A finite element method for degenerate two-phase flow in porous media. Part I: Well-posedness

Vivette Girault, Beatrice Riviere, Loic Cappanera

DE GRUYTER J. Numer. Math. 2021; 29(2):1–21

Vivette Girault, Beatrice Riviere*, and Loic Cappanera

A finite element method for degenerate two-phase flow in porous media. Part I: Well-posedness

https://doi.org/10.1515/jnma-2020-0004
Received January 24, 2020; revised July 27, 2020; accepted December 23, 2020

Abstract: A finite element method with mass-lumping and flux upwinding is formulated for solving the immiscible two-phase flow problem in porous media. The method approximates directly the wetting phase pressure and saturation, which are the primary unknowns. The discrete saturation satisfies a maximum principle. Stability of the scheme and existence of a solution are established.

Keywords: stability, compactness, maximum principle, pressure-saturation

Classification: 65M60, 65M12

1 Introduction

This work discretizes on a suitable mesh a degenerate two-phase flow system set in a polyhedral domain by a finite element scheme that directly approximates the wetting phase pressure and saturation, similar to the formulation proposed in [19]. Mass lumping is used to compute the integrals and a suitable upwinding is used to compute the flux, guaranteeing that the discrete saturation satisfies a maximum principle. The resulting system of discrete equations is a finite element analogue of the finite volume scheme introduced and analyzed by Eymard et al. in the seminal work [16].

Finite volume methods are popular discretization methods for solving porous media flow problems because they approximate the unknowns by piecewise constants, they are locally mass conservative and they satisfy the maximum principle. From the point of view of implementation, the advantage of finite elements is that they only use nodal values and a single simplicial mesh. In particular, no orthogonality property is required between the faces and the lines joining the centers of control volumes, as is the case with finite volume methods.

From a theoretical point of view, owing that the finite element scheme is based on functions, some steps in its numerical analysis are simpler, but nevertheless the major difficulty in the analysis consists in proving sufficient a priori estimates in spite of the degeneracy. By following closely [16], the degeneracy is remediated by reintroducing in the proofs discrete artificial pressures. But the complete analysis is intricate and lengthy and because of its length it is split into two parts. This paper is part one, dedicated to well-posedness of this discrete scheme: stability and existence. The second part, see [20], establishes the convergence of the numerical solutions via a compactness argument.

Incompressible two-phase flow is a popular and important multiphase flow model in reservoirs for the oil and gas industry. Based on conservation laws at the continuum scale, the model assumes the existence of a representative elementary volume. Each wetting phase and non-wetting phase saturation satisfies a mass balance equation and each phase velocity follows the generalized Darcy law [4, 26]. The equations of the

Vivette Girault, Sorbonne-Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions, F-75005 Paris, France; Rice University, Houston, TX 77005, USA
***Corresponding author: Beatrice Riviere,** Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA. Email: riviere@rice.edu
Loic Cappanera, Department of Mathematics, University of Houston, TX 77204, USA.

mathematical model read

$$
\begin{aligned}
\partial_t(\varphi s_w) - \nabla\cdot(\eta_w(s_w)\nabla p_w) &= f_w(s_{\text{in}})\overline{q} - f_w(s_w)\underline{q}\\
\partial_t(\varphi s_o) - \nabla\cdot(\eta_o(s_w)\nabla p_o) &= f_o(s_{\text{in}})\overline{q} - f_o(s_w)\underline{q}\\
p_c(s_w) = p_o - p_w, &\quad s_w + s_o = 1
\end{aligned}
\tag{1.1}
$$

complemented by initial and boundary conditions. Here p_w, s_w, η_w, f_w (respectively, p_o, s_o, η_o, f_o) are the pressure, saturation, mobility, and fractional flow of the wetting (respectively non-wetting) phase, φ is the porosity, s_{in} is a given input saturation, and $\overline{q}$, $\underline{q}$ are given flow rates. The capillary pressure, p_c, is a given function that depends nonlinearly on the saturation. This problem is referred to as the degenerate two-phase flow problem because the coefficients (phase mobilities) are allowed to vanish in some regions of the domain. This degeneracy makes the theoretical analysis problematic because it creates a loss of ellipticity in these regions. As the phase mobilities are degenerate when they are evaluated at certain values of the saturation (see (1.8)) and moreover the derivative of the capillary pressure may be unbounded, this system of two coupled nonlinear partial differential equations requires not only a carefully designed discretization preserving the maximum principle, but also a delicate analysis to circumvent the loss of ellipticity and the unboundedness of some coefficients. The discretization relies on mass lumping and upwinding. The use of mass lumping and upwinding with finite elements of degree one was introduced in [19] for porous media flows. Under the assumption that the pressure is known (which simplifies the problem to one equation with saturation as unknown), the maximum principle is proved for the saturation but no convergence analysis is obtained in [19]. The effects of gravity have been neglected in problem (1.1) as the gravity term further complicates the numerical analysis of the scheme.

At the continuous level, problem (1.1) has several equivalent formulations, linked to the choice of primary unknowns selected among wetting phase and non-wetting phase pressure and saturation, or capillary pressure [5, 22]. A good state of the art can be found in the reference [2]. Up to our knowledge, the mathematical analysis of the system of equations was first done in [1, 23]. A formulation of the model, based on Chavent's global pressure [7] that removes the degeneracy, was analyzed in [9, 10]. Since then, the global pressure formulation has been discretized and analyzed in many references [11, 24, 25], but unfortunately, this formulation is not equivalent to the original problem and it is not used in engineering practice because the global pressure is not a physical quantity that can be measured. Otherwise, with one exception, the numerical analysis of the discrete version of (1.1), has always been done under unrealistic assumptions that cannot be checked at the discrete level [14, 15]. Related to this line of work, the discretization of a degenerate parabolic equation has been studied in the literature [3, 17, 27, 28]. As far as we know, the only publication that performs the complete numerical analysis of the discrete degenerate two-phase flow system written as above (i.e., in the form used by engineers) is the analysis on finite volumes done in reference [16]. This motivates our extension of this work to finite elements.

The remaining part of this introduction makes precise problem (1.1) by introducing notation and the weak variational formulation. The numerical scheme is developed in Section 2 and is written in two equivalent forms: the first one is discrete and directly involves the nodal values of the unknowns and the second one is variational and uses the finite element test and trial functions. Because of the nonlinearity and degeneracy of its equations, existence of a discrete solution requires that the discrete wetting phase saturation satisfies a maximum principle. This is the first object of Section 3, the second one being basic a priori pressure estimates, after which existence is shown in Section 4. Numerical results are presented in Section 5. The basic a priori pressure estimates in Section 3.2 are not strong enough to show convergence of the numerical solution to the weak solution. Tighter bounds are obtained in the following work [20].

1.1 Model problem

Let $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, be a bounded connected Lipschitz domain with boundary $\partial\Omega$ and unit exterior normal $\boldsymbol{n}$, and let T be a final time. The primary unknowns are the wetting phase pressure and saturation. With the last relation in (1.1), s_w is the only unknown saturation; so we set $s = s_w$, and rewrite (1.1) almost everywhere in $\Omega\times]0, T[$ as

$$\partial_t(\varphi s) - \nabla \cdot (\eta_w(s)\nabla p_w) = f_w(s_{\rm in})\overline{q} - f_w(s)\underline{q} \tag{1.2}$$

$$-\partial_t(\varphi s) - \nabla \cdot (\eta_o(s)\nabla p_o) = f_o(s_{\rm in})\overline{q} - f_o(s)\underline{q} \tag{1.3}$$

complemented by a natural boundary condition almost everywhere on $\partial\Omega\times]0, T[$:

$$\eta_w(s)\nabla p_w \cdot \mathbf{n} = 0, \quad \eta_o(s)\nabla p_o \cdot \mathbf{n} = 0 \tag{1.4}$$

and an initial condition almost everywhere in Ω:

$$s(\cdot, 0) = s^0 := s_w^0, \quad 0 \leqslant s_w^0 \leqslant 1. \tag{1.5}$$

The fractional flows are related to the mobilities by

$$\forall\, 0 \leqslant s \leqslant 1, \qquad f_w(s) = \frac{\eta_w(s)}{\eta_w(s) + \eta_o(s)}, \quad f_o(s) = 1 - f_w(s). \tag{1.6}$$

Recall that the phase saturations sum up to 1 and the phase pressures are related by the capillary pressure, p_c, which is a function of the saturation:

$$\forall\, 0 \leqslant s \leqslant 1, \quad p_c(s) = p_o - p_w. \tag{1.7}$$

This work is done under the following basic assumptions.

Assumption 1.1.

- The porosity φ is piecewise constant in space, independent of time, positive, bounded, and uniformly bounded away from zero.
- The mobility of the wetting phase $\eta_w \geqslant 0$ is continuous and increasing on the interval $[0, 1]$. The mobility of the non-wetting phase $\eta_o \geqslant 0$ is continuous and decreasing on the interval $[0, 1]$. This implies that the function f_w is increasing and the function f_o is decreasing on $[0, 1]$. We also recall that these functions are degenerate, indeed they satisfy:

$$\eta_w(0) = 0, \quad \eta_o(1) = 0. \tag{1.8}$$

- There is a positive constant η_* such that

$$\eta_w(s) + \eta_o(s) \geqslant \eta_* \quad \forall s \in [0, 1]. \tag{1.9}$$

- The capillary pressure p_c is a continuous, strictly decreasing function in $W^{1,1}(0, 1)$.
- The flow rates at the injection and production wells, $\overline{q}, \underline{q} \in L^2(\Omega\times]0, T[)$ satisfy

$$\overline{q} \geqslant 0, \quad \underline{q} \geqslant 0, \quad \int_\Omega \overline{q} = \int_\Omega \underline{q}. \tag{1.10}$$

- The prescribed input saturation $s_{\rm in}$ satisfies almost everywhere in $\Omega\times]0, T[$

$$0 \leqslant s_{\rm in} \leqslant 1. \tag{1.11}$$

Since $p_c, \eta_\alpha, f_\alpha, \alpha = w, o$ are bounded above and below, it is convenient to extend them continuously by constants to $\mathbb{R}$.

