16}$$

*cf.* Figure 1. We assume moreover that the parametrization is non-degenerate in the sense that $\overline{s} + \overline{p}$ is a strictly increasing function.

Then the knowledge of the unphysical variable $w$ allows to reconstruct both $s$ and $p$, so that on can use $w$ as a primary variable in the computations. There are an infinity of possible choices for the parametrization $(\overline{s}, \overline{p})$ of $\pi$. One example is the resolvents of $\pi$ and $\pi^{-1}$, *i.e.*, $\overline{s} = (\mathrm{id} + \pi)^{-1}$ and $\overline{p} = (\mathrm{id} + \pi^{-1})^{-1}$. The idea in [34] is to take advantage of this flexibility and to choose the parametrization in order to optimize the convergence of the Newton's method.

Before discretizing the nonlinear transient problems (1)–(4) or (9), we need to introduce some material concerning the discretization of elliptic equations.

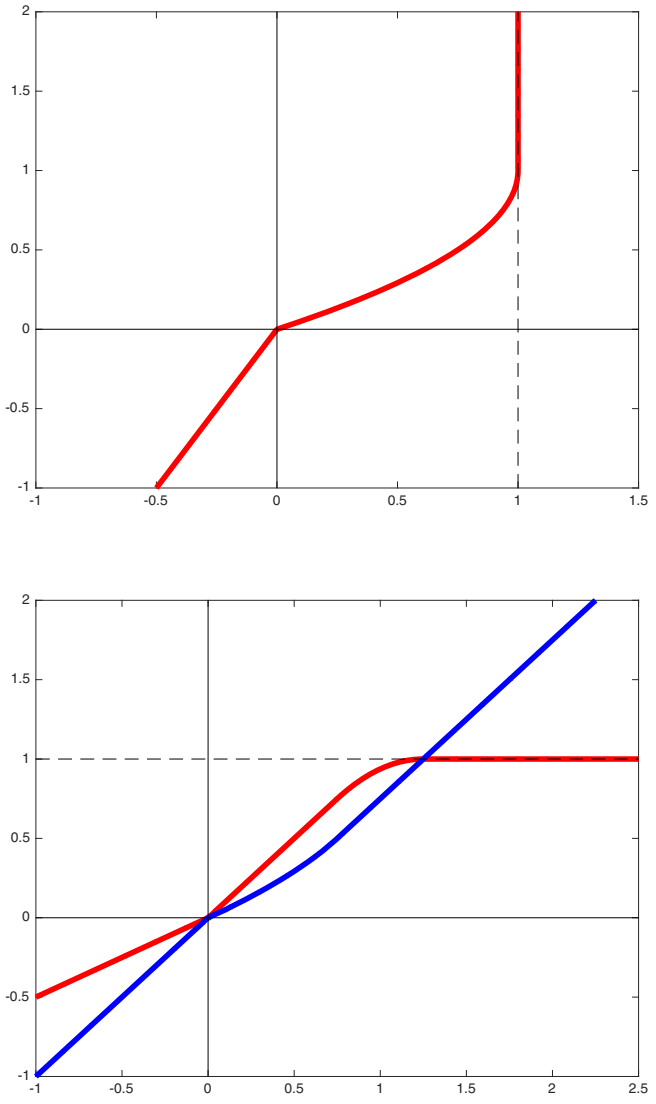

**Fig. 1.** The maximal monotone graph $\pi$ depicted on top is parametrized as $(\overline{s}(w), \overline{p}(w))$ with some monotone functions $\overline{s}$ (red) and $\overline{p}$ (blue) depicted in the bottom picture. Here, $\overline{s}$ and $\overline{p}$ satisfy $\max(\overline{s}'(w), \overline{p}'(w)) = 1$.

### 1.3 A diffusion building block for numerical approximation

The goal of this section is to introduce a rather general framework that can fit to many numerical methods to solve the elliptic equation

$$-\nabla \cdot \mathbb{K} \nabla u = f \,\text{in}\, \Omega \quad \text{with} \int_\Omega f \,\mathrm{d}\boldsymbol{x} = 0 \tag{17}$$

subject to no-flux boundary conditions. There is a huge literature about the numerical resolution of the above problem. Besides classical conforming and non-conforming finite elements (see for instance [36, 37]), let us mention some approaches based on mixed finite elements [38, 39], Multi-Point Flux Approximation (MPFA) finite volumes [40–43], or Discontinuous Galerkin method

[44–46]. The list of aforementioned methods and related publications is of course far from being exhaustive, and some other numerical methods will be discussed in what follows.

Let $\mathcal{U}$ be the set of geometrical entities to which correspond the unknowns (*i.e.*, the cells for cell centered methods, the edges for hybrid methods, or the vertices for vertex centered methods), then we are interested in numerical methods for solving (17) that reduce into a linear system of the form

$$\sum_{L\in\mathcal{U}} a_{KL}(u_K - u_L) = m_K f_K, \quad \forall K \in \mathcal{U}, \tag{18}$$

where $(m_K)_{K\in\mathcal{U}} \subset \mathbb{R}_+$ are such that $\sum_K m_K = |\Omega|$. The compatibility condition on the right-hand side $\boldsymbol{f} = (f_K)_{K\in\mathcal{U}}$ reduces to $\sum_K m_K f_K = 0$. We denote by $\boldsymbol{u} = (u_K)_{K\in\mathcal{U}}$ the vector unknown and by $\mathbb{A}$ the matrix corresponding to the system (18), *i.e.*, $\mathbb{A}\boldsymbol{u} = \mathbf{b}$ where $b_K = m_K f_K$. We require the transmissivity coefficients $a_{KL}$ (and thus the matrix $\mathbb{A}$) to be symmetric

$$a_{KL} = a_{LK}, \quad \forall (K,L) \in \mathcal{U}^2,\ K \neq L. \tag{19}$$

In what follows, $\mathcal{S}$ denotes the set of the couples $(K,L) \in \mathcal{U}^2$ such that the transmissivity $a_{KL}$ is different from 0. Thanks to the symmetry property (19), the scheme (18) is equivalent to: $\forall \boldsymbol{v} = (v_K)_{K\in\mathcal{U}} \in \mathbb{R}^{\mathcal{U}}$,

$$\sum_{(K,L)\in\mathcal{S}} a_{KL}(u_K - u_L)(v_K - v_L) = \sum_{K\in\mathcal{U}} m_K f_K v_K =: \langle \boldsymbol{f}, \boldsymbol{v}\rangle_{0,\mathcal{U}}. \tag{20}$$

Many numerical schemes can write as (18) and (19). This is for instance the case of the classical Two-Point Flux Approximation (TPFA) finite volume scheme [47–49], where

$$a_{KL} = \kappa_{KL}\frac{|\sigma_{KL}|}{\mathrm{dist}(\boldsymbol{x}_K, \boldsymbol{x}_L)} \geq 0 \tag{21}$$

as soon as the cells $K$ and $L$ share an edge $\sigma_{KL}$. In (21), $\boldsymbol{x}_K$ (resp. $\boldsymbol{x}_L$) denotes the center of the cell $K$, whereas the coefficient $m_K$ in (18) is the Lebesgue measure of the cell $K$. The TPFA scheme requires strong assumptions for its consistency. The permeability tensor $\mathbb{K} = \kappa\mathbb{I}$ must be isotropic – $\kappa_{KL}$ in (21) is the harmonic mean of the permeabilities $\kappa_K$ and $\kappa_L$ associated to the cells $K$ and $L$ –, and the edge $\sigma_{KL}$ must be orthogonal to the straight line $[\boldsymbol{x}_K, \boldsymbol{x}_L]$ joining the cell centers $K$ and $L$.

Another classical example of scheme satisfying (18) is the classical conformal $\mathbb{P}_1$ finite elements with mass lumping, which can be seen as a particular box scheme [50]. In this case, the unknowns are located at the vertices $\mathcal{U}$ of a simplicial mesh $\mathcal{T}$ of $\Omega$. Denoting by $\phi_K$ the Lagrange basis function associated to the vertex $K \in \mathcal{U}$, the transmissivities $a_{KL}$ are defined by

$$a_{KL} = -\int_\Omega \mathbb{K}\nabla\phi_K \cdot \nabla\phi_L \mathrm{d}\boldsymbol{x}, \quad (K,L) \in \mathcal{U}^2,\ K \neq L \tag{22}$$

and the lumped mass associated to the vertex $K$ is $m_K = \int_\Omega \phi_K \mathrm{d}\boldsymbol{x}$. Note that in this context, the transmissivities $a_{KL}$ may be negative (for instance in the case where $\mathcal{T}$ does not satisfy the Delaunay condition if $\mathbb{K} = \mathbb{I}$).

The Hybrid-Mixed-Mimetic (HMM) methods [51] containing Hybrid Finite Volumes (HFV, see [52]), Mixed Finite Volumes [53], and Mimetic Finite Differences [54, 55] also enter the framework (18)–(19), as well as Discrete Duality Finite Volumes (DDFV) [56]. We refer to Droniou's review paper [57] and to the book [58] for a more complete overview of the methods entering our framework.

In the case where the transmissivities are nonnegative

$$a_{KL} > 0, \quad \forall (K,L) \in \mathcal{S}, \tag{23}$$

the scheme (18) is monotone and it fulfills the maximum principle. The property (23) is lost as soon as the mesh and the anisotropy tensor do not fulfill restrictive conditions. But for many methods, the transmissivities remain blockwise positive. This means that there exist geometrical entities $M \in \mathcal{M}$ and positive definite symmetric matrices $\mathbb{A}^M \in \mathbb{R}^{\ell_M\times\ell_M}$, $M \in \mathcal{M}$, such that the scheme (20) rewrites

$$\sum_{M\in\mathcal{M}} \boldsymbol{\delta}^M\boldsymbol{u} \cdot \mathbb{A}^M\boldsymbol{\delta}^M\boldsymbol{v} = \langle \boldsymbol{f}, \boldsymbol{v}\rangle_{0,\mathcal{U}}. \tag{24}$$

The vector $\boldsymbol{\delta}^M\boldsymbol{u}$ represents the inner variations of $u$ inside $M \in \mathcal{M}$. Let us illustrate formula (24) on some classical methods. First, for the TPFA scheme, since $a_{KL} > 0$, then one can choose $\mathcal{M} = \mathcal{S}$, as the set of the edges, $\ell_M = 1$, and $\mathbb{A}^M = a_{KL}$ where $M$ is the diamond cell corresponding to the edge $(K,L)$. The case conformal $\mathbb{P}_1$ finite elements is more interesting. In this case, $\mathcal{M} = \mathcal{T}$ denotes the set of the simplices, whereas $\ell_M$ is equal to the space dimension $d$. The vertices of the simplex $M \in \mathcal{T}$ are denoted by $K_0^M, \ldots, K_d^M$, and

$$\boldsymbol{\delta}^M\boldsymbol{v} = \begin{pmatrix} v_{K_1^M} - v_{K_0^M} \\ \vdots \\ v_{K_d^M} - v_{K_0^M} \end{pmatrix} \in \mathbb{R}^d, \quad \forall \boldsymbol{v} \in \mathbb{R}^{\mathcal{U}},$$

whereas the matrix $\mathbb{A}^M = \left(a_{ij}^M\right)_{1\leq i,j\leq d}$ is defined by

$$a_{ij}^M = \int_M \mathbb{K}\nabla\phi_{K_i^M} \cdot \nabla\phi_{K_j^M} \mathrm{d}\boldsymbol{x} = a_{ji}^M, \quad \forall (i,j) \in \{1,\ldots,d\}^2. \tag{25}$$

A crucial property of the local diffusion matrices $\mathbb{A}^M$ is that the condition number of $\mathbb{A}^M$ can be bounded by a quantity depending only on the condition number of $\mathbb{K}^M = \frac{1}{|M|}\int_M \mathbb{K}\mathrm{d}\boldsymbol{x}$ and on the regularity (in Ciarlet's sense [36], see also [37]) of the simplex $M$, *i.e.*,

$$\mathrm{Cond}(\mathbb{A}^M) \leq C, \ \forall M \in \mathcal{M}. \tag{26}$$

The HFV (or SUSHI [52]) also enter the framework (24). In this case, $\mathcal{M}$ is the set of the control volumes that can be quite general (non-convex cells having various number of edges). To each cell $M$, there is one cell unknown

$u_{M,0}$ and $\ell_M$ edge unknowns $u_{M,1},\dots,u_{M,\ell_M}$ where $\ell_M$ denotes the number of edges of the element $M$, and the inner variation vector $\boldsymbol{\delta}^M\boldsymbol{v}$ is given by

$$\delta^M\boldsymbol{v} = \begin{pmatrix} v_{M,1} - v_{M,0} \\ \vdots \\ v_{M,\ell_M} - v_{M,0} \end{pmatrix} \in \mathbb{R}^{\ell_M},\ \forall \boldsymbol{v} \in \mathbb{R}^{\mathcal{U}}.$$

The matrix $\mathbb{A}^M$ is built thanks to a formula similar to (25) but with an approximate gradient that is piecewise constant on the half diamonds. Here again, the conditioning of the matrix only depends on the one of $\mathbb{K}$ and on the regularity of the mesh. The Vertex Approximate Gradient (VAG) scheme [59, 60] has a very similar structure and also enters our framework (see [61]). The main difference with SUSHI is that the unknowns are located at the vertices of the cells $M \in \mathcal{M}$ instead of the edges. A last example concerns the DDFV method that, at least in 2D, also enters the framework (24) as highlighted in [62, 63]. In this context, $\mathcal{M}$ is the set of the diamond cells and once again, the local matrices $\mathbb{A}^M$ have bounded conditioning numbers if $\mathbb{K}$ has a bounded conditioning number and if the mesh fulfills some weak regularity assumption.

