universiteit hasselt

UHasselt Computational Mathematics Preprint Series

A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities

Florin A. Radu, Kundan Kumar, Jan Martin Nordbotten, Iuliu Sorin Pop

UHasselt Computational Mathematics Preprint Nr. UP-16-04

November 17, 2016

A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities

Florin A. Radu[1], Kundan Kumar[1], Jan M. Nordbotten[1], Iuliu S. Pop[1,2]

[1] Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020 Bergen, Norway
[2] Faculty of Sciences, Hasselt University, Campus Diepenbeek, Agoralaan building D, BE3590 Diepenbeek, Belgium

e-mails: {*florin.radu, jan.nordbotten, kundan.kumar*}*@math.uib.no, sorin.pop@uhasselt.be*

Abstract. In this work we present a mass conservative numerical scheme for two-phase flow in porous media. The model for flow consists on two fully coupled, non-linear equations: a degenerate parabolic equation and an elliptic one. The proposed numerical scheme is based on backward Euler for the temporal discretization and mixed finite element method (MFEM) for the spatial one. *A priori* stability and error estimates are presented to prove the convergence of the scheme. A monotone increasing, Hölder continuous saturation is considered. The convergence of the scheme is naturally depending on the Hölder exponent. The non-linear systems within each time step are solved by a robust linearization method, called the L-method. This iterative method does not involve any regularization step. The convergence of the L-scheme is rigorously proved under the assumption of a Lipschitz continuous saturation. For the Hölder continuous case, a numerical convergence is established. Numerical results are presented to sustain the theoretical findings.

Keywords: linearization, two-phase flow, mixed finite element method, convergence analysis, a priori error estimates, porous media, Richards' equation, degenerate parabolic problems, coupled problems, Hölder continuity.

1 Introduction

Two-phase porous media flow models are widely encountered in real-life applications of utmost societal relevance, including water and soil pollution, oil recovery, geological carbon dioxide sequestration, or nuclear waste management [34, 24]. Such complex problems admit only in very simplified situations analytical solutions, therefore numerical methods for solving multiphase flow in porous media are playing a determining role in understanding and solving the problems. Nevertheless, the design and analysis of robust, accurate and efficient numerical schemes is a very challenging task.

Here we discuss a numerical scheme for a two-phase porous media flow model. The fluids are assumed immiscible and incompressible and the solid matrix is non-deformable. The adopted formulation uses the global pressure and a complementary pressure, obtained by using the Kirchhoff transformation, as primary

unknowns (see [12, 3, 14]). This leads to a system of two coupled non-linear partial differential equations, a degenerate elliptic - parabolic one and an elliptic one.

Numerical methods for two-phase flow have been the object of intensive research in the last decades. The major challenge in developing efficient schemes is related to the degenerate nature of the problem. Due to this, the solution typically lacks regularity, which makes lower order finite elements or finite volumes a natural choice for the spatial discretization. In this respect, we refer to [23, 35] for Galerkin finite elements, to [22, 37, 33] for finite volumes, to [19, 15, 16] for methods combining Galerkin finite elements combined with the mixed finite element method (MFEM), and to [20, 47] for the discontinuous Galerkin method. In all cases, the convergence of the numerical schemes is proved rigorously either by compactness arguments, or by obtaining *a priori* error estimates. *A posteriori* error estimates are obtained e.g. in [10]. Furthermore, similar issues appear for the Richards equation, which is a simplified model for saturated/unsaturated flow in the case when the pressure of one phase is supposed to be constant. In this context we mention Galerkin finite elements [35, 39], MFEM based works [4, 41, 44, 54, 56], multipoint flux approximation (MPFA) [31] or finite volume - MFEM combined methods [21].

In this paper we propose a mass conservative scheme based on MFEM (lowest order Raviart-Thomas elements [7]) and backward Euler for numerical simulation of the two-phase flow in porous media. Continuous, semi-discrete (continuous in space) and fully discrete mixed variational formulations are defined. Existence and uniqueness of solutions is discussed. We show the convergence of the numerical scheme and provide explicit order of convergence estimates. The estimates were obtained for a Hölder continuous, not necessary strictly increasing saturation. The order of convergence is depending on the Hölder exponent of the saturation (confirmed by numerical experiments as well). The analysis is inspired by similar results in [4, 16, 41, 44].

Typical problems involving flow in porous media, like e.g. water and soil pollution or nuclear waste management are spread over decades or even centuries, so that the use of relatively large time steps is a necessity. Due to this, implicit methods are strongly recommended (our choice here being the first-order backward Euler method, due to the low regularity of the considered problem). Since the original model is non-linear, at each time step one needs to solve non-linear algebraic systems. In this work we propose a robust linearization scheme for the systems appearing at each time step, as a valuable alternative to modified Picard method [11] or Newton's method [5, 38, 42, 36, 30] or iterative IMPES [25, 26]. Although the applicability of Newton's method for parabolic equations is well recognized, its convergence is not straightforward for degenerate equations, where the Jacobian might become singular. A possible way to overcome this is to regularize the problem. However, even in this case convergence is guaranteed only under a severe stability condition for the discretization parameters, see [42]. This has motivated the alternative, robust linearization scheme proposed in this work. The new scheme, called $L-$scheme from now on, does not involve the calculations of any derivatives and does not need a regularization step. The L-scheme combines the idea of a classical Picard method and the scheme presented in [40] for MFEM or [55, 51] for Galerkin finite elements. The L-scheme was proposed for two-phase flow in combination with the MPFA method in [45], the proof of convergence there being only sketched and not made completely rigorous. We show here that, in the case of Lipschitz continuous saturation (not necessarily strictly increasing), the L-scheme for MFEM based discretizations converges linearly if the time step satisfies a mild condition. This robustness is the main advantage of the scheme when compared to the quadratic, but locally convergent Newton method.

All the papers quoted above are considering Lipschitz continuous nonlinearities, like the dependency of the saturation on the complementary pressure. This is due to the fact that the L-scheme as proposed there involves the constants that need to be larger than the Lipschitz constants of the non-linear functions in the models. If the nonlinearities are only Hölder continuous but not Lipschitz, the derivatives become unbounded. Then the convergence proof for the L-scheme, as presented in [40] for the MFEM discretization of the Richards equation, or in [55, 51] for the Galerkin finite elements also for the Richards equation, or in

[45] for MPFA/two-phase flow, is not valid anymore. Commonly, one is regularizing first the problem by approximating the non-Lipschitz nonlinearities by Lipschitz ones, and then iterative methods like Newton, Picard, or the above mentioned L-scheme is applied. In this paper we show that the L-scheme can be applied for the Hölder continuous case as well. We prove the numerical convergence of the scheme for two-phase flow and a Hölder continuous saturation. We refer to [46] for the case of Richards' equation.

Finally, we mention that the L-scheme can be interpreted as a non-linear preconditioner, because the linear systems to be solved within each iteration are much better conditioned than the corresponding systems in the case of modified Picard or Newton's method. We refer to [29] for illustrative examples concerning the Richards equation, which is a particular case of the more general model considered in the present work.

To summarize, the main new contributions of this paper are

- We present and analyze a MFEM based numerical scheme for two-phase flow in porous media. A Hölder continuous saturation is assumed. Order of convergence estimates, depending on the Hölder exponent are obtained.
- We present and analyze rigorously a robust, first-order convergent linearization method for MFEM based schemes for two-phase flow in porous media.
- The paper is the first to apply the L-scheme to models involving non-Lipschitz nonlinearities.

The paper is structured as follows. In Section 2, we present the model equations for two-phase flow in porous media and we define the discretization and linearization schemes. In Section 3 we analyze the convergence of the discretization scheme based on *a priori* error estimates. We also discuss the existence and uniqueness for the problem involved, and give *a priori* (or stability) estimates. The analysis of the linearization scheme is presented in Section 4. Section 5 provides numerical examples confirming the theoretical results. The paper is ending with concluding remarks in Section 6.

2 Mathematical model and discretization

In this section we introduce the mathematical model for two-phase flow used in this work, the proposed MFEM/Euler implicit discretization (Problems P, P^n and P_h^n) and a new linearization scheme (Problem $P_h^{n,i}$) to solve the non-linear systems appearing at each time step.

In what follows we let $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) be a bounded domain having a Lipschitz continuous boundary Γ and $T > 0$ be an upper bound for the time. The two-phase porous media flow model considered here assumes that the fluids are immiscible and incompressible, and that the solid matrix is non-deformable. By denoting with $\alpha = w, n$ the wetting and non-wetting phases, $s_\alpha, p_\alpha, \mathbf{q}_\alpha, \rho_\alpha$ the saturation, pressure, flux and density of phase α, respectively, the two-phase model under consideration reads (see e.g. [6, 12, 24, 34])

$$\frac{\partial(\phi\rho_\alpha s_\alpha)}{\partial t} + \nabla\cdot(\rho_\alpha \mathbf{q}_\alpha) = 0, \qquad \alpha = w, n, \tag{1}$$

$$\mathbf{q}_\alpha = -k\frac{k_{r,\alpha}}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha \mathbf{g}), \quad \alpha = w, n, \tag{2}$$

$$s_w + s_n = 1, \tag{3}$$

$$p_n - p_w = p^{cap}(s_w), \tag{4}$$

where $\mathbf{g}$ denotes the constant gravitational vector. Equation (1) is a mass balance, (2) is the Darcy law, (3) is an algebraic evidence expressing that all pores in the medium are filled by a mixture of the two fluid

phases and (4) is the capillary pressure relationship, with $p^{cap}(\cdot)$ supposed to be known. The porosity ϕ, permeability k, the viscosities μ_α are given constants (scaled to 1 when making the model dimensionless) and the relative permeabilities $k_{r,\alpha}(\cdot)$ are given functions of s_w. We consider here a scalar permeability, but the results can be easily extended to the case when the permeability is a positive-definite tensor.

In this paper we adopt a global/complementary pressure formulation [3, 12, 14]. The global pressure (denoted by p) was introduced in [12] and the complementary pressure in [3]. For a given water saturation s_w, these are defined as

$$p(s_w) := p_n - \int_0^{s_w} f_w(\xi) \frac{\partial p^{cap}}{\partial \xi}(\xi) d\,\xi, \tag{5}$$

$$\Theta(s_w) := -\int_0^{s_w} f_w(\xi)\lambda_n(\xi) \frac{\partial p^{cap}}{\partial \xi}(\xi) d\,\xi, \tag{6}$$

where $\lambda_\alpha := \dfrac{k_{r,\alpha}}{\mu_\alpha}$ $(\alpha = w, n)$ stands for the mobility of phase α and $f_w := \dfrac{\lambda_w}{\lambda_w + \lambda_n}$ is the fractional flow function. Observe the use of Kirchhoff transformation above. In the new unknowns, the resulting system consists of two coupled non-linear partial differential equations, a degenerate parabolic one and an elliptic one. For more details on the modelling we refer to [14], where the existence and uniqueness of a weak solution is proved for a Galerkin-MFEM formulation. In the new unknowns the system (1)-(4) becomes

$$\partial_t s(\Theta) + \nabla \cdot \mathbf{q} = 0, \tag{7}$$

$$\mathbf{q} = -\nabla\Theta + f_w(s)\mathbf{u} + \mathbf{f}_1(s), \tag{8}$$

$$\nabla \cdot \mathbf{u} = f_2(s), \tag{9}$$

$$a(s)\mathbf{u} = -\nabla p - \mathbf{f}_3(s). \tag{10}$$

with $s := s_w$, $a(s) := \dfrac{1}{k(\lambda_w(s) + \lambda_o(s))}$, $\mathbf{q}$ the (wetting) flux, and $\mathbf{u}$ the total flux. The equations hold true in $\Omega \times (0, T]$. The coefficient functions $s(\cdot), a(\cdot), f_w(\cdot), \mathbf{f}_1(\cdot), f_2(\cdot), \mathbf{f}_3(\cdot)$ are given and satisfy the assumptions listed below (see [14] for an exact calculation of these functions). The system is completed by initial and boundary conditions,

$$\Theta(0, \cdot) = \Theta_I \text{ in } \Omega, \quad \Theta = 0,\ p = 0 \text{ on } (0, T] \times \Gamma. \tag{11}$$

Observe that the present formulation has the complementary pressure Θ as unknown and not the saturation s. Hence the initial value is provided for Θ and not for s. However, given an initial saturation s_I, obtaining the initial value for Θ is straightforward. Further, for simplicity, we have restricted our attention to homogeneous Dirichlet boundary conditions, but other kinds of conditions can be considered.

Remark 2.1. *Adding source/well terms $f_{s,\alpha}$ in the mass balance laws* (7) *can be done without major complications, at least if these have the standard regularity (see e.g. [3]). Further, if $f_{s,\alpha}$ depend on the phase saturation, to obtain the same results on the convergence of the discretization or of the linearization one needs to assume that $f_{s,\alpha}(\cdot)$ satisfies a condition similar to the one in Assumption (A3) below. For the ease of presentation and due to the analogy with the model considered in [44], such terms are left out here. This also simplifies the presentation of the convergence proof in Section 3.*

Notations. In the following we use common notations from functional analysis, e.g. $L^\infty(\Omega)$ is the space of essentially bounded functions on Ω, $L^2(\Omega)$ the space of square integrable functions on Ω, or $H^1(\Omega)$ is the subspace of $L^2(\Omega)$ containing functions which have also first order distributional derivatives in $L^2(\Omega)$.

We denote by $H_0^1(\Omega)$ the space of $H^1(\Omega)$ functions with a vanishing trace on Γ and by $H^{-1}(\Omega)$ its dual. $\langle\cdot,\cdot\rangle$ denotes the inner product in $L^2(\Omega)$, or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Further, $\|\cdot\|$, $\|\cdot\|_1$, $\|\cdot\|_p (p>1, p\neq 2)$ and $\|\cdot\|_\infty$ stand for the norms in $L^2(\Omega)$, $H^1(\Omega)$, $L^p(\Omega)$ and $L^\infty(\Omega)$, respectively. The functions in $H(\mathrm{div};\Omega)$ are vector valued having a L^2 divergence. The norm in $H(\mathrm{div};\Omega)$ is denoted by $\|\cdot\|_{div}$. $L^2(0,T;X)$ denotes the Bochner space of X-valued functions defined on $(0,T)$, where X is a Banach space. Similarly, $C(0,T;X)$ are X-valued functions continuous (w. r. t. X norm) on $[0,T]$. By C we mean a generic positive constant, not depending on the unknowns or the discretization parameters and we denote by L_f the Lipschitz constant of a (Lipschitz continuous) function $f(\cdot)$.

Further, we will denote by $N \geq 1$ an integer giving the time step $\tau = T/N$. For a given $n \in \{1,2,\ldots,N\}$, the nth time point is $t_n = n\tau$. We will also use the following notation for the mean over a time interval. Given the function $g \in L^2(0,T;X)$ (X being a Banach space like $L^2(\Omega)$, or $H^1(\Omega)$), its time averaged over the interval $(t_{n-1},t_n]$ is defined as

$$\bar{g}^n := \frac{1}{\tau}\int_{t_{n-1}}^{t_n} g(t)dt.$$

Clearly, this is an element in X as well.

Problem P: Continuous mixed variational formulation. Find $\Theta, p \in L^2(0,T;L^2(\Omega))$, $\mathbf{q} \in L^2(0,T;L^2(\Omega)^d)$, $\mathbf{u} \in L^2(0,T;H(\mathrm{div};\Omega))$ such that there holds $s(\Theta) \in L^\infty(\Omega\times(0,T))$, $\int_0^t \mathbf{q}(y)\,dy \in C(0,T;H(\mathrm{div};\Omega))$, and

$$\langle s(\Theta(t)) - s(\Theta^0), w\rangle + \langle \nabla\cdot\int_0^t \mathbf{q}(y)\,dy, w\rangle = 0, \tag{12}$$

$$\langle\int_0^t \mathbf{q}(y)\,dy, \mathbf{v}\rangle - \langle\int_0^t \Theta(y)\,dy, \nabla\cdot\mathbf{v}\rangle$$

$$-\langle\int_0^t f_w(s(\Theta(y)))\mathbf{u}(y)\,dy, \mathbf{v}\rangle = \langle\int_0^t \mathbf{f}_1(s(\Theta(y)))\,dy, \mathbf{v}\rangle, \tag{13}$$

$$\langle\nabla\cdot\mathbf{u}(t), w\rangle = \langle f_2(s(\Theta(t))), w\rangle, \tag{14}$$

$$\langle a(s(\Theta(t)))\mathbf{u}(t), \mathbf{v}\rangle - \langle p(t), \nabla\cdot\mathbf{v}\rangle + \langle\mathbf{f}_3(s(\Theta(t))), \mathbf{v}\rangle = 0 \tag{15}$$

for all $t\in(0,T]$, $w\in L^2(\Omega)$ and $\mathbf{v}\in H(\mathrm{div};\Omega)$, with $\Theta(0)=\Theta_I\in L^2(\Omega)$.

