Iterative linearisation schemes for doubly degenerate parabolic equations

Jakub Wiktor Both[1,3], Kundan Kumar[1], Jan Martin Nordbotten[1,2], Iuliu Sorin Pop[3,1], and Florin Adrian Radu[1]

[1] University of Bergen, Department of Mathematics, Bergen, Norway, {jakub.both, kundan.kumar, jan.nordbotten, florin.radu}@uib.no
[2] Princeton University, Department of Civil and Environmental Engineering, Princeton, NJ, USA
[3] Hasselt University, Faculty of Sciences, Diepenbeek, Belgium, sorin.pop@uhasselt.be

Abstract. Mathematical models for flow and reactive transport in porous media often involve non-linear, degenerate parabolic equations. Their solutions have low regularity, and therefore lower order schemes are used for the numerical approximation. Here the backward Euler method is combined with a mixed finite element method, which results in a stable and locally mass-conservative scheme. At each time step one has to solve a non-linear algebraic system, for which one needs adequate iterative solvers. Finding robust ones is particularly challenging here, since the problems considered are double degenerate (i.e. two type of degeneracies are allowed: parabolic-elliptic and parabolic-hyperbolic).

Commonly used schemes, like Newton and Picard, are defined either for non-degenerate problems, or after regularising the problem in the case of degenerate ones. Convergence is guaranteed only if the initial guess is sufficiently close to the solution, which translates into severe restrictions on the time step. Here we discuss an iterative linearisation scheme which builds on the L-scheme, and does not employ any regularisation. We prove its rigorous convergence, which is obtained for Hölder type non-linearities. Finally, we present numerical results confirming the theoretical ones, and compare the behaviour of the proposed scheme with schemes based on a regularisation step.

1 Introduction

We consider the following non-linear, degenerate parabolic equation

$$\partial_t b(u(t,\mathbf{x})) - \nabla \cdot (\nabla u(t,\mathbf{x})) = f(t,\mathbf{x}), \qquad t \in (0,T], \mathbf{x} \in \Omega, \tag{1}$$

with given functions $b : \mathbb{R} \to \mathbb{R}$ and $f : (0,T] \times \Omega \to \mathbb{R}$. Ω is a bounded domain in $\mathbb{R}^d, d \in \{1,2,3\}$, having a Lipschitz continuous boundary $\partial\Omega$ and T is the final time. Initial and boundary conditions (for simplicity the latter are assumed to be of homogeneous Dirichlet type) complete the problem.

Equation (1) is the transformed Richards equation after applying the Kirchhoff transformation in the absence of gravity (see e.g. [21]) or a diffusion equation with equilibrium sorption modelled by a Freundlich isotherm.

Solving (1) is of interest for many applications of societal relevance, like environmental pollution, CO_2 storage or geothermal energy extraction.

A particular feature of (1) is that the problem may become degenerate, namely change its type from parabolic into elliptic or hyperbolic. One consequence of this is that the solutions typically lack regularity. Here we assume that $b(\cdot)$ is monotone increasing and Hölder continuous, which means that two types of degeneracy are allowed in (1). The first is when the derivative of $b(\cdot)$ vanishes (*fast diffusion*) and the second when it blows up (*slow diffusion*). In particular, here the vanishing of $b'(\cdot)$ may occur on intervals.

Since solutions to degenerate parabolic equations have low regularity (see [1]), low order discretisation methods are well suited for the numerical approximation of the solution. Here we combine the backward Euler (BE) method for the time discretisation with the mixed finite element method (MFEM). For the rigorous convergence analysis of the method we refer to [21] and the references therein. The resulting is a scheme that is both stable and locally mass-conservative.

In this paper we discuss iterative solvers for the non-linear algebraic systems arising at each time step after the complete discretisation of (1). Observe that although referring specifically to the MFEM approach, the non-linear solvers presented here can be also applied to other spatial discretisations, like finite volumes, conforming or discontinuous Galerkin finite elements.

The literature on non-linear solvers for (1) is very extensive, but covers in particular non-degenerate problems, or the case when $b(\cdot)$ is Lipschitz continuous. We refer to [4,18] for Newton's scheme, and to [6] for the modified Picard scheme. A combination of both is discussed in [12,17]. Also, the Jäger-Kačur scheme was introduced in [11]. We refer to [20] for the analysis of the Newton, modified Picard and the Jäger-Kačur schemes for BE/MFEM discretisations. Recently, in [5] the capillary pressure and the saturation are expressed both in terms of a new variable, by respecting the original saturation-capillary pressure dependency. If the new variable is properly chosen, the Richards equation receives a character that is more suited for Newton's scheme, in the sense that all non-linearities are Lipschitz continuous. We refer to [10] for a review detailing on such aspects.