Before coming back to our nonlinear parabolic problem, there is a last point that we need to point out here. The coercivity of the numerical method (20) amounts to claiming that the matrix $\mathbb{A}$ is a symmetric definite positive matrix, the lowest eigenvalue of $\mathbb{A}$ being bounded away from 0 uniformly w.r.t. the mesh size:

$$\mathbb{A}\boldsymbol{v}\cdot\boldsymbol{v} = \sum_{(K,L)\in\mathcal{S}} a_{KL}(v_K - v_L)^2 \geq \alpha|\boldsymbol{v}|^2$$

where

$$|\boldsymbol{v}|^2 = \sum_{K\in\mathcal{U}} m_K|v_K|^2, \qquad \forall \boldsymbol{v} \in \mathbb{R}^{\mathcal{U}}.$$

Combining this information with (26), it was proven in [64] that

$$\mathbb{A}\boldsymbol{v}\cdot\boldsymbol{v} \leq \sum_{(K,L)\in\mathcal{S}} |a_{KL}|(v_K - v_L)^2 \leq C\mathbb{A}\boldsymbol{v}\cdot\boldsymbol{v},\ \forall \boldsymbol{v} \in \mathcal{U}, \quad (27)$$

for some $C$ depending only on the mesh regularity and on the anisotropy of the permeability tensor. A similar property was derived for the SUSHI scheme in [52]. The abstract framework of polytopal toolboxes of Droniou *et al.* [58] also allows to derive inequalities of this type.

### 1.4 About the paper content

The (physical) energy stability of numerical methods appears to be a secondary point for a large part of the mathematical literature concerning porous media flows with capillary diffusion. This originates probably from the fact that the mathematical analysis for such problems often relies on the use of the so-called *Kirchhoff transform* and *global pressure* (see for instance [65–68]) which may mask the considerations regarding the physical energy. The question of the convergence of numerical methods is more often addressed. For degenerate parabolic scalar equations, we refer for instance to [69–79], whereas for multiphase porous media flows, we refer to [80–85]. Here again, the list is of course far from being exhaustive.

In the remaining of this paper, we will focus on the construction of schemes for (1)–(4) or more simply for (9) that make the discrete counterpart of the energy decay with time. The goal is to design schemes that capture in an accurate way the long-time behavior of the continuous model, and in particular the equilibriums (8) and (15). Two strategies will be discussed. The first one is detailed in Section 2 and consists in working with the formulation (20) and to use an appropriate upwinding of the mobilities. The second one, which is presented in Section 3, rather uses the formulation (24) and aims at preserving a similar structure even in the presence of a degenerate mobility.

For the sake of simplicity, we do not discuss here about possible source terms or other boundary conditions. Adding these terms to the problem is possible but the energy of the system would no longer decrease in general. However, under reasonable assumptions, the problem remains energy stable. We refer to [86] for a detailed treatment of source terms and inhomogeneous boundary conditions in the context mass-lumped finite elements.

## 2 Upstream mobility schemes

Before addressing the two-phase flow problem (1)–(4), let us first focus on the simpler scalar problem (9).

### 2.1 The scheme for the simplified problem (9)

The goal is to tune the numerical method (18) to approximate the solution to (9) while preserving a good energetic behavior at the discrete level. Concerning the time discretization, we stick to the backward Euler scheme. Let $\overline{s},\overline{p}$ be parametrizations of the graph $\pi$ as in (16), then the numerical scheme writes: $\forall K \in \mathcal{U}$, $\forall n \geq 1$,

$$\frac{\overline{s}(w_K^n) - s_K^{n-1}}{\tau_n} m_K + \sum_{L\in\mathcal{N}_K} a_{KL}\eta_{KL}^n\left(\overline{p}(w_K^n) - \overline{p}(w_L^n) + \Psi_K - \Psi_L\right) = 0, \quad (28)$$

where we have used the notation $\mathcal{N}_K = \{L\,|\,(K,L)\in\mathcal{S}\}$ for the neighbors of $K$. The time step $\tau_n$ is not necessarily uniform and can depend on $n \geq 1$. The initial saturation $\left(s_K^0\right)_{K\in\mathcal{U}}$ is assumed to be given and once $w_K^n$ has been computed for $n \geq 1$, one sets

$$s_K^n = \overline{s}(w_K^n).$$

Following [64] and [87], the mobility $\eta_{KL}^n$ is chosen thank to an upwinding that takes the sign of $a_{KL}$ into account:

$$\eta_{KL}^n = \begin{cases} \eta(\overline{s}(w_K^n)) & \text{if } a_{KL}\left(\overline{p}(w_K^n) - \overline{p}(w_L^n) + \Psi_K - \Psi_L\right) \geq 0, \\ \eta(\overline{s}(w_L^n)) & \text{otherwise.} \end{cases} \quad (29)$$

Note that $\eta^n_{KL} = \eta^n_{LK}$, so that one can write the scheme (28) under the equivalent form

$$\sum_{K\in\mathcal{U}} m_K \frac{\overline{s}(w^n_K) - s^{n-1}_K}{\tau_n} v_K + \sum_{(K,L)\in\mathcal{S}} a_{KL}\eta^n_{KL}\big(\overline{p}(w^n_K) - \overline{p}(w^n_L) + \Psi_K - \Psi_L\big)(v_K - v_L) = 0. \tag{30}$$

In the above formula and as in (20), $\boldsymbol{v}$ is an arbitrary element of $\mathbb{R}^{\mathcal{U}}$.

## 2.2 Some properties of the scheme

As a consequence of the choice (29) for the mobility $\eta^n_{KL}$, the method preserves the non-negativity. More precisely, this means that if $\boldsymbol{w}^n = \big(w^n_K\big)_K$ is a solution of the scheme (28), then so does its projection $\tilde{\boldsymbol{w}}^n = \big(\max(w^n_K, w_\star)\big)_K$ on $[w,\infty)^{\mathcal{U}}$, where $w \in \mathbb{R}\cup\{-\infty\}$ is such that $\overline{s}(w_\star) = 0$ and $\overline{p}(w_\star) = \max\{\pi(0)\}$ (*cf.* Fig. 1). So without loss of generality, we can assume that

$$w^n_K \geq w_\star, \quad \forall K\in\mathcal{U}, \quad \forall n \geq 1. \tag{31}$$

This property still holds for transmissivities $a_{KL}$ violating the monotonicity condition (23). Note that the inequality (31) is of interest only if $w_\star > -\infty$.

Let us now turn to the energy estimate. Choosing $\boldsymbol{v} = \big(\overline{p}(w^n_K) + \Psi_K\big)_K$ in the weak formulation (30) and using the convexity inequality

$$\big(s - \check{s}\big)\overline{p} \geq \Pi(s) - \Pi(\check{s}),$$

for all $(s,\check{s}) \in D(\Pi)^2$ and for all $p \in \partial\Pi(s) = \pi(s)$, we obtain the following discrete counterpart of (14):

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) + \tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL}\eta^n_{KL}\big(\overline{p}(w^n_K) + \Psi_K - \overline{p}(w^n_L) - \Psi_L\big)^2 \leq \mathcal{E}_{\mathcal{U}}\big(\boldsymbol{s}^{n-1}\big), \tag{32}$$

where

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{v}) = \sum_{K\in\mathcal{U}} m_K(\Pi(v_K) + v_K\Psi_K), \quad \forall \boldsymbol{v}\in\mathbb{R}^{\mathcal{U}}. \tag{33}$$

If the monotonicity condition (23) holds (like for instance for TPFA finite volume schemes), the inequality (32) is enough to ensure that the second term is non-negative, thus the discrete energy is diminishing along time, *i.e.*,

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) \leq \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^{n-1}).$$

Moreover, summing over $n$ provides a control on the numerical dissipation

$$0 \leq \sum_{n\geq 1}\tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL}\eta^n_{KL}\big(\overline{p}(w^n_K) + \Psi_K - \overline{p}(w^n_L) - \Psi_L\big)^2 \leq \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^0) - \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) < \infty. \tag{34}$$

This estimate together with (31) is enough to prove the existence of one solution $(\boldsymbol{w}^n)$ to the scheme for all $n \geq 1$. Thanks to the monotonicity of the scheme and the non-degeneracy of the parametrization $(\overline{s},\overline{p})$, the solution is unique (see [34]). Finally, the convergence of the scheme can be proved if the approximation of the diffusion operator is consistent. This requires some classical conditions on the mesh and on the anisotropy tensor for the TPFA scheme [49].

The case where the condition (23) is not satisfied is more intricate. Indeed, the estimate (32) still hold but the fact that 0 is a lower bound in (34) is no longer true in general, deducing a control on the energy and on its dissipation from (32) is not straightforward. Indeed, the second term does not have an obvious sign and we are not able to claim that the dissipation is positive along time. However, it is possible to bound from below the dissipation rate. To do so, let us first remark that the definition (29) of the upstream mobilities implies that

$$\eta^n_{KL} = \begin{cases} \max\limits_{w\in I^n_{KL}} \eta(\overline{s}(w)) & \text{if } a_{KL}\big(\overline{p}(w^n_K) + \Psi_K - \overline{p}(w^n_L) - \Psi_L\big) \\ & \quad \times \big(\overline{\varphi}(w^n_K) - \overline{\varphi}(w^n_L)\big) > 0, \\ \min\limits_{w\in I^n_{KL}} \eta(\overline{s}(w)) & \text{if } a_{KL}\big(\overline{p}(w^n_K) + \Psi_K - \overline{p}(w^n_L) - \Psi_L\big) \\ & \quad \times \big(\overline{\varphi}(w^n_K) - \overline{\varphi}(w^n_L)\big) < 0, \end{cases} \tag{35}$$

whatever the non-decreasing function $\overline{\varphi}$. In the above relations, $I^n_{KL}$ denotes the interval with extremal points $w^n_K$ and $w^n_L$ and we have used the monotonicity of $\eta$. In the case where $\eta\circ\overline{s}$ is bounded (this is natural in the porous media flow context since $\overline{s}$ is), we can take advantage of this formulation with $\overline{\varphi} = \overline{p}$, of the inequality (27), on the Lipschitz regularity of the external potential $\Psi$ to show (*cf.* [88], Sect. 3.2) that

$$\sum_{n=1}^{N}\tau_n \sum_{(K,L)\in\mathcal{S}} |a_{KL}|\eta^n_{KL}\big(\overline{p}(w^n_K) + \Psi_K - \overline{p}(w^n_L) - \Psi_L\big)^2 \leq C\left(1 + \sum_{n=1}^{N}\tau_n\right). \tag{36}$$

A similar inequality can be obtained even in the case where $\eta\circ\overline{s}$ is not bounded under suitable assumptions on the nonlinearities, see for instance ([61], Sect. 3.1). Inequality (36) allows to show that the discrete energy grows at most linearly with time, as well as the existence of (at least) one solution to the scheme thanks to a topological degree argument. Moreover, it allows one to show the convergence of the scheme towards the unique solution to the continuous problem if the mobility function $\eta$ is bounded and if $\pi^{-1}$ is a function (*i.e.*, when there are no hyperbolic degeneracy in the problem (9)) when the mesh size and the time step $\tau_n$ tend to 0. We refer to [88] for the details concerning the convergence analysis of the scheme (28).