In the present formulation the initial condition is given for Θ. Moreover, the initial condition is imposed explicitly and not through integrating in time. In fact, (12) - (15) hold for every t, this being justified as follows. By (12), since $\int_0^t \mathbf{q}(y)dy$ is continuous in time, it follows that $s(\Theta)\in C(0,T;L^2(\Omega))$. Further, $s(\cdot)$ and $f_2(\cdot)$ are assumed continuous (see (A1) and (A3) below), so from (14) one obtains that $\mathbf{u}\in C(0.T;H(\mathrm{div};\Omega))$. Similarly, p is continuous in time as well, so (13)-(15) hold for all $t\in(0,T]$.

We proceed with the time discretization for Problem P, which is achieved by the Euler implicit scheme. For a given $n\in\{1,2,\ldots,N\}$, we define the time discrete mixed variational problem at time t_n.

Problem P^n: Semi-discrete variational formulation. Let Θ^{n-1} be given. Find $\Theta^n, p^n \in L^2(\Omega)$ and $\mathbf{u}^n, \mathbf{q}^n \in H(\mathrm{div};\Omega)$ such that

$$\langle s^n - s^{n-1}, w\rangle + \tau\langle\nabla\cdot\mathbf{q}^n, w\rangle = 0, \tag{16}$$

$$\langle\mathbf{q}^n, \mathbf{v}\rangle - \langle\Theta^n, \nabla\cdot\mathbf{v}\rangle - \langle f_w(s^n)\mathbf{u}^n, \mathbf{v}\rangle = \langle\mathbf{f}_1(s^n), \mathbf{v}\rangle, \tag{17}$$

$$\langle\nabla\cdot\mathbf{u}^n, w\rangle = \langle f_2(s^n), w\rangle, \tag{18}$$

$$\langle a(s^n)\mathbf{u}^n, \mathbf{v}\rangle - \langle p^n, \nabla\cdot\mathbf{v}\rangle + \langle\mathbf{f}_3(s^n), \mathbf{v}\rangle = 0 \tag{19}$$

for all $w \in L^2(\Omega)$, and $\mathbf{v} \in H(\mathrm{div};\Omega)$. Initially we take $\Theta^0 = \Theta_I \in L^2(\Omega)$. Throughout this paper s^k stands for $s(\Theta^k)$, $k \in \mathbb{N}$, making the presentation easier.

We can now proceed with the spatial discretization. For this let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size (see [17]). Here we assume $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$, hence Ω is polygonal. Thus we neglect the errors caused by an approximation of a non-polygonal domain and avoid an excess of technicalities (a complete analysis in this sense can be found in [35]).

The discrete subspaces $W_h \times V_h \subset L^2(\Omega) \times H(\mathrm{div};\Omega)$ are defined as

$$\begin{aligned} W_h &:= \{p \in L^2(\Omega) |\ p \text{ is constant on each element } T \in \mathcal{T}_h\}, \\ V_h &:= \{\mathbf{q} \in H(\mathrm{div};\Omega) | \mathbf{q}_{|T}(\mathbf{x}) = \mathbf{a_T} + b_T\mathbf{x}, \mathbf{a_T} \in \mathbb{R}^d, b_T \in \mathbb{R} \text{ for all } T \in \mathcal{T}_h\}. \end{aligned} \tag{20}$$

So W_h denotes the space of piecewise constant functions, while V_h is the RT_0 space (see [7]).

The fully discrete (non-linear) scheme can now be given. To simplify the presentation we use in the following the notation $s_h^n := s(\Theta_h^n)$, $n \in \mathbb{N}$.

Problem P_h^n: Fully discrete (non-linear) variational formulation. Let $n \in \mathbb{N}, n \geq 1$, and assume Θ_h^{n-1} is known. Find $\Theta_h^n, p_h^n \in W_h$ and $\mathbf{q}_h^n, \mathbf{u}_h^n \in V_h$ such that there holds

$$\langle s_h^n - s_h^{n-1}, w_h\rangle + \tau\langle \nabla\cdot\mathbf{q}_h^n, w_h\rangle = 0, \tag{21}$$

$$\langle \mathbf{q}_h^n, \mathbf{v}_h\rangle - \langle \Theta_h^n, \nabla\cdot\mathbf{v}_h\rangle - \langle f_w(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle = \langle \mathbf{f}_1(s_h^n), \mathbf{v}_h\rangle, \tag{22}$$

$$\langle \nabla\cdot\mathbf{u}_h^n, w_h\rangle = \langle f_2(s_h^n), w_h\rangle, \tag{23}$$

$$\langle a(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle - \langle p_h^n, \nabla\cdot\mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^n), \mathbf{v}_h\rangle = 0 \tag{24}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$.

The fully discrete scheme (21) – (24) is non-linear, and an iterative method is required for solving it. Moreover, as will follow from the assumptions below, here we consider the degenerate model, which means that the derivatives of the function $s(\cdot)$ may vanish or blow up for some arguments. This is, indeed, the situation encountered in two-phase porous media flow models. In such cases, usual schemes such as the Newton method may not converge without performing a regularization step (see [42] for a proof obtained for a similar model, the Richards equation). However, the regularization may affect the mass balance. To avoid this, we follow the ideas in [40, 51, 55], and propose a robust, first order convergent linearization scheme for solving (21) – (24). In particular, the scheme is not requiring any regularization. It is defined for the case of a MFEM discretization, but similar ideas can be applied for any other spatial discretization method, see e.g. [45] for MPFA.

Let $n \in N, n \geq 1$ be fixed. With $L > 0$ being a constant that will be specified below, an iterative scheme for solving the non-linear problem (21) – (24) is introduced through

Problem $P_h^{n,i}$: Linearization scheme (L-scheme). Let $L > 0$, $i \in \mathbb{N}, i \geq 1$ and let $\Theta_h^{n,i-1} \in W_h$ be given. Find $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ such that

$$\langle L(\Theta_h^{n,i} - \Theta_h^{n,i-1}) + s_h^{n,i-1}, w_h\rangle + \tau\langle \nabla\cdot\mathbf{q}_h^{n,i}, w_h\rangle = \langle s_h^{n-1}, w_h\rangle, \tag{25}$$

$$\langle \mathbf{q}_h^{n,i}, \mathbf{v}_h\rangle - \langle \Theta_h^{n,i}, \nabla\cdot\mathbf{v}_h\rangle - \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i}, \mathbf{v}_h\rangle = \langle \mathbf{f}_1(s_h^{n,i-1}), \mathbf{v}_h\rangle, \tag{26}$$

$$\langle \nabla\cdot\mathbf{u}_h^{n,i}, w_h\rangle = \langle f_2(s_h^{n,i-1}), w_h\rangle, \tag{27}$$

$$\langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i}, \mathbf{v}_h\rangle - \langle p_h^{n,i}, \nabla\cdot\mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^{n,i-1}), \mathbf{v}_h\rangle = 0 \tag{28}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. We use the notation $s_h^{n,i} := s(\Theta_h^{n,i})$, $n \in \mathbb{N}$ and, as previously, $s_h^n := s(\Theta_h^n)$, $n \in \mathbb{N}$. For starting the iterations, a natural choice is $\Theta_h^{n,0} := \Theta_h^{n-1}$ and, correspondingly, $s_h^{n,0} := s_h^{n-1}$.

Throughout this paper we make the following assumptions.

(A1) The function $s(\cdot) : \mathbb{R} \to \mathbb{R}$, $s(0) = 0$ is strictly increasing and Hölder continuous, with exponent $\alpha \in (0, 1]$. There exists $L_s > 0$ s.t. there holds

$$|s(\Theta_1) - s(\Theta_2)| \leq L_s |\Theta_1 - \Theta_2|^\alpha \quad \text{for all } \Theta_1, \Theta_2 \in \mathbb{R}.$$

(A2) $a(\cdot)$ satisfies the following growth condition

$$|a(s(\Theta_1) - a(s(\Theta_2)|^2 \leq C\langle s_1 - s_2, \Theta_1 - \Theta_2\rangle, \quad \forall \Theta_1, \Theta_2 \in \mathbb{R},$$

and there exists $a_\star, a^\star > 0$ such that for all $y \in \mathbb{R}$ one has

$$0 < a_\star \leq a(y) \leq a^\star < \infty.$$

(A3) The functions $\mathbf{f}_1(\cdot)$, $f_2(\cdot)$, $\mathbf{f}_3(\cdot)$ and $f_w(\cdot)$ satisfy $F(0) = 0$ and the growth condition

$$|F(s(\Theta_1)) - F(s(\Theta_2))|^2 \leq C\langle s_1 - s_2, \Theta_1 - \Theta_2\rangle,$$

for some generic constant $C > 0$ and where F is any of the functions above. Additionally, $f_w(\cdot)$ is uniformly bounded.

(A4) There exits a constant $M_{\mathbf{u}} < \infty$ such that $\|\mathbf{u}\|_\infty \leq M_{\mathbf{u}}$, $\|\mathbf{u}_h^n\|_\infty \leq M_{\mathbf{u}}$, and $\|\mathbf{u}_h^{n,i}\|_\infty \leq M_{\mathbf{u}}$ for all $n \in \mathbb{N}$, the last two being uniformly in h and i. Here $\mathbf{u}$, $\mathbf{u}_h^n$ and $\mathbf{u}_h^{n,i}$ are the solution components in Problems P, P_h^n and $\mathrm{P}_h^{n,i}$ respectively.

(A5) The function Θ_I is in $L^2(\Omega)$.

(A6) The solutions $\mathbf{q}^n, \mathbf{u}^n$ of the semi-discrete problem P^n satisfy

$$\sum_{n=1}^{N} \tau \|\mathbf{u}^n\|_1^2 + \sum_{n=1}^{N} \tau \|\mathbf{q}^n\|_1^2 \leq C\tau^{\frac{2(\alpha-1)}{1+\alpha}},$$

where α is the Hölder exponent of $s(\cdot)$.

Remark 2.2. *The Hölder continuous assumption covers many cases of practical interest. For example, when considering Brooks-Corey type models one has $p^{cap}(S) \propto S^{-1/\lambda}$ with some $\lambda > 1$, S being the water saturation. Assuming now that close to $S = 0$ the water mobility behaves asymptotically as $\lambda_w(S) \propto S^\alpha$, then $s(\cdot)$ is Hölder continuous if $\alpha \geq \dfrac{\lambda + 1}{\lambda}$.*

Remark 2.3. *We assume that $s(\cdot)$ is strictly increasing only in order to ensure the existence and uniqueness of a solution, as already have been proved in the literature (we discuss this in Section 3). The a priori estimates in Section 3, as well as the results concerning the convergence of the linearization scheme in Section 4 are obtained without this. Nevertheless, the existence of a solution is not proved rigorously for this case.*

Remark 2.4. *The growth conditions in (A2) and (A3) are, in the case of a Lipschitz continuous $s(\cdot)$, weaker assumptions as the Lipschitz continuity w.r.t. to $s(\cdot)$. Although in the Hölder continuous case the growth conditions are not necessarily stronger assumptions as the Lipschitz continuity (one can not order the two in this case), they seem to be indispensable for the analysis and can not be replaced by Lipschitz continuity requirements.*

Remark 2.5. *Concerning (A4), for* $\mathbf{u}$ *this is practically the outcome of the assumptions (A1) and (A3), which guarantee that for every* $t \in [0,T]$ *one has* $f_2(s(\Theta(t))) \in L^\infty(\Omega)$*, and that the* L^∞ *norm is bounded uniformly w.r.t. time. Now, without being rigorous, we observe that by* (14) *one obtains* $\mathbf{u}(t) = -\nabla w(t)$*, where* w *satisfies* $-\Delta w(t) = f_2(s(\Theta(t)))$*. Classical regularity theory (see e.g. [32], Thm. 15.1 in Chapter 3) guarantees that* ∇w *is continuous on the compact* $\bar{\Omega}$*, and that the* L^∞ *norm can be bounded uniformly in time. For the approximation* $\mathbf{u}_h^n$*, one can reason in the same manner, and observe that* $\mathbf{u}_h^n$ *becomes the projection* $\Pi_h(-\nabla w(t_n))$*. Since* $\|\nabla w(t_n)\|_\infty$ *is bounded uniformly in time, the construction of the projector* Π_h *(see e. g. [48], Chapter 7.2) guarantees that* $\mathbf{u}_h^n$ *satisfies the same bounds as* ∇w*. Finally, case of* $\mathbf{u}_h^{n,i}$ *is similar. We also refer to [44] for a similar situation but in the case of a one phase flow model, where conditions ensuring the validity of (A4) are provided.*

Remark 2.6. *The assumption (A6) is inspired by the stability estimates in Proposition 3.1. We point out the negative exponent of* τ *on the right hand side and that in the one dimensional case the inequality in (A6) is proved in Proposition 3.1.*

The following two technical lemmas will be used in Sections 3 and 4. Their proofs can be found e.g. in [44] and [53] or [18], respectively.

Lemma 2.1. *Given a* $w \in L^2(\Omega)$*, there exists a* $\mathbf{v} \in H(\mathrm{div};\Omega)$ *such that*

$$\nabla \cdot \mathbf{v} = w \text{ and } \|\mathbf{v}\| \leq C_\Omega \|w\|,$$

with $C_\Omega > 0$ *not depending on* w*.*

Lemma 2.2. *Given a* $w_h \in W_h$*, there exits a* $\mathbf{v}_h \in V_h$ *satisfying*

$$\nabla \cdot \mathbf{v}_h = w_h \text{ and } \|\mathbf{v}_h\| \leq C_{\Omega,d} \|w_h\|,$$

with $C_{\Omega,d} > 0$ *not depending on* w_h *or mesh size.*

Also, the following elementary results will be used

Proposition 2.1. *Let* $\mathbf{a}_k \in \mathbb{R}^d$ $(k \in \{1,\dots,N\}, d \geq 1)$ *be a set of* N *vectors. It holds*

$$\sum_{n=1}^{N} \langle \mathbf{a}_n, \sum_{k=1}^{n} \mathbf{a}_k \rangle \;=\; \frac{1}{2} \left\| \sum_{n=1}^{N} \mathbf{a}_n \right\|^2 + \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{a}_n\|^2 \,. \tag{29}$$

Proposition 2.2. *(Young's inequality) Let* $a, b \in \mathbb{R}$*,* $\epsilon > 0$ *and* $p, q > 1$ *s.t.* $\dfrac{1}{p} + \dfrac{1}{q} = 1$*. Then,*

$$|ab| \leq \epsilon \frac{|a|^p}{p} + \epsilon^{-\frac{q}{p}} \frac{|b|^q}{q} \,. \tag{30}$$

Lemma 2.3. *(***Discrete Gronwall Lemma** *(see e.g. [48]) Let* a_n, λ_n *positive numbers for all integers* $n \in \mathbb{N}$ *and* $B \geq 0$*. If* $\lambda_n < 1 \, \forall \, n \in \mathbb{N}$ *and*

$$a_n \leq B + \sum_{k=0}^{n} \lambda_k a_k \quad \forall \, n \in \mathbb{N},$$

then there holds

$$a_n \leq B e^{\sum_{k=0}^{n} \frac{\lambda_k}{1-\lambda_k}} \quad \forall \, n \in \mathbb{N}.$$

3 Analysis of the discretization

In this section we analyze the discretization of Problems P, introduced through the problems P_n and P_h^n in Section 2. The convergence of the numerical scheme is shown by deriving *a priori* stability and error estimates. The main result is given in Theorem 3.1. Note that the convergence result involves the exact solution of the non-linear, fully discrete systems (21) - (24). However, in general only an approximation of this solution is available, and this is obtained by means of an iterative scheme. Here we use the iterations introduced through Problem $P_h^{n,i}$, the convergence of this iterative method being analyzed in Section 4.