Although the numerical scheme studied below does not discretize the global pressure, following [16], its convergence proof uses a number of auxiliary functions related to the global pressure. First, we introduce the primitive g_c of p_c,

$$\forall x \in [0, 1], \quad g_c(x) = \int_x^1 p_c(s)\,\mathrm{d}s. \tag{1.12}$$

Since p_c is a continuous function on $[0,1]$, the function g_c belongs to $\mathcal{C}^1([0,1])$. Next, we introduce the auxiliary pressures p_{wg}, p_{wo}, and g,

$$\forall x \in [0,1], \quad p_{wg}(x) = \int_0^x f_o(s) p_c'(s)\, \mathrm{d}s, \quad p_{og}(x) = \int_0^x f_w(s) p_c'(s)\, \mathrm{d}s \tag{1.13}$$

$$\forall x \in [0,1], \quad g(x) = -\int_0^x \frac{\eta_w(s)\eta_o(s)}{\eta_w(s)+\eta_o(s)} p_c'(s)\, \mathrm{d}s. \tag{1.14}$$

Owing to (1.6),

$$\forall x \in [0,1], \quad p_{wg}(x) + p_{og}(x) = \int_0^x p_c'(s)\, \mathrm{d}s = p_c(x) - p_c(0). \tag{1.15}$$

Moreover, the derivative of g satisfies formally the identities

$$\forall x \in [0,1], \quad \eta_\alpha(x) p_{\alpha g}'(x) + g'(x) = 0, \quad \alpha = w, o. \tag{1.16}$$

1.2 Weak variational formulation

By multiplying (1.2) and (1.3) with a smooth function v, say $v \in \mathcal{C}^1(\Omega \times [0,T])$ that vanishes at $t = T$, applying Green's formula in time and space, and using the boundary and initial conditions (1.4) and (1.5), we formally derive a weak variational formulation

$$-\int_0^T \int_\Omega \varphi\, s\, \partial_t v + \int_0^T \int_\Omega \eta_w(s) \nabla p_w \cdot \nabla v = \int_\Omega \varphi\, s^0 v(0) + \int_0^T \int_\Omega (f_w(s_{\text{in}})\overline{q} - f_w(s)\underline{q}) v$$

$$\int_0^T \int_\Omega \varphi\, s\, \partial_t v + \int_0^T \int_\Omega \eta_o(s) \nabla p_o \cdot \nabla v = -\int_\Omega \varphi\, s^0 v(0) + \int_0^T \int_\Omega (f_o(s_{\text{in}})\overline{q} - f_o(s)\underline{q}) v.$$

But in general, the pressures are not sufficiently smooth to make this formulation meaningful and following [8], by using (1.16), it is rewritten in terms of the artificial pressures,

$$\begin{aligned} -\int_0^T \int_\Omega \varphi\, s\, \partial_t v + \int_0^T \int_\Omega (\eta_w(s)\nabla(p_w + p_{wg}(s)) + \nabla g(s)) \cdot \nabla v &= \int_\Omega \varphi\, s^0 v(0) \\ &+ \int_0^T \int_\Omega (f_w(s_{\text{in}})\overline{q} - f_w(s)\underline{q}) v \\ \int_0^T \int_\Omega \varphi\, s\, \partial_t v + \int_0^T \int_\Omega (\eta_o(s)\nabla(p_o - p_{og}(s)) - \nabla g(s)) \cdot \nabla v &= -\int_\Omega \varphi\, s^0 v(0) \\ &+ \int_0^T \int_\Omega (f_o(s_{\text{in}})\overline{q} - f_o(s)\underline{q}) v. \end{aligned} \tag{1.17}$$

With the above assumptions, problem (1.17) has been analyzed in reference [1], where it is shown that it has a solution s in $L^\infty(\Omega\times]0,T[)$ with $g(s)$ in $L^2(0,T;H^1(\Omega))$, p_α, $\alpha = w, o$, in $L^2(\Omega\times]0,T[)$ with both $p_w + p_{wg}(s)$ and $p_o - p_{og}(s)$ in $L^2(0,T;H^1(\Omega))$.

2 Scheme

From now on, we assume that Ω is a polygon ($d = 2$) or Lipschitz polyhedron ($d = 3$) so it can be entirely meshed.

2.1 Meshes and discretization spaces

The mesh $\mathcal{T}_h$ is a regular family of simplices K, with a constraint on the angle that will be used to enforce the maximum principle: each angle is not larger than $\pi/2$, see [6]. This is easily constructed in 2D. In 3D, since we only investigate convergence we can embed the domain in a triangulated box. Moreover, since the porosity φ is a piecewise constant, to simplify we also assume that the mesh is such that φ is a constant per element. The parameter h denotes the mesh size, i.e., the maximum diameter of the simplices. On this mesh, we consider the standard finite element space of order one

$$X_h = \{v_h \in \mathcal{C}^0(\overline{\Omega});\ \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{P}_1\}. \tag{2.1}$$

Thus the dimension of X_h is the number of nodes, say M, of $\mathcal{T}_h$. Let φ_i be the Lagrange basis function, that is piecewise linear, and takes the value 1 at node i and the value 0 at all other nodes. As usual, the Lagrange interpolation operator $I_h \in \mathcal{L}(\mathcal{C}^0(\overline{\Omega}); X_h)$ is defined by

$$\forall v \in \mathcal{C}^0(\overline{\Omega}), \quad I_h(v) = \sum_{i=1}^{M} v_i \varphi_i \tag{2.2}$$

where v_i is the value of v at the node of index i. It is easy to see that under the mesh condition, we have

$$\forall K, \quad \int_K \nabla\varphi_i \cdot \nabla\varphi_j \leqslant 0 \quad \forall i \neq j. \tag{2.3}$$

For a given node i, we denote by Δ_i the union of elements sharing the node i and by $\mathcal{N}(i)$ the set of indices of all the nodes in Δ_i. In the spirit of [21], we define

$$c_{ij} = \int_{\Delta_i \cap \Delta_j} |\nabla\varphi_i \cdot \nabla\varphi_j| \quad \forall i, j. \tag{2.4}$$

Recall that the trapezoidal rule on a triangle or a tetrahedron K is

$$\int_K f \approx \frac{1}{d+1}|K| \sum_{\ell=1}^{d+1} f_{i_\ell}$$

where f_{i_ℓ} is the value of the function f at the ℓth node (vertex), with global number i_ℓ, of K. For any region $\mathcal{O}$, the notation $|\mathcal{O}|$ means the measure (volume) of $\mathcal{O}$.

We define

$$m_i = \frac{1}{d+1} \sum_{K \in \Delta_i} |K| = \frac{1}{d+1}|\Delta_i|$$

and taking into account the porosity φ, we define more generally

$$\widetilde{m}_i(\varphi) = \frac{1}{d+1} \sum_{K \in \Delta_i} \varphi|_K |K|$$

so that $m_i = \widetilde{m}_i(1)$. It is well-known that the trapezoidal rule defines a norm on X_h, $\|\cdot\|_h$, uniformly equivalent to L^2 norm. Let $U_h \in X_h$ and write

$$U_h = \sum_{i=1}^{M} U^i \varphi_i.$$

The discrete L^2 norm associated with the trapezoidal rule is

$$\|U_h\|_h = \left(\sum_{i=1}^{M} m_i |U^i|^2\right)^{1/2}.$$

There exist positive constants $\underline{C}$ and $\overline{C}$, independent of h and M, such that

$$\forall U_h \in X_h, \quad \underline{C}\,\|U_h\|^2_{L^2(\Omega)} \leqslant \|U_h\|_h^2 \leqslant \overline{C}\,\|U_h\|^2_{L^2(\Omega)}. \tag{2.5}$$

This is also true for other piecewise polynomial functions, but with possibly different constants. The scalar product associated with this norm is denoted by $(\cdot,\cdot)_h$,

$$\forall U_h, V_h \in X_h, \quad (U_h, V_h)_h = \sum_{i=1}^{M} m_i U^i V^i. \tag{2.6}$$

By analogy, we introduce the notation

$$\forall U_h, V_h \in X_h, \quad (U_h, V_h)_h^{\varphi} = \sum_{i=1}^{M} \widetilde{m}_i(\varphi) U^i V^i. \tag{2.7}$$

The assumptions on the porosity φ imply that (2.7) defines a weighted scalar product associated with the weighted norm $\|\cdot\|_h^{\varphi}$,

$$\forall U_h \in X_h, \quad \|U_h\|_h^{\varphi} = \big((U_h, U_h)_h^{\varphi}\big)^{1/2}$$

that satisfies the analogue of (2.5), with the same constants $\underline{C}$ and $\overline{C}$,

$$\forall U_h \in X_h, \quad \underline{C}\,(\min_{\Omega} \varphi)\,\|U_h\|_{L^2(\Omega)}^2 \leqslant \big(\|U_h\|_h^{\varphi}\big)^2 \leqslant \overline{C}\,(\max_{\Omega} \varphi)\,\|U_h\|_{L^2(\Omega)}^2. \tag{2.8}$$

2.2 Motivation of the space discretization

While discretizing the time derivative is fairly straightforward, discretizing the space derivatives is more delicate because we need a scheme that is consistent and satisfies the maximum principle for the saturation. For the moment, we freeze the time variable and focus on consistency in space. First, we recall a standard property of functions of X_h on meshes satisfying (2.3).

Proposition 2.1. Under condition (2.3), the following identities holds for all U_h and V_h in X_h, with c_{ij} defined in (2.4):

$$\int_{\Omega} \nabla U_h \cdot \nabla V_h = -\sum_{i=1}^{M} U^i \sum_{j\neq i, j\in \mathcal{N}(i)} c_{ij}(V^j - V^i) = \frac{1}{2}\sum_{i=1}^{M} \sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}(U^j - U^i)(V^j - V^i). \tag{2.9}$$

Proof. The first equality is obtained by using (2.3), (2.4) and the fact that

$$\sum_{j=1}^{M} \varphi_j = 1$$

as in [18, Sect. 12.1].

For the second part, we use the symmetry of c_{ij} and the anti-symmetry of $V^j - V^i$ to deduce that

$$-\sum_{i=1}^{M} U^i \sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}(V^j - V^i) = \frac{1}{2}\sum_{i=1}^{M} \sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}(U^j - U^i)(V^j - V^i)$$

which is the desired result. □

Note that c_{ij} vanishes when $j \notin \mathcal{N}(i)$. Therefore, when there is no ambiguity it is convenient to write the above double sums on i and j with i and j running from 1 to M.

As an immediate consequence of Proposition 2.1, we have, by taking $V_h = U_h$,

$$\forall U_h \in X_h, \quad \|\nabla U_h\|_{L^2(\Omega)} = \frac{1}{\sqrt{2}}\left(\sum_{i,j=1}^{M} c_{ij}|U^j - U^i|^2\right)^{1/2}. \tag{2.10}$$

Now, we consider the case of the product of the gradients by a third function. Beforehand, we introduce the following notation: for indices i and j of two neighboring interior nodes, $\Delta_i \cap \Delta_j$ in two dimensions is the union of two triangles and in three dimensions the union of a number of tetrahedra bounded by a fixed constant, say L, determined by the regularity of the mesh. We shall use the following notation

$$c_{ij,K} = \int_K |\nabla \varphi_i \cdot \nabla \varphi_j|, \quad w_K = \frac{1}{|K|}\int_K w. \tag{2.11}$$

Note that

$$\sum_{K\subset \Delta_i\cap\Delta_j} c_{ij,K} = c_{ij}. \tag{2.12}$$

Then we have the following proposition.