The upstream mobility scheme can be tested by other quantities of the form $\overline{\varphi}(\boldsymbol{w}^n)$ thanks to (35). For $s\in[0,1]$, we denote by $\overline{s}^{-1}(s) = [\underline{w}(s),\overline{w}(s)]$ the largest interval of $[-\infty,+\infty]$ such that $\overline{s}(w) = s$ for all

$w \in \overline{s}^{-1}(s)$. Then given a non-decreasing onto function $\varphi : \mathbb{R} \to \mathbb{R}$, there corresponds a convex and coercive function $\Phi : [0,1] \to \mathbb{R} \cup \{+\infty\}$ such that $\partial\Phi(s) = \overline{\varphi} \circ \overline{s}^{-1}(s)$. In particular, there holds

$$(\overline{s}(w) - \overline{s}(\check{w}))\overline{\varphi}(w) \geq \Phi \circ s(w) - \Phi \circ s(\check{w}) \tag{37}$$

for all $w, \check{w} \in [-\infty, +\infty]$.

We can then take benefits of the characterization (35) of the upstream mobility to get enhanced regularity estimates. For instance, choosing

$$\overline{\varphi}(w) = \int^{w} \frac{\overline{s}'(a)}{\eta \circ \overline{s}(a)} \mathrm{d}a, \tag{38}$$

one gets that

$$\sum_{n=1}^{N} \tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL} \left(\overline{p}(w_K^n) - \overline{p}(w_L^n)\right)\left(\overline{s}(w_K^n) - \overline{s}(w_L^n)\right) \leq C\left(1 + \sum_{n=1}^{N} \tau_n\right), \tag{39}$$

provided $\Psi$ is regular enough and the initial entropy is finite:

$$\sum_{K\in\mathcal{U}} m_K \Phi(s_K^0) < \infty.$$

Details for estimate (39) are provided in Appendix A.1. Such an estimate is the corner-stone in the study [89] where the convergence of an upstream mobility scheme was established for the Dupuit approximation of multiphase porous media flows. It was also used in [8], or in [90] where a degenerate Cahn-Hilliard system with a very similar mathematical structure to (1)–(4) (see [91, 92]) was considered.

### 2.3 Convergence of the scheme

Let us illustrate the convergence of the scheme (28) when the mesh size and the time step tend to 0. As a model problem, we consider the very simple convection diffusion equation

$$\partial_t s - \boldsymbol{\nabla} \cdot (s\mathbb{K}\boldsymbol{\nabla}(\log(s) - x_2)) = 0 \quad \text{in } \boldsymbol{\Omega} \times (0, t_\mathrm{f}) \tag{40}$$

where $\boldsymbol{\Omega} = (0,1)^2$, $\boldsymbol{x} = (x_1, x_2)^t$, and $t_\mathrm{f} = 0.05$. The permeability tensor is assumed to be diagonal of the form

$$\mathbb{K} = \begin{pmatrix} \kappa & 0 \\ 0 & 1 \end{pmatrix}.$$

The equation boils down into the very simple linear equation

$$\partial_t s + \boldsymbol{\nabla} \cdot (s\boldsymbol{e}_2 - \mathbb{K}\boldsymbol{\nabla} s) = 0, \tag{41}$$

where $\boldsymbol{e}_2 = (0,1)^t$. We choose the initial condition and the no-flux boundary condition in accordance with the exact solution

$$s_{\mathrm{ex}}(\boldsymbol{x}, t) = \exp\left(-\alpha t + \frac{x_2}{2}\right)\left(\pi\cos(\pi x_2) + \frac{1}{2}\sin(\pi x_2)\right) + \pi\exp\left(x_2 - \frac{1}{2}\right), \tag{42}$$

with $\alpha = \pi^2 + 1/4$. Note that $s^{\mathrm{ini}}(\boldsymbol{x}) = s_{\mathrm{ex}}(\boldsymbol{x}, 0)$ vanishes when $x_2 = 1$.

The solution is computed thanks to a Control Volume Finite Element scheme [87, 88]. This scheme requires conformal triangulations of $\boldsymbol{\Omega}$. We use successively refined Delaunay grids from the FVCA5 benchmark [93]. In the isotropic case (corresponding to $\kappa = 1$), the scheme is monotone, *i.e.*, condition (23) is fulfilled. This is no longer true in the anisotropic case $\kappa = 20$.

We also compute the solution to our scheme with an anisotropy ratio $\kappa = 20$. The numerical solutions are compared with those computed with the following linear scheme with centered convection: $\forall K \in \mathcal{U}$,

$$\frac{s_K^n - s_K^{n-1}}{\Delta t} m_K + \sum_{L\in\mathcal{N}_K} a_{KL}\left(s_K^n - s_L^n + \frac{s_K^n + s_L^n}{2}(\Psi_K - \Psi_L)\right) = 0. \tag{43}$$

Here, $\Psi_K = \boldsymbol{x}_K \cdot \boldsymbol{e}_2$. The scheme (43) is second order accurate in space. In the isotropic case and for fine enough grids (leading to small enough Peclet numbers), the scheme turns out to be monotone. This is no longer the case in the anisotropic case $\kappa = 20$ even though the scheme remains second order accurate in space.

In order to make sure that the error corresponding to the time discretization remains small when compared to the error corresponding to the space discretization, we divide the time step by 4 when the mesh size is divided by 2. We present on Figure 2 the $L^2(\boldsymbol{\Omega} \times (0, t_\mathrm{f}))$ error corresponding to the numerical solutions produced by the schemes (28) and (43) for Delaunay triangular grids from [93].

The method (28) preserves positivity whatever the anisotropy ratio. This is not the case of the linear scheme (43) that produces undershoots in the anisotropic case $\kappa = 20$. As expected because of the upwinding procedure, the scheme (28) is first order accurate in space, *i.e.*,

$$||u - u_h||_{L^2(\boldsymbol{\Omega}\times(0,t_\mathrm{f}))} \leq Ch, \tag{44}$$

where $u_h$ denotes the piecewise constant reconstruction on the dual mesh and where $h$ denotes the mesh size. But it appears on Figure 2 that the constant $C$ appearing in (44) strongly depends on the anisotropy ratio.

To sum up, the method (28) enjoys a very strong stability when used in the non-monotone context, but this is mainly due to the excessive numerical diffusion. Even though the method is still first order accurate w.r.t. space, the constant strongly depends on the anisotropy ratio. This makes the method inefficient in the case of large anisotropy ratios or of poor quality meshes. However, the upstream mobility finite volume scheme remains a very robust method for solving complex but isotropic problems, like for instance some degenerate Cahn-Hilliard systems [90] or multiphase porous media flows [8, 89, 94].

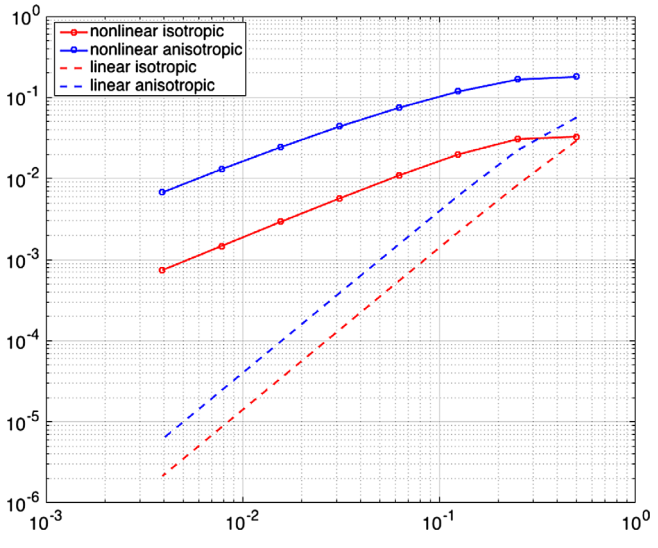


**Fig. 2.** $L^2(\Omega \times (0,t_f))$ error as a function of the mesh size for the scheme (28) in the isotropic case $\kappa = 1$ (solid red) and anisotropic case $\kappa = 20$ (solid blue) corresponding to upstream mobilities (29). Comparison with the solution to the linear scheme (43) in the isotropic case (dashed red) and anisotropic case (dashed blue).

### 2.4 Long-time behavior of the scheme

It is natural to wonder if the numerical scheme is able to reproduce in an accurate way the long-time behavior (15) of the continuous problem. This question is indeed of broad interest for porous media flows, in particular in the context where the time scales are long like for instance in basin modeling or for nuclear waste repository management. Therefore, the study of the long-time behavior of Finite Volume schemes for convection-diffusion problems has been the purpose of several contributions (see, *e.g.*, [13, 62, 95–97]).

We go back to the two-phase flow model presented in Section 1.1 that we discretize on a Delaunay mesh thanks to a classical fully implicit upstream mobility finite volume scheme (see for instance [94] or [8]). In particular, the monotonicity relation (23) is fulfilled and the discrete saturations remain bounded between 0 and 1.

At the continuous level, the relation (8) prescribing the long-time limit boils down to the following alternative:

$$\text{either } \mathrm{s}_{\mathrm{n}}^{\infty} \in \{0,1\} \text{ or } \pi(\mathrm{s}_{\mathrm{n}}^{\infty}) = (\rho_{\mathrm{n}} - \rho_{\mathrm{w}})\mathbf{g}\cdot\boldsymbol{x} + \gamma \quad (45)$$

for some $\gamma \in \mathbb{R}$. The parameter $\gamma$ is determined by the conservation of mass

$$\int_\Omega \phi s_{\mathrm{n}}^{\infty} \mathrm{d}\boldsymbol{x} = \int_\Omega \phi s_{\mathrm{n}}^{\mathrm{ini}} \mathrm{d}\boldsymbol{x}.$$

The steady state (45) can be discretized directly into

$$0 \in \pi(s_{\mathrm{n},K}^{\infty}) + (\rho_{\mathrm{w}} - \rho_{\mathrm{n}})\mathbf{g}\cdot\boldsymbol{x}_K + \gamma', \ \forall K \in \mathcal{U} \quad (46)$$

with $\gamma'$ fixed so that

$$\sum_{K\in\mathcal{U}} m_K \phi_K s_{\mathrm{n},K}^{\infty} = \sum_{K\in\mathcal{U}} m_K \phi_K s_{\mathrm{n},K}^{0}. \quad (47)$$

In the above relation, $\phi_K = \phi(\boldsymbol{x}_K)$ is the discrete porosity. In particular, in the classical case where $\pi : (0,1) \to \mathbb{R}$ is a (single-valued) function, the following alternative holds:

$$\text{either } \mathrm{s}_{\mathrm{n,K}}^{\infty} \in \{0,1\} \quad \text{or} \quad \pi(\mathrm{s}_{\mathrm{n,K}}^{\infty}) = (\rho_{\mathrm{n}} - \rho_{\mathrm{w}})\mathbf{g}\cdot\boldsymbol{x}_{\mathrm{K}} - \gamma'$$

for all $K \in \mathcal{U}$.

Let $\mathcal{E}_{\mathcal{U}}$ be the discrete counterpart of the energy (5), *i.e.*,

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}_{\mathrm{n}}^n) = \sum_{K\in\mathcal{U}} \Big(\Pi(s_{\mathrm{n},K}^n) + s_{\mathrm{n},K}^n(\rho_{\mathrm{w}} - \rho_{\mathrm{n}})\mathbf{g}\cdot\boldsymbol{x}_K\Big) m_K, \quad (48)$$

then the energy is decreasing along time, *i.e.*,

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}_{\mathrm{n}}^n) \le \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}_{\mathrm{n}}^{n-1}), \qquad n \ge 1.$$

One can show (*cf.* Appendix A.2) that $\left(s_{\mathrm{n},K}^{\infty}\right)_{K\in\mathcal{U}}$ is a minimizer of $\mathcal{E}_{\mathcal{U}}$ under the constraint (47). Therefore the relative energy $\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}_{\mathrm{n}}^n) - \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}_{\mathrm{n}}^{\infty})$ is non-negative. If the capillary pressure $\pi$ is an increasing function on $(0,1)$, then the relative energy vanishes if and only if $\boldsymbol{s}_{\mathrm{n}}^n = \boldsymbol{s}_{\mathrm{n}}^{\infty}$. This quantity can be used to illustrate the convergence of $\boldsymbol{s}_{\mathrm{n}}^n$ towards $\boldsymbol{s}_{\mathrm{n}}^{\infty}$ as $n$ tends to $\infty$. In the case depicted on Figure 3, we set $\Omega = (-1/2,1/2)^2$, $\mathbb{K} = \mathbb{I}$, $k_{\mathrm{r},\alpha}(s) = s$, $\mu_{\mathrm{n}} = 10$, $\mu_{\mathrm{w}} = 1$, $\rho_{\mathrm{n}} = 0.87$, $\rho_{\mathrm{w}} = 1$, $\mathbf{g} = (0,-9.81)^T$, $\phi \equiv 1$, $s^{\mathrm{ini}}(\boldsymbol{x}) = e^{-4|\boldsymbol{x}|^2}$, and

$$\pi(s) = \begin{cases} (-\infty, 0] & \text{if } s = 0, \\ s/10 & \text{if } s \in (0,1), \\ [1/10, +\infty) & \text{if } s = 1. \end{cases}$$

We see on Figure 3 that the relative energy converges exponentially fast towards 0 until it reaches the machine precision. This shows both that the energy is effectively dissipated along time and that $\boldsymbol{s}_n^{\infty}$ given by (45) and (46) is a steady solution to the scheme.