When carrying out the analysis announced above we assume that the all problems introduced before in the variational formulation (continuous, semi-discrete, or fully discrete) have a unique solution. A rigorous proof of the existence and uniqueness for these problems is nevertheless beyond the scope of this work.

We mention that existence and uniqueness results for the two phase flow model has been studied intensively in the past. Closest to the framework considered here are [3, 14, 49] (see also [16]). We refer to [3], where the existence of a weak solution is proved under assumptions that are similar to (A1)-(A3), by employing a Galerkin approach. This result has been extended in [14] by a time-discretization method. In [49], the existence of a solution is the outcome of the convergence result for a combined finite volume-nonconforming finite volume scheme, and applies to the case considered here when the fluids are considered incompressible. Also, the growth condition in (A3) is required in [14] for proving the uniqueness of a solution.

Other relevant references for the existence and uniqueness are [1, 2, 23, 27]. Also, we refer to [9] for the existence of a solution in heterogeneous media, where the phase pressure differences may become discontinuous at the interface separating two homogeneous blocks.

Observe that all papers mentioned above deal with the conformal formulation of the two-phase porous media flow model. However, the numerical scheme discussed here uses a mixed formulation. For such methods, existence results are generally using the *LBB condition* [7]. Alternatively, one can use the existence and uniqueness results for the conformal formulation, and prove the equivalence between the two formulations. This idea is being adopted in [41, 44, 46]. Finally, we mention [50], where existence for a general class of degenerate evolution problems in mixed formulation is obtained in an abstract framework.

For the time discrete problems P^n we refer again to [41, 44] (see also [46], where the equivalence between semi-discrete conformal and mixed formulations is established for Lipschitz continuous problems). Another approach for obtaining the existence of a solution is by considering the limit of Galerkin approximations, as done in [3] (also see [14]). The uniqueness follows from the *a priori* estimates proved in Proposition 3.1 below.

Finally, for the fully discrete non-linear problems P_h^n, if $s(\cdot)$ is Lipschitz continuous, i.e. $\alpha = 1$ in (A1), the existence and uniqueness follows from Theorem 4.1 in Section 4. This is, in fact a consequence of the Banach fixed point theorem. When $s(\cdot)$ is only Hölder continuous, Theorem 4.1 does not provide a contraction argument anymore. In this case, assuming additionally that $s(\cdot)$ is strictly increasing, the existence can be proved by using Brouwer's fixed point theorem, see e.g. [43] or the recent paper [13] for a similar approach applied to MFEM. We also refer to [14] for a similar situation, the only difference being in the fact that there a standard conformal Galerkin formulation is adopted for the pressure equation, whereas here we use the MFEM. The uniqueness can be established without difficulties by the same techniques used in the present work to obtain error estimates. Nevertheless, the existence and uniqueness for the case when the saturation is not strictly continuous remains an open problem.

3.1 *A priori* (stability) estimates

In this section we prove stability estimates for the semi-discrete Problem P^n. One can prove estimates also for the fully discrete Problem P_h^n, but these estimates are not needed for proving the convergence of the scheme and are therefore skipped here. The techniques used are similar with the ones in [44] (Lemma 3.2, pg. 293).

Proposition 3.1. *Assume that (A1)-(A5) hold true. Let* $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$, $n \in \mathbb{N}, n \geq 1$ *be the solution of Problem* P^n. *Then there holds*

$$\sum_{n=1}^{N} \tau \langle s^n, \Theta^n \rangle + \tau \sum_{n=1}^{N} \|p^n\|_1^2 + \tau \sum_{n=1}^{N} \|\mathbf{u}^n\|_{div}^2 + \tau \sum_{n=1}^{N} \|\Theta^n\|_1^2 + \tau \sum_{n=1}^{N} \|\mathbf{q}^n\|^2 \leq C, \tag{31}$$

$$\sum_{n=1}^{N} \langle s^n - s^{n-1}, \Theta_n - \Theta_{n-1} \rangle + \sum_{n=1}^{N} \|s^n - s^{n-1}\|_{1+1/\alpha}^{1+1/\alpha} + \tau \sum_{n=1}^{N} \|\mathbf{q}^n - \mathbf{q}^{n-1}\|^2 \leq C\tau, \tag{32}$$

$$\tau \sum_{n=1}^{N} \|\nabla \cdot \mathbf{q}^n\|^2 \leq C\tau^{\frac{2(\alpha-1)}{1+\alpha}} \tag{33}$$

where the constants $C > 0$ *do not depend on the discretization parameters and* $\alpha \in (0, 1]$ *is the Hölder exponent of the saturation.*

Proof. We prove first that there holds

$$\|p^n\|^2 + \|\mathbf{u}^n\|^2 \leq C \langle s^n, \Theta^n \rangle. \tag{34}$$

Testing (18) with $w = p^n \in W$ and (19) with $\mathbf{v} = \mathbf{u}^n \in V$ and adding the results we obtain

$$\langle a(s^n)\mathbf{u}^n, \mathbf{u}^n \rangle = \langle f_2(s^n), p^n \rangle - \langle \mathbf{f}_3(s^n), \mathbf{u}^n \rangle.$$

Using (A2), (A3) and Young's inequality, one immediately obtains from the above

$$\|\mathbf{u}^n\|^2 \leq \frac{C}{\delta} \langle s^n, \Theta^n \rangle + \delta \|p^n\|^2, \tag{35}$$

for any $\delta > 0$. Using now Lemma 2.1, there exists $\mathbf{v} \in H(\mathrm{div}; \Omega)$, such that $\nabla \cdot \mathbf{v} = p^n$ and $\|\mathbf{v}\| \leq C\|p^n\|$. Testing (19) with this $\mathbf{v}$ and using (A2), (A3) and Young's inequality we get

$$\|p^n\|^2 \leq C(\|\mathbf{u}^n\|^2 + \langle s^n, \Theta^n \rangle). \tag{36}$$

The inequality (34) follows now immediately from (35) and (36), by choosing δ properly. We proceed by showing that there holds $\sum_{n=1}^{N} \tau \langle s^n, \Theta^n \rangle \leq C$. Before doing it, we point out that this inequality, together with (36), the estimate $\|\nabla \cdot \mathbf{u}^n\|^2 \leq C \langle s^n, \Theta^n \rangle$ (which is immediately from (18), using (A3)) and the fact that p^n is actually in $H_0^1(\Omega)$ and $\|\nabla p^n\|^2 \leq \|\mathbf{u}^n\|^2 + \langle s^n, \Theta^n \rangle$ (which follows from (19), using again (A3)) implies also $\tau \sum_{n=1}^{N} \|p^n\|_1^2 \leq C$ (we used Poincare inequality as well). The same procedure was used in [44] at pg. 294 to estimate $\|p^n\|_1^2$.

Writing (16) for $n = k$ and summing up from $k = 1$ to n, one gets for all $w \in W$

$$\langle s^n, w \rangle + \tau \langle \sum_{k=1}^{n} \nabla \cdot \mathbf{q}^k, w \rangle = \langle s^0, w \rangle. \tag{37}$$

Testing (17) with $\mathbf{v} = \sum_{k=1}^{n} \mathbf{q}^k \in V$ we get

$$\langle \mathbf{q}^n, \sum_{k=1}^{n} \mathbf{q}^k \rangle - \langle \Theta^n, \nabla \cdot \sum_{k=1}^{n} \mathbf{q}^k \rangle - \langle f_w(s^n)\mathbf{u}^n, \sum_{k=1}^{n} \mathbf{q}^k \rangle = \langle \mathbf{f}_1(s^n), \sum_{k=1}^{n} \mathbf{q}^k \rangle.$$

Adding the above multiplied by τ to (37) tested with $w = \Theta^n \in W$, and summing up the result from $n = 1$ to N we obtain

$$\sum_{n=1}^{N} \langle s^n, \Theta^n \rangle + \tau \sum_{n=1}^{N} \langle \mathbf{q}^n, \sum_{k=1}^{n} \mathbf{q}^k \rangle = \langle s^0, \sum_{n=1}^{N} \Theta^n \rangle + \tau \sum_{n=1}^{N} \langle f_w(s^n)\mathbf{u}^n, \sum_{k=1}^{n} \mathbf{q}^k \rangle + \tau \sum_{n=1}^{N} \langle \mathbf{f}_1(s^n), \sum_{k=1}^{n} \mathbf{q}^k \rangle.$$

By using the identity (29), Young's inequality and the assumption (A3), the above further implies

$$\sum_{n=1}^{N} \langle s^n, \Theta^n \rangle + \tau \sum_{n=1}^{N} \|\mathbf{q}^n\|^2 + \tau \| \sum_{n=1}^{N} \mathbf{q}^n \|^2 \leq \frac{\|s^0\|^2}{2\delta\tau} + \frac{\delta\tau}{2} \| \sum_{n=1}^{N} \Theta^n \|^2 + C\tau^2 \sum_{n=1}^{N} \| \sum_{k=1}^{n} \mathbf{q}^k \|^2. \tag{38}$$

Due to Lemma 2.1, there exists $\mathbf{v} \in H(\mathrm{div}; \Omega)$, such that $\nabla \cdot \mathbf{v} = \sum_{n=1}^{N} \Theta^n$ and $\|\mathbf{v}\| \leq C \| \sum_{n=1}^{N} \Theta^n \|$. Summing up (17) from $n = 1$ to N testing with this $\mathbf{v}$, and applying Young's, triangle's and Cauchy-Schwarz's inequalities together with (36) gives

$$\| \sum_{n=1}^{N} \Theta^n \|^2 \leq C \| \sum_{n=1}^{N} \mathbf{q}^n \|^2 + C \frac{1}{\tau} \sum_{n=1}^{N} \langle s^n, \Theta^n \rangle. \tag{39}$$

The result $\sum_{n=1}^{N} \tau \langle s^n, \Theta^n \rangle \leq C$ follows now from (38) and (39), by choosing appropriately δ and applying Gronwall's lemma 2.3. In order to complete the proof of (31), we still have to show the boundedness of $\tau \sum_{n=1}^{N} \|\Theta^n\|_1^2 + \tau \sum_{n=1}^{N} \|\mathbf{q}^n\|^2$. Testing (16) with $w = \Theta^n \in W$ and (17) with $\mathbf{v} = \tau \mathbf{q}^n$, adding the results and summing up from $n = 1$ to N we obtain

$$\sum_{n=1}^{N} \langle s^n - s^{n-1}, \Theta^n \rangle + \tau \sum_{n=1}^{N} \|\mathbf{q}^n\|^2 = \tau \sum_{n=1}^{N} \langle f_w(s^n)\mathbf{u}^n, \mathbf{q}^n \rangle + \tau \sum_{n=1}^{N} \langle \mathbf{f}_1(s^n), \mathbf{q}^n \rangle. \tag{40}$$

Following [44], pg. 293 there holds for the first term above

$$\begin{aligned}
\sum_{n=1}^{N} \langle s^n - s^{n-1}, \Theta^n \rangle &\geq \sum_{n=1}^{N} \int_\Omega \int_{\Theta^{n-1}}^{\Theta^n} \Theta s'(\Theta) d\Theta dx = \int_\Omega \int_0^{\Theta^n} \Theta s'(\Theta) d\Theta dx - \int_\Omega \int_0^{\Theta^0} \Theta s'(\Theta) d\Theta dx \\
&\geq -\int_\Omega \int_0^{\Theta^0} \Theta s'(\Theta) d\Theta dx = \int_\Omega \int_0^{\Theta^0} s(\Theta) d\Theta dx - \int_\Omega \Theta^0 s(\Theta^0) dx \\
&\geq -C \int_\Omega |\Theta^0|^{1+\alpha} \geq -C.
\end{aligned}$$

The two terms in the right hand side of (40) can be estimated by using Young's inequality, (A3) and (36). This, together with the above and (40) implies

$$\tau \sum_{n=1}^{N} \|\mathbf{q}^n\|^2 \leq C + C \sum_{n=1}^{N} \tau \langle s^n, \Theta^n \rangle \leq C. \tag{41}$$

Considering now that $\Theta^n \in H_0^1(\Omega)$, using (17), (A3), (36), again $\sum_{n=1}^N \tau\langle s^n, \Theta^n\rangle \leq C$ and (41) gives

$$\tau \sum_{n=1}^{N} \|\nabla\Theta^n\|^2 \leq C.$$

This, together with the Poincare inequality and (41) gives (31). Let us now prove (32). Subtracting (19) at $n-1$ from the one at n, testing with $\tau\mathbf{u}^n$ and adding the result to the difference between (19) at n and at $n-1$, tested with $p^n - p^{n-1}$ we get

$$\langle a(s^n)\mathbf{u}^n - a(s^{n-1})\mathbf{u}^{n-1}, \mathbf{u}^n - \mathbf{u}^{n-1}\rangle = -\langle \mathbf{f}_3(s^n) - \mathbf{f}_3(s^{n-1}), \mathbf{u}^n - \mathbf{u}^{n-1}\rangle + \langle f_2(s^n) - f_2(s^{n-1}), p^n - p^{n-1}\rangle.$$

By using now (A2), (A3), (A4) and Young's inequality, the above implies

$$\|\mathbf{u}^n - \mathbf{u}^{n-1}\|^2 \leq (\frac{C}{\delta} + C)\langle s^n - s^{n-1}, \Theta^n - \Theta^{n-1}\rangle + \delta\|p^n - p^{n-1}\|^2 \tag{42}$$

for all $\delta > 0$. Using Lemma 2.1, and (19) we obtain

$$\|p^n - p^{n-1}\|^2 \leq C\langle s^n - s^{n-1}, \Theta^n - \Theta^{n-1}\rangle + C\|\mathbf{u}^n - \mathbf{u}^{n-1}\|^2. \tag{43}$$

From (42) and (43) follows

$$\|\mathbf{u}^n - \mathbf{u}^{n-1}\|^2 + \|p^n - p^{n-1}\|^2 \leq C\langle s^n - s^{n-1}, \Theta^n - \Theta^{n-1}\rangle. \tag{44}$$

Proceeding similarly with (17) and (16) one obtains

$$\begin{aligned}\sum_{n=1}^{N}\langle s^n - s^{n-1}, \Theta^n - \Theta^{n-1}\rangle + \tau\sum_{n=1}^{N}\langle \mathbf{q}^n - \mathbf{q}^{n-1}, \mathbf{q}^n\rangle = \tau\sum_{n=1}^{N}\langle f_w(s^n)\mathbf{u}^n - f_w(s^{n-1})\mathbf{u}^{n-1}, \mathbf{q}^n\rangle \\ +\tau\sum_{n=1}^{N}\langle \mathbf{f}_1(s^n) - \mathbf{f}_1(s^{n-1}), \mathbf{q}^n\rangle.\end{aligned} \tag{45}$$

The above equation implies (use (44), (A2), (A3) and Young's inequality)

$$\sum_{n=1}^{N}\langle s^n - s^{n-1}, \Theta^n - \Theta^{n-1}\rangle + \frac{\tau}{2}\sum_{n=1}^{N}\|\mathbf{q}^n - \mathbf{q}^{n-1}\|^2 + \frac{\tau}{2}\|\mathbf{q}^N\|^2 \leq \frac{\tau}{2}\|\mathbf{q}^0\|^2 + C\tau^2\sum_{n=1}^{N}\|\mathbf{q}^n\|^2.$$

The result (32) follows now from above by using the Gronwall lemma 2.3 and that $\mathbf{q}^0 \in L^2(\Omega)$ (see Proposition 3.5 in [41]). It remains to show (33). By testing (16) with $w = \nabla\cdot\mathbf{q}^n \in V$ we have

$$\tau\|\nabla\cdot\mathbf{q}^n\|^2 = \langle s^n - s^{n-1}, \nabla\cdot\mathbf{q}^n\rangle.$$

This and (A1) immediately gives

$$\tau\sum_{n=1}^{N}\|\nabla\cdot\mathbf{q}^n\|^2 \leq \frac{C}{\tau}\sum_{n=1}^{N}\|s^n - s^{n-1}\|^2$$

Following [44], pg. 295, the L^2-norms can be estimated by using the $L^{1+\frac{1}{\alpha}}$ bounds. By using Young's inequality there holds

$$\sum_{n=1}^{N}\left\|s^n - s^{n-1}\right\|^2 \leq \frac{1}{\tau^r}\sum_{n=1}^{N}\left(\int_\Omega \frac{\tau}{p}^{rp}dx + \frac{1}{q}\|s^n - s^{n-1}\|^{2q}_{L^{2q}(\Omega)}\right),$$

where $r = \dfrac{2(1-\alpha)}{1+\alpha}$. With $p = \dfrac{\alpha+1}{1-\alpha}$ and $q = \dfrac{\alpha+1}{2\alpha}$ and recalling (32) this gives

$$\sum_{n=1}^{N} \left\| s^n - s^{n-1} \right\|^2 \leq C\tau^{1-r},$$

and the rest of the proof is straightforward. □

3.2 *A priori* error estimates

With the stability estimates obtained above we can focus now on the convergence of the scheme (21)-(24). This will be done by deriving *a priori* error estimates. We point out that the convergence of the scheme will be proved without assuming that the saturation or its inverse are Lipschitz continuous, which makes the analysis very challenging. We assume that the fully discrete non-linear problem (21)-(24) is solved exactly. The proofs of this section follow the lines in [44] and [16]. The following two propositions will quantify the error between the continuous and the semi-discrete formulations, and between the semi-discrete and discrete ones, respectively. Finally the two propositions will be put together to obtain the main convergence result given in Theorem 3.1.