Proposition 2.2. Let (2.3) hold. With the notation (2.11), the following identity holds for all w in $L^1(\Omega)$:

$$\forall U_h, V_h \in X_h,\ \int_\Omega w(\nabla U_h \cdot \nabla V_h) = -\sum_{i=1}^{M} U^i \sum_{j=1}^{M} \Big(\sum_{K\subset \Delta_i\cap\Delta_j} c_{ij,K} w_K \Big)(V^j - V^i). \tag{2.13}$$

Proof. It is easy to prove that

$$\int_\Omega w(\nabla U_h \cdot \nabla V_h) = \sum_{i,j=1}^{M} d_{ij} U^i V^j \tag{2.14}$$

where

$$d_{ij} = \int_{\Delta_i\cap\Delta_j} w(\nabla \varphi_i \cdot \nabla \varphi_j) = \int_\Omega w(\nabla \varphi_i \cdot \nabla \varphi_j). \tag{2.15}$$

Again, we have for any i,

$$\sum_{j=1}^{M} d_{ij} = 0, \qquad d_{ii} = -\sum_{1\leqslant j\leqslant M, j\neq i} d_{ij}$$

and by substituting this equality into (2.14), we obtain

$$\int_\Omega w(\nabla U_h \cdot \nabla V_h) = \sum_{i,j=1}^{M} U^i d_{ij}(V^j - V^i). \tag{2.16}$$

But, in view of (2.11) and (2.15), and since $\nabla \varphi_i \cdot \nabla \varphi_j$ is a constant in each element K contained in $\Delta_i \cap \Delta_j$,

$$d_{ij} = -\sum_{K\subset\Delta_i\cap\Delta_j} c_{ij,K} w_K, \tag{2.17}$$

and (2.13) follows by substituting this equation into (2.16). □

Note that $d_{ij} = d_{ji}$ owing to (2.17). The first consequence of Proposition 2.2 is that the right-hand side of (2.13) is a consistent approximation of $(w, \nabla u \cdot \nabla v)$.

Proposition 2.3. Let (2.3) hold, let u and v belong to $H^2(\Omega)$ and w to $L^\infty(\Omega)$, and let $U_h = I_h u$, $V_h = I_h v$ be defined by (2.2). Then, there exists a constant C, independent of h, M, u, v, and w, such that

$$\Big| \int_\Omega w\nabla u \cdot \nabla v + \sum_{i,j=1}^{M} U^i \Big(\sum_{K\subset\Delta_i\cap\Delta_j} c_{ij,K} w_K \Big)(V^j - V^i) \Big| \leqslant C\, h\, \|w\|_{L^\infty(\Omega)} \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)}. \tag{2.18}$$

Proof. In view of the identity (2.13), the left-hand side of (2.18) is bounded as follows:

$$\Big| \int_\Omega w(\nabla u \cdot \nabla v - \nabla U_h \cdot \nabla V_h) \Big| \leqslant \|w\|_{L^\infty(\Omega)} \Big(\|\nabla(u - U_h)\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|\nabla(v - V_h)\|_{L^2(\Omega)} \|\nabla U_h\|_{L^2(\Omega)} \Big).$$

From here, (2.18) is a consequence of standard finite element interpolation error. □

Now, if w is in $W^{1,\infty}(\Omega)$, then again, standard finite element approximation shows that there exists a constant C, independent of h, $K \subset \Delta_i \cap \Delta_j$, and w, such that

$$\|w_K - w\|_{L^\infty(K)} \leqslant C\, h\, |w|_{W^{1,\infty}(K)} \leqslant C\, h\, |w|_{W^{1,\infty}(\Omega)}. \tag{2.19}$$

As a consequence, we will show that in the error formula (2.18), the average w_K can be replaced by any value of w in K. Since all K in $\Delta_i \cap \Delta_j$ share the edge, say e_{ij}, whose end points are the nodes with indices i and j, then we can pick the value of w at any point, say $\widetilde{W}^{i,j}$, of e_{ij}. At this stage, we choose this value freely, but we prescribe that it be symmetrical with respect to i and j, i.e.,

$$\widetilde{W}^{i,j} = \widetilde{W}^{j,i}. \tag{2.20}$$

Then we have the following approximation result.

Theorem 2.1. *With the assumption and notation of Proposition 2.3, there exists a constant C, independent of h and M, such that for all u, and v in $H^2(\Omega)$ and w in $W^{1,\infty}(\Omega)$,*

$$\int_\Omega w\nabla u \cdot \nabla v = -\sum_{i,j=1}^{M} U^i c_{ij}\widetilde{W}^{i,j}(V^j - V^i) + R \tag{2.21}$$

for any arbitrary value $\widetilde{W}^{i,j}$ of w in the common edge e_{ij} satisfying (2.20), and the remainder R satisfies

$$|R| \leqslant C\,h\,|w|_{W^{1,\infty}(\Omega)}\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}. \tag{2.22}$$

Proof. We infer from (2.12) and (2.13) that

$$\int_\Omega w(\nabla U_h \cdot \nabla V_h) = -\sum_{i,j=1}^{M} U^i(V^j - V^i)\sum_{K\subset\Delta_i\cap\Delta_j} c_{ij,K}(w_K - \widetilde{W}^{i,j}) - \sum_{i,j=1}^{M} U^i c_{ij}(V^j - V^i)\widetilde{W}^{i,j}.$$

Let

$$R_{ij} = \sum_{K\subset\Delta_i\cap\Delta_j} c_{ij,K}(w_K - \widetilde{W}^{i,j})$$

which is symmetric in i and j by assumption (2.20). As in Proposition 2.1, the symmetry of R_{ij} and the anti-symmetry of $V^j - V^i$, imply

$$-\sum_{i,j=1}^{M} U^i R_{ij}(V^j - V^i) \leqslant \frac{1}{2}\Big(\sum_{i,j=1}^{M}|R_{ij}|(U^j - U^i)^2\Big)^{1/2}\Big(\sum_{i,j=1}^{M}|R_{ij}|(V^j - V^i)^2\Big)^{1/2}. \tag{2.23}$$

From the nonnegativity of $c_{ij,K}$, (2.12), and (2.19), we infer that

$$|R_{ij}| \leqslant \Big(\sum_{K\subset\Delta_i\cap\Delta_j} c_{ij,K}\Big)C\,h\,|w|_{W^{1,\infty}(\Omega)} = c_{ij}C\,h\,|w|_{W^{1,\infty}(\Omega)}.$$

Hence, with (2.10) and standard finite element approximation,

$$\Big|\sum_{i,j=1}^{M} U^i R_{ij}(V^j - V^i)\Big| \leqslant C\,h\,|w|_{W^{1,\infty}(\Omega)}\|\nabla U_h\|_{L^2(\Omega)}\|\nabla V_h\|_{L^2(\Omega)} \leqslant C\,h\,|w|_{W^{1,\infty}(\Omega)}\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}.$$

The result follows by combining this inequality with (2.18). □

The above considerations show that

$$-\sum_{i,j=1}^{M} U^i c_{ij}\widetilde{W}^{i,j}(V^j - V^i) \text{ is a consistent approximation of order one of } \int_\Omega w\nabla u \cdot \nabla v$$

for any symmetric choice of $\widetilde{W}^{i,j}$ in e_{ij}, the common edge of $\Delta_i \cap \Delta_j$. This will lead to the upwinded space discretization in the next subsection (see also [24]). Furthermore, for all real numbers V^i and $\widetilde{W}^{i,j}$ satisfying (2.20), $1 \leqslant i, j \leqslant M$, the symmetry of c_{ij} and anti-symmetry of $V^j - V^i$ imply

$$\sum_{i,j=1}^{M} c_{ij}\widetilde{W}^{i,j}(V^j - V^i) = 0. \tag{2.24}$$

2.3 Fully discrete scheme

Let $\tau = T/N$ be the time step, $t_n = n\tau$, the discrete times, $0 \leqslant n \leqslant N$. Regarding time, we shall use the standard L^2 projection ρ_τ defined on $]t_{n-1}, t_n]$, for any function f in $L^1(0, T)$, by

$$\rho_\tau(f)^n := \rho_\tau(f)|_{]t_{n-1},t_n]} := \frac{1}{\tau}\int_{t_{n-1}}^{t_n} f. \tag{2.25}$$

Regarding space, we shall use a standard element-by-element L^2 projection ρ_h as well as a nodal approximation operator r_h defined at each node $\boldsymbol{x}_i$ for any function $g \in L^1(\Omega)$ by

$$r_h(g)(\boldsymbol{x}_i) = \frac{1}{|\Delta_i|}\int_{\Delta_i} g, \quad 1 \leqslant i \leqslant M \tag{2.26}$$

and extended to Ω by $r_h(g) \in X_h$. The operator ρ_h is defined for any f in $L^1(\Omega)$ by $\rho_h(f)|_K = \rho_K(f)$ where, in any element K,

$$\rho_K(f) = \frac{1}{|K|}\int_K f. \tag{2.27}$$

The initial saturation s^0 is approximated by the operator r_h,

$$S_h^0 = r_h(s^0). \tag{2.28}$$

The input saturation $s_{\rm in}$ is approximated in space and time by

$$s_{\rm in,h,\tau} = \rho_\tau(r_h(s_{\rm in})) \tag{2.29}$$

with space-time nodal values denoted by $s_{\rm in}^{n,i}$. Clearly, (1.11) implies in space and time

$$0 \leqslant s_{\rm in,h,\tau} \leqslant 1.$$

In order to preserve (1.10), the functions $\overline{q}$ and $\underline{q}$ are approximated by the functions $\overline{q}_{h,\tau}$ and $\underline{q}_{h,\tau}$ defined with r_h and corrected as follows:

$$\overline{q}_{h,\tau} = \rho_\tau\left(r_h(\overline{q}) - \frac{1}{|\Omega|}\int_\Omega (r_h(\overline{q}) - \overline{q})\right), \quad \underline{q}_{h,\tau} = \rho_\tau\left(r_h(\underline{q}) - \frac{1}{|\Omega|}\int_\Omega (r_h(\underline{q}) - \underline{q})\right). \tag{2.30}$$