## 3 Schemes with local positive dissipation tensors

The upstream mobility numerical schemes presented in the previous section enjoy very nice properties but they are merely first order accurate in space and lack robustness w.r.t. the anisotropy ratio (or to the mesh regularity) as illustrated in Figure 2. This motivates the development of alternative numerical methods with the following specifications.

(i) *No Kirchhoff transform:* the scheme should only involve quantities with a clear physical meaning.
(ii) *Nonlinear stability:* the scheme must fulfill a discrete counterpart of the energy/energy dissipation relation (14).
(iii) *Convergence:* when the mesh size tends to 0 and under mild regularity assumptions, the approximate solution produced by the numerical scheme must converge towards a solution to the PDE (9).
(iv) *Second order accuracy:* if the solution to (9) is smooth enough (say $C^2$), the error between the exact

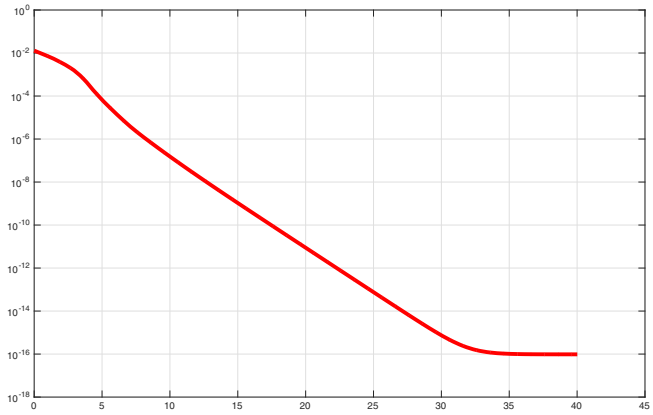

**Fig. 3.** Evolution of the relative energy $\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) - \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^\infty)$ as a function of time. We observe the exponential convergence towards 0 until the machine precision is reached.

and the approximate solution must behave as $h^2$ where $h$ stands for the mesh size.

(v) *Robustness:* the scheme should allow general grids and general anisotropy tensors. The accuracy should not be excessively impacted by the anisotropy ratio or the mesh regularity.

### 3.1 Presentation of the methodology

Assume that the building block (18) is second order accurate, then a natural scheme to fulfill the specifications (i) and (iv) of the previous list for problem (9) is

$$\frac{\overline{s}(w_K^n) - s_K^{n-1}}{\tau_n} m_K + \sum_{L \in \mathcal{N}_K} a_{KL} \eta_{KL}^n \big(\overline{p}(w_K^n) + \Psi_K - \overline{p}(w_L^n) - \Psi_L\big) = 0, \tag{49}$$

with a centered choice for the mobility:

$$\eta_{KL}^n = \frac{\eta\big(\overline{s}\big(w_K^n\big)\big) + \eta\big(\overline{s}\big(w_L^n\big)\big)}{2}, \quad (K,L) \in \mathcal{S}.$$

Multiplying the scheme by $\tau_n\big(\overline{p}(w_K^n) + \Psi_K\big)$ and summing over $K \in \mathcal{U}$ leads once again to

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) + \tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL} \eta_{KL}^n \big(\overline{p}(w_K^n) + \Psi_K - \overline{p}(w_L^n) - \Psi_L\big)^2 \leq \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^{n-1}). \tag{50}$$

Already in the case developed in Section 2 where $\eta_{KL}^n$ was chosen thanks to upwinding, the sign of the second term of the left-hand side was unclear. It was however possible to get a sufficient control to claim that the energy was growing at most linearly, *cf.* (36). This conclusion does not hold any longer in general for a centered choice of the mobilities and no control on the energy can be deduced from (50). Hence the specification (ii) is not satisfied by scheme (49).

In order to correct this, we propose a scheme based on the formalism (24), that is

$$\left\langle \frac{\overline{s}(\boldsymbol{w}^n) - \boldsymbol{s}^{n-1}}{\tau_n}, \boldsymbol{v} \right\rangle_{0,\mathcal{U}} + \sum_{M\in\mathcal{M}} \boldsymbol{\delta}^M\big(\overline{p}(\boldsymbol{w}^n) + \boldsymbol{\Psi}\big) \cdot \mathbb{B}^M(\boldsymbol{w}^n)\boldsymbol{\delta}^M \boldsymbol{v} = 0. \tag{51}$$

The above relation must hold for any $\boldsymbol{v} \in \mathcal{U}$. The matrix $\mathbb{B}^M(\boldsymbol{w}^n) \in \mathbb{R}^{\ell_M \times \ell_M}$ is called the *local dissipation tensor*. It must incorporate the local diffusion $\mathbb{A}^M$ but also the mobilities $\eta(\overline{s}(\boldsymbol{w}^n))$. In order to ensure the dissipation property for the scheme, we want $\mathbb{B}^M(\boldsymbol{w}^n)$ to be symmetric semi-definite positive. In [61], we proposed to choose

$$\mathbb{B}^M(\boldsymbol{v}) = \mathbb{H}^M(\boldsymbol{v})\mathbb{A}^M\mathbb{H}^M(\boldsymbol{w}), \tag{52}$$

with

$$\mathbb{H}^M(\boldsymbol{v}) = \mathrm{diag}\left\{\sqrt{\eta_i^M(\boldsymbol{v})}, 1 \leq i \leq \ell_M\right\}, \quad \forall \boldsymbol{v} \in \mathcal{U}, \tag{53}$$

for the particular choice

$$\eta_i^M(\boldsymbol{v}) = \frac{\eta(\overline{s}(v_{K_i^M})) + \eta(\overline{s}(v_{K_0^M}))}{2}, \ 1 \leq i \leq \ell_M. \tag{54}$$

The fact that the energy is diminishing along time is obtained by choosing $\boldsymbol{v} = \overline{p}(\boldsymbol{w}^n) + \boldsymbol{\Psi}$ and by applying a simple convexity inequality, leading to

$$\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) + \tau_n \sum_{M\in\mathcal{M}} \boldsymbol{\delta}^M\big(\overline{p}(\boldsymbol{w}^n) + \boldsymbol{\Psi}\big) \cdot \mathbb{B}^M(\boldsymbol{w}^n)\boldsymbol{\delta}^M\big(\overline{p}(\boldsymbol{w}^n) + \boldsymbol{\Psi}\big) \leq \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^{n-1}), \tag{55}$$

the second term being non-negative since $\mathbb{B}^M(\boldsymbol{w}^n)$ is semi-definite positive.

Additionally, the method is globally mass conservative, *i.e.*,

$$\sum_{K\in\mathcal{U}} s_K^n m_K = \sum_{K\in\mathcal{U}} s_K^{n-1} m_K = \sum_{K\in\mathcal{U}} s_K^0 m_K, \tag{56}$$

where $s_K^n = \overline{s}(w_K^n)$. This estimate is obtained by choosing $\boldsymbol{v} = \mathbf{1}$ in (50). Let us stress that in the particular cases of the schemes studied in [61, 62, 86], the scheme is also locally conservative since fluxes can be constructed.

In general, the quantity $\eta_i^M(\boldsymbol{v})$ appearing in (52) is chosen as a convex combination of $\Big(\eta(\overline{s}(v_{K_j^M}))\Big)_{0\leq j\leq \ell_M}$. In order to assess that the scheme (50)–(52) converges, the numerical mobilities have to satisfy an additional coercivity condition, that is

$$\eta_i^M(\boldsymbol{v}) \geq \alpha \max\Big(\eta(\overline{s}(v_{K_i^M})), \eta(\overline{s}(v_{K_0^M}))\Big) \tag{57}$$

for some uniform $\alpha > 0$. The condition (56) is clearly satisfied by the choice (53) with $\alpha = 1/2$, but it prohibits the choice $\eta_i^M(\boldsymbol{v}) = \eta(\overline{s}(v_{K_0^M}))$ that would have been quite natural in the context of the SUSHI [52] or VAG [59] schemes.

The implementation of the scheme (50)–(53) can appear to be too involved. An easy way to simplify it is to choose $\eta_i^M(\boldsymbol{v})$ independent on $i$, *i.e.*,

**Table 1.** Choice (a) of mobility and pressure functions, convergence towards (64).

| $h$ | $\#\mathcal{V}$ | $\Delta t_{\text{init}}$ | $\Delta t_{\max}$ | $\text{err}_{L^2}$ | rate | $\text{err}_{L^1}$ | Rate | $\text{err}_{L^\infty}$ | 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 |

$$\eta_i^M(\boldsymbol{v}) = \eta_j^M(\boldsymbol{v}) =: \eta^M(\boldsymbol{v}), \quad 1 \le i,j \le \ell_M.$$

The matrix $\mathbb{H}^M(\boldsymbol{v})$ then reduces to $\sqrt{\eta^M(\boldsymbol{v})}\mathbb{I}_{\ell_M}$ and commutes with $\mathbb{A}^M$, leading to the simpler formula $\mathbb{B}^M(\boldsymbol{v}) = \eta^M(\boldsymbol{v})\mathbb{A}^M$. In view of the constraint (56), a natural choice for $\eta^M(\boldsymbol{v})$ is

$$\eta^M(\boldsymbol{v}) = \frac{1}{\ell_M+1}\sum_{j=0}^{\ell_M}\eta(\overline{s}(v_{K_j^M})) \ge \frac{1}{\ell_M+1}\max\Big(\eta(\overline{s}(v_{K_i^M})),\eta(\overline{s}(v_{K_0^M}))\Big). \tag{58}$$

This choice was successfully used in [86] for a method based on conformal $\mathbb{P}_1$ finite elements with mass lumping (here $\ell_M = d$), and in [62, 63] for a nonlinear DDFV method (here $d = 2$ and $\ell_M = 3$).

### 3.2 Conditional positivity preservation and convergence w.r.t. the grid

The scheme (50) amounts at each time step to a system of nonlinear equations of the form

$$\mathcal{F}_n(\boldsymbol{w}^n) = \boldsymbol{0}, \quad \text{with} \quad \mathcal{F}_n : \mathbb{R}^{\mathcal{U}} \to \mathbb{R}^{\mathcal{U}}. \tag{59}$$

The functions $\eta$, $\overline{s}$ and $\overline{p}$ are uniformly continuous on $\mathbb{R}$, then $\mathcal{F}$ is also uniformly continuous. In order to ensure the existence of a solution $\boldsymbol{w}^n$ to the system, we need some bounds on $\boldsymbol{w}^n$ (that might depend on the mesh and the time step).

In the case where (10) holds, the estimate (55) (in particular the dissipation term) provides a sufficient bound on $\boldsymbol{w}^n$ in order to apply a topological degree argument and to claim that there exists (at least) one solution to the scheme (51). In the more intricate situation where (10) is no longer satisfied, corresponding to the situation where

$$\lim_{s\to 0^+}\pi(s) = -\infty, \tag{60}$$

(with a slight abuse of notation), then one can show that

$$\overline{s}(\boldsymbol{w}^n) \ge \zeta > 0 \tag{61}$$

for some $\zeta$ depending on the time step $\tau_n$ and on the mesh. This is done for instance in [61] of Lemma 3.7 in the context of the VAG scheme, or in [63] of Lemma 3.5 for a DDFV scheme. In both case, the proof strongly relies on the coercivity assumption (57).