Recalling the definition $\bar{g}^n := \frac{1}{\tau}\int_{t_{n-1}}^{t_n} g(t)dt \in X$, for any $g \in L^2(0,T;X)$ and X a Banach space, we have

Proposition 3.2. *Let $(\Theta, \mathbf{q}, p, \mathbf{u})$ be the solution of Problem P and $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$ be the solution of $P^n, n \in \mathbb{N}, n \geq 1$. Assuming (A1)-(A5) and that the time step τ is small enough there holds*

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta^n), \Theta(t) - \Theta^n \rangle \, dt + \| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\mathbf{q} - \mathbf{q}^n) \, dt \|^2$$

$$+ \sum_{n=1}^{N} \| \int_{t_{n-1}}^{t_n} (\mathbf{q} - \mathbf{q}^n) \, dt \|^2 \leq C\tau, \tag{46}$$

$$\tau \|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 + \tau \|\overline{p}^n - p^n\|^2 \leq C \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta^n), \Theta(t) - \Theta^n \rangle dt, \tag{47}$$

$$\| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\Theta(t) - \Theta^n) \, dt \|^2 \leq C\tau, \tag{48}$$

with the constants C not depending on the discretization parameters.

Proof. We start with proving (47). By integrating (14), (15) from t_{n-1} to t_n one obtains

$$\langle \nabla \cdot \overline{\mathbf{u}}^n, w \rangle = \langle \overline{f_2(s)}^n, w \rangle, \tag{49}$$

$$\langle \overline{a(s)\mathbf{u}}^n, \mathbf{v} \rangle - \langle \overline{p}^n, \nabla \cdot \mathbf{v} \rangle + \langle \overline{\mathbf{f}_3(s)}^n, \mathbf{v} \rangle = 0, \tag{50}$$

for all $w \in L^2(\Omega)$ and $\mathbf{v} \in H(\mathrm{div};\Omega)$. By subtracting now (18) and (19) from (49) and (50), respectively, we get for all $w \in L^2(\Omega)$ and $\mathbf{v} \in H(\mathrm{div};\Omega)$

$$\langle \nabla \cdot (\overline{\mathbf{u}}^n - \mathbf{u}^n), w \rangle = \langle \overline{f_2(s)}^n - f_2(s^n), w \rangle, \tag{51}$$

$$\langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \mathbf{v} \rangle - \langle \overline{p}^n - p^n, \nabla \cdot \mathbf{v} \rangle = -\langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \mathbf{v} \rangle, \tag{52}$$

Taking $w = \overline{p}^n - p^n \in L^2(\Omega)$ in (51) and $\mathbf{v} = \overline{\mathbf{u}}^n - \mathbf{u}^n \in H(\mathrm{div};\Omega)$ in (52), and adding the results we obtain

$$\langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle = \langle \overline{f_2(s)}^n - f_2(s^n), \overline{p}^n - p^n\rangle - \langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle.$$

By Young's inequality and some algebraic manipulation, this further implies

$$\begin{aligned}\langle \frac{1}{\tau}\int_{t^{n-1}}^{t_n} (a(s) - a(s^n))\mathbf{u}\, dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle + \langle \frac{a(s^n)}{\tau}\int_{t^{n-1}}^{t_n} \mathbf{u} - \mathbf{u}^n\, dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle \\ \leq \frac{1}{2\delta}\|\overline{f_2(s)}^n - f_2(s^n)\|^2 + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2 + \frac{1}{a_\star}\|\overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n)\|^2 + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2,\end{aligned}$$

for all $\delta > 0$. Using now (A2)-(A4) and Young's inequality, the above leads to

$$\frac{a_\star}{2}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 \leq \frac{C(\delta)}{\tau}\int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta^n), \Theta(t) - \Theta^n\rangle dt + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2, \tag{53}$$

with $C(\delta) > 0$ not depending on the discretization parameters. To estimate the last term above, one uses Lemma 2.1, ensuring the existence of a $\mathbf{v} \in H(\mathrm{div};\Omega)$ such that $\nabla\cdot\mathbf{v} = \overline{p}^n - p^n$ and $\|\mathbf{v}\| \leq C_\Omega\|\overline{p}^n - p^n\|$. Using this as test function in (52) gives

$$\begin{aligned}\|\overline{p}^n - p^n\|^2 &= \langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \mathbf{v}\rangle + \langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \mathbf{v}\rangle \\ &\leq C_\Omega^2\|\overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n\|^2 + C_\Omega^2\|\overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n)\|^2 + \frac{1}{2}\|\overline{p}^n - p^n\|^2.\end{aligned} \tag{54}$$

Proceeding as for (53), (54) further implies

$$\|\overline{p}^n - p^n\|^2 \leq \frac{C}{\tau}\int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta^n), \Theta(t) - \Theta^n\rangle dt + C\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2, \tag{55}$$

with the constant C not depending on the discretization parameters. Using (55) and (53), and choosing δ properly, one obtains (47).

To prove (46) we follow the steps in the proof of Lemma 3.3, pg. 296 in [44]. By summing up (16) for $k = 1$ to n and subtracting (12) from the resulting we get for all $w \in L^2(\Omega)$

$$\langle s(\Theta(t_n)) - s^n, w\rangle + \tau\sum_{k=1}^{n}\langle \nabla\cdot(\overline{\mathbf{q}}^k - \mathbf{q}^k), w\rangle = 0. \tag{56}$$

Further, subtracting (13) at $t = t_{k-1}$ from (13) at $t = t_k$, dividing by the time step size τ and subtracting from the result the equation (17) we obtain for all $\mathbf{v} \in H(\mathrm{div};\Omega)$

$$\langle \overline{\mathbf{q}}^n - \mathbf{q}^n, \mathbf{v}\rangle - \langle \overline{\Theta}^n - \Theta^n, \nabla\cdot\mathbf{v}\rangle - \langle \overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \mathbf{v}\rangle = \langle \overline{\mathbf{f}_1(s)}^n - \mathbf{f}_1(s^n), \mathbf{v}\rangle. \tag{57}$$

By testing (56) with $w = \overline{\Theta}^n - \Theta^n \in L^2(\Omega)$ and (57) with $\mathbf{v} = \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k) \in H(\mathrm{div};\Omega)$, adding the results and summing up from $n = 1$ to N we get

$$\begin{aligned}\sum_{n=1}^{N}\langle s(\Theta(t_n)) - s^n, \overline{\Theta}^n - \Theta^n\rangle + \sum_{n=1}^{N}\tau\langle \overline{\mathbf{q}}^n - \mathbf{q}^n, \sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle \\ - \sum_{n=1}^{N}\langle \overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle = \sum_{n=1}^{N}\langle \overline{\mathbf{f}_1(s)}^n - \mathbf{f}_1(s^n), \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle.\end{aligned} \tag{58}$$

We estimate separately each term in (58), which are denoted by T_1, T_2, T_3 and T_4. For T_1 we proceed as in the proof of Lemma 3.3, pg. 296 in [44], as the term here is identical to the one there and obtain

$$T_1 = \frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle s(\Theta)-s^n, \Theta-\Theta^n\rangle\, dt + \frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle s(\Theta(t_n))-s(\Theta), \Theta-\Theta^n\rangle\, dt \tag{59}$$

with the first term above being positive to remain on the left hand side of (46). Using the regularity of the solutions (both continuous and semi-discrete) and the stability estimates in Proposition 3.1 one can follow the steps in estimating T_{11} in [44] to obtain for the second term above

$$|\frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle s(\Theta(t_n))-s(\Theta), \Theta-\Theta^n\rangle| \leq C, \tag{60}$$

with C not depending on the discretization parameters. Moreover, if the data is such that both phases are present at any time and everywhere in the system, the estimate in (60) can be improved to $C\tau$, as discussed in Remark 3.3 below. Such an estimate would be optimal.

For the second term in (58) one uses the algebraic identity (29) to obtain

$$T_2 = \frac{\tau}{2}\|\sum_{n=1}^{N}(\overline{\mathbf{q}}^n-\mathbf{q}^n)\|^2 + \sum_{n=1}^{N}\frac{\tau}{2}\|\overline{\mathbf{q}}^n-\mathbf{q}^n\|^2. \tag{61}$$

The two terms above will remain on the left hand side of (46). We proceed by estimating T_3 in (58). By the Young inequality there holds

$$\begin{aligned} |T_3| &= |\sum_{n=1}^{N}\langle\overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k-\mathbf{q}^k)\rangle| \\ &\leq \frac{\delta}{2}\sum_{n=1}^{N}\|\overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n\|^2 + \frac{\tau^2}{2\delta}\sum_{n=1}^{N}\|\sum_{k=1}^{n}(\overline{\mathbf{q}}^k-\mathbf{q}^k)\|^2 \end{aligned} \tag{62}$$

Observe that in the above, the second term on the right involves two sums. The terms in the second sum are similar to the first term on the right in (61) and therefore one can deal with it by using the discrete Gronwall lemma 2.3. For the first term, denoted T_{31}, one uses (A3), (A4) and (47) to obtain

$$\begin{aligned} |T_{31}| &= \frac{\delta}{2}\sum_{n=1}^{N}\int_\Omega\left(\frac{1}{\tau}\int_{t_{n-1}}^{t_n} f_w(s)\mathbf{u}-f_w(s^n)\mathbf{u}^n\, dt\right)^2 dx \\ &\leq \frac{\delta}{\tau^2}\sum_{n=1}^{N}\int_\Omega\left(\int_{t_{n-1}}^{t_n}(f_w(s)-f_w(s^n))\mathbf{u}\, dt\right)^2 dx + \frac{\delta}{\tau^2}\sum_{n=1}^{N}\int_\Omega f_w^2(s^n)\left(\int_{t_{n-1}}^{t_n}(\mathbf{u}-\mathbf{u}^n)\, dt\right)^2 dx \\ &\leq \frac{C\delta}{\tau}\int_{t_{n-1}}^{t_n}\langle s(\Theta(t))-s(\Theta^n), \Theta(t)-\Theta^n\rangle dt, \end{aligned} \tag{63}$$

for all $\delta > 0$ and with a constant C not depending on the discretization parameters. In the same manner one can bound the last term in (58). Using again Young's inequality and (A5), one gets

$$|T_4| \leq \frac{C\delta'}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle s(\Theta(t))-s(\Theta^n), \Theta(t)-\Theta^n\rangle dt + \frac{\tau^2}{2\delta'}\sum_{n=1}^{N}\|\sum_{k=1}^{n}(\overline{\mathbf{q}}^k-\mathbf{q}^k)\|^2 \tag{64}$$

for all $\delta' > 0$ and with a constant C not depending on the discretization parameters. Putting now together (58) - (64), choosing δ and δ' properly and applying the discrete Gronwall lemma 2.3 gives (46).

Finally, to prove (48) one follows the step in Lemma 3.9, pg. 300 in [44]. By Lemma 2.1, there exists a function $\mathbf{v} \in H(\mathrm{div};\Omega)$ which satisfies $\nabla \cdot \mathbf{v} = \sum_{n=1}^{N}(\overline{\Theta}^n - \Theta^n)$ and $\|\mathbf{v}\| \leq C_\Omega \|\sum_{n=1}^{N}(\overline{\Theta}^n - \Theta^n)\|$. We use this as test function in (57), summed up from $n = 1$ to N. Now (48) follows from (46). □

Remark 3.1. *We point out that* (46) *and* (47) *imply immediately that there holds*

$$\sum_{n=1}^{N} \tau \|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 + \tau \|\overline{p}^n - p^n\|^2 \leq C\tau.$$

We proceed by deriving error estimates for the fully discrete approximations. We need first to introduce the following projectors (see [7])

$$P_h : L^2(\Omega) \to W_h, \qquad \langle P_h w - w, w_h \rangle = 0, \tag{65}$$

and

$$\Pi_h : H(\mathrm{div};\Omega) \cap [L^s(\Omega)]^d \to V_h, \qquad \langle \nabla \cdot (\Pi_h \mathbf{v} - \mathbf{v}), w_h \rangle = 0, \tag{66}$$

for all $w \in L^2(\Omega)$, $\mathbf{v} \in H(\mathrm{div};\Omega) \cap [L^s(\Omega)]^d$ and $w_h \in W_h$, where $s > 2$ is fixed arbitrary. For these operators we have

$$\|w - P_h w\| \leq Ch\|w\|_1, \qquad \text{respectively} \qquad \|\mathbf{v} - \Pi_h \mathbf{v}\| \leq Ch\|\mathbf{v}\|_1 \tag{67}$$

for any $w \in H^1(\Omega)$ and $\mathbf{v} \in (H^1(\Omega))^d$.

Remark 3.2. *Observe that instead of* $H(\mathrm{div};\Omega)$*, in* (66) *we consider* $\mathbf{v} \in H(\mathrm{div};\Omega) \cap [L^s(\Omega)]^d$*, whereas the estimates in* (67) *hold for* $\mathbf{v} \in H^1(\Omega)$*. In fact, one can follow e.g. the procedure in [48], p. 237 to construct a projector with similar properties like* Π_h*, but now defined on* $H(\mathrm{div};\Omega)$*. However, it is not clear whether the estimates in* (67) *remain valid and what the impact is for the convergence order of the scheme as estimated in Section 3.*

The next proposition quantifies the error between the semi-discrete solution and the fully discrete one. Recall the notations $s^k = s(\Theta^k)$ and $s_h^k = s(\Theta_h^k)$, $k \in \mathbb{N}$.

Proposition 3.3. *Let* $n \in \mathbb{N}, n \geq 1$ *and let* $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$ *be the solution of* P^n*, and* $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n)$ *be the solution of* P_h^n*. Assuming (A1)-(A6) and that the time step* τ *is small enough, there holds*

$$\begin{aligned} &\|\mathbf{u}^n - \mathbf{u}_h^n\|_{div}^2 + \|p^n - p_h^n\|^2 \\ &\quad \leq C(\|\mathbf{u}^n - \Pi_h \mathbf{u}^n\|^2 + \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle + \|p^n - P_h p^n\|^2) \\ &\quad \leq Ch^2 \tau^{\frac{\alpha-3}{\alpha+1}} \end{aligned} \tag{68}$$

and

$$\begin{aligned} &\sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle + \sum_{n=1}^{N} \|s^n - s_h^n\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}} + \tau \|\sum_{n=1}^{N}(\mathbf{q}^n - \mathbf{q}_h^n)\|^2 + \tau \|\sum_{n=1}^{N}(\Theta^n - \Theta_h^n)\|^2 \\ &\quad \leq C \sum_{n=1}^{N} (\|\mathbf{q}^n - \Pi_h \mathbf{q}^n\|^2 + \|\Theta^n - P_h \Theta^n\|^2 + \|\mathbf{u}^n - \Pi_h \mathbf{u}^n\|^2 + \|p^n - P_h p^n\|^2), \\ &\quad \leq Ch^2 \tau^{\frac{\alpha-3}{\alpha+1}} \end{aligned} \tag{69}$$

with the constants C *above not depending on the discretization parameters.*

Proof. The proof of (68) is following the lines of [16], where a MFEM was applied for the discretization of the pressure equation, but the Galerkin FEM for the saturation equation.