Since $\overline{q}_{h,\tau}$ and $\underline{q}_{h,\tau}$ are piecewise linears in space, they are exactly integrated by the trapezoidal rule and we easily derive from (1.10) and (2.30) that we have for all n,

$$(\overline{q}_h^n, 1)_h = (\underline{q}_h^n, 1)_h. \tag{2.31}$$

The set of primary unknowns is the discrete wetting phase saturation and the discrete wetting phase pressure, S_h^n and $P_{w,h}^n$, defined pointwise at time t_n by:

$$S_h^n = \sum_{i=1}^M S^{n,i}\varphi_i, \quad P_{w,h}^n = \sum_{i=1}^M P_w^{n,i}\varphi_i, \quad 1 \leqslant n \leqslant N.$$

Then the discrete non-wetting phase pressure $P_{o,h}^n$ defined by

$$P_{o,h}^n = \sum_{i=1}^M P_o^{n,i}\varphi_i, \quad 1 \leqslant n \leqslant N$$

is a secondary unknown. The upwind scheme we propose for discretizing (1.2)–(1.3) is inspired by the control volume finite element approach in [19] and by the finite volume scheme in [16]. For each time step n, $1 \leqslant n \leqslant N$, the lines of the discrete equations are

$$\frac{\widetilde{m}_i(\varphi)}{\tau}(S^{n,i} - S^{n-1,i}) - \sum_{j=1}^M c_{ij}\eta_w(S_w^{n,ij})(P_w^{n,j} - P_w^{n,i}) = m_i\left(f_w(s_{\rm in}^{n,i})\overline{q}^{n,i} - f_w(S^{n,i})\underline{q}^{n,i}\right) \tag{2.32}$$

$$-\frac{\widetilde{m}_i(\varphi)}{\tau}(S^{n,i} - S^{n-1,i}) - \sum_{j=1}^M c_{ij}\eta_o(S_o^{n,ij})(P_o^{n,j} - P_o^{n,i}) = m_i\left(f_o(s_{\rm in}^{n,i})\overline{q}^{n,i} - f_o(S^{n,i})\underline{q}^{n,i}\right) \tag{2.33}$$

$$P_o^{n,i} - P_w^{n,i} = p_c(S^{n,i}), \quad 1 \leqslant i \leqslant M \tag{2.34}$$

$$\sum_{i=1}^M m_i P_w^{n,i} = 0. \tag{2.35}$$

Here i runs from 1 to $M-1$ in (2.32) and from 1 to M in (2.33); the upwind values $S_w^{n,ij}$, $S_o^{n,ij}$ are defined by

$$S_w^{n,ij} = \begin{cases} S^{n,i}, & P_w^{n,i} > P_w^{n,j} \\ S^{n,j}, & P_w^{n,i} < P_w^{n,j} \\ \max(S^{n,i}, S^{n,j}), & P_w^{n,i} = P_w^{n,j} \end{cases} \tag{2.36}$$

$$S_o^{n,ij} = \begin{cases} S^{n,i}, & P_o^{n,i} > P_o^{n,j} \\ S^{n,j}, & P_o^{n,i} < P_o^{n,j} \\ \min(S^{n,i}, S^{n,j}), & P_o^{n,i} = P_o^{n,j}. \end{cases} \tag{2.37}$$

We observe that

$$S_w^{n,ij} = S_w^{n,ji}, \qquad S_o^{n,ij} = S_o^{n,ji}$$

so that, if we interpret in (2.32) (respectively, (2.33)) $\eta_w(S_w^{n,ij})$ (respectively, $\eta_o(S_o^{n,ij})$) as $\widetilde{W}^{i,j}$, then (2.20) and hence (2.24) hold.

Remark 2.1. Before setting (2.32)–(2.35) in variational form, observe that:

1. The scheme (2.32)–(2.35) forms a square system in the primary unknowns, S_h^n and $P_{w,h}^n$.
2. Formula (2.32) is also valid for $i = M$. Indeed, we pass to the left-hand side the right-hand side of (2.32) and set A^i the resulting line of index i. Let $\widetilde{A}^M$ denote what should be the line of index M, i.e.,

$$\widetilde{A}^M = \frac{\widetilde{m}_M(\varphi)}{\tau}(S^{n,M} - S^{n-1,M}) - \sum_{j=1}^{M} c_{Mj}\eta_w(S_w^{n,Mj})(P_w^{n,j} - P_w^{n,M})$$
$$- m_M(f_w(s_{\text{in}}^{n,M})\overline{q}^{n,M} - f_w(S^{n,M})\underline{q}^{n,M}).$$

   Then, in view of (2.24),

$$\widetilde{A}^M = \sum_{i=1}^{M-1} A^i + \widetilde{A}^M = \sum_{i=1}^{M} \frac{\widetilde{m}_i(\varphi)}{\tau}(S^{n,i} - S^{n-1,i}) - \sum_{i=1}^{M} m_i(f_w(s_{\text{in}}^{n,i})\overline{q}^{n,i} - f_w(S^{n,i})\underline{q}^{n,i}).$$

   By summing in the same fashion the lines of (2.33), we obtain

$$\sum_{i=1}^{M} \frac{\widetilde{m}_i(\varphi)}{\tau}(S^{n,i} - S^{n-1,i}) = -\sum_{i=1}^{M} m_i(f_o(s_{\text{in}}^{n,i})\overline{q}^{n,i} - f_o(S^{n,i})\underline{q}^{n,i}).$$

   A combination of these two equations yields

$$\widetilde{A}^M = -\sum_{i=1}^{M} m_i\Big((f_w(s_{\text{in}}^{n,i}) + f_o(s_{\text{in}}^{n,i}))\overline{q}^{n,i} - (f_w(S^{n,i}) + f_o(S^{n,i}))\underline{q}^{n,i}\Big) = -\sum_{i=1}^{M} m_i(\overline{q}^{n,i} - \underline{q}^{n,i}) = 0$$

   by virtue of (1.6), the definition (2.25), and (1.10).
3. In (2.32) (respectively, (2.33)), any constant can be added to P_w (respectively, P_o), but in view of (2.34), the constant must be the same for both pressures. The last equation (2.35) is added to resolve this constant.

As usual, it is convenient to associate time functions $S_{h,\tau}$, $P_{\alpha,h,\tau}$ with the sequences indexed by n. These are piecewise constant in time in $]0, T[$, for instance

$$P_{\alpha,h,\tau}(t,x) = P_{\alpha,h}^n(x), \ \alpha = w, o \quad \forall (t,x) \in \Omega \times]t_{n-1}, t_n]. \tag{2.38}$$

In view of the material of the previous subsection, we introduce the following form:

$$\forall W_h, U_h, V_h, Z_h \in X_h, \quad [Z_h, W_h; V_h, U_h]_h = \sum_{i,j=1}^{M} U^i c_{ij} \widetilde{W}^{ij}(V^j - V^i) \tag{2.39}$$

where the first argument Z_h indicates that the choice of $\widetilde{W}^{ij}$ depends on Z_h. Such dependence, used for the upwinding, will be specified further on, but *it is assumed from now on* that $\widetilde{W}^{ij}$ satisfies (2.20). Considering (2.24), the form satisfies the following properties,

$$\forall Z_h, W_h, V_h \in X_h, \quad [Z_h, W_h; V_h, 1]_h = 0 \tag{2.40}$$

$$\forall Z_h, W_h, V_h \in X_h, \quad [Z_h, W_h; V_h, V_h]_h = -\frac{1}{2}\sum_{i,j=1}^{M} c_{ij}\widetilde{W}_{ij}(V^i - V^j)^2. \tag{2.41}$$

This last property is derived by the same argument as in proving (2.9).

With the above notation, and taking into account that (2.32) extends to $i = M$, the scheme (2.32)–(2.35) has the equivalent variational form. Starting from S_h^0 (see (2.28)): Find S_h^n, $P_{w,h}^n$, and $P_{o,h}^n$ in X_h, for $1 \leqslant n \leqslant N$, solution of, for all ϑ_h in X_h,

$$\frac{1}{\tau}(S_h^n - S_h^{n-1}, \vartheta_h)_h^{\varphi} - [P_{w,h}^n, I_h(\eta_w(S_h^n)); P_{w,h}^n, \vartheta_h]_h = (I_h(f_w(s_{\mathrm{in},h}^n))\overline{q}_h^n - I_h(f_w(S_h^n))\underline{q}_h^n, \vartheta_h)_h \tag{2.42}$$

$$-\frac{1}{\tau}(S_h^n - S_h^{n-1}, \vartheta_h)_h^{\varphi} - [P_{o,h}^n, I_h(\eta_o(S_h^n)); P_{o,h}^n, \vartheta_h]_h = (I_h(f_o(s_{\mathrm{in},h}^n))\overline{q}_h^n - I_h(f_o(S_h^n))\underline{q}_h^n, \vartheta_h)_h \tag{2.43}$$

$$P_{o,h}^n - P_{w,h}^n = I_h(p_c(S_h^n)) \tag{2.44}$$

$$(P_{w,h}^n, 1)_h = 0 \tag{2.45}$$

where the choice of $\eta_w(S_h^n)$ in the left-hand side of (2.42) (respectively, $\eta_o(S_h^n)$ in the left-hand side of (2.43)) is given by (2.36) (respectively (2.37)). Strictly speaking, the interpolation operator I_h is introduced in (2.42) and (2.43) because the forms are defined for functions of X_h, but for the sake of simplicity, since only nodal values are used, it may be dropped further on.

We shall see that under the above basic hypotheses, the discrete problem (2.42)–(2.45) has at least one solution. In the sequel, we shall use the following discrete auxiliary pressures:

$$U_{w,h,\tau} = P_{w,h,\tau} + I_h(p_{wg}(S_{h,\tau})), \qquad U_{o,h,\tau} = P_{o,h,\tau} - I_h(p_{og}(S_{h,\tau})). \tag{2.46}$$

3 A priori bounds

The present section is devoted to basic a priori bounds used in proving existence of a discrete solution. Existence is fairly technical and will be postponed till Section 4. The first step is a key bound on the discrete saturation. In the second step, this bound will lead to a pressure estimate and in particular to a bound on the discrete analogue of auxiliary pressures.

3.1 Maximum principle

The scheme (2.32)–(2.35) satisfies the maximum principle property. The proof given below uses a standard argument as in [16].