As a consequence of (61), the scheme (51) preserves the positivity as soon as (60) holds. This property is unfortunately lost in general when (10) is satisfied. To illustrate this fact, we show results of [61] where the solution of the porous medium equation

$$\partial_t s - \boldsymbol{\nabla}\cdot\left(\mathbb{K}\boldsymbol{\nabla}s^2\right) = 0 \tag{62}$$

is rewritten under the form

$$\partial_t s - \boldsymbol{\nabla}\cdot(\mathbb{K}\eta(s)\boldsymbol{\nabla}\pi(s)) = 0 \tag{63}$$

for three different choices of nonlinearities, that are

(a) $\eta(s) = 1$ and $\pi(s) = |s|s$ (recall that we need to extend the nonlinearities for negative saturations if (10) holds);
(b) $\eta(s) = 2|s|$ and $\pi(s) = s$;
(c) $\eta(s) = 2s^2$ and $\pi(s) = \log(s)$.

The case (a) does not enter our framework since $\eta(0) = 1 \neq 0$, but it is interesting since it corresponds to the most natural approach to solve (62). The condition (10) holds in cases (a) and (b), whereas (60) holds in case (c). Therefore, the positivity of the solutions should be guaranteed only in this last case. To illustrate this fact, let us choose $\boldsymbol{\Omega} = \{(x,y) \in (0,1)^2\}$, $\mathbb{K} = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$ and let us approximate the exact solution to (62) defined by

$$s(x,y,t) = \max(0, 2t - x) \tag{64}$$

thanks to the VAG scheme [61] on successively refined triangular meshes from the FVCA5 benchmark [93]. The problem is here complemented with Dirichlet boundary conditions.

We observe in Tables 1–3 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 (c) remains positive because condition (60) is verified. This is no longer the case for the choices (a) and (b) and undershoots are observed. But the choice (b) appears to be both cheaper and more accurate than the choice (a), whereas the amplitude of the undershoots is reduced.

Let us illustrate again the ability of the approach. We consider the linear and isotropic convection diffusion equation

$$\partial_t s + \boldsymbol{\nabla}\cdot(s\boldsymbol{e}_2 - \boldsymbol{\nabla}s) = 0,$$

**Table 2.** Choice (b) of mobility and pressure functions, convergence towards (63).

| $h$ | $\#\mathcal{V}$ | $\Delta t_{\text{init}}$ | $\Delta t_{\max}$ | $\text{err}_{L^2}$ | Rate | $\text{err}_{L^1}$ | Rate | $\text{err}_{L^\infty}$ | 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 3.** Choice (c) of mobility and pressure functions, convergence towards (63).

| $h$ | $\#\mathcal{V}$ | $\Delta t_{\text{init}}$ | $\Delta t_{\max}$ | $\text{err}_{L^2}$ | Rate | $\text{err}_{L^1}$ | Rate | $\text{err}_{L^\infty}$ | 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 |

that we rewrite under the nonlinear form

$$\partial_t s + \boldsymbol{\nabla} \cdot (s \boldsymbol{\nabla}(\log(s) - x_2)) = 0. \tag{65}$$

We aim to approximate the exact solution (42) thanks to the nonlinear DDFV method proposed in [62, 63]. We compute the approximate solution corresponding to the sequence of Kershaw meshes from [93], see Figure 4. The numerical results are presented in Table 4.

As expected, since $\pi(s) = \log(s)$ fulfills (60), the solution remains positive although the grid is very irregular. The convergence at order 2 on u is not affected by the poor mesh regularity. We observe a super-convergence for the gradient.

Finally, let us show that the methodology presented in this section is often much more robust w.r.t. anisotropy than the methodology presented in Section 2. To this end, we stick to the test-case of the linear Fokker-Planck equation written in a nonlinear form described in Section 2.3. We discretize it with the energy stable $\mathbb{P}_1$ finite element scheme with mass lumping proposed in [86]. In this scheme, the mobilities $\eta^M(\boldsymbol{v})$ are chosen according to formula (58). The final time is set to $t_{\text{f}} = 0.25$. The $L^2((0, t_{\text{f}}) \times \Omega)$ error as a function of the mesh size is plotted on Figure 5. As expected, the method is of order 2 whatever the anisotropy ratio. But it is worth noticing that the accuracy is almost not affected by the anisotropy ratio $\kappa$.

### 3.3 About the long-time behavior of the scheme

We aim now to illustrate the long-time behavior of the scheme. One discretizes the equation (65) complemented with no-flux boundary conditions following the methodology of [61]. In particular, we take $\overline{s} = \text{Id}$, hence we denote by $\boldsymbol{s}^n = \boldsymbol{w}^n$ the vector of the unknown saturations. The scheme (51) then rewrites

$$\left\langle \frac{\boldsymbol{s}^{\text{n}} - \boldsymbol{s}^{n-1}}{\tau}, \boldsymbol{v} \right\rangle_{0,\mathcal{U}} + \sum_{M \in \mathcal{M}} \boldsymbol{\delta}^M(\log(\boldsymbol{s}^n) + \boldsymbol{\Psi}) \cdot \mathbb{B}^M(\boldsymbol{s}^n) \delta^M \boldsymbol{v} = 0, \tag{66}$$

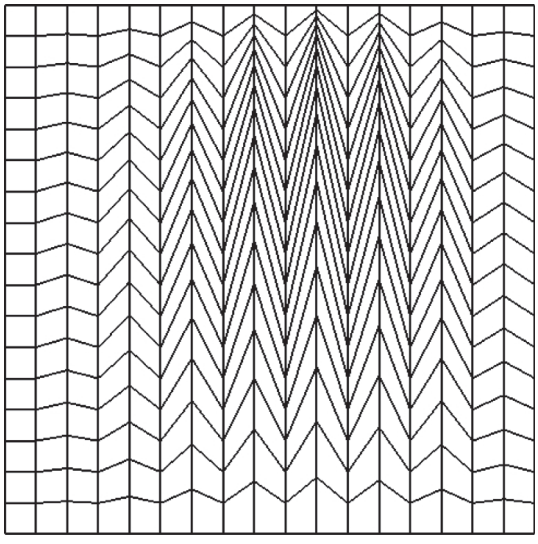

**Fig. 4.** The Kershaw meshes are highly deformed topologically cartesian grids. The coarsest mesh of the family is depicted here.

for any $\boldsymbol{v} \in \mathcal{U}$. The mobilities are discretized thanks to formula (54). The mass is conserved along time, *i.e.*,

$$\sum_{K \in \mathcal{U}} m_K s_K^n = \sum_{K \in \mathcal{U}} m_K s_K^0, \qquad \forall n \geq 1. \tag{67}$$

For any $\rho > 0$, the vector $\boldsymbol{s}^\infty = \left(s_K^\infty\right)_{K \in \mathcal{U}}$ of $\mathbb{R}^{\mathcal{U}}$ defined by

$$s_K^\infty = \rho e^{-\Psi_K}, \qquad \forall K \in \mathcal{U}, \tag{68}$$

is a steady solution to the scheme (66) since $\delta^M(\log(\boldsymbol{s}^n) + \boldsymbol{\Psi}) = \boldsymbol{0}$ for all $M \in \mathcal{M}$. Then the expected long-time limit as $n \to \infty$ is $\boldsymbol{s}^\infty$ where $\rho$ has been tuned so that

$$\sum_{K \in \mathcal{U}} m_K s_K^\infty = \sum_{K \in \mathcal{U}} m_K s_K^0.$$

**Table 4.** Approximation of (42) with a nonlinear DDFV scheme [62, 63] on the Kershaw mesh family from [93], final time $T = 0.25$. $M$ is the mesh index, $\tau$ is the time step, errgs and errs respectively stand for the $L^2(\Omega \times (0, T))$ error on $\nabla s$ and the $L^\infty((0, T); L^2(\Omega))$ error on $s$, whereas ordgs and ords are the corresponding convergence orders. $N_{\max}$ and $N_{\text{mean}}$ are the maximal and mean numbers of Newton iterations, and $s_{\min}$ minimal value of the approximation of $s$ during the whole simulation.

| $M$ | $\tau$ | errgs | ordgs | errs | ords | $N_{\max}$ | $N_{\text{mean}}$ | $s_{\min}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.0E-03 | 6.693E-02 | – | 7.254E-03 | – | 9 | 2.15 | 1.010E-01 |
| 2 | 5.0E-04 | 2.353E-02 | 1.54 | 1.751E-03 | 2.09 | 8 | 2.02 | 2.582E-02 |
| 3 | 1.25E-04 | 1.235E-02 | 1.61 | 7.237E-04 | 2.20 | 7 | 1.49 | 6.488E-03 |
| 4 | 3.125E-05 | 7.819E-03 | 1.60 | 3.962E-04 | 2.11 | 7 | 1.07 | 1.628E-03 |
| 5 | 3.125E-05 | 5.507E-03 | 1.58 | 2.556E-04 | 1.98 | 7 | 1.04 | 1.628E-03 |

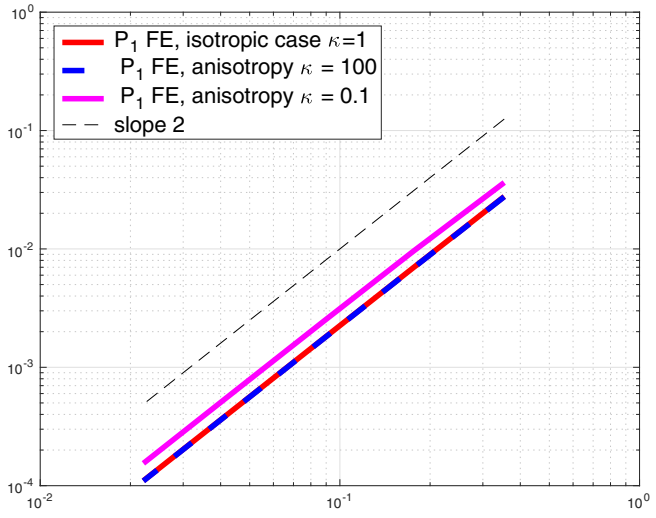


**Fig. 5.** Convergence towards the analytical solution (42) of the Fokker-Planck equation (41) for different values of the anisotropy ratio $\kappa$.

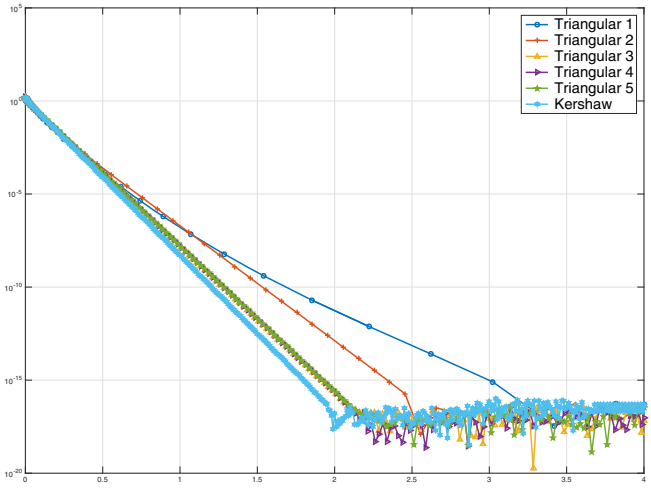


**Fig. 6.** Plot of the log of the relative energy $\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) - \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^\infty)$ (in log scale) as a function of time.

To illustrate this fact, we plot on Figure 6 the evolution of the relative energy $\mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^n) - \mathcal{E}_{\mathcal{U}}(\boldsymbol{s}^\infty)$ as a function of the discrete time $n\tau$ for the sequence of successively refined triangular meshes used in Tables 1–3 and for the Kershaw mesh depicted on Figure 4. The relative energy is proved to be decreasing along time and vanishes if and only if $\boldsymbol{s}^n = \boldsymbol{s}^\infty$.

We show in Figure 6 that the relative energy converges exponentially fast towards 0, providing a discrete counterpart to the relation

$$\left| s_{\text{ex}}(\boldsymbol{x}, t) - e^{-\Psi(\boldsymbol{x})} \right| \leq Ce^{-\alpha t}$$

that is deduced from (42). We observe on Figure 6 that the scheme preserves exactly (up to machine precision) the long-time behavior of the equation, *i.e.*, the long-time limit $\boldsymbol{s}^\infty$ is a discretization through (67) of the exact long-time behavior of (64). Figure 6 suggests that the convergence speed is sensitive to mesh regularity (the convergence is slightly too fast on the Kershaw mesh) and to mesh size (the convergence is too slow on coarse triangular meshes). Note that the relative energy is defined only for non-negative $\boldsymbol{s}^n$. For a behavior like the one depicted on Figure 6, the scheme must (at least) preserve positivity and admit $\boldsymbol{s}^\infty$ defined by (67) as a steady state.