By subtracting (23) and (24) from (18) and (19) and using the properties of the projectors one gets

$$\begin{aligned} \langle \nabla \cdot (\Pi_h \mathbf{u}^n - \mathbf{u}_h^n), w_h \rangle &= \langle f_2(s^n) - f_2(s_h^n), w_h \rangle, & (70) \\ \langle a(s^n)\mathbf{u}^n - a(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h \rangle - \langle P_h p^n - p_h^n, \nabla \cdot \mathbf{v}_h \rangle &= \langle \mathbf{f}_3(s_h^n) - \mathbf{f}_3(s^n), \mathbf{v}_h \rangle, & (71) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. Testing (70) above with $w_h = P_h p^n - p_h^n \in W_h$ and (71) with $\mathbf{v}_h = \Pi_h \mathbf{u}^n - \mathbf{u}_h^n \in V_h$ and adding the results we obtain

$$\langle a(s^n)\mathbf{u}^n - a(s_h^n)\mathbf{u}_h^n, \Pi_h \mathbf{u}^n - \mathbf{u}_h^n \rangle = \langle f_2(s^n) - f_2(s_h^n), P_h p^n - p_h^n \rangle + \langle \mathbf{f}_3(s_h^n) - \mathbf{f}_3(s^n), \Pi_h \mathbf{u}^n - \mathbf{u}_h^n \rangle.$$

This further implies, by using (A2), (A3), (A4) and Young's inequality

$$\|\mathbf{u}^n - \mathbf{u}_h^n\|^2 \leq C\|\mathbf{u}^n - \Pi_h \mathbf{u}^n\|^2 + \frac{C}{\delta_1}\langle s(\Theta^n) - s(\Theta_h^n), \Theta^n - \Theta_h^n \rangle + \delta_1 \|P_h p^n - p_h^n\|^2, \quad (72)$$

for all $\delta_1 > 0$. Using Lemma 2.2, there exits a $\mathbf{v}_h \in V_h$ such that $\nabla \cdot \mathbf{v}_h = \Pi_h p - p_h^n$ and $\|\mathbf{v}_h\| \leq C\|P_h p^n - p_h^n\|$. Taking this $\mathbf{v}_h$ as test function in (71) one gets

$$\|P_h p^n - p_h^n\|^2 = \langle a(s^n)\mathbf{u}^n - a(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h \rangle + \langle \mathbf{f}_3(s^n) - \mathbf{f}_3(s_h^n), \mathbf{v}_h \rangle.$$

Using (A2)-(A4) and Young's inequality this further implies

$$\|P_h p^n - p_h^n\|^2 \leq C\langle s(\Theta^n) - s(\Theta_h^n), \Theta^n - \Theta_h^n \rangle + C\|\mathbf{u}^n - \mathbf{u}_h^n\|^2. \quad (73)$$

From (72) and (73) we immediately get by choosing δ_1 properly

$$\|\mathbf{u}^n - \mathbf{u}_h^n\|^2 + \|p^n - p_h^n\|^2 \leq C\langle s(\Theta^n) - s(\Theta_h^n), \Theta^n - \Theta_h^n \rangle. \quad (74)$$

Testing now (70) with $w_h = \Pi_h \mathbf{u}^n - \mathbf{u}_h^n \in W_h$, using (A3) and Young's inequality one gets

$$\|\nabla \cdot (\Pi_h \mathbf{u}^n - \mathbf{u}_h^n)\|^2 \leq C\langle s(\Theta^n) - s(\Theta_h^n), \Theta^n - \Theta_h^n \rangle.$$

This, together with (74), (A6) and the triangle inequality gives (68).

We give now a proof of (69). By subtracting (21) and (22) from (16) and (17), summing up from $k = 1$ to n and using the properties of the projectors, one gets

$$\begin{aligned} \langle s^n - s_h^n, w_h \rangle + \tau \sum_{k=1}^{n} \langle \nabla \cdot (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k), w_h \rangle &= 0, & (75) \\ \langle \mathbf{q}^n - \mathbf{q}_h^n, \mathbf{v}_h \rangle - \langle P_h \Theta^n - \Theta_h^n, \nabla \cdot \mathbf{v}_h \rangle - \langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h \rangle &= \langle \mathbf{f}_1(s^n) - \mathbf{f}_1(s_h^n), \mathbf{v}_h \rangle & (76) \end{aligned}$$

for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$. Taking $w_h = P_h \Theta^n - \Theta_h^n \in W_h$ and $\mathbf{v}_h = \tau \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \in V_h$ in (75) and (76), respectively, adding the results and summing up from $n = 1$ to N we obtain

$$\begin{aligned} &\sum_{n=1}^{N} \langle s^n - s_h^n, P_h \Theta^n - \Theta_h^n \rangle + \tau \sum_{n=1}^{N} \langle \mathbf{q}^n - \mathbf{q}_h^n, \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \rangle \\ &\quad - \sum_{n=1}^{N} \langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \tau \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \rangle = \sum_{n=1}^{N} \langle \mathbf{f}_1(s^n) - \mathbf{f}_1(s_h^n), \tau \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \rangle. \end{aligned} \quad (77)$$

Denoting the terms above by $\hat{T}_1$, $\hat{T}_2$, $\hat{T}_3$ and $\hat{T}_4$, we proceed by estimating them separately. For $\hat{T}_1$ there holds

$$\hat{T}_1 = \sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle + \sum_{n=1}^{N} \langle s^n - s_h^n, P_h\Theta^n - \Theta^n \rangle \tag{78}$$

with the first part above being positive due to the monotonicity of $s(\cdot)$. It even holds, due to (A1)

$$\sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle \geq \sum_{n=1}^{N} \frac{1}{L_s^{\frac{1}{\alpha}}} \|s^n - s_h^n\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}. \tag{79}$$

By Young's inequality, for the second term in (78) one gets

$$\sum_{n=1}^{N} \langle s^n - s_h^n, P_h\Theta^n - \Theta^n \rangle \leq \frac{\delta_1^{1+\frac{1}{\alpha}}}{1+\frac{1}{\alpha}} \sum_{n=1}^{N} \|s^n - s_h^n\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}} + \frac{1}{(1+\alpha)\delta_1^{1+\alpha}} \sum_{n=1}^{N} \|P_h\Theta^n - \Theta^n\|_{1+\alpha}^{1+\alpha}, \tag{80}$$

for all $\delta_1 > 0$. Using the algebraic identity (29), for $\hat{T}_2$ there holds

$$\begin{aligned} \hat{T}_2 &= \tau \sum_{n=1}^{N} \langle \mathbf{q}^n - \Pi_h\mathbf{q}^n, \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \rangle + \tau \sum_{n=1}^{N} \langle \Pi_h\mathbf{q}^n - \mathbf{q}_h^n, \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \rangle \\ &= \hat{T}_{21} + \frac{\tau}{2} \| \sum_{n=1}^{N} (\Pi_h\mathbf{q}^n - \mathbf{q}_h^n) \|^2 + \frac{\tau}{2} \sum_{n=1}^{N} \|\Pi_h\mathbf{q}^n - \mathbf{q}_h^n\|^2. \end{aligned} \tag{81}$$

Further, we use Young's inequality to estimate $\hat{T}_{21}$. We have

$$|\hat{T}_{21}| \leq \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{q}^n - \Pi_h\mathbf{q}^n\|^2 + \frac{\tau^2}{2} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \|^2. \tag{82}$$

In estimating $\hat{T}_3$ we use (A2)-(A4), Young's inequality and (68). There holds

$$\begin{aligned} |\hat{T}_3| &= |\sum_{n=1}^{N} \langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \tau \sum_{k=1}^{n} \Pi_h\mathbf{q}^k - \mathbf{q}_h^k \rangle| \\ &\leq \frac{\delta_2}{2} \sum_{n=1}^{N} \|f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n\|^2 + \frac{\tau^2}{2\delta_2} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \|^2 \\ &\leq C\delta_2 \sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle + C \sum_{n=1}^{N} \|\mathbf{u}^n - \Pi_h\mathbf{u}^n\|^2 + C \sum_{n=1}^{N} \|p^n - P_h p^n\|^2 \\ &\quad + \frac{\tau^2}{2\delta_2} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \|^2 \end{aligned} \tag{83}$$

for all $\delta_2 > 0$. In a similar manner, by using (A3) we can also bound the last term $\hat{T}_4$. There holds for all $\delta_3 > 0$

$$|\hat{T}_4| \leq C\delta_3 \sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle + \frac{\tau^2}{\delta_3} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \|^2. \tag{84}$$

Finally, putting together (77) - (84), choosing $\delta_1 - \delta_3$ properly, and using the discrete Gronwall lemma we obtain the result (69) (except the bound for $\tau\|\sum_{n=1}^{N}(\Theta^n - \Theta_h^n)\|^2$ which is but immediately by using Lemma 2.2, (76), (A3) and Young's inequality). The explicit order of convergence is obtained by using the properties of the projectors and (A6). □

The main result below is a straightforward consequence of Proposition 3.2 and Proposition 3.3, the properties of the projectors, the stability estimates in Proposition 3.1 and the regularity of the solution.

Theorem 3.1. *Let $(\Theta, \mathbf{q}, p, \mathbf{u})$ be the solution of Problem P and let $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n)$ be the solution of P_h^n, $n \in \{1, \dots, N\}$. Assuming (A1)-(A6) and that the time step τ is small enough, there holds*

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s(\Theta(t)) - s(\Theta_h^n)\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}dt + \|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(\mathbf{q}-\mathbf{q}_h^n)\,dt\|^2 + \|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(\Theta-\Theta_h^n)\,dt\|^2 \leq C(\tau + h^2\tau^{\frac{2(\alpha-1)}{1+\alpha}}), \tag{85}$$

$$\sum_{n=1}^{N}\tau\|\overline{\mathbf{u}}^n - \mathbf{u}_h^n\|^2 + \sum_{n=1}^{N}\tau\|\overline{p}^n - p_h^n\|^2 \leq C(\tau + h^2\tau^{\frac{2(\alpha-1)}{1+\alpha}}), \tag{86}$$

where $\alpha \in (0, 1]$ denotes the Hölder exponent of the saturation and the constants $C > 0$ not depending on the discretization parameters.

Remark 3.3. *The error estimates presented above are optimal in space for the Lipschitz continuous case, i.e. $\alpha = 1$. Moreover, if additionally to Lipschitz continuity one assumes no-degeneracy (disappearance of phases is not allowed), one can prove optimal error estimates in space and time (similar to Corollary 3.6, pg. 299 in [44])*

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s(\Theta(t)) - s(\Theta_h^n)\|^2dt + \|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(\mathbf{q}-\mathbf{q}_h^n)\,dt\|^2 + \|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(\Theta-\Theta_h^n)\,dt\|^2 \leq (\tau^2 + h^2), \tag{87}$$

$$\sum_{n=1}^{N}\tau\|\overline{\mathbf{u}}^n - \mathbf{u}_h^n\|^2 + \sum_{n=1}^{N}\tau\|\overline{p}^n - p_h^n\|^2 \leq C(\tau^2 + h^2). \tag{88}$$

4 Linearization scheme

In this section we analyze the convergence of the (fully discrete) linearization scheme (25)–(28). We show that the scheme is robust and it converges linearly. The analysis covers the case of a Hölder continuous saturation. In this case, since the Jacobian matrix becomes singular, the Newton method can not be applied without performing a regularization of the problem. Such an approach is not needed for the L-scheme proposed here.

The L-scheme has been considered previously as an alternative to methods based on the Picard or the Newton iteration, in particular when dealing with degenerate parabolic models. In this sense we mention [51, 40, 29, 45] for the fast diffusion case, or [55] for the slow diffusion case, but in connection with a regularization step. To the best of our knowledge, the case when $s(\cdot)$ is only Hölder continuous has not been discussed in the literature so far. Here we show how to use the L-scheme for this case as well. We mention that the having a Hölder exponent below 1 is decreasing the convergence of the scheme.

As the scheme is used to solve the non-linear systems within one time step, throughout this section $n \in N, n \geq 1$ is fixed. Referring to Problem $P_h^{n,i}$ introduced in Section 2, we use

$$\begin{aligned} e_\Theta^{n,i} &= \Theta_h^{n,i} - \Theta_h^{n,i-1}, & e_{\mathbf{q}}^{n,i} &= \mathbf{q}_h^{n,i} - \mathbf{q}_h^{n,i-1}, \\ e_p^{n,i} &= p_h^{n,i} - p_h^{n,i-1}, & e_{\mathbf{u}}^{n,i} &= \mathbf{u}_h^{n,i} - \mathbf{u}_h^{n,i-1}, \\ e_s^{n,i} &= s_h^{n,i} - s_h^{n,i-1} := s(\Theta_h^{n,i}) - s(\Theta_h^{n,i-1}). \end{aligned}$$

which are the errors between two consecutive iterations i and $i-1$.

The convergence of the scheme (25) - (28) is obtained by guaranteeing that the errors e_Θ and $e_{\mathbf{q}}$ can be made arbitrarily small, completed by estimates for e_p and $e_{\mathbf{u}}$. Note that, whenever the scheme converges, the term involving the factor L will vanish and hence the limit is a solution of Problem P_h^n. As will follow from below, if s is Lipschitz continuous, the iteration is a contraction and the existence and uniqueness of a solution for the problem (21) - (24) follows immediately. For the linear problems (25) - (28), the existence and uniqueness are obtained in

Proposition 4.1. *For any $n \geq 1$ and $i \geq 1$, Problem $P_h^{n,i}$ has a unique solution.*

Proof. Note that since Problem $P_h^{n,i}$ is linear and finite dimensional, it is enough to prove the uniqueness of a solution. This immediately implies the existence of a solution.