Theorem 3.1. *The following bounds hold:*

$$0 \leqslant S_{h,\tau} \leqslant 1. \tag{3.1}$$

Proof. As $0 \leqslant s^0 \leqslant 1$ almost everywhere, by construction (2.28), we immediately have

$$0 \leqslant \min_{\Omega} s^0 \leqslant S_h^0 \leqslant \max_{\Omega} s^0 \leqslant 1.$$

Now, the proof proceeds by contradiction. Assume that there is an index $n \geqslant 1$ such that

$$S_h^{n-1} \leqslant 1$$

and that there is a node i such that

$$S^{n,i} = \|S_h^n\|_{L^\infty(\Omega)} > 1$$

and thus

$$S^{n,i} > S^{n-1,i}.$$

Dropping the index n in the rest of the proof, (2.32) and (2.33) imply

$$\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_w(S_w^{ij})(P_w^j - P_w^i) + m_i\left(f_w(s_{\text{in}}^i)\overline{q}^i - f_w(S^i)\underline{q}^i\right) > 0 \tag{3.2}$$

$$-\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_o(S_o^{ij})(P_o^j - P_o^i) - m_i\left(f_o(s_{\text{in}}^i)\overline{q}^i - f_o(S^i)\underline{q}^i\right) > 0. \tag{3.3}$$

We first show that (3.2) holds true with S_w^{ij} replaced by S^i. Indeed if $P_w^i > P_w^j$, then $S_w^{ij} = S^i$. If $P_w^i < P_w^j$, then $S_w^{ij} = S^j$, and as η_w is increasing and by assumption, $S^j \leqslant S^i$,

$$\eta_w(S_w^{ij})(P_w^j - P_w^i) \leqslant \eta_w(S^i)(P_w^j - P_w^i).$$

Finally, the term vanishes when $P_w^i = P_w^j$. Therefore we have in all cases

$$\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_w(S^i)(P_w^j - P_w^i) + m_i\left(f_w(s_{\text{in}}^i)\overline{q}^i - f_w(S^i)\underline{q}^i\right) > 0. \tag{3.4}$$

A similar argument gives

$$-\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_o(S^i)(P_o^j - P_o^i) - m_i\left(f_o(s_{\text{in}}^i)\overline{q}^i - f_o(S^i)\underline{q}^i\right) > 0. \tag{3.5}$$

The substitution of (2.34) into (3.5) yields

$$-\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_o(S^i)\left((P_w^j - P_w^i) + (p_c(S^j) - p_c(S^i))\right) - m_i\left(f_o(s_{\text{in}}^i)\overline{q}^i - f_o(S^i)\underline{q}^i\right) > 0. \tag{3.6}$$

Since p_c is decreasing and $S^i \geqslant S^j$, the second term in the above sum is negative. This implies that

$$-\sum_{j\neq i, j\in\mathcal{N}(i)} c_{ij}\eta_o(S^i)(P_w^j - P_w^i) - m_i\left(f_o(s_{\text{in}}^i)\overline{q}^i - f_o(S^i)\underline{q}^i\right) > 0. \tag{3.7}$$

The sum on j cancels by multiplying (3.4) by $\eta_o(S^i)$, (3.7) by $\eta_w(S^i)$, and adding the two. The sign is unchanged because either $\eta_o(S^i)$ or $\eta_w(S^i)$ is strictly positive. Hence,

$$m_i\eta_o(S^i)\left(f_w(s_{\text{in}}^i)\overline{q}^i - f_w(S^i)\underline{q}^i\right) - m_i\eta_w(S^i)\left(f_o(s_{\text{in}}^i)\overline{q}^i - f_o(S^i)\underline{q}^i\right) > 0.$$

By definition of f_w and f_o, this reduces to

$$\eta_o(S^i)f_w(s_{\text{in}}^i) - \eta_w(S^i)f_o(s_{\text{in}}^i) > 0. \tag{3.8}$$

Now consider the function:

$$r(s) = \eta_o(s)f_w(s_{\text{in}}^i) - \eta_w(s)f_o(s_{\text{in}}^i). \tag{3.9}$$

It is decreasing and $r(s_{\text{in}}^i) = 0$. Then, since $S^i > 1 \geqslant s_{\text{in}}^i$, see (1.11), we have

$$r(S^i) \leqslant r(s_{\text{in}}^i) = 0$$

which contradicts (3.8). The proof of the lower bound in (3.1) follows the same lines. □

3.2 Pressure bounds

The following properties will be used frequently.

Lemma 3.1. *The fact that p_c is strictly decreasing and* (2.34) *yield the following:*

$$P_w^i > P_w^j, \text{ and } P_o^i \leqslant P_o^j \text{ implies } S^i \geqslant S^j; \tag{3.10}$$

$$\text{if } P_w^i = P_w^j, \text{ then } P_o^i \geqslant P_o^j \text{ if and only if } S^i \leqslant S^j; \tag{3.11}$$

$$\text{if } P_o^i = P_o^j, \text{ then } P_w^i \leqslant P_w^j, \text{ if and only if } S^i \leqslant S^j. \tag{3.12}$$

Let us start with a lower bound that removes the degeneracy caused by the mobilities when they multiply the discrete pressures.

Lemma 3.2. *Let $U_{w,h}$ be defined by* (2.46) *with p_{wg} defined in* (1.13). *We have for all n and any i and j*

$$\eta_*(U_w^{n,j} - U_w^{n,i})^2 \leqslant \eta_w(S_w^{n,ij})(P_w^{n,j} - P_w^{n,i})^2 + \eta_o(S_o^{n,ij})(P_o^{n,j} - P_o^{n,i})^2. \tag{3.13}$$

Proof. To simplify the notation, we drop the superscript n. The second mean formula for integrals gives

$$p_{wg}(S^j) - p_{wg}(S^i) = \int_{S^i}^{S^j} f_o(s)p_c'(s)\,\mathrm{d}s = f_o(\xi)(p_c(S^j) - p_c(S^i)) \tag{3.14}$$

for some ξ between S^i and S^j. Using (2.34) we write

$$U_w^j - U_w^i = (1 - f_o(\xi))(P_w^j - P_w^i) + f_o(\xi)(P_o^j - P_o^i) = f_w(\xi)(P_w^j - P_w^i) + f_o(\xi)(P_o^j - P_o^i).$$

Therefore since $f_w + f_o = 1$, we have

$$(U_w^j - U_w^i)^2 \leqslant \frac{\eta_w(\xi)}{\eta_w(\xi) + \eta_o(\xi)}(P_w^j - P_w^i)^2 + \frac{\eta_o(\xi)}{\eta_w(\xi) + \eta_o(\xi)}(P_o^j - P_o^i)^2. \tag{3.15}$$

We now consider the following six cases.

1. If $P_w^i > P_w^j$ and $P_o^i \leqslant P_o^j$, then $\eta_w(S_w^{ij}) = \eta_w(S^i)$ and $\eta_o(S_o^{ij}) = \eta_o(S^j)$ when $P_o^i < P_o^j$; when $P_o^i = P_o^j$, the value of η_o does not matter. From (3.10) we then have $S^i \geqslant S^j$. Since η_w is increasing, $\eta_w(\xi) \leqslant \eta_w(S^i)$ and since η_o is decreasing, $\eta_o(\xi) \leqslant \eta_o(S^j)$. Thus we have

$$(U_w^j - U_w^i)^2 \leqslant \frac{\eta_w(S_w^{ij})}{\eta_w(\xi) + \eta_o(\xi)}(P_w^j - P_w^i)^2 + \frac{\eta_o(S_o^{ij})}{\eta_w(\xi) + \eta_o(\xi)}(P_o^j - P_o^i)^2$$

and with (1.9)

$$(U_w^j - U_w^i)^2 \leqslant \frac{1}{\eta_*}\left(\eta_w(S_w^{ij})(P_w^j - P_w^i)^2 + \eta_o(S_o^{ij})(P_o^j - P_o^i)^2\right). \tag{3.16}$$

2. If $P_w^i > P_w^j$ and $P_o^i > P_o^j$, then $\eta_w(S_w^{ij}) = \eta_w(S^i)$ and $\eta_o(S_o^{ij}) = \eta_o(S^i)$. From

$$\eta_o(S^i)(p_c(S^j) - p_c(S^i)) = (\eta_o(S^i) + \eta_w(S^i))\int_{S^i}^{S^j} f_o(S^i)p_c'(s)\,\mathrm{d}s$$

and (3.14), we derive

$$\begin{aligned}\eta_o(S^i)(p_c(S^j) - p_c(S^i)) &- (\eta_o(S^i) + \eta_w(S^i))(p_{wg}(S^j) - p_{wg}(S^i))\\ &= (\eta_o(S^i) + \eta_w(S^i))\int_{S^i}^{S^j}(f_o(S^i) - f_o(s))p_c'(s)\,\mathrm{d}s.\end{aligned}$$

As p_c and f_o are decreasing, the above right-hand side is negative. Hence

$$\eta_o(S^i)(p_c(S^j) - p_c(S^i)) - (\eta_o(S^i) + \eta_w(S^i))(p_{wg}(S^j) - p_{wg}(S^i)) \leqslant 0. \tag{3.17}$$

We multiply (3.17) by $(P_o^j - P_o^i) + (P_w^j - P_w^i) < 0$ and use (2.34),

$$\big(\eta_o(S^i)(p_c(S^j) - p_c(S^i)) - (\eta_o(S^i) + \eta_w(S^i))(p_{wg}(S^j) - p_{wg}(S^i))\big)\left(2(P_w^j - P_w^i) + p_c(S^j) - p_c(S^i)\right) \geqslant 0.$$

By expanding and using the next inequality implied by (3.14), if $f_o(\xi) \neq 0$,

$$(p_{wg}(S^j) - p_{wg}(S^i))(p_c(S^j) - p_c(S^i)) \geqslant (p_{wg}(S^j) - p_{wg}(S^i))^2$$

we obtain

$$\eta_o(S^i)(p_c(S^j)-p_c(S^i))^2 + 2\eta_o(S^i)(p_c(S^j)-p_c(S^i))(P^j_w - P^i_w)$$
$$\geq (\eta_o(S^i)+\eta_w(S^i))(p_{wg}(S^j)-p_{wg}(S^i))\Big(2(P^j_w - P^i_w) + p_{wg}(S^j) - p_{wg}(S^i)\Big).$$

When $(\eta_o(S^i)+\eta_w(S^i))(P^j_w - P^i_w)^2$ is added to both sides, this becomes

$$\eta_w(S^i)(P^j_w - P^i_w)^2 + \eta_o(S^i)(P^j_o - P^i_o)^2 \geq (\eta_o(S^i)+\eta_w(S^i))(U^j_w - U^i_w)^2$$

and (1.9) implies the desired result. It remains to consider the case $f_o(\xi) = 0$, i.e., $p_{wg}(S^j) = p_{wg}(S^i)$. If $\eta_o(S^i) \neq 0$, then (3.17) yields

$$p_c(S^j) - p_c(S^i) \leq 0, \quad \text{which implies } P^i_o - P^j_o \geq P^i_w - P^j_w$$

and we deduce immediately

$$\eta_w(S^i)(P^j_w - P^i_w)^2 + \eta_o(S^i)(P^j_o - P^i_o)^2 \geq (\eta_w(S^i)+\eta_o(S^i))(P^j_w - P^i_w)^2 \geq \eta_*(P^j_w - P^i_w)^2.$$