### 3.4 Application to two-phase porous media flows

We show now preliminary results obtained thanks to the methodology of this section to simulate two-phase porous media flows. We consider an anisotropic quarter five spot problem with injection in a top-left corner and extraction in the bottom-right corner. We neglect gravity, so that the equations boil down into

$$\phi \partial_t s_{\text{n}} + \boldsymbol{\nabla} \cdot (\eta_{\text{n}}(s_{\text{n}}) \mathbb{K} \boldsymbol{\nabla} (p_{\text{w}} + \pi(s_{\text{n}}))) = q_{\text{n}}(s_{\text{n}}, \boldsymbol{x}), \tag{69}$$

$$-\phi \partial_t s_{\text{n}} + \boldsymbol{\nabla} \cdot (\eta_{\text{w}}(s_{\text{n}}) \mathbb{K} \boldsymbol{\nabla} p_{\text{w}}) = q_{\text{w}}(s_{\text{n}}, \boldsymbol{x}). \tag{70}$$

In the above system, we eliminated the unknowns $p_{\text{n}}$ and $s_{\text{w}}$. They can be deduced from $p_{\text{w}}$ and $s_{\text{n}}$ by the

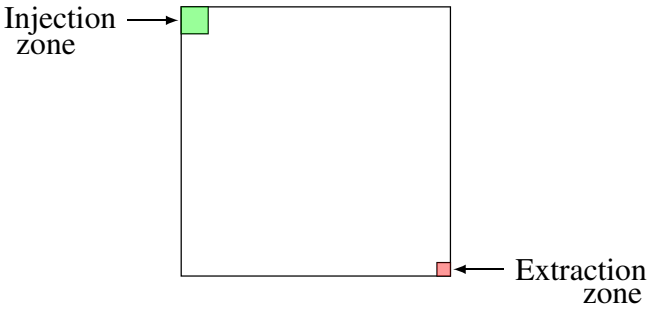


**Fig. 7.** Schematic representation of the injection and production wells.

relations $s_w = 1 - s_n$ and $p_n = p_w + \pi(s_n)$. We fix the nonlinearities $\eta_n$, $\eta_w$, $\pi$, $f_n$, and $f_w$ as in [98]:

$$\eta_n(s) = \begin{cases} s^3(2-s) & \text{if } s \in [0,1), \\ s^2 & \text{if } s < 0, \end{cases} \qquad \eta_w(s) = 2(1-s)^4,$$

and

$$\pi(s) = \begin{cases} (1-s)^{-1/2} & \text{if } s \in [0,1), \\ 0.5\, s & \text{if } s \le 0. \end{cases}$$

The pure wetting phase is injected in $\omega_{\text{inj}} = (0, 0.05) \times (0.95, 1)$ near the top-left corner, while the fluid occupying the area $\omega_{\text{prod}} = (0.98, 1) \times (0, 0.02)$ near the bottom-right corner is extracted (see Fig. 7). Defining the fractional flow functions by

$$f_n(s) = \frac{\eta_n(s)}{\eta_n(s) + \eta_w(s)} \quad \text{and} \quad f_w(s) = \frac{\eta_w(s)}{\eta_n(s) + \eta_w(s)},$$

then the source terms $q_n$ and $q_w$ are defined by

$$q_\alpha(s, \boldsymbol{x}) = f_\alpha(0)\mathbf{1}_{\omega_{\text{inj}}}(\boldsymbol{x}) - f_\alpha(s)\mathbf{1}_{\omega_{\text{prod}}}(\boldsymbol{x}), \quad \alpha \in \{n, w\}.$$

The permeability tensor $\mathbb{K} = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$ is anisotropic and we take a constant saturation profile $\phi \equiv 1$.

The system (69)–(70) is complemented with no-flux boundary conditions and with the initial condition $s_n(\cdot, 0) \equiv 0.9$.

For the discretization, we use conforming $\mathbb{P}_1$ finite elements with mass lumping on a structured triangulation $\mathcal{T}$ made of 5000 triangles and a constant time step $\tau = 10^{-3}$. For the phase mobilities, we make a choice of type (58). As explained in [86], the method is locally conservative: one can build equilibrated fluxes $\boldsymbol{\phi}^n_{\alpha,h}$ in the Raviart-Thomas-Nedelec space $\mathbf{RTN}_1(\mathcal{T})$ such that

$$\frac{s^n_{\alpha,h} - s^{n-1}_{\alpha,h}}{\tau} + \nabla \cdot \boldsymbol{\phi}^n_\alpha = q^n_{\alpha,h}, \quad \alpha \in \{n, w\},$$

where $s^n_{\alpha,h}$ denotes the $\mathbb{P}_1$ approximation of the saturation $s_\alpha$ at time $t_n = \sum_{k=1}^n \tau_k$.

We plot snapshots of the saturation $s_n$ on Figure 8.

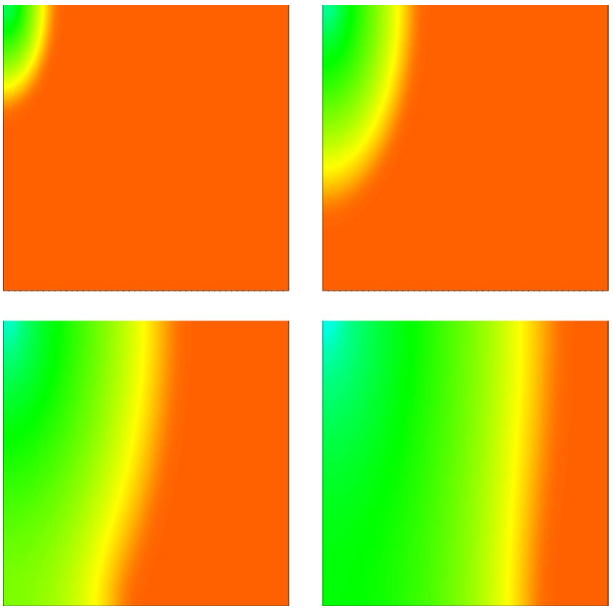

**Fig. 8.** Approximate saturation $s_{n,h}$ at time $t = 0.2$ (top left), $t = 0.5$ (top right), $t = 1.5$ (bottom left) and $t = 2.5$ (bottom right).

*Acknowledgments.* The author warmly thanks the anonymous referees for their precious feedback. He overall wants to thank his numerous collaborators on this topics, namely Ahmed Ait Hammou Oulhaj, Konstantin Brenner, Claire Chainais-Hillairet, Thomas O. Gallouët, Cindy Guichard, Stella Krell, Maxime Laborde, Léonard Monsaingeon, Flore Nabet, and Martin Vohralík. They actively contributed to the realisation of this work. This research was supported by the French National Research Agency through project GEOPOR (ANR-13-JS01-0007-01) and Labex CEMPI (ANR-11-LABX-0007-01).

## References

1 van Duijn C.J., Molenaar J., de Neef M.J. (1995) The effect of capillary forces on immiscible two-phase flows in heterogeneous porous media, *Transp. Porous Media* **21**, 71–93.
2 Bertsch M., Dal Passo R., van Duijn C.J. (2003) Analysis of oil trapping in porous media flow, *SIAM J. Math. Anal.* **35**, 1, 245–267. ISSN 0036-1410.
3 Buzzi F., Lenzinger M., Schweizer B. (2009) Interface conditions for degenerate two-phase flow equations in one space dimension, *Analysis* **29**, 299–316.
4 Cancès C., Gallouët T., Porretta A. (2009) Two-phase flows involving capillary barriers in heterogeneous porous media, *Interfaces Free Bound.* **11**, 2, 239–258.
5 Cancès C., Pierre M. (2012) An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field, *SIAM J. Math. Anal.* **44**, 2, 966–992. doi: 10.1137/11082943X. URL http://hal.archives-ouvertes.fr/hal-00518219.
6 Cancès C., Gallouët T.O., Monsaingeon L. (2015) The gradient flow structure of immiscible incompressible two-phase flows in porous media, *C. R. Acad. Sci. Paris Ser. I Math.* **353**, 985–989.

7 Cancès C., Gallouët T.O., Monsaingeon L. (2017) Incompressible immiscible multiphase flows in porous media: a variational approach, *Anal. PDE* **10**, 8, 1845–1876.

8 Cancès C., Gallouët T.O., Laborde M., Monsain-Geon L. (2018) *Simulation of multiphase porous media flows with minimizing movement and finite volume schemes*, HAL, hal-01700952. URL https://hal.archives-ouvertes.fr/hal-01700952/document.

9 Murphy T.J., Walkington N.J. Control volume approximation of degenerate two-phase porous media flows, submitted for publication.

10 Mielke A. (2011) A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems, *Nonlinearity* **24**, 4, 1329–1346, ISSN 0951-7715. doi: 10.1088/0951-7715/24/4/016. URL http://dx.doi.org/10.1088/0951-7715/24/4/016.

11 Otto F. (2001) The geometry of dissipative evolution equations: the porous medium equation, *Comm. PDE* **26**, 1–2, 101–174, ISSN 0360-5302.

12 Ambrosio L., Gigli N., Savaré G. (2008) *Gradient flows in metric spaces and in the space of probability measures*, 2nd edn, Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, ISBN 978-3-7643-8721-1.

13 Bessemoulin-Chatard M. (2012) Développement et analyse de schémas volumes finis motivés par la préservation de comportements asymptotiques. Application à des modèles issus de la physique et de la biologie, *PhD Thesis*, Université Blaise Pascal – Clermont-Ferrand II, 2012. URL http://tel.archives-ouvertes.fr/tel-00836514

14 Bear J., Bachmat Y. (1990) *Introduction to modeling of transport phenomena in porous media*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

15 Maury B., Roudneff-Chupin A., Santambrogio F. (2010) A macroscopic crowd motion model of gradient flow type, *Math. Models Methods Appl. Sci.* **20**, 10, 1787–1821. ISSN 0218-2025. doi: 10.1142/S0218202510004799. URL http://dx.doi.org/10.1142/S0218202510004799.

16 Kumar K., Pop I.S., Radu F.A. (2013) Convergence analysis of mixed numerical schemes for reactive flow in a porous medium, *SIAM J. Numer. Anal.* **51**, 4, 2283–2308.

17 Zarba R.L., Bouloutas E.T., Celia M. (1990) General massconservative numerical solution for the unsaturated flow equation, *Water Resour. Res.* **26**, 7, 1483–1496.

18 Jäger W., Kacur J. (1991) Solution of porous medium type systems by linear approximation schemes, *Numer. Math.* **60**, 3, 407–427.

19 Jäger W., Kacur J. (1995) Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes, *RAIRO Modél. Math. Anal. Numér* **29**, 5, 605–627.

20 Pop I.S., Radu F.A., Knabner P. (2004) Mixed finite elements for the Richards equation: linearization procedure, *J. Comput. Appl. Math.* **168**, 1, 365–373.

21 Radu F.A., Nordbotten J.M., Pop I.S., Kumar K. (2015) A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media, *J. Comput. Appl. Math.* **289**, 134–141, ISSN 0377-0427. URL https://doi.org/10.1016/j.cam.2015.02.051.

22 Radu F.A., Kumar K., Nordbotten J.M., Pop I.S. (2018) A robust, mass conservative scheme for two-phase flow in porous media including hlder continuous nonlinearities, *IMA J. Numer. Anal.* **38**, 2, 88420. doi: 10.1093/imanum/drx032. URL http://dx.doi.org/10.1093/imanum/drx032

23 Casulli V., Zanolli P. (2010) A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form, *SIAM J. Sci. Comp.* **32**, 4, 2255–2273. doi: 10.1137/100786320. URL https://doi.org/10.1137/100786320.

24 Younis R., Tchelepi H.A., Aziz K. (2010) Adaptively localized continuation-Newton method-nonlinear solvers that converge all the time, *SPE J.* **15**, 02, 526–544.

25 Wang X., Tchelepi H.A. (2013) Trust-region based solver for nonlinear transport in heterogeneous porous media, *J. Comput. Phys.* **253**, 114–137.

26 Lehmann F., Ackerer P.H. (1998) Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media, *Transp. Porous Media.* **31**, 3, 275–292.