Assuming that Problem $P_h^{n,i}$ has two solutions, $(\Theta_{h,k}^{n,i}, p_{h,k}^{n,i}, \mathbf{q}_{h,k}^{n,i}, \mathbf{u}_{h,k}^{n,i}) \in W_h \times W_h \times V_h \times V_h$ $(k = 1, 2)$, their difference satisfies

$$\begin{aligned} \langle L(\Theta_{h,1}^{n,i} - \Theta_{h,2}^{n,i}), w_h \rangle + \tau \langle \nabla \cdot (\mathbf{q}_{h,1}^{n,i} - \mathbf{q}_{h,2}^{n,i}), w_h \rangle &= 0, & (89) \\ \langle \mathbf{q}_{h,1}^{n,i} - \mathbf{q}_{h,2}^{n,i}, \mathbf{v}_h \rangle - \langle \Theta_{h,1}^{n,i} - \Theta_{h,2}^{n,i}, \nabla \cdot \mathbf{v}_h \rangle - \langle f_w(s_h^{n,i-1})(\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i}), \mathbf{v}_h \rangle &= 0, & (90) \\ \langle \nabla \cdot (\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i}), w_h \rangle &= 0, & (91) \\ \langle a(s_h^{n,i-1})(\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i}), \mathbf{v}_h \rangle - \langle p_{h,1}^{n,i} - p_{h,2}^{n,i}, \nabla \cdot \mathbf{v}_h \rangle &= 0 & (92) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. Testing (91) with $w_h = \nabla \cdot (\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i}) \in W_h$ gives $\nabla \cdot (\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i}) = 0$. Then, using the assumption (A2) and testing (92) with $\mathbf{u}_{h,1}^{n,i} - \mathbf{u}_{h,2}^{n,i} \in V_h$ immediately gives that $\mathbf{u}_{h,1}^{n,i} = \mathbf{u}_{h,2}^{n,i}$. By Lemma 2.2, it follows also that $p_{h,1}^{n,i} = p_{h,2}^{n,i}$. Testing now (89) with $w_h = \Theta_{h,1}^{n,i} - \Theta_{h,2}^{n,i} \in W_h$ and (90) with $\mathbf{v}_h = \tau(\mathbf{q}_{h,1}^{n,i} - \mathbf{q}_{h,2}^{n,i}) \in V_h$, and then adding the results leads to

$$L\|\Theta_{h,1}^{n,i} - \Theta_{h,2}^{n,i}\|^2 + \tau\|\mathbf{q}_{h,1}^{n,i} - \mathbf{q}_{h,2}^{n,i}\|^2 = 0,$$

since $\mathbf{u}_{h,1}^{n,i} = \mathbf{u}_{h,2}^{n,i}$. This implies the uniqueness of the solution. □

The main result, the convergence of the scheme (25) – (28), is stated in the following

Theorem 4.1. *Let $n \in \mathbb{N}$ be fixed, and $\Theta_h^{n-1}, p_h^{n-1} \in W_h$ and $\mathbf{q}_h^{n-1}, \mathbf{u}_h^{n-1} \in V_h$ be given, solving P_h^{n-1}. Further, let $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ solving $P_h^{n,i}$ for any $i \geq 1, i \in \mathbb{N}$. Assuming (A2)–(A5), the following hold for all $i \geq 2$, and for i-independent constants that will be specified below.*

• *If s is Lipschitz continuous ($\alpha = 1$), then*

$$\|e_\Theta^{n,i}\|^2 + \frac{R(L,\tau)(1-2\tau C)}{L}\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle + (\frac{1}{L_s} - \frac{1}{L})\|e_s^{n,i-1}\|^2 + \frac{\tau R(L,\tau)}{4L}\|e_{\mathbf{q}}^{n,i}\|^2 \leq R(L,\tau)\|e_\Theta^{n,i-1}\|^2. \quad (93)$$

• *If s is only Hölder continuous ($\alpha \in (0,1)$)*

$$\|e_\Theta^{n,i}\|^2+\frac{R(L,\tau)(1-2\tau C)}{L}\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{\tau R(L,\tau)}{4L}\|e_{\mathbf{q}}^{n,i}\|^2 \leq R(L,\tau)\|e_\Theta^{n,i-1}\|^2+C_1R(L,\tau)L^{2/(\alpha-1)}. \tag{94}$$

The positive constants C and C_1 do not depend on the discretization parameters or the iteration index. Further, with C_{Ω_d} introduced in Lemma 2.2,

$$R(L,\tau)=\frac{L}{L+\frac{\tau}{32C_{\Omega_d}^2}}<1. \tag{95}$$

Using this, and with τ sufficiently small, the convergence of the scheme (25) -(28) *results if L is chosen as follows:*

• *In the Lipschitz continuous case ($\alpha = 1$), taking*

$$L \geq L_s \tag{96}$$

ensures the linear convergence of the iterative linearization scheme.

• *In the Hölder continuous case ($\alpha \in (0,1)$), taking L large enough guarantees that the error can be made sufficiently small ('numerical' convergence).*

The convergence concept and the condition on L stated above for the Hölder continuous case are made explicit at the end of the proof of Theorem 4.1 and in Remark 4.3 below. To prove Theorem 4.1 we use Lemma 4.1 and Lemma 4.2 below.

Lemma 4.1. *Let $n \in \mathbb{N}$ be fixed, and Θ_h^{n-1} be given. Let $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ solving $P_h^{n,i}$ for any $i \geq 1, i \in \mathbb{N}$. Assuming (A2)–(A4), there holds for $i \geq 2$*

$$\|e_{\mathbf{u}}^{n,i}\|_{div}^2+\|e_p^{n,i}\|^2 \leq C\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle \tag{97}$$

where the constant $C > 0$ does not depend on the discretization parameters or the iteration index.

Proof. We prove first that $\|e_{\mathbf{u}}^{n,i}\|^2 \leq C\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle$, $i \geq 2$. Subtracting (27) and (28) for i and $i-1$ respectively, one obtains

$$\langle \nabla\cdot e_{\mathbf{u}}^{n,i}, w_h\rangle = \langle f_2(s_h^{n,i-1})-f_2(s_h^{n,i-2}), w_h\rangle, \tag{98}$$

$$\langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i}-a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1},\mathbf{v}_h\rangle-\langle e_p^{n,i},\nabla\cdot\mathbf{v}_h\rangle = \langle \mathbf{f}_3(s_h^{n,i-2})-\mathbf{f}_3(s_h^{n,i-1}),\mathbf{v}_h\rangle \tag{99}$$

for all $w_h \in W_h$, $\mathbf{v}_h \in V_h$. Taking now $w_h = e_p^{n,i} \in W_h$ in (98) and $\mathbf{v}_h = e_{\mathbf{u}}^{n,i} \in V_h$ in (99), and adding the results one gets

$$\langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i}-a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1},e_{\mathbf{u}}^{n,i}\rangle = \langle f_2(s_h^{n,i-1})-f_2(s_h^{n,i-2}),e_p^{n,i}\rangle+\langle \mathbf{f}_3(s_h^{n,i-2})-\mathbf{f}_3(s_h^{n,i-1}),e_{\mathbf{u}}^{n,i}\rangle.$$

Using (A2) - (A4), together with Young's inequality, the above implies that for any $\epsilon_1 > 0$ there holds

$$\frac{a_\star}{2}\|e_{\mathbf{u}}^{n,i}\|^2 \leq \left(\frac{M_{\mathbf{u}}^2L_a+L_{\mathbf{f}_3}}{a_\star}+\frac{L_{f_2}}{2\epsilon_1}\right)\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{\epsilon_1}{2}\|e_p^{n,i}\|^2. \tag{100}$$

Recalling Lemma 2.2, a $\mathbf{v}_h \in V_h$ exists such that $\nabla\cdot\mathbf{v}_h = e_p^{n,i}$ and $\|\mathbf{v}_h\| \leq C_{\Omega,d}\|e_p^{n,i}\|$. Taking this $\mathbf{v}_h$ as test function in (99), using (A2)-(A4) and Young's inequality gives

$$\begin{aligned}\|e_p^{n,i}\|^2 &= \langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i}-a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1},\mathbf{v}_h\rangle+\langle \mathbf{f}_3(s_h^{n,i-2})-\mathbf{f}_3(s_h^{n,i-1}),\mathbf{v}_h\rangle\\ &\leq C_{\Omega,d}^2(a^\star)^2\|e_{\mathbf{u}}^{n,i}\|^2+C_{\Omega,d}^2(M_{\mathbf{u}}^2L_a+L_{\mathbf{f}_3})\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{3}{4}\|e_p^{n,i}\|^2.\end{aligned}$$

This allows estimating $e_p^{n,i}$ in terms of $e_s^{n,i}$, $e_\Theta^{n,i-1}$ and $e_{\mathbf{u}}^{n,i}$,

$$\|e_p^{n,i}\|^2 \leq 4C_{\Omega,d}^2(a^\star)^2\|e_{\mathbf{u}}^{n,i}\|^2 + 4C_{\Omega,d}^2(M_{\mathbf{u}}^2 L_a + L_{\mathbf{f}_3})\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle. \tag{101}$$

Choosing now ϵ_1 properly, from (100) and (101) one obtains

$$\|e_{\mathbf{u}}^{n,i}\|^2 \leq C\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle. \tag{102}$$

Further, $\|e_p^{n,i}\|^2 \leq C\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle$ follows immediately from (101) and (102). Finally, $\|\nabla \cdot e_{\mathbf{u}}^{n,i}\|^2 \leq C\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle$ is a straightforward consequence of (98) and (A3). □

Remark 4.1. *The constants C in Lemma 4.1 and Lemma 4.2, as well as in Theorem 4.1 can be calculated exactly as function of the data. To simplify the presentation, we avoid doing it here, but refer to [46] where these constants are determined exactly in the Lipschitz continuous case.*

Lemma 4.2. *Let $n \in \mathbb{N}$ be fixed, and Θ_h^{n-1} be given. Let $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ solving $P_h^{n,i}$ for any $i \geq 1, i \in \mathbb{N}$. Assuming (A2)–(A4), there holds for $i \geq 2$*

$$\|e_{\mathbf{q}}^{n,i}\|^2 \geq \frac{1}{8C_\Omega^2}\|e_\Theta^{n,i}\|^2 - C\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle, \tag{103}$$

where the constant $C > 0$ is not depending on the discretization parameters or the iteration index.

Proof. Subtracting (25) and (26) for i and $i-1$ respectively gives

$$\langle L(e_\Theta^{n,i} - e_\Theta^{n,i-1}) + e_s^{n,i-1}, w_h\rangle + \tau\langle \nabla \cdot e_{\mathbf{q}}^{n,i}, w_h\rangle = 0, \tag{104}$$

$$\langle e_{\mathbf{q}}^{n,i}, \mathbf{v}_h\rangle - \langle e_\Theta^{n,i}, \nabla \cdot \mathbf{v}_h\rangle - \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h\rangle = \langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), \mathbf{v}_h\rangle. \tag{105}$$

By Lemma 2.2, there exists a $\mathbf{v}_h \in V_h$ such that $\nabla \cdot \mathbf{v}_h = e_\Theta^{n,i}$ and $\|\mathbf{v}_h\| \leq C_{\Omega,d}\|e_\Theta^{n,i}\|$. Taking this $\mathbf{v}_h$ as test function in (105) and using (A3) and (A4) gives

$$\begin{aligned}
\|e_\Theta^{n,i}\|^2 &= \langle e_{\mathbf{q}}^{n,i}, \mathbf{v}_h\rangle + \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h\rangle + \langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), \mathbf{v}_h\rangle \\
&\leq C_{\Omega,d}\|e_{\mathbf{q}}^{n,i}\|\|e_\Theta^{n,i}\| + C_{\Omega,d}M_{\mathbf{u}}\|f_w(s_h^{n,i-1}) - f_w(s_h^{n,i-2})\|\|e_\Theta^{n,i}\| \\
&\quad + C_{\Omega,d}M_{f_w}\|e_{\mathbf{u}}^{n,i}\|\|e_\Theta^{n,i}\| + C_{\Omega,d}\|\mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2})\|\|e_\Theta^{n,i}\|.
\end{aligned}$$

From the above, (103) follows by using Young's inequality, the estimate (97) and (A3)-(A4). □

We can now come back and prove the Theorem 4.1.

Proof. We take $w_h = e_\Theta^{n,i} \in W_h$ in (104) and $\mathbf{v}_h = \tau e_{\mathbf{q}}^{n,i} \in V_h$ in (105), and add the results to obtain

$$\begin{aligned}
&L\langle (e_\Theta^{n,i} - e_\Theta^{n,i-1}) + e_s^{n,i-1}, e_\Theta^{n,i}\rangle + \tau\|e_{\mathbf{q}}^{n,i}\|^2 = \\
&\qquad \tau\langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-2}, e_{\mathbf{q}}^{n,i}\rangle + \tau\langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), e_{\mathbf{q}}^{n,i}\rangle.
\end{aligned}$$

Using the identity $2\langle x, x-y\rangle = \|x\|^2 + \|x-y\|^2 - \|y\|^2$, we rewrite the above as

$$\begin{aligned}
&\frac{L}{2}\|e_\Theta^{n,i}\|^2 + \frac{L}{2}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle + \tau\|e_{\mathbf{q}}^{n,i}\|^2 = \frac{L}{2}\|e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i-1} - e_\Theta^{n,i}\rangle \\
&\quad + \tau\langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-2}, e_{\mathbf{q}}^{n,i}\rangle + \tau\langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), e_{\mathbf{q}}^{n,i}\rangle.
\end{aligned} \tag{106}$$

The last two terms in the equality above can be estimated by using the assumptions (A3)-(A4), the estimate (97) and Young's inequality. There holds

$$\tau\langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i}-f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-2},e_{\mathbf{q}}^{n,i}\rangle+\tau\langle\mathbf{f}_1(s_h^{n,i-1})-\mathbf{f}_1(s_h^{n,i-2}),e_{\mathbf{q}}^{n,i}\rangle\leq C\tau\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{3\tau}{4}\|e_{\mathbf{q}}^{n,i}\|^2.$$

Using this estimate in (106) and Young's inequality for the term $\langle e_s^{n,i-1},e_\Theta^{n,i-1}-e_\Theta^{n,i}\rangle$ we further get

$$\frac{L}{2}\|e_\Theta^{n,i}\|^2+(1-C\tau)\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{\tau}{4}\|e_{\mathbf{q}}^{n,i}\|^2\leq\frac{L}{2}\|e_\Theta^{n,i-1}\|^2+\frac{1}{2L}\|e_s^{n,i-1}\|^2. \tag{107}$$

Due to (A1), there holds for any $\alpha\in(0,1]$

$$\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle\geq\frac{1}{L_s^{\frac{1}{\alpha}}}\|e_s^{n,i-1}\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}. \tag{108}$$

Using (108) and the estimate (103) in (107), after multiplying with 2 and doubling C in (107) we obtain

$$(L+\frac{\tau}{32C_\Omega^2})\|e_\Theta^{n,i}\|^2+(1-C\tau)\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{1}{L_s^{\frac{1}{\alpha}}}\|e_s^{n,i-1}\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}+\frac{\tau}{4}\|e_{\mathbf{q}}^{n,i}\|^2\leq L\|e_\Theta^{n,i-1}\|^2+\frac{1}{L}\|e_s^{n,i-1}\|^2. \tag{109}$$

If $\alpha=1$, (93) follows immediately from the above. In this case, taking $L\leq L_s$ and the time step τ small enough, since $\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle\geq 0$ one uses the Banach fixed point theorem to obtain the convergence for $\Theta^{n,i}$, and immediately for $\mathbf{q}^{n,i}$. The convergence for $\mathbf{u}^{n,i}$ and $p^{n,i}$ then follows then from the estimates in Lemma 4.1.

In the case $\alpha\in(0,1)$, with $\sigma(\Omega)$ denoting the area of Ω, for any $f\in L^2(\Omega)$ it holds

$$\|f\|^2\leq\sigma(\Omega)^{\frac{1-\alpha}{1+\alpha}}\|f\|_{1+\frac{1}{\alpha}}^2.$$

Further, letting $C(\alpha)=\dfrac{1-\alpha}{1+\alpha}\left(\dfrac{4\alpha}{1+\alpha}\right)^{\frac{2\alpha}{1-\alpha}}$, Young's inequality (30) with $p=\dfrac{1+\alpha}{2\alpha}$ and $q=\dfrac{1+\alpha}{1-\alpha}$ gives

$$A_1^2A_2\leq\frac{1}{2L_s^{\frac{1}{\alpha}}}A_1^{1+\frac{1}{\alpha}}+C(\alpha)A_2^{\frac{1+\alpha}{1-\alpha}}L_s^{\frac{2}{1-\alpha}},\qquad\text{for all } A_1,A_2\geq 0.$$

Using the above, the last term in (109) is estimated as

$$\begin{aligned}\frac{1}{L}\|e_s^{n,i-1}\|^2&\leq\frac{\sigma(\Omega)^{\frac{1-\alpha}{1+\alpha}}}{L}\|e_s^{n,i-1}\|_{1+\frac{1}{\alpha}}^2\\&\leq\frac{1}{2L_s^{\frac{1}{\alpha}}}\|e_s^{n,i-1}\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}+C(\alpha)\sigma(\Omega)L_s^{\frac{2}{1-\alpha}}L^{\frac{1+\alpha}{\alpha-1}}.\end{aligned} \tag{110}$$

With with $C_1=C(\alpha)\sigma(\Omega)L_s^{\frac{2}{1-\alpha}}$, using (110) in (109) leads to

$$(L+\frac{\tau}{32C_\Omega^2})\|e_\Theta^{n,i}\|^2+(1-2C\tau)\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\frac{1}{2L_s^{\frac{1}{\alpha}}}\|e_s^{n,i-1}\|_{1+\frac{1}{\alpha}}^{1+\frac{1}{\alpha}}+\frac{\tau}{4}\|e_{\mathbf{q}}^{n,i}\|^2\leq L\|e_\Theta^{n,i-1}\|^2+C_1L^{\frac{1+\alpha}{\alpha-1}}. \tag{111}$$

Dividing the above by $L+\dfrac{\tau}{32C_\Omega^2}$ gives (94).