When $\eta_o(S^i) = 0$, we have trivially

$$\eta_w(S^i)(P^j_w - P^i_w)^2 + \eta_o(S^i)(P^j_o - P^i_o)^2 = \eta_w(S^i)(P^j_w - P^i_w)^2 \geq \eta_*(P^j_w - P^i_w)^2.$$

3. If $P^i_w \leq P^j_w$ and $P^i_o > P^j_o$, then $\eta_w(S^{ij}_w) = \eta_w(S^j)$ and $\eta_o(S^{ij}_o) = \eta_o(S^i)$ in the case of a strict inequality; also $S^i \leq S^j$. Then (3.15) and the monotonic properties of η_w and η_o yield (3.13). If $P^i_w = P^j_w$, then according to (3.11), $S^i \leq S^j$ and the same conclusion holds.
4. If $P^i_w \leq P^j_w$ and $P^i_o = P^j_o$, then from (3.12), we have $S^i \leq S^j$ and with (3.15):

$$(U^j_w - U^i_w)^2 \leq \frac{\eta_w(\xi)}{\eta_w(\xi)+\eta_o(\xi)}(P^j_w - P^i_w)^2 \leq \frac{\eta_w(S^{ij}_w)}{\eta_w(\xi)+\eta_o(\xi)}(P^j_w - P^i_w)^2$$

which is the desired result.
5. Similarly, if $P^i_w = P^j_w$ and $P^i_o < P^j_o$, then from (3.11), we have $S^j \leq S^i$ and with (3.15):

$$(U^j_w - U^i_w)^2 \leq \frac{\eta_o(\xi)}{\eta_w(\xi)+\eta_o(\xi)}(P^j_o - P^i_o)^2 \leq \frac{\eta_o(S^{ij}_o)}{\eta_w(\xi)+\eta_o(\xi)}(P^j_o - P^i_o)^2.$$

6. If $P^i_w < P^j_w$ and $P^i_o < P^j_o$, (3.13) follows from the second case by switching i and j.

This completes the proof. □

The pressure bound in the next theorem is the one that arises naturally from the left-hand side of (2.42) and (2.43).

Theorem 3.2. *There exists a constant C, independent of h and τ, such that*

$$\tau \sum_{n=1}^{N} \sum_{i,j=1}^{M} c_{ij}\Big(\eta_w(S^{n,ij}_w)(P^{n,i}_w - P^{n,j}_w)^2 + \eta_o(S^{n,ij}_o)(P^{n,i}_o - P^{n,j}_o)^2\Big) \leq C. \tag{3.18}$$

Proof. We test (2.42) by $P^n_{w,h}$, (2.43) by $P^n_{o,h}$, add the two equations, multiply by τ and sum over n from 1 to N. By using (2.44) and (2.41), we obtain

$$-\sum_{n=1}^{N}(S^n_h - S^{n-1}_h, p_c(S^n_h))^\varphi_h + \frac{1}{2}\sum_{n=1}^{N}\tau\sum_{\alpha=w,o}\sum_{i,j=1}^{M} c_{ij}\eta_\alpha(S^{n,ij}_\alpha)(P^{n,i}_\alpha - P^{n,j}_\alpha)^2$$
$$= \sum_{n=1}^{N}\tau\sum_{\alpha=w,o}(f_\alpha(s^n_{\text{in},h})\overline{q}^n_h - f_\alpha(S^n_h)\underline{q}^n_h, P^n_{\alpha,h})_h. \tag{3.19}$$

Following [16], the first term in (3.19) is treated with the primitive g_c of p_c, see (1.12). Indeed, by the mean-value theorem, there exists ξ between $S^{n,i}$ and $S^{n-1,i}$ such that

$$g_c(S^{n,i}) - g_c(S^{n-1,i}) = -(S^{n,i} - S^{n-1,i})p_c(\xi).$$

As the function p_c is decreasing, then $p_c(\xi) \geqslant p_c(S^{n,i})$ when $S^{n,i} \geqslant S^{n-1,i}$ and $p_c(\xi) \leqslant p_c(S^{n,i})$ when $S^{n,i} \leqslant S^{n-1,i}$. In both cases, we have

$$g_c(S^{n,i}) - g_c(S^{n-1,i}) \leqslant -(S^{n,i} - S^{n-1,i})p_c(S^{n,i})$$

and owing that φ is positive and constant in time, (3.19) can be replaced by the inequality

$$(g_c(S_h^N) - g_c(S_h^0), 1)_h^{\varphi} + \frac{1}{2}\sum_{n=1}^{N} \tau \sum_{\alpha=w,o} \sum_{i,j=1}^{M} c_{ij}\eta_\alpha(S_\alpha^{n,ij})(P_\alpha^{n,i} - P_\alpha^{n,j})^2 \leqslant \sum_{n=1}^{N} \tau \sum_{\alpha=w,o} (f_\alpha(s_{\mathrm{in},h}^n)\overline{q}_h^n - f_\alpha(S_h^n)\underline{q}_h^n, P_{\alpha,h}^n)_h. \tag{3.20}$$

As the first term in the above left-hand side is bounded, owing to the continuity of g_c and boundedness of $S_{h,\tau}$, it suffices to handle the right-hand side. Let us drop the superscript n and treat one term in the time sum. Following again [16], in view of Lemma 3.2 we use the auxiliary pressures p_{wg} and p_{wo}, defined in (1.13). Clearly, (1.15) and (2.34) imply

$$P_w^i + p_{wg}(S^i) + p_{og}(S^i) + p_c(0) = P_o^i \quad \forall i. \tag{3.21}$$

Using this, a generic term, say Y, in the right-hand side of (3.20) can be expressed as

$$Y = (\overline{q}_h - \underline{q}_h, U_{w,h})_h + (f_o(s_{\mathrm{in},h})\overline{q}_h - f_o(S_h)\underline{q}_h, p_c(0))_h + (f_o(s_{\mathrm{in},h})\overline{q}_h - f_o(S_h)\underline{q}_h, p_{og}(S_h))_h - (f_w(s_{\mathrm{in},h})\overline{q}_h - f_w(S_h)\underline{q}_h, p_{wg}(S_h))_h = T_1 + \cdots + T_4.$$

We now bound each term T_i. For T_1, (2.31) implies that any constant β can be added to $U_{w,h}$, in particular β can be chosen so that the sum has zero mean value in Ω. Hence, considering the generalized Poincaré inequality

$$\forall v \in H^1(\Omega), \quad \|v\|_{L^2(\Omega)} \leqslant C\left(\left|\int_\Omega v\right| + \|\nabla v\|_{L^2(\Omega)}\right) \tag{3.22}$$

with a constant C, depending only on the domain Ω, we have

$$\|U_{w,h} + \beta\|_h \leqslant C\|U_{w,h} + \beta\|_{L^2(\Omega)} \leqslant C\|\nabla U_{w,h}\|_{L^2(\Omega)}$$

with another constant C. Then Young's inequality yields

$$|T_1| \leqslant \frac{C^2}{2\eta_*}\|\overline{q}_h - \underline{q}_h\|_h^2 + \frac{\eta_*}{4}\|\nabla U_{w,h}\|_{L^2(\Omega)}^2$$

and with Lemma 3.2, this becomes

$$|T_1| \leqslant \frac{C^2}{2\eta_*}\|\overline{q}_h - \underline{q}_h\|_h^2 + \frac{1}{4}\sum_{i,j=1}^{M} c_{ij}\left(\eta_w(S^{ij})(P_w^j - P_w^i)^2 + \eta_o(S^{ij})(P_o^j - P_o^i)^2\right).$$

The term T_2 is easily bounded since $p_c(0)$ is a number, and so are the terms T_3 and T_4, in view of the boundedness of the saturation and the continuity of p_{og} and p_{wg}. We thus have

$$|T_2 + T_3 + T_4| \leqslant C(\|\overline{q}_h\|_{L^1(\Omega)} + \|\underline{q}_h\|_{L^1(\Omega)}).$$

Then substituting these bounds for each n into (3.20), we obtain

$$\frac{1}{4}\tau\sum_{n=1}^{N}\sum_{i,j=1}^{M} c_{ij}(\eta_w(S_w^{n,ij})(P_w^{n,i} - P_w^{n,j})^2 + \eta_o(S_o^{n,ij})(P_o^{n,i} - P_o^{n,j})^2) \leqslant C\left(\|\overline{q}_{h,\tau} - \underline{q}_{h,\tau}\|_{L^2(\Omega\times]0,T[)}^2 + \|\overline{q}_{h,\tau}\|_{L^1(\Omega\times]0,T[)} + \|\underline{q}_{h,\tau}\|_{L^1(\Omega\times]0,T[)}\right)$$

thus proving (3.18). □

By combining Theorem 3.2 with Lemma 3.2, we immediately derive a bound on the discrete auxiliary pressures. The bound (3.23) with $\alpha = o$ follows from the same with $\alpha = w$, (1.15), and (2.34).

Theorem 3.3. *For $\alpha = w, o$ we have*

$$\eta_* \|\nabla U_{\alpha,h,\tau}\|^2_{L^2(\Omega\times]0,T[)} \leqslant C \tag{3.23}$$

with the constant C of (3.18).

4 Existence of numerical solution

We fix $n \geqslant 1$ and assume there exists a solution $(S_h^{n-1}, P_{w,h}^{n-1})$ at time t^{n-1} with $0 \leqslant S_h^{n-1} \leqslant 1$. We want to show existence of a solution $(S_h^n, P_{w,h}^n)$ by means of the topological degree [12, 13].