27 Bergamaschi L., Putti M. (1999) Mixed finite elements and Newton-type linearizations for the solution of Richards' equation, *Int. J. Numer. Meth. Eng.* **45**, 8, 1025–1046.

28 Radu F.A., Pop I.S., Knabner P. (2006) *Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations. Numerical mathematics and advanced applications*, Springer.

29 List F., Radu F.A. (2016) A study on iterative methods for solving Richards' equation, *Comput. Geosci.* 1–13.

30 Marchand E., Müller T., Knabner P. (2012) Fully coupled generalised hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part II: numerical scheme and numerical results, *Comput. Geosci.* **16**, 3, 691–708. doi: 10.1007/s10596-012-9279-1. URL https://doi.org/10.1007/s10596-012-9279-1.

31 Marchand E., Müller T., Knabner P. (2013) Fully coupled generalized hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part I: Formulation and properties of the mathematical model, *Comput. Geosci.* **17**, 2, 431–442, ISSN 1573-1499. doi: 10.1007/s10596-013-9341-7. URL https://doi.org/10.1007/s10596-013-9341-7.

32 Ben Gharbia I. (2012) Résolution de problèmes de complémentarité : application à un écoulement diphasique dans un milieu poreux, *Theses*, Université Paris Dauphine - Paris IX, December 2012. URL https://tel.archives-ouvertes.fr/tel-00776617

33 Diersch H.-J.G., Perrochet P. (1999) On the primary variable switching technique for simulating unsaturated-saturated flows, *Adv. Water Resour.* **23**, 3, 271–301.

34 Brenner K., Cancès C. (2017) Improving Newton's method performance by parametrization: The case of the Richards equation, *SIAM J. Numer. Anal.* **55**, 4, 1760–1785. doi: 10.1137/16M1083414. URL https://doi.org/10.1137/16M1083414

35 Brenner K., Groza M., Jeannin L., Masson R., Pellerin J. (2017) Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media, *Comput. Geosci.* **21**, 5–6, 1075–1094.

36 Ciarlet P.G. (1978) *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, ISBN 0-444-85028-7. Studies in Mathematics and its Applications, Vol. 4.

37 Ern A., Guermond J.L. (2004) *Theory and Practice of Finite Elements, volume 159 of Applied Mathematical Series*, Springer, New York.

38 Franco Brezzi and Michel Fortin (1991) *Mixed and hybrid finite element methods, volume 15 of Springer Series in Computational Mathematics*, Springer-Verlag, New York. ISBN 0-387-97582-9

39 Arbogast T., Wheeler M.F., Yotov I. (1997) Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences, *SIAM J. Numer. Anal.* **34**, 2, 828–852. doi: 10.1137/S0036142994262585. URL https://doi.org/10.1137/S0036142994262585.

40 Aavatsmark I., Barkve T., Bøe Ø., Mannseth T. (1998) Discretization on unstructured grids for inhomogeneous, anisotropic media. I. Derivation of the methods, *SIAM J. Sci. Comput.* **19**, 5, 1700–1716. doi: 10.1137/S1064827595293582. URL http://dx.doi.org/10.1137/S1064827595293582.

41 Edwards M.G., Rogers C.F. (1998) Finite volume discretization with imposed flux continuity for the general tensor pressure equation, *Comput. Geosci.* **2**, 4, 259–290. doi: 10.1023/A:1011510505406. URL http://dx.doi.org/10.1023/A:1011510505406.

42 Edwards M.G. (2002) Unstructured, control-volume distributed, full- tensor finite-volume schemes with flow based grids, *Comput. Geosci.* **6**, 3–4, 433–452, ISSN 1420-0597. doi: 10.1023/A:1021243231313. URL https://doi.org/10.1023/A:1021243231313.

43 Agelas L., Guichard C., Masson R. (2010) Convergence of finite volume MPFA O type schemes for heterogeneous anisotropic diffusion problems on general meshes, *Int. J. Finite* **7**, 2, 33.

44 Arnold D., Brezzi F., Cockburn B., Marini L. (2002) Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM J. Numer. Anal.* **39**, 5, 1749–1779. doi: 10.1137/S0036142901384162. URL https://doi.org/10.1137/S0036142901384162.

45 Rivière B. (2008) Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations, *SIAM*. doi: 10.1137/1.9780898717440. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717440.

46 Di Pietro D.A., Ern A. (2012) *Mathematical aspects of discontinuous Galerkin methods, volume 69 of Mathématiques & Applications (Berlin) [Mathematics & Applications]*, Springer, Heidelberg, ISBN 978-3-642-22979-4. doi: 10.1007/978-3-642-22980-0. URL http://dx.doi.org/10.1007/978-3-642-22980-0.

47 Herbin R. (1995) An error estimate for a finite volume scheme for a diffusiononvection problem on a triangular mesh, *Numer. Methods Partial Differ. Equ.* **11**, 2, 165–173. doi: 10.1002/num.1690110205. URL https://doi.org/10.1002/num.1690110205.

48 Eymard R., Gallouët T., Herbin R. (2000) Finite volume methods, in: Ciarlet P.G., *et al.* (eds), Handbook of numerical analysis, North-Holland: Amsterdam, p. 7131020.

49 Eymard R., Gallouët T., Guichard C., Herbin R., Masson R. (2014) TP or not TP, that is the question, *Comput. Geosci.* **18**, 285–296.

50 Hackbusch W. (1989) On first and second order box schemes, *Computing* **41**, 4, 277–296, ISSN 0010-485X. doi: 10.1007/BF02241218. URL https://doi.org/10.1007/BF02241218.

51 Droniou J., Eymard R., Gallouët T., Herbin R. (2010) A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods, *Math. Models Methods Appl. Sci.* **20**, 2, 265–295.

52 Eymard R., Gallouët T., Herbin R. (2010) Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes sushi: a scheme using stabilization and hybrid interfaces, *IMA J. Numer. Anal.* **30**, 4, 1009–1043.

53 Droniou J., Eymard R. (2006) A mixed finite volume scheme for anisotropic diffusion problems on any grid, *Numer. Math.* **105**, 35–71.

54 Brezzi F., Lipnikov K., Simoncini V. (2005) A family of mimetic finite difference methods on polygonal and polyhedral meshes, *Math. Models Methods Appl. Sci.* **15**, 10, 1533–1551.

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

56 Domelevo K., Omnes P. (2005) A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids, *M2AN: Math. Model. Numer. Anal.* **39**, 6, 1203–1249.

57 Droniou J. (2014) Finite volume schemes for diffusion equations: introduction to and review of modern methods, *Math. Models Methods Appl. Sci.* **24**, 8, 1575–1620.

58 Droniou J., Eymard R., Gallouët T., Guichard C., Herbin R. (2018) *The gradient discretisation method*, Vol. 42, Mathématiques et Applications, Springer International Publishing, https://doi.org/10.1007/978-3-319-79042-8.

59 Eymard R., Guichard C., Herbin R. (2012) Small-stencil 3D schemes for diffusive flows in porous media, *ESAIM: Math. Model. Numer. Anal.* **46**, 2, 265–290. doi: 10.1051/m2an/2011040. URL http://dx.doi.org/10.1051/m2an/2011040.

60 Eymard R., Guichard C., Herbin R. (2011) Benchmark 3D: the VAG scheme, in: Fort J., Fürst J., Halama J., Herbin R., Hubert F. (eds), Finite Volumes for Complex Applications VI Problems & Perspectives, volume 4 of Springer Proceedings in Mathematics, Springer, Berlin Heidelberg, pp. 1013–1022. ISBN 978-3-642-20670-2. doi: 10.1007/978-3-642-20671-9_99. URL http://dx.doi.org/10.1007/978-3-642-20671-9_99.

61 Cancès C., Guichard C. (2017) Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure, *Found. Comput. Math.* **17**, 6, 1525–1584. doi: 10.1007/s10208-016-9328-6.

62 Cancès C., Chainais-Hillairet C., Krell S. (2017) A nonlinear Discrete Duality Finite Volume Scheme for convection- diffusion equations, in: Cancès C., Omnes P. (eds), *FVCA8 2017 – International Conference on Finite Volumes for Complex Applications VIII, volume 199 of Springer Proceedings in Mathematics & Statistics,* Lille, France, Springer International Publishing, pp. 439–447. URL https://hal.archives-ouvertes.fr/hal-01468811.

63 Cancès C., Chainais-Hillairet C., Krell S. (2017) Numerical analysis of a nonlinear free-energy diminishing Discrete Duality Finite Volume scheme for convection diffusion equations, *Comput Methods Appl. Math.* doi: 10.1515/cmam-2017-0043. URL https://hal.archives-ouvertes.fr/hal-01529143. Special issue on "Advanced numerical methods: recent developments, analysis and application".

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

65 Chavent G., Jaffré J. (1986), *Mathematical Models and Finite Elements for Reservoir Simulation*, Vol. 17, Stud. Math. Appl. edition, North-Holland, Amsterdam.

66 Antontsev S.N., Kazhikhov A.V., Monakhov V.N. (1990) *Boundary value problems in mechanics of nonhomogeneous fluids, vol. 22 of Studies in Mathematics and its Applications*, North-Holland Publishing Co., Amsterdam, ISBN 0-444-88382-7. Translated from the Russian.

67 Gagneux G., Madaune-Tort M. (1996) *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière, vol. 22 of Mathématiques & Applications (Berlin) [Mathematics & Applications]*, Springer-Verlag, Berlin, ISBN 3-540-60588-6.

68 Chen Z. (2001) Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution, *J. Diff. Equ.* **171**, 2, 203–232.

69 Nochetto R.H., Verdi C. (1988) Approximation of degenerate parabolic problems using numerical integration, *SIAM J. Numer. Anal.* **25**, 4, 784–814. doi: 10.1137/0725046. URL https://doi.org/10.1137/0725046.

70 Arbogast T., Wheeler M.F., Zhang N.-Y. (1996) A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, *SIAM J. Numer. Anal.* **33**, 4, 1669–1687. doi: 10.1137/S0036142994266728. URL http://dx.doi.org/10.1137/S0036142994266728.

71 Eymard R., Gallouët T., Hilhorst D., Naït Slimane Y. (1998) Finite volumes and nonlinear diffusion equations, *RAIRO Modél. Math. Anal. Numér.* **32**, 6, 747–761.

72 Eymard R., Gutnic M., Hilhorst D. (1999) The finite volume method for Richards equation, *Comput. Geosci.* **3**, 3–4, 259–294. doi: 10.1023/A:1011547513583. URL http://dx.doi.org/10.1023/A%3A1011547513583.

73 Woodward C.S., Dawson C.N. (2000) Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media, *SIAM J. Numer. Anal.* **37**, 3, 701–724. doi: 10.1137/S0036142996311040. URL https://doi.org/10.1137/S0036142996311040.

74 Eymard R., Gallouët T., Herbin R., Michel A. (2002) Convergence of finite volume schemes for parabolic degenerate equations, *Numer. Math.* **92**, 41–82.

75 Pop I.S. (2002) Error estimates for a time discretization method for the Richards' equation, *Comput. Geosci.* **6**, 141–160.

76 Radu F.A., Pop I.S., Knabner P. (2004) Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation, *SIAM J. Numer. Anal.* **42**, 4, 1452–1478.

77 Eymard R., Hilhorst D., Vohralík M. (2006) A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems, *Numer. Math.* **105**, 1, 73–131. doi: 10.1007/s00211-006-0036-z. URL http://dx.doi.org/10.1007/s00211-006-0036-z.

78 Radu F.A., Pop I.S., Knabner P. (2008) Error estimates for a mixed finite element discretization of some degenerate parabolic equations, *Numer. Math.* **109**, 2, 285–311. doi: 10.1007/s00211-008-0139-9. URL http://dx.doi.org/10.1007/s00211-008-0139-9.

79 Angelini O., Brenner K., Hilhorst D. (2013) A finite volume method on general meshes for a degenerate parabolic convection-reaction-diffusion equation, *Numer. Math.* **123**, 219–257, ISSN 0029-599X. doi: 10.1007/s00211-012-0485-5. URL http://dx.doi.org/10.1007/s00211-012-0485-5.

80 Chen Z., Ewing R.E. (1997) Fully discrete finite element analysis of multiphase flow in groundwater hydrology, *SIAM J. Numer. Anal.* **34**, 6, 2228–2253.