Note that the presence of the term $C_1R(L,\tau)L^{2/(\alpha-1)}$ on the right hand side does not allow concluding from (94) that the linearization scheme is convergent. However, in practical situations this term is very small

and therefore do not affect the convergence, see e.g. the discussion in [46] for the case of Richards equations. Moreover, for the convergence of the scheme in the Hölder continuous case we proceed as follows.

Assuming the existence of a solution for the nonlinear, fully discrete problem (21) – (24), proceeding as above one obtains that

$$\begin{aligned}\|\Theta_h^{n,i}-\Theta_h^n\|^2+\frac{R(L,\tau)(1-\tau C)}{L}\langle s(\Theta_h^{n,i-1})-s(\Theta_h^n),\Theta_h^{n,i-1}-\Theta_h^n\rangle+\frac{\tau R(L,\tau)}{4L}\|\mathbf{q}_h^{n,i}-\mathbf{q}_h^n\|^2\\ \leq R(L,\tau)\|\Theta_h^{n,i-1}-\Theta_h^n\|^2+C_1R(L,\tau)L^{2/(\alpha-1)}.\end{aligned} \tag{112}$$

From this, one obtains

$$\|\Theta_h^{n,i}-\Theta_h^n\|^2\leq\big(R(L,\tau)\big)^i\|\Theta_h^{n,0}-\Theta_h^n\|^2+C_1\frac{R(L,\tau)}{1-R(L,\tau)}L^{2/(\alpha-1)}. \tag{113}$$

Observe that for $\alpha\in(0,1)$, the power of L in the last term above is negative. This means that with L sufficiently large, this term can be made small enough, namely smaller than a threshold value. Further, after a sufficiently large number of iterations, the first term on the right an be made arbitrarily small. This means that, from practical point of view, the proposed linearization scheme can be used to provide an approximation of the solution to Problem P_h^n that falls within the desired range of accuracy.

□

Remark 4.2. *Observe that the estimates* (93) *and* (94) *can only be used for proving the convergence if all terms on the left are positive. Since s is increasing, this provides a restriction on the time step,*

$$\tau\leq\frac{1}{C}. \tag{114}$$

Note that C depends only on the data, but not on the mesh size or the iteration index. C can be even determined exactly, see [46] for the Lipschitz continuous case. Therefore one can say that the restriction on τ is mild. It is superior to the one guaranteeing the stability of an explicit discretization in time, or to the typical conditions guaranteeing the convergence of the Newton method for degenerate parabolic problems (see e.g. [42]).

Remark 4.3. *As follows from* (113)*, in the Hölder continuous case the error at the i^{th} iteration is bounded by the sum of two terms. From practical point of view, for a specified tolerance TOL one can choose L large enough to guarantee that the last term in* (113) *is less than $TOL/2$. With this choice, one can iterate i times to reduce the first term on the right below $TOL/2$. In this way, $\|\Theta_h^{n,i}-\Theta_h^n\|^2<TOL$, and similar estimates can be obtained for the other solution components. This is why in Theorem 4.1 the convergence is called 'numerical'. We refer to [46], Remark 4.4. for a discussion related to this aspect for the Richards equation.*

Remark 4.4. *One can use the L-scheme* (25)–(28) *in combination with the Newton method. The goal is to combine the robustness of the L-scheme, which converges regardless of the starting point, with the quadratic convergence of Newton's method, which requires a starting point that is sufficiently close to the solution. Specifically, at each time step one can perform a few L-scheme iterations, followed by Newton iterations. In this way one enhances the robustness of the Newton method and relaxes the severe restriction on the time step, by which convergence is guaranteed. We refer to [29], where this strategy is studied for solving the Richards equation. In the same context, a similar idea is proposed in [28], but there the robustness of the Newton method is improved by a modified Picard method.*

Remark 4.5. *Another possible strategy when dealing with a Hölder continuous saturation function s is based on regularization. Specifically, with $\epsilon > 0$ being a small parameter, one can approximate $s(\cdot)$ by a Lipschitz continuous function $s_\epsilon(\cdot)$. Clearly, its Lipschitz constant L_{s_ϵ} cannot be uniform in ϵ, but s_ϵ can be chosen s.t. $L_{s_\epsilon} = \frac{1}{\epsilon}$). Then one can adapt the results for the Lipschitz case in Theorem 4.1 to prove the convergence of the linearization scheme. Observe that negative powers of ϵ will appear in* (109)*. To finish the proof in this case one needs to estimate the error between the solution of* (25)–(28) *and its regularized variant.*

5 Numerical results

In this section we present two numerical studies, one evidencing the convergence of the backward Euler/MFEM discretization analyzed here, and the other focussing on the convergence of the linearization scheme. In all calculations we have used the linearization scheme (25)–(28), and iterations are concluded whenever the L^2 error for Θ decreases below a certain threshold: 10^{-8}.

In the first example we test the convergence of the MFEM scheme and the influence of the Hölder/Lipschitz coefficient $\alpha \in (0, 1]$. Therefore we consider an academic problem defined in a two-dimensional domain $\Omega = (0, 1) \times (0, 1)$ and for $t \in (0, T]$, for which an explicit solution can be found. Specifically, we let the coefficient functions satisfy

$$s(\Theta) = \Theta^\alpha, \quad \lambda_w(s) = s^\alpha, \quad \lambda_o(s) = 1-s^\alpha, \quad \mathbf{f}_1 = 0, \quad f_2 = 2x(1-x)+2y(1-y), \text{ and } \mathbf{f}_3 = 0, \tag{115}$$

and add a source term

$$\begin{aligned} f(t,x,y) =& \alpha t^{\alpha-1}\big(x(1-x)y(1-y)\big)^\alpha + 2t\Big(x(1-x)+y(1-y)\Big) \\ &+ 2t\Big(x(1-x)+y(1-y)\Big)x(1-x)y(1-y) - tx^2(1-x)^2(1-2y)^2 - ty^2(1-y)^2(1-2x)^2 \end{aligned}$$

on the right of (7).

Observe that the functions λ_w and λ_o are defined in terms of the saturation s, which is in line with the way such models are written in practice. Clearly, in the present context s itself is a function of Θ. Also note that $s(\cdot)$ is Lipschitz continuous only when $\alpha = 1$. For $\alpha < 1$, $s(\cdot)$ is only Hölder continuous, $C^{0,\alpha}$. Finally, the particular choice of the functions f_i $(i = 1, 2, 3)$ and of the source f ensures that

$$p = x(1-x)y(1-y), \quad \Theta = tx(1-x)y(1-y), \tag{116}$$

are the first two components of the solution of the two-phase flow model, the fluxes being computed from (8) and (10). The initial and Dirichlet boundary conditions are matching the exact solution given above.

Knowing the exact solution in this case, and with the numerical solutions computed as mentioned below we calculate

$$\begin{aligned} &E_p = \textstyle\sum_{n=1}^N \tau \|p(t_n) - p_h^n\|^2, && E_\Theta = \textstyle\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \|\Theta(t) - \Theta_h^n\|^2 \, dt, \\ &E_s = \textstyle\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \|s(t) - s_h^n\|^2 \, dt, && E_{s\Theta} = \textstyle\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta_h^n), \Theta(t) - \Theta_h^n \rangle \, dt. \end{aligned}$$

We use such expressions to estimate the convergence order of the scheme. More precisely, the numerical solutions are computed on four rectangular and uniform meshes, and with a uniform time stepping. We compute the numerical solution on four meshes, $h_i = 1/2^{(i+1)}$ $(i = 1, \ldots, 4)$, and take the time step in

accordance with the estimate in Theorem 3.1, namely $\tau_i = h_i^{((2\alpha+2)/(3-\alpha))}$. With this, the convergence rate is estimated as

$$\text{conv rate}(i) = \frac{1}{2}\frac{\log E_z^{i+1} - \log E_z^i}{\log h_{i+1} - \log h_i},$$

where z stands for either p, Θ, s, or $s\Theta$. The factor $1/2$ in the above is due to the fact that E_z are, in fact, squared errors.

The results are presented in Tables 1–4. As follows from Table 1, the convergence is linear and thus optimal for the Lipschitz continuous case. Since $\alpha = 1$, the function $s(\cdot)$ is the identity, which explains why E_s, E_Θ and $E_{s\Theta}$ are identical. For the Hölder continuous case we performed experiments for $\alpha = 0.9, 0.8, 0.7$, the results being presented in Tables 2–4. Observe that the convergence rate decays with the exponent α, as expected.

Having seen the convergence order of the numerical scheme, we proceed by studying the behavior of the linearization scheme, as influenced by the Hölder exponent α. With the stopping criterion mentioned above, Figure 1 gives the number of iterations for the cases $\alpha = 1$, 0.8 and 0.6, at $T = 0.15$ and for the different meshes. The calucations were done with $L = 1$, 3 and 6, respectively. Observe that the number of iterations is relatively robust with respect to the mesh size, but depends on the Hölder exponent. This is again in line with the theoretical findings.

h	τ	E_p	conv rate	$E_{s\Theta}$	conv rate	E_Θ	conv rate	E_s	conv rate
$\frac{1}{8}$	$1.56E-2$	$4.48E-5$	–	$1.44E-5$	–	$1.44E-5$	–	$1.44E-5$	–
$\frac{1}{16}$	$3.91E-3$	$1.09E-5$	1.01	$3.61E-6$	1.00	$3.61E-6$	1.00	$3.61E-6$	1.00
$\frac{1}{32}$	$9.76E-4$	$2.72E-6$	1.00	$9.04E-7$	1.00	$9.04E-7$	1.00	$9.04E-7$	1.00
$\frac{1}{64}$	$2.44E-4$	$6.78E-7$	1.00	$2.26E-7$	1.00	$2.26E-7$	1.00	$2.26E-7$	1.00

Table 1: Convergence rates for the manufactured solution (116) for the Lipschitz continuous case, $\alpha = 1$.

h	τ	E_p	conv rate	$E_{s\Theta}$	conv rate	E_Θ	conv rate	E_s	conv rate
$\frac{1}{4}$	$2.32E-2$	$4.58E-5$	–	$2.14E-5$	–	1.49E-5	–	$3.10E-5$	–
$\frac{1}{8}$	$6.62E-3$	$1.09E-5$	1.03	$5.13E-6$	1.03	3.58E-6	1.03	$7.44E-6$	1.03
$\frac{1}{16}$	$1.89E-3$	$2.72E-6$	1.00	$1.33E-6$	0.97	9.32E-7	0.97	$1.92E-6$	0.97
$\frac{1}{32}$	$5.39E-4$	$6.78E-7$	1.00	$3.68E-7$	0.92	2.62E-7	0.91	$5.27E-7$	0.93

Table 2: Convergence rates for the manufactured solution (116) for the Hölder exponent $\alpha = 0.9$.

Next, we discuss a three dimensional example. We consider rectangular grids of different sizes and study the convergence of the linearization scheme. The computational domain is now the unit cube (in meters), $\Omega = (0m, 1m)^3$. We use the following constitutive relationships

$$k_{rw} = s^2,\ k_{ro} = (1-s)^2,\ s(\Theta) = \Theta^2,$$

h	τ	E_p	conv rate	$E_{s\Theta}$	conv rate	E_Θ	conv rate	E_s	conv rate
$\frac{1}{8}$	$3.32E-2$	$4.62E-5$	–	$3.09E-5$	–	1.51E-5	–	$6.63E-5$	–
$\frac{1}{16}$	$1.07E-2$	$1.10E-5$	1.03	$7.65E-6$	1.00	3.75E-6	1.00	$1.65E-5$	1.00
$\frac{1}{32}$	$3.44E-3$	$2.72E-6$	1.00	$2.19E-6$	0.90	1.11E-6	0.87	$4.59E-6$	0.92
$\frac{1}{64}$	$1.11E-3$	$6.78E-7$	1.00	$8.29E-7$	0.70	4.56E-7	0.92	$1.61E-6$	0.76

Table 3: Convergence rates for the manufactured solution (116) for the Hölder exponent $\alpha = 0.8$.

h	τ	E_p	conv rate	$E_{s\Theta}$	conv rate	E_Θ	conv rate	E_s	conv rate
$\frac{1}{8}$	$4.62E-2$	$4.56E-5$	–	$4.25E-5$	–	$9.14E-5$	–	$1.39E-4$	–
$\frac{1}{16}$	$1.66E-2$	$1.11E-5$	1.02	$1.23E-5$	0.89	$1.66E-5$	1.23	$4.00E-5$	0.90
$\frac{1}{32}$	$5.96E-3$	$2.72E-6$	1.01	$4.55E-6$	0.72	$3.9E-6$	1.04	$1.32E-5$	0.79
$\frac{1}{64}$	$2.13E-3$	$6.78E-7$	1.00	$3.14E-6$	0.27	$1.45E-6$	0.71	$7.51E-6$	0.40

Table 4: Convergence rates for the manufactured solution (116) for the Hölder exponent $\alpha = 0.7$.

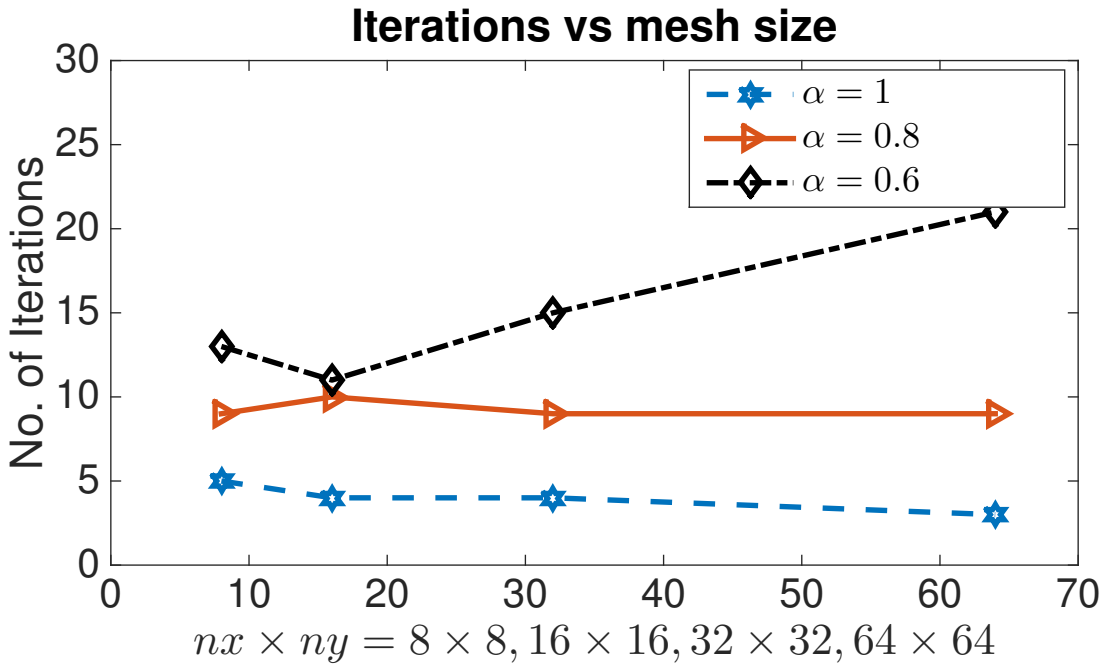

Figure 1: Number of iterations for different Hölder exponents, $\alpha = 1, 0.8$ and 0.6 and for different grids.

and the parameters

$$T = 50 \text{ days }, \ \tau = 0.5 \text{ day }, \ L = 2, k = 10^{-6} \text{ m}^2, \ \mu_w = 1 \text{ cP }, \mu_o = 10 \text{ cP}.$$

Observe that in this case $s(\cdot)$ is locally Lipschitz. For the formulation given in (7) - (10), these choices correspond to

$$a(s) = 10^6 \frac{1}{s^2 + (1-s)^2}, \; f_w(s) = \frac{s^2}{s^2 + (1-s)^2}, \; \mathbf{f_1} = \mathbf{0}, \; \mathbf{f_3} = \mathbf{0}.$$

For the pressure, we use Dirichlet boundary conditions at the left and right sides ($p = 0$ at $x = 0$, respectively $p = 10$ at $x = 1$) and homogeneous Neumann at all other boundaries. For the saturation, we use no flow boundary conditions and consider an injection at the center of the cells $f_2 = 10^{-5}$ m^3/s. The initial value is taken such that the initial saturation is $s = 0.05$.