The scheme analysed here builds on the L-scheme, a robust fixed point scheme, which does not involve the computations of any derivatives or a regularisation step. The convergence, proved rigorously in [19,25,27], holds in the H^1 norm and regardless of the initial guess, but is linear. To improve this convergence, a combination between the $L-$ and Newton schemes was discussed recently in [13]. By performing first a number of L-scheme iterations, one obtains an approximation that is close enough to the solution. After a switch to the Newton iterations, the convergence becomes quadratic.

Compared to the literature cited above, here we adopt a more challenging setting: $b(\cdot)$ is only Hölder continuous and not necessarily strictly increasing. Whenever $b'(\cdot)$ is unbounded, neither Newton nor Picard schemes can be applied directly. The common way to overcome this is to regularise $b(\cdot)$

(see [14]), e.g. to approximate it by a Lipschitz continuous function $b_\varepsilon(\cdot)$. Nevertheless, a regularisation will also imply a perturbation of the solution, which affects the accuracy of the method. Here, we propose an L-scheme for the degenerate equation (1), which is adapted to the Hölder continuous non-linearity. The linear convergence of the scheme is proved rigorously, and its performance is compared with the ones of the standard L- and Newton schemes, applied for the regularised problems.

The paper is organised as follows. In the next section the fully discrete variational approximation of (1) is given and the assumptions are stated. Section 3 discusses different iterative schemes. First the modified L-scheme together with the convergence proof are given. Then the approach based on regularisation is discussed, with particular emphasis on the Newton scheme. Finally, in Section 5 a comprehensive comparison between the L-schemes and the Newton scheme are presented. The paper is concluded with final remarks.

2 The fully discrete approximation

Throughout this paper we will use common notations in the functional analysis. By $L^p(\Omega)$ we mean the p-integrable functions with the norm $\|f\|_p := \left(\int_\Omega f(\mathbf{x})\, d\mathbf{x}\right)^{1/p}$, whereas $H(\mathrm{div};\Omega) := \{\mathbf{f} \in (L^2(\Omega))^d | \nabla \cdot \mathbf{f} \in L^2(\Omega)\}$. Further, we denote by $\langle \cdot, \cdot \rangle$ the inner product on $L^2(\Omega)$ and by $\sigma(\Omega)$ the volume of Ω. Similarly, by $H^1(\Omega)$ we mean the $L^2(\Omega)$ functions having the first order weak derivatives in L^2.

To define the discretisation we let $\mathcal{T}_h$ be a regular decomposition of the domain Ω (h is the mesh size) and $0 = t_0 < t_1 < ... < t_N = T$, $N \in \mathbb{N}$, is a partition of the time interval $[0, T]$ with constant time step size $\tau = t_{k+1} - t_k$, $k \geq 0$. The lowest-order Raviart-Thomas elements (see e.g. [2]) are used for the discretisation in space. The spaces $W_h \times V_h \subset L^2(\Omega) \times H(\mathrm{div};\Omega)$ are defined as

$$W_h := \{p \in L^2(\Omega) |\ p_{|T}(\mathbf{x}) = p_T \in \mathbb{R} \text{ for all } T \in \mathcal{T}_h\},$$
$$V_h := \{\mathbf{q} \in H(\mathrm{div};\Omega) | \mathbf{q}_{|T}(\mathbf{x}) = \mathbf{a_T} + b_T \mathbf{x}, \mathbf{a_T} \in \mathbb{R}^d, b_T \in \mathbb{R} \text{ for all } T \in \mathcal{T}_h\}.$$

The lemma below (see [9]) will be used in the proof of Theorem 4.

Lemma 1. *There exists a constant $C_\Omega > 0$ not depending on the mesh size h, such that given an arbitrary $w_h \in W_h$ there exists $\mathbf{v}_h \in V_h$, satisfying $\nabla \cdot \mathbf{v}_h = w_h$ and $\|\mathbf{v}_h\| \leq C_\Omega \|w_h\|$.*

As mentioned, (1) is completed with homogeneous Dirichlet boundary conditions, and with the initial condition $u(0, \mathbf{x}) := u_0(\mathbf{x})$, with $u_0 \in L^2(\Omega)$. Furthermore, the source term is $f \in L^2(\Omega)$. We make the following assumptions on $b(\cdot)$.

(A1) The function $b : \mathbb{R} \to \mathbb{R}$, with $b(0) = 0$, is non-decreasing and Hölder continuous: there exist $L_b > 0$ and $\alpha \in (0, 1]$ such that

$$|b(x) - b(y)| \leq L_b |x - y|^\alpha \qquad \text{for all} \quad x, y \in \mathbb{R}. \tag{2}$$

Remark 2. The case $\alpha = 1$ corresponds to a Lipschitz continuous $b(\cdot)$, a case which is relatively well-understood [4,11–13,18–20,25,27]. The case $\alpha \in (0,1)$ is encountered for the Richards equation under physically relevant parametrisations (the van Genuchten curves [15], see Remark 1.1 in [21]). Also, if Freundlich rates are used for modelling reactive transport, one has $b(u) = u + \phi(u)$, with ϕ increasing but non-Lipschitz. Then there exists an $m \in \mathbb{R}$ such that $b' \geq m > 0$, which simplifies the analysis of the iterative schemes.

Remark 3. Non-linear convection $\mathbf{q}(\cdot)$ can be added, however, if being Lipschitz continuous. The numerical schemes can be then easily modified to include such changes: one can deal with such non-linearities by using either the outcome at the last iteration, or by including this term in the Newton iteration, depending on the method used. For the ease of presentation, such cases are not considered here.

In view of the lacking regularity, the solutions to (1) are weak. We refer to [1,16] for existence and uniqueness results. Also, the equivalence between the conforming and mixed formulation, for both time continuous and time discrete problems, is being discussed in [21] (see also [22] for the case of a two-phase flow model). Such results provide the existence and uniqueness of a solution for the mixed formulation, and can be used for obtaining the rigorous convergence of the discretisation. Finally, for each time step, the backward Euler-MFEM discretisation of (1) reduces to a non-linear, fully discrete variational problem ($n \geq 1$).

Problem P_h^n (The non-linear fully discrete problem).
Let $u_h^{n-1} \in W_h$ be given. Find $u_h^n \in W_h$ and $\mathbf{q}_h^n \in V_h$ such that for any $w_h \in W_h$ and $\mathbf{v}_h \in V_h$ there holds

$$\langle b(u_h^n) - b(u_h^{n-1}), w_h \rangle + \tau \langle \nabla \cdot \mathbf{q}_h^n, w_h \rangle = \tau \langle f, w_h \rangle, \tag{3}$$

$$\langle \mathbf{q}_h^n, \mathbf{v}_h \rangle - \langle u_h^n, \nabla \cdot \mathbf{v}_h \rangle = 0. \tag{4}$$

Clearly, for $n = 1$, u_h^0 can be taken as the L^2-projection of the initial condition u_0 onto W_h (see also [21]).

Here we assume that a solution to Problem P_h^n exists and is unique. For $\alpha = 1$, i.e. when b is Lipschitz continuous, Theorem 4 below guarantees that the iterative scheme (5)–(6) is H^1-contractive. This immediately provides the existence of a solution. For $\alpha \in (0,1)$, the existence can be proved by using Brouwer's fixed point theorem (see e.g. Lemma 1.4, p. 140 in [26]). We refer to [3,7,8,23] for similar results in the context of two-phase porous media flow models. Finally, since b is monotone, uniqueness can be proved by comparison.

The main challenge in solving the non-linear Problem P_h^n is to construct a linearisation scheme that is converging also for the case when $b(\cdot)$ is only Hölder continuous, implying that $b'(\cdot)$ may become unbounded. The scheme is discussed in the section below. Typically, iterative approaches like

the Newton, (modified) Picard, or the L-schemes are applied to the regularised problem, with a Lipschitz continuous approximation b_ε replacing b (see [4,6,13,18,19,24,25]). This will be detailed in Section 4.

3 A robust iterative scheme

Below we define a robust iterative scheme for (3)–(4), which does not involve regularisation, or computing any derivatives. We let the time step $n \geq 1$ be fixed and assume $u_h^{n-1} \in W_h$ be given. Also, let $L = \frac{1}{\delta}$, where $\delta > 0$ is a small parameter that will be chosen later to guarantee that the error decreases below a prescribed threshold. With $i \in \mathbb{N}$, $i > 0$ being the iteration index