81 Chen Z., Ewing R.E. (2001) Degenerate two-phase incompressible flow. III. Sharp error estimates, *Numer. Math.* **90**, 2, 215–240, ISSN 0029-599X. doi: 10.1007/s002110100291. URL http://dx.doi.org/10.1007/s002110100291.

82 Michel A. (2003) A finite volume scheme for two-phase immiscible flow in porous media, *SIAM J. Numer. Anal.* **41**, 4, 1301–1317.

83 Epshteyn Y., Rivière B. (2009) Analysis of *hp* discontinuous Galerkin methods for incompressible two-phase flow, *J. Comput. Appl. Math.* **225**, 2, 487–509. doi: 10.1016/j.cam.2008.08.026. URL https://doi.org/10.1016/j.cam.2008.08.026.

84 Brenner K., Masson R. (2013) Convergence of a vertex centered discretization of two-phase Darcy flows on general meshes, *Int. J. Finite* **10**, 1–37.

85 Cancès C., Pop I.S., Vohralík M. (2014) An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow, *Math. Comp.* **83**, 285, 153–188. doi: 10.1090/S0025-5718-2013-02723-8. URL http://dx.doi.org/10.1090/S0025-5718-2013-02723-8.

86 Cancès C., Nabet F., Vohralík M. Convergence and a posteriori error analysis for energy stable finite element approximations of degenerate parabolic equations, in preparation.

87 Forsyth P.A. (1991) A control volume finite element approach to NAPL groundwater contamination, *SIAM J. Sci. Statist. Comput.* **12**, 5, 1029–1057.

88 Ait Hammou Oulhaj A., Cancès C., Chainais-Hillairet C. (2018) Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy, *ESAIM Math. Model. Numer. Anal.* **52**, 1532—1567.

89 Ait Hammou Oulhaj A. (2018) Numerical analysis of a finite volume scheme for a seawater intrusion model with cross-diffusion in an unconfined aquifer, *Numer. Methods Partial Differ. Equ.* **34**, 3, 857–880. doi: 10.1002/num.22234. URL https://doi.org/10.1002/num.22234.

90 Cancès C., Nabet F. (2017) Finite volume approximation of a degenerate immiscible two-phase flow model of Cahn-Hilliard type, in: Cancès C., Omnes P. (eds), Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects : FVCA 8, Lille, France, June 2017, number 199 in Proceedings in Mathematics and Statistics, Cham, Springer International Publishing, pp. 431–438, ISBN 978-3-319-57397-7. doi: 10.1007/978-3-319-57397-7_36. http://dx.doi.org/10.1007/978-3-319-57397-7_36.

91 Otto F., Weinan E. (1997) Thermodynamically driven incompressible fluid mixtures, *J. Chem. Phys.* **107**, 23, 10177–10184.

92 Cancès C., Matthes D., Nabet F. (2017) *A two-phase two-fluxes degenerate Cahn-Hilliard model as constrained Wasserstein gradient flow*, HAL, hal-01665338, December 2017. URL https://hal.archives-ouvertes.fr/hal-01665338.

93 Herbin R., Hubert F. (2008) Benchmark on discretization schemes for anisotropic diffusion problems on general grids, in: Eymard R., Herard J.-M. (eds), Finite Volumes for Complex Applications V, Wiley, pp. 659–692. https://www.latp.univ-mrs.fr/fvca5/benchmark/

94 Eymard R., Herbin R., Michel A. (2003) Mathematical study of a petroleum-engineering scheme, *M2AN: Math. Model. Numer. Anal.* **37**, 6, 937–972.

95 Chainais-Hillairet C., Filbet F. (2007) Asymptotic behaviour of a finite-volume scheme for the transient drift-diffusion model, *IMA J. Numer. Anal.* **27**, 4, 689–716. doi: 10.1093/imanum/drl045. URL http://dx.doi.org/10.1093/imanum/drl045.

96 Bessemoulin-Chatard M., Chainais-Hillairet C. (2017) Exponential decay of a finite volume scheme to the thermal equilibrium for drift-diffusion systems, *J. Numer. Math.* **25**, 3. https://hal.archives-ouvertes.fr/hal-01250709.

97 Filbet F., Herda M. (2017) A finite volume scheme for boundary-driven convection-diffusion equations with relative entropy structure, *Numer. Math.* URL https://hal.archives-ouvertes.fr/hal-01326029.
98 Ganis B., Kumar K., Pencheva G., Wheeler M., Yotov I. (2014) A global Jacobian method for mortar discretizations of a fully implicit two-phase flow model, *Multiscale Model. Simul.* **12**, 4, 1401–1423. doi: 10.1137/140952922. URL https://doi.org/10.1137/140952922.

## A Some technical details

### A.1 About Estimate (39)

We now give some details on the derivation of estimate (39). Note that the derivation of estimate (36) relies on similar arguments.

Let $\overline{\varphi}$ be defined by (38), then multiplying (28) by $\tau_n \overline{\varphi}(w_K^n)$ and summing over $K \in \mathcal{U}$ leads to

$$A_n + B_n = 0 \tag{71}$$

where

$$A_n = \sum_{K\in\mathcal{U}} m_K \big(s_K^n - s_K^{n-1}\big)\overline{\varphi}(s_K^n),$$

$$B_n = \tau_n \sum_{(K,L)\in\mathcal{S}} \eta_{KL}^n a_{KL}(\overline{p}(w_K^n) + \Psi_K - \overline{p}(w_L^n) - \Psi_L) \times (\overline{\varphi}(w_K^n) - \overline{\varphi}(w_L^n)).$$

The term $A_n$ can be underestimated thanks to the convexity inequality (37):

$$A_n \geq \sum_{K\in\mathcal{U}} m_K \big(\Phi(s_K^n) - \Phi(s_K^{n-1})\big). \tag{72}$$

Thanks to the characterization (35) of the upstream mobility, the term $B_n$ can be underestimated by

$$B_n = \tau_n \sum_{(K,L)\in\mathcal{S}} \tilde{\eta}_{KL}^n a_{KL}(\overline{p}(w_K^n) + \Psi_K - \overline{p}(w_L^n) - \Psi_L) \times (\overline{\varphi}(w_K^n) - \overline{\varphi}(w_L^n)),$$

whatever the mean value $\tilde{\eta}_{KL}^n$ between $\eta(s_K^n)$ and $\eta(s_L^n)$. Choosing

$$\tilde{\eta}_{KL}^n = \begin{cases} \frac{s_K^n - s_L^n}{\overline{\varphi}(w_K^n) - \overline{\varphi}(w_L^n)} & \text{if } s_K^n \neq s_L^n, \\ s_K^n & \text{otherwise,} \end{cases}$$

which lies between $\eta(s_K^n)$ and $\eta(s_L^n)$ because of the definition (38) of $\overline{\varphi}$, one gets that

$$B_n \geq B_n^{(1)} + B_n^{(2)},$$

where

$$B_n^{(1)} = \tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL}(\overline{p}(w_K^n) - \overline{p}(w_L^n))(s_K^n - s_L^n)$$

is the quantity that we want to bound from above, and

$$B_n^{(2)} = \tau_n \sum_{(K,L)\in\mathcal{S}} a_{KL}(\Psi_K - \Psi_L)(s_K^n - s_L^n) = \tau_n \sum_{K\in\mathcal{U}} m_K s_K^n \left(\frac{1}{m_K} \sum_{L\in\mathcal{N}_K} a_{KL}(\Psi_K - \Psi_L)\right).$$

The quantity

$$\frac{1}{m_K} \sum_{L\in\mathcal{N}_K} a_{KL}(\Psi_K - \Psi_L),$$

which is an approximation of $-\boldsymbol{\nabla}\cdot(\mathbb{K}\boldsymbol{\nabla}\Psi)$ at $\boldsymbol{x}_K$, is supposed to be bounded from below by some quantity $M$ depending only on the regularity of the mesh. Since the method preserves the positivity of the saturations $s_K^n$, one obtains that

$$B_n^{(2)} \geq -M\tau_n \sum_{K\in\mathcal{U}} m_K s_K^n = -M\tau_n \sum_{K\in\mathcal{U}} m_K s_K^0 \tag{73}$$

thanks to the global conservativity of the scheme (28). As a consequence, one gets that

$$\sum_{n=0}^{N} B_n^{(1)} \leq \sum_{K\in\mathcal{U}} m_K \big(\Phi(s_K^0) - \Phi(s_K^N)\big) + C\sum_{n=1}^{N} \tau_n.$$

One concludes by noticing that $\Phi$ is bounded from below, hence

$$\sum_{K\in\mathcal{U}} m_K \Phi(s_K^N) \geq -m_\Omega \inf \Phi.$$

### A.2 Relation (45) as an optimality condition

Before justifying why (45) can be seen as an optimality condition for the discrete energy under the mass contraint (47), let us first notice that since

$$s_{\mathrm{n},K}^n + s_{\mathrm{w},K}^n = 1, \quad \forall K \in \mathcal{U}, \forall n \geq 0,$$

the discrete counterpart

$$\mathcal{E}_\mathcal{U}(\boldsymbol{s}^n) = \sum_{K\in\mathcal{U}} \phi_K m_K \left(\Pi(s_{\mathrm{n},K}^n) - \sum_{\alpha\in\{\mathrm{n},w\}} s_{\alpha,K}^n \rho_\alpha \mathbf{g}\cdot\boldsymbol{x}_K\right)$$

of the energy (5) can be rewritten as

$$\mathcal{E}_\mathcal{U}(\boldsymbol{s}^n) = \sum_{K\in\mathcal{U}} \phi_K m_K \Big(\Pi(s_{\mathrm{n},K}^n) + s_{\mathrm{n},K}^n(\rho_\mathrm{w} - \rho_\mathrm{n})\mathbf{g}\cdot\boldsymbol{x}_K\Big) + \sum_{K\in\mathcal{U}} \phi_K m_K \rho_\mathrm{w} \mathbf{g}\cdot\boldsymbol{x}_K.$$

The second term in the above equality does not depend on $\boldsymbol{s}^n$, hence it can be omitted in (48) and we can write the discrete energy as a function of $\boldsymbol{s}_\mathrm{n}^n$ only, *i.e.*, $\mathcal{E}_\mathcal{U}(\boldsymbol{s}_\mathrm{n}^n)$.

A second preliminary remark is the following: if $\phi_K m_K = 0$ for some $K \in \mathcal{U}$, then $s_{\mathrm{n},K}^n$ has no influence on

the energy $\mathcal{E}_\mathcal{U}(\boldsymbol{s}_\mathrm{n}^n)$. Its value can be fixed arbitrarily, for instance by (45). We assume for simplicity that $\phi_K m_K > 0$ for all $K \in \mathcal{U}$ even though this assumption can be easily bypassed.

Let us now go to the constrained optimization problem

$$\boldsymbol{s}_\mathrm{n}^\infty \in \operatorname*{argmin}_{\boldsymbol{s}_\mathrm{n} \in X} \mathcal{E}_\mathcal{U}(\boldsymbol{s}_\mathrm{n}^n),$$

where, setting $\boldsymbol{m} = (\phi_K m_K) \in \mathbb{R}^\mathcal{U}$, we denoted

$$X = \left\{ \boldsymbol{s}_\mathrm{n} \in \mathbb{R}^\mathcal{U} \mid \boldsymbol{s}_\mathrm{n} \cdot \boldsymbol{m} = \boldsymbol{s}_\mathrm{n}^0 \cdot \boldsymbol{m} \right\}.$$

Note that $\boldsymbol{s}^n$ necessarily belongs to $[0,1]^\mathcal{U}$ otherwise $\mathcal{E}_\mathcal{U}$ would be infinite. The problem is equivalent to the saddle-point problem

$$\min_{\boldsymbol{s}_\mathrm{n} \in \mathbb{R}^\mathcal{U}} \max_{\gamma' \in \mathbb{R}} \left\{ \mathcal{E}_\mathcal{U}(\boldsymbol{s}_\mathrm{n}) + \gamma'(\boldsymbol{s}_\mathrm{n} - \boldsymbol{s}_\mathrm{n}^0) \cdot \boldsymbol{m} \right\}.$$

We can swap the min and the max and optimize w.r.t. $\boldsymbol{s}_\mathrm{n}$, so that we get the optimality condition

$$\boldsymbol{0} \in \partial \mathcal{E}_\mathcal{U}(\boldsymbol{s}_\mathrm{n}^\infty) + \gamma' \boldsymbol{m} \subset \mathbb{R}^\mathcal{U},$$

which is equivalent to (45).