Let ϑ be a constant parameter in $[0, 1]$. For any continuous function $f : [0, 1] \to \mathbb{R}$ and any $t \in [0, 1]$, we define the transformed function $\widetilde{f} : [0, 1] \to \mathbb{R}$ by

$$\forall s \in [0, 1], \quad \widetilde{f}(s) = f(ts + (1 - t)\vartheta).$$

Since ϑ is fixed, when $t = 0$, $\widetilde{f}(s) = f(\vartheta)$, a constant independent of s. Now, (2.45) implies that any solution $P_{w,h,\tau}$ of (2.42)–(2.45) belongs to the following subspace $X_{0,h}$ of X_h,

$$X_{0,h} = \left\{\Lambda_h \in X_h;\ \int_\Omega \Lambda_h = 0\right\}. \tag{4.1}$$

This suggests to define the mapping $\mathcal{F} : [0, 1] \times X_h \times X_{0,h} \to X_h \times X_{0,h}$ by

$$\mathcal{F}(t, \zeta, \Lambda) = (A_h, A_h + B_h)$$

where A_h, respectively B_h, solves for all $\Theta_h \in X_h$,

$$\begin{aligned}(A_h, \Theta_h) = {} & \frac{1}{\tau}(\zeta_h - S_h^{n-1}, \Theta_h)_h^{\varphi} - [\Lambda_h, I_h(\widetilde{\eta_w}(\zeta_h)); \Lambda_h, \Theta_h]_h \\ & - (I_h(\widetilde{f_w}(s^n_{\mathrm{in},h}))t\bar{q}_h^n - I_h(\widetilde{f_w}(\zeta_h))t\underline{q}_h^n, \Theta_h)_h \end{aligned} \tag{4.2}$$

$$\begin{aligned}(B_h, \Theta_h) = {} & -\frac{1}{\tau}(\zeta_h - S_h^{n-1}, \Theta_h)_h^{\varphi} - [P_{o,h}, I_h(\widetilde{\eta_o}(\zeta_h)); P_{o,h}, \Theta_h]_h \\ & - (I_h(\widetilde{f_o}(s^n_{\mathrm{in},h}))t\bar{q}_h^n - I_h(\widetilde{f_o}(\zeta_h))t\underline{q}_h^n, \Theta_h)_h \end{aligned} \tag{4.3}$$

and $P_{o,h}$ is defined by

$$P_{o,h} = \Lambda_h - I_h(\widetilde{p_c}(\zeta_h)). \tag{4.4}$$

The choice of $\widetilde{\eta_w}(\zeta_h)$ in (4.2) (respectively $\widetilde{\eta_o}(\zeta_h)$ in (4.3)) is given by (2.36) (respectively (2.37)) where Λ_h plays the role of $P_{w,h}$ and $P_{o,h}$ is defined in (4.4). As in (2.36) and (2.37), it leads us to introduce the variables ζ_w^{ij} and ζ_o^{ij} for all $1 \leqslant i, j \leqslant M$. Clearly, (4.2)–(4.4) determine uniquely A_h and B_h, and it is easy to check that $A_h + B_h$ belongs to $X_{0,h}$.

The mapping $t \mapsto \mathcal{F}(t, \zeta_h, \Lambda_h)$ is continuous. Indeed, since the space has finite dimension, we only need to check continuity of the upwinding. By splitting x into its positive and negative part, $x = x^+ + x^-$, the upwind term, say $\widetilde{\eta_w}(\zeta_w^{ij})(P_w^j - P_w^i)$ reads

$$\widetilde{\eta_w}(\zeta_w^{ij})(P_w^j - P_w^i) = \eta_w(t\zeta^i + (1 - t)\vartheta)((P_w^j - P_w^i)_-) + \eta_w(t\zeta^j + (1 - t)\vartheta)((P_w^j - P_w^i)_+)$$

which is continuous with respect to t.

We remark that $\mathcal{F}(1, \zeta_h, \Lambda_h) = \mathbf{0}$ implies that (ζ_h, Λ_h) solves (2.42)–(2.45). Conversely, if (ζ_h, Λ_h) solves (2.42)–(2.45) then $\mathcal{F}(1, \zeta_h, \Lambda_h) = \mathbf{0}$. Thus, showing existence of a solution to the problem (2.42)–(2.45) is equivalent to showing existence of a zero of $\mathcal{F}(1, \zeta_h, \Lambda_h)$. Before proving existence of a zero, we use the estimates established in the previous section to determine an a priori bound of any zero (ζ_h, Λ_h) of $\mathcal{F}(1, \zeta_h, \Lambda_h)$.

4.1 A priori bounds on (ζ_h, Λ_h)

In the following we consider $t \in [0,1]$ and $(\zeta_h, \Lambda_h) \in X_h \times X_{0,h}$ that satisfy

$$\mathcal{F}(t, \zeta_h, \Lambda_h) = \mathbf{0}. \tag{4.5}$$

We first show that ζ_h satisfies a maximum principle.

Proposition 4.1. The following bounds hold for all (t, ζ_h, Λ_h) satisfying (4.5):

$$0 \leqslant \zeta_h \leqslant 1. \tag{4.6}$$

Proof. Either $t \in]0,1]$ or $t = 0$. The proof for $t \in]0,1]$ follows closely the argument used in proving Theorem 3.1 and is left to the reader. For $t = 0$ we proceed again by contradiction. Assume first that $\|\zeta_h\|_{L^\infty(\Omega)} > 1$, i.e., there is a node i such that

$$\zeta^i = \|\zeta_h\|_{L^\infty(\Omega)} > 1 \geqslant S^{n-1,i}.$$

As $t = 0$, (4.5) reduces to

$$\sum_{j \neq i} c_{ij} \eta_w(\vartheta)(\Lambda^i - \Lambda^j) > 0, \qquad -\sum_{j \neq i} c_{ij} \eta_o(\vartheta)(\Lambda^i - \Lambda^j) > 0 \quad \forall 1 \leqslant i \leqslant M.$$

Since η_o and η_w are non-negative functions satisfying (1.9), the inequalities above yield a contradiction. A similar argument is used to show that $\zeta_h \geqslant 0$. □

Next we show the following bound on Λ_h.

Proposition 4.2. There is a constant C such that for all $t \in [0,1]$ we have

$$\eta_* \sum_{i,j=1}^{M} c_{ij} \left(\Lambda^j - \Lambda^i + p_{wg}(t\zeta^j + (1-t)\vartheta) - p_{wg}(t\zeta^i + (1-t)\vartheta)\right)^2 \leqslant C. \tag{4.7}$$

Proof. The proof follows closely that of Theorem 3.2. First we show there exists a constant C_1 independent of t such that

$$\sum_{i,j=1}^{M} c_{ij}\Big(\eta_w(t\zeta_w^{ij} + (1-t)\vartheta)(\Lambda^j - \Lambda^i)^2 + \eta_o(t\zeta_o^{ij} + (1-t)\vartheta)(P_{o,h}^j - P_{o,h}^i)^2\Big) \leqslant C_1$$

with $P_{o,h}$ defined in (4.4). This bound is obtained via arguments similar to those used in proving Theorem 3.2. The main difference is that the formula is neither summed over n nor multiplied by the time step τ. As a consequence, the constant C_1 includes a term of the form $\tau^{-1}\|g_c\|_{L^\infty(\Omega)}$ arising from the bound of the discrete time derivative. To finish the proof we must show that

$$\begin{aligned}\eta_*\Big(\Lambda^j - \Lambda^i + p_{wg}(t\zeta^j + (1-t)\vartheta) - p_{wg}(t\zeta^i + (1-t)\vartheta)\Big)^2 \\ \leqslant \eta_w(t\zeta_w^{ij} + (1-t)\vartheta)(\Lambda^j - \Lambda^i)^2 + \eta_o(t\zeta_o^{ij} + (1-t)\vartheta)(P_o^j - P_o^i)^2.\end{aligned}$$

By (1.9), this is trivially satisfied when $t = 0$. When $t \in]0,1]$, the argument is the same as in the proof of Lemma 3.2. □

Propositions 4.1 and 4.2 are combined to obtain a bound on $\|\zeta_h\|_h + \|\Lambda_h\|_h$.

Proposition 4.3. There exists a constant $R_1 > 0$, independent of $t \in [0,1]$, such that any solution (ζ_h, Λ_h) of (4.5) satisfies

$$\|\zeta_h\|_h + \|\Lambda_h\|_h \leqslant R_1. \tag{4.8}$$

Proof. According to Proposition 4.1, there exists a constant C_1 independent of t such that

$$\|\zeta_h\|_h \leqslant C_1.$$

To establish a bound on $\|\Lambda_h\|_h$, we infer from (1.13) that the function $|p_{wg}|$ is bounded by $p_c(0) - p_c(1)$ because f_o is bounded by one and p_c is a decreasing function. Thus (4.7) implies that there exists a constant C_2 independent of t that satisfies

$$\sum_{i,j=1}^{M} c_{ij}\left(\Lambda^j - \Lambda^i\right)^2 \leqslant C_2, \qquad \text{i.e., } \|\nabla \Lambda_h\|_{L^2(\Omega)} \leqslant \frac{\sqrt{C_2}}{\sqrt{2}} \tag{4.9}$$

owing to (2.10). As $\Lambda_h \in X_{0,h}$, the generalized Poincaré inequality (3.22) shows there exists a constant C_3 independent of t such that

$$\|\Lambda_h\|_{L^2(\Omega)} \leqslant C_3.$$

Then the equivalence of norm (2.5) yields

$$\|\Lambda_h\|_h \leqslant C_4$$

and (4.8) follows by setting $R_1 = C_1 + C_4$, a constant independent of t. □

4.2 Proof of existence

For any $R > 0$, let B_R denote the ball

$$B_R = \left\{(\zeta_h, \Lambda_h) \in X_h \times X_{0,h};\ \|\zeta_h\|_h + \|\Lambda_h\|_h \leqslant R\right\} \tag{4.10}$$

and let $R_0 = R_1 + 1$, where R_1 is the constant of (4.8). Since all solutions (ζ_h, Λ_h) of (4.5) are in the ball B_{R_1}, this function has no zero on the boundary ∂B_{R_0}. Existence of a solution of (2.42)–(2.45) follows from the following result.

Theorem 4.1. *The equation* $\mathcal{F}(1, \zeta_h, \Lambda_h) = \mathbf{0}$ *has at least one solution* $(\zeta_h, \Lambda_h) \in B_{R_0}$.

Proof. The proof proceeds in two steps. First, we show that the system with $t = 0$ has a solution:

$$\mathcal{F}(0, \zeta_h, \Lambda_h) = 0.$$

This is a square linear system in finite dimension, so existence is equivalent to uniqueness. Thus we assume that it has two solutions, and for convenience, we still denote by (ζ_h, Λ_h) the difference between the two solutions. The system reads

$$\frac{\widetilde{m}_i}{\tau}\zeta_h^i - \sum_{j \neq i, j \in \mathcal{N}(i)} c_{ij}\eta_w(\vartheta)(\Lambda^j - \Lambda^i) = 0, \quad 1 \leqslant i \leqslant M \tag{4.11}$$

$$-\frac{\widetilde{m}_i}{\tau}\zeta_h^i - \sum_{j \neq i, j \in \mathcal{N}(i)} c_{ij}\eta_o(\vartheta)(\Lambda^j - \Lambda^i) = 0, \quad 1 \leqslant i \leqslant M \tag{4.12}$$

$$\sum_i m_i \Lambda^i = 0. \tag{4.13}$$

We add the first two equations, multiply by Λ^i, and sum over i. Then (2.10) and (2.41) imply that Λ_h is a constant and finally (4.13) shows that this constant is zero. This yields $\zeta_h = 0$.