We performed experiments for the case of a Hölder continuous saturation function $s(\cdot)$, namely for $\alpha = 0.9, 0.7, 0, 5$ and 0.3. The constant in the linearization scheme is $L = 1.2, 2, 8$ and 225 respectively. Observe that L increases as α decreases. The time step was $\tau = 0.25$ (days), the final time $T = 0.5$ (days), and the mesh size $nx = 20, ny = 20, nz = 20$. The convergence results are presented in Figure 2, where the normalized error (i.e. the iteration error divided to the error after the first iteration) is being plotted. As predicted by the theory, the L-scheme converges relatively fast as long as the Hölder coefficient does not become to small. For more results concerning the convergence of the L-scheme for the Lipschitz continuous case (same 3D example) we refer to [46].

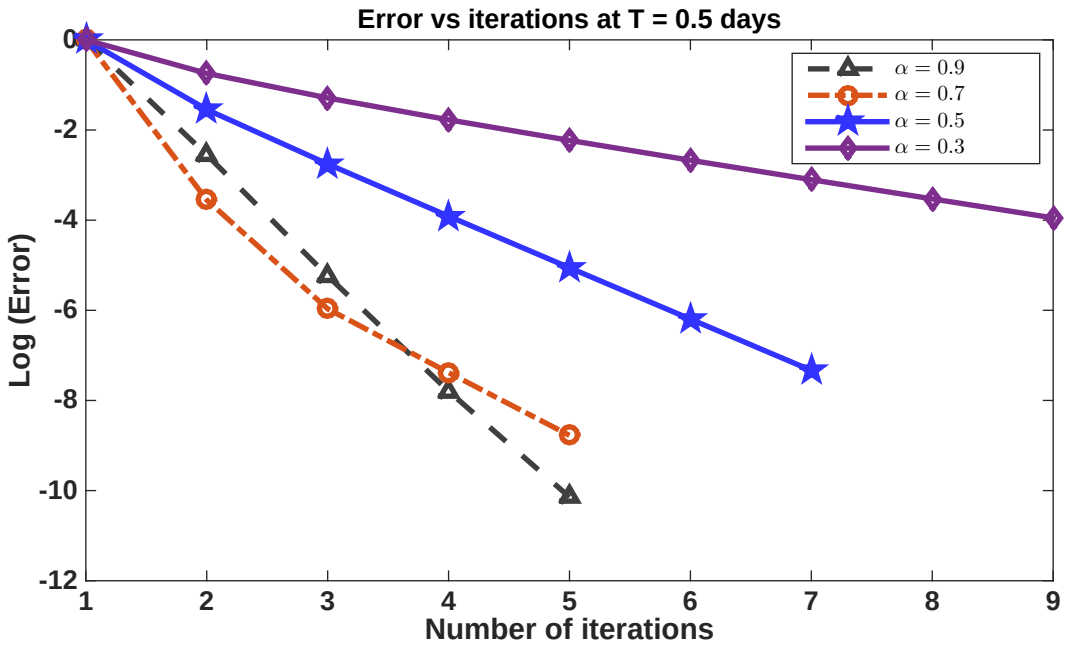

Figure 2: Convergence of the L-scheme for the Hölder continuous case.

6 Conclusions

A fully implicit, mass conservative numerical scheme for approximating the the solution of a two-phase porous media flow model is discussed. We use a mixed formulation of the model, involving the global and the complementary pressure, and leading to a system of nonlinear and possibly degenerate partial differential equations. In contrast with previous results requiring that the saturation function $s(\cdot)$ is Lipschitz continuous, the present work covers the case when $s(\cdot)$ is Hölder continuous.

The proposed numerical scheme is based on the backward Euler time stepping and the lowest order Raviart-Thomas mixed finite element discretization in space. For this scheme, *a priori* stability and error estimates are obtained, and the convergence is being proved. The estimates obtained theoretically are confirmed by numerical experiments, showing clearly the dependence of the convergence order on the Hölder exponent in the saturation function. In particular, the results for the Lipschitz continuous case are optimal.

Since the numerical scheme is implicit in time, at each time step one has to solve a non-linear algebraic system. Here a robust, first order convergent linearization method is proposed. This scheme, which does not require the computation of any derivatives and is therefore very easy to implement, involves a single parameter ($L > 0$) and is therefore called the L-method. Its convergence is proved rigorously for the case of a Lipschitz continuous saturation. The proof includes the case of a not necessarily strictly increasing saturation (i.e. the inverse is not assumed to be Lipschitz continuous). For the Hölder continuous case, it is shown that the parameter L can be chosen in such a way that the iteration error can be decreased below any given threshold. In this way, a 'numerical' convergence is established. Moreover, in either case the convergence is guaranteed under a mild restriction in the time step and is not conditioned by a good choice of the initial iteration. Finally, the convergence rate of the method is robust w.r.t. the mesh size.

The L-method is of particular relevance for the case of a Hölder continuous $s(\cdot)$. In this case the Newton method can not be applied directly, but only after regularizing the model, which can affect properties like mass conservation. Such a regularization is not needed for the L-method. The L-method can be used also to enhance the robustness of Newton's method, as it was shown in [29]. The provided numerical examples in both two and three spatial dimensions are in good agreement with the theoretical results.

Acknowledgments. F. A. Radu and I. S. Pop acknowledge the support of Statoil through the Akademia agreement. J.M. Nordbotten acknoweldges the NWO support through the Visitors Grant 040.11.351. Part of this work was done during F.A. Radu's sabbatical in Eindhoven, we acknowledge for this the support of Meltzer foundation, of University of Bergen and the NWO Visitors Grant 040.11.499. I.S. Pop is supported by the Research Foundation - Flanders FWO through the the Odysseus programme grant G0G1316N. Finally the authors want to thank the anonymous referees for helping to improve the paper substantially.

References

[1] H. W. ALT AND E. DI BENEDETTO, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scu. Norm. Sup. Pisa Cl. Sci. 12 (1985), pp. 335-392.

[2] B. AMAZIANE, M. JURAK AND Ž. KEKO, *An existence result for a coupled system modeling a fully equivalent global pressure formulation for immiscible compressible two-phase flow in porous media*, J. Differ. Equ. 250 (2011), pp. 1685-1718.

[3] T. ARBOGAST, *The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow*, J. Non-linear Analysis: Theory, Methods & Applications 19 (1992), pp. 1009-1031.

[4] T. ARBOGAST, M. F. WHEELER AND N. Y. ZHANG, *A non-linear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal. 33 (1996), pp. 1669-1687.

[5] L. BERGAMASCHI AND M. PUTTI, *Mixed finite elements and Newton-type linearizations for the solution of Richards' equation*, Int. J. Num. Meth. Engng. 45 (1999), pp. 1025-1046.

[6] J. BEAR AND Y. BACHMAT, *Introduction to Modelling of Transport Phenomena in Porous Media*, Kluwer Academic, Dordrecht, 1991.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] C. CANCÈS AND T. GALLOUËT, *On the time continuity of entropy solutions*, J. Evol. Equ. 11 (2011), pp. 43-55.

[9] C. CANCÈS AND M. PIERRE, *An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field*, SIAM J. Math. Anal. 44 (2012), pp. 966-992.

[10] C. CANCÈS, I. S. POP AND M. VOHRALIK, *An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*, Math. Comp. 83 (2014), pp. 153-188.

[11] M. CELIA, E. BOULOUTAS AND R. ZARBA, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res. 26 (1990), pp. 1483–1496.

[12] G. CHAVENT AND J. JAFFRE, *Mathematical models and finite elements for reservoir simulation*, Elsevier, 1991.

[13] L. CHERFILS, C. CHOQUET AND M. M. DIEDHIOU, *Numerical validation of an upscaled sharp-diffuse interface model for stratified miscible flows*, Math. Comput. Simulation (2016), pp. 1-34.

[14] Z. CHEN, *Degenerate two-phase incompressible flow. Existence, uniqueness and regularity of a weak solution*, J. Diff. Eqs. 171 (2001), pp. 203-232.

[15] Z. CHEN AND R. EWING, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology*, SIAM J. Numer. Anal. (1997), pp. 2228-2253.

[16] Z. CHEN AND R. EWING, *Degenerate two-phase incompressible flow III. Sharp error estimates*, Numer. Mat. 90 (2001), pp. 215-240.

[17] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[18] J.DOUGLAS JR. AND J.ROBERTS, *Global estimates for mixed methods for second order elliptic problems*, Math. Comp. 45 (1985), pp. 3952.

[19] L.J. DURLOFSKY, *A triangle based mixed finite element-finite volume technique for modelling two-phase flow through porous media*, J. of Comput. Phys. Appl. Math. 105 (1993), pp. 252-266.

[20] Y. EPSHTEYN AND B. RIVIERE, *Analysis of hp discontinuous Galerkin methods for incompressible two-phase flow*, J. of Comput. and Appl. Math. 225 (2009), pp. 487-509.

[21] R. EYMARD, D. HILHORST AND M. VOHRALIK, *A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems*, Numer. Math. 105 (2006), pp. 73-131.

[22] R. EYMARD, R. HERBIN AND A. MICHEL, *Mathematical study of a petroleum-engineering scheme*, Math. Modell. and Numer. Anal. 37 (2003), pp. 937-962.

[23] K. B. FADIMBA, *On existence and uniqueness for a coupled system modeling immiscible flow through a porous medium*, J. Math. Anal. Appl. 328 (2007), pp. 1034-1056.

[24] R. HELMIG, *Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems*, Springer Verlag, 1997.

[25] J. KOU AND S. SUN, *A new treatment of capillarity to improve the stability of IMPES two-phase ow formulation*, Computers and Fluids 39 (2010), pp. 1923-1931.

[26] J. KOU AND S. SUN, *On iterative IMPES formulation for two phase ow with capillarity in heterogeneous porous media*, Inter. J. of Numer. Anal. and Modeling, Series B 1 (2010), pp. 20-40.

[27] D. KRÖNER AND S. LUCKHAUS, *Flow of oil and water in a porous medium*, J. Differ. Equ. 55 (1984), pp. 276-288.

[28] F. LEHMANN AND PH. ACKERER, *Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media*, Transport in porous media 31 (1998), pp. 275–292.

[29] F. LIST AND F.A. RADU, *A study on iterative methods for Richards' equation*, Comput. Geosci. 20 (2016), pp. 341-353.

[30] S. KRÄUTLE, *The semismooth Newton method for multicomponent reactive transport with minerals*, Adv. Water Resources 34 (2011), pp. 137-151.

[31] R. KLAUSEN, F.A. RADU AND G. EIGESTAD, *Convergence of MPFA on triangulations and for Richards' equation*, Intern. J. for Numerical Methods in Fluids (2008), pp. 1327-1351.

[32] O. A LADYZHENSKAYA AND N. N. URALTSEVA, *Linear and quasilinear elliptic equations*, Academic Press, New York-London, 1968.

[33] A. MICHEL, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal. 41 (2003), pp. 1301-1317.

[34] J. M. NORDBOTTEN AND M. A. CELIA, *Geological Storage of CO2. Modeling Approaches for Large-Scale Simulation*, John Wiley & Sons, 2012.

[35] R. H. NOCHETTO AND C. VERDI, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal. 25 (1988), pp. 784–814.

[36] R. NEUMANN, P. BASTIAN AND O. IPPISCH, *Modeling and simulation of two-phase two-component flow with disappearing nonwetting phase*, Comput. Geosci.17 (2013), pp. 139-149.

[37] M. OHLBERGER, *Convergence of a mixed finite element- finite volume method for two phase flow in porous media*, East-West J- Numer. Math. 5 (1997), pp. 183-210.

[38] E. J. PARK, *Mixed finite elements for non-linear second-order elliptic problems*, SIAM J. Numer. Anal. 32 (1995), pp. 865-885.

[39] I. S. POP, *Error estimates for a time discretization method for the Richards' equation*, Comput. Geosci. 6 (2002), pp. 141-160.

[40] I. S. POP, F.A. RADU AND P. KNABNER, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. and Appl. Math. 168 (2004), pp. 365-373.

[41] F. A. RADU, I. S. POP AND P. KNABNER, *Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation*, SIAM J. Numer. Anal. 42 (2004), pp. 1452-1478.

[42] F.A. RADU, I.S. POP AND P. KNABNER, *On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation*, In Numerical Mathematics and Advanced Applications. A. Bermudez de Castro et al. (editors), Springer, 1194-1200, 2006.

[43] F. A. RADU, I. S. POP AND S. ATTINGER, *Analysis of an Euler implicit - mixed finite element scheme for reactive solute transport in porous media*, Numer. Meth. PDE's 26 (2010), pp. 320-344.

[44] F. A. RADU, I. S. POP AND P. KNABNER, *Error estimates for a mixed finite element discretization of some degenerate parabolic equations*, Numer. Math. 109 (2008), pp. 285-311.

[45] F. A. RADU, J. M. NORDBOTTEN, I. S. POP AND K. KUMAR, *A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media*, J. Comput. and Appl. Math. 289 (2015), pp. 134-141.

[46] F. A. RADU, K. KUMAR, J. M. NORDBOTTEN AND I. S. POP, *A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media*, arXiv:1512.08387, 2015.

[47] B. RIVIERE AND N. WALKINGTON, *Convergence of a discontinuous Galerkin method for the miscible displacement equation under low regularity*, SIAM J. Numer. Anal. 49 (2011), pp. 1085-1110.

[48] A. QUARTERONI AND A. VALLI, *Numerical approximations of partial differential equations*, Springer-Verlag, 1994.

[49] B. SAAD AND M. SAAD, *A combined finite volume–nonconforming finite element scheme for compressible two phase flow in porous media*, Numer. Math. 129 (2015), pp. 691-722.

[50] R.E. SHOWALTER, *Nonlinear degenerate evolution equations in mixed formulation*, SIAM J. Math. Anal. 42 (2010), pp.2114-2131.

[51] M. SLODICKA, *A robust and efficient linearization scheme for doubly non-linear and degenerate parabolic problems arising in flow in porous media*, SIAM J. Sci. Comput. 23 (2002), pp.1593-1614.

[52] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, Vol. 343. American Mathematical Soc., 2001.

[53] J.M. THOMAS, *Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes*, These d'Etat, University Pierre et Marie Curie (Paris 6), 1977.

[54] C. WOODWARD AND C. 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), pp. 701-724.

[55] W.A. YONG AND I.S. POP, *A numerical approach to porous medium equations*, Preprint 95-50, SFB 359, IWR, University of Heidelberg, 1996.

[56] I. YOTOV, *A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arizing in porous media flow*, EastWest J. Numer. Math. 5 (1997), pp. 211-230.

universiteit hasselt

UHasselt Computational Mathematics Preprint Series

UP-16-01 *Jochen Schütz and Vadym Aizinger*, **A hierarchical scale separation approach for the hybridized discontinuous Galerkin method**, 2016

UP-16-02 *Klaus Kaiser, Jochen Schütz, Ruth Schöbel and Sebastian Noelle*, **A new stable splitting for the isentropic Euler equations**, 2016

UP-16-03 *Sergey Alyaev, Eirik Keilegavlen, Jan Martin Nordbotten, Iuliu Sorin Pop*, **Fractal structures in freezing brine**, 2016

UP-16-04 *Florin A. Radu, Kundan Kumar, Jan Martin Nordbotten, Iuliu Sorin Pop*, **A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities**, 2016