Next, we argue on the topological degree. Since the topological degree of a linear map is the sign of its determinant, we have, by denoting d the degree,

$$d(\mathcal{F}(0, \zeta_h, \Lambda_h), B_{R_0}, 0) \neq 0.$$

We also know that $d(\mathcal{F}(t, \zeta_h, \Lambda_h), B_{R_0}, 0)$ is independent of t since the mapping $t \mapsto \mathcal{F}(t, \zeta_h, \Lambda_h)$ is continuous and for every $t \in [0, 1]$, if $\mathcal{F}(t, \zeta_h, \Lambda_h) = 0$, then (ζ_h, Λ_h) does not belong to ∂B_{R_0}. Therefore we have

$$d(\mathcal{F}(1, \zeta_h, \Lambda_h), B_{R_0}, 0) = d(\mathcal{F}(0, \zeta_h, \Lambda_h), B_{R_0}, 0) \neq 0.$$

This implies that $\mathcal{F}(1, \zeta_h, \Lambda_h)$ has a zero $(\zeta_h, \Lambda_h) \in B_{R_0}$. □

5 Numerical validation

The present section proposes a numerical validation of our algorithm with a two dimensional finite difference code. Details on the algorithm implemented are given. A problem with manufactured solutions is then considered to study the convergence properties of our algorithm.

5.1 Implementation of the model

The scheme developed in Section 2.3 is linearized by time lagging the saturation, by using (2.34) to eliminate P_o and by approximating p_c^{n+1} by a first order Taylor expansion. More precisely, p_c^{n+1} is approximated by

$$p_c^{*,n+1} = p_c^n + \left(\frac{\partial p_c}{\partial S}\right)^n (S^{n+1} - S^n). \tag{5.1}$$

Thus, for each node $1 \leqslant i \leqslant M$, the unknowns $(S^{n+1,i}, P_w^{n+1,i})$ are computed as the solution of the following problem:

$$\frac{\widetilde{m}_i}{\tau}(S^{n+1,i} - S^{n,i}) - \sum_{j\neq i, j\in N(i)} c_{ij}\eta_w(S_w^{n,ij})(P_w^{n+1,j} - P_w^{n+1,i}) = m_i f_1^{n+1,i}, \quad 1 \leqslant i \leqslant M$$

$$-\frac{\widetilde{m}_i}{\tau}(S^{n+1,i} - S^{n,i}) - \sum_{j\neq i, j\in N(i)} c_{ij}\eta_o(S_o^{n,ij})(P_w^{n+1,j} - P_w^{n+1,i}) - \sum_{j\neq i, j\in N(i)} c_{ij}\eta_o(S_o^{n,ij})(p_c^{*,n+1,j} - p_c^{*,n+1,i}) = m_i f_2^{n+1,i}, \quad 1 \leqslant i \leqslant M$$

We note that to facilitate the implementation of this algorithm in a two dimensional finite difference code, the source terms of the equations (2.32)–(2.33) have been replaced by functions denoted by f_1 and f_2.

5.2 Numerical test with a manufactured solution

The numerical validation of the algorithm is done by approximating the analytical solutions defined by

$$P_w(t,x,y) = 2 + x^2 y - y^2 + x^2 \sin(t+y) \tag{5.2}$$

$$S(t,x,y) = 0.2(2 + 2xy + \cos(t+x)) \tag{5.3}$$

on the computational domain $\Omega = [0,1]^2$. Dirichlet boundary conditions are applied on $\partial\Omega$ on both unknowns P_w and S. The initial conditions of the problem satisfy (5.2)–(5.3). The porosity of the domain is set to:

$$\varphi(t,x,y) = 0.2(1+xy). \tag{5.4}$$

The mobilities η_w and η_o, introduced in Section 1.1, are defined as follows:

$$\eta_w(s) = 4s^2, \qquad \eta_o(s) = 0.4(1-s)^2. \tag{5.5}$$

The capillary pressure is based on the Brooks–Corey model, it reads:

$$p_c(s) = \begin{cases} 50 s^{-1/2} & \text{if } s > 0.05 \\ 25(0.05)^{-1/2}(3 - s/0.05) & \text{otherwise.} \end{cases} \tag{5.6}$$

The term sources f_1 and f_2 are computed accordingly. The convergence tests are performed on a set of six structured grids. The coarsest grid is made of 5×5 squares and each square is divided into 2 triangles. Then,

L^2-norm of error		Water pressure P_w		Water saturation S	
$h/\sqrt{2}$	n_{df}	Error	Rate	Error	Rate
0.2	36	8.50E−3	—	4.21E−3	—
0.1	121	4.15E−3	1.03	2.30E−3	0.87
0.05	441	2.08E−3	1.00	1.14E−4	1.01
0.025	1681	1.04E−3	1.00	5.57E−4	1.03
0.0125	6561	5.23E−4	0.99	2.75E−4	1.02

Tab. 1: Results of convergence tests where the mesh size is denoted by h and the number of degrees of freedom per unknown by n_{df}. The time step τ is set to h and errors are computed at final time $T = 1$.

we uniformly refine the mesh by dividing each into four triangles to obtain the second structured grid. We continue this process until all the six grids have been constructed. The convergence properties are evaluated by using a time step τ set to the mesh size h with a final time $T = 1$. As the time derivatives and the saturations $S_w^{n+1,ij}$, $S_o^{n+1,ij}$ are computed with first order time approximation, we expect the convergence rate in the L^2 norm to be of order one.

The results of the convergence tests are presented in Table 1. The theoretical order of convergence, equal to one, is recovered for both unknowns which confirms the correct behavior of the algorithm.

6 Conclusions

This paper formulates a $\mathbb{P}_1$ finite element method to solve the immiscible two-phase flow problem in porous media. The unknowns are the phase pressure and saturation, which are the preferred unknowns in industrial reservoir simulators. The numerical method employs mass lumping for integration and an upwind flux technique. In this paper, we prove existence of the numerical solutions and some stability bounds. We also show that the numerical saturation is bounded between zero and one. The convergence analysis is to be presented in the second part of the paper.

Funding: The work of the second author was supported in part by NSF-DMS 1913291.

References

[1] H. W. Alt and E. Di Benedetto, Nonsteady flow of water and oil through inhomogeneous porous media, *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* **12** (1985), No. 3, 335–392.
[2] T. Arbogast, The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow, *Nonlinear Analysis, Theory, Models & Applications* **19** (1992), No. 11, 1009–1031.
[3] T. Arbogast and M. F Wheeler, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, *SIAM J. Numer. Anal.* **33** (1996), No. 4, 1669–1687.
[4] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Appl. Sci. Publ. Ltd., London, 1979.
[5] P. Bastian, A fully-coupled discontinuous Galerkin method for two-phase flow in porous media with discontinuous capillary pressure, *Comput. Geosci.* **18** (2014), No. 5, 779–796.
[6] J. Casado-Diaz, T. Chacón Rebollo, V. Girault, M. Gomez Marmol, and F. Murat, Finite elements approximation of second order linear elliptic equations in divergence form with right-hand side in L^1, *Numerische Mathematik* **105** (2007), No. 3, 337–374.
[7] G. Chavent, A new formulation of diphasic incompressible flows in porous media, In: *Applications of Methods of Functional Analysis to Problems in Mechanics, Vol. 503*, Springer, Berlin–Heidelberg, 1976, pp. 258–270.
[8] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation: Single Phase, Multiphase and Multicomponent Flows Through Porous Media*, Elsevier, 1986.
[9] Z. Chen, Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution, *J. Differ. Equ.*, **171** (2001), No. 2, 203–232.
[10] Z. Chen and R. Ewing, Mathematical analysis for reservoir models, *SIAM J. Numer. Anal.*, **30** (1999), No. 2, 431–453.

[11] Z. Chen and R.E. Ewing, Degenerate two-phase incompressible flow, III. Sharp error estimates, *Numerische Mathematik* **90** (2001), No. 2, 215–240.
[12] K. Deimling, *Nonlinear Functional Analysis*, Dover Publications, Mineola, New York, 1985.
[13] G. Dinca and J. Mawlin, *Brouwer Degree and Applications*, Laboratoire Jacques-Louis Lions, Université Paris VI, Report, 2009.
[14] J. Douglas, Jr., Finite difference methods for two-phase incompressible flow in porous media, *SIAM J. Numer. Anal.* **20** (1983), No. 4, 681–696.
[15] Y. Epshteyn and B. Riviere, Analysis of hp discontinuous Galerkin methods for incompressible two-phase flow, *J. Comput. Appl. Math.*, **225** (2009), 487–509.
[16] R. Eymard, R. Herbin, and A. Michel, Mathematical study of a petroleum-engineering scheme, *ESAIM: Math. Modelling Numer. Anal.*, **37** (2003), No. 6, 937–972.
[17] R. Eymard, D. Hilhorst, and M. Vohralík, A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems, *Numerische Mathematik*, **105** (2006), No. 1, 73–131.
[18] R. Eymard, T. Gallouët, and R. Herbin, Finite volume methods, *Handbook of Numerical Analysis*, **7** (2000), 713–1018.
[19] P. A. Forsyth, A control volume finite element approach to NAPL groundwater contamination, *SIAM J. Sci. Stat. Comp.*, **12** (1991), No. 5, 1029–1057.
[20] V. Girault, B. Riviere, and L. Cappanera, A finite element method for degenerate two-phase flow in porous media. Part II: Convergence, *J. Numer. Math.*, **29** (2021), No. 3 (to appear).
[21] J.-L. Guermond and B. Popov, Invariant domains and first-order continuous finite element approximations for hyperbolic systems, *SIAM J. Numer. Anal.* **54** (2016), No. 4, 2466–2489.
[22] R. Helmig, *Multiphase Flow and Transport Processes in the Subsurface: a Contribution to the Modeling of Hydrosystems*, Springer-Verlag, 1997.
[23] D. Kroener and S. Luckhaus, Flow of oil and water in a porous medium, *J. Differ. Equ.*, **55** (1984), 276–288.
[24] A. Michel, A finite volume scheme for two-phase incompressible flow in porous media, *SIAM J. Numer. Anal.*, **41** (2003), No. 4, 1301–1317.
[25] M. Ohlberger, Convergence of a mixed finite element: Finite volume method for the two phase flow in porous media, *East West J. Numer. Math.*, **5** (1997), 183–210.
[26] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, 2000.
[27] C. S. Woodward and C. N. Dawson, Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media, *SIAM J. Numer. Anal.*, **37** (2000), No. 3, 701–724.
[28] I. Yotov, A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arising in porous media flow, *East West J. Numer. Math.* **5** (1997), 211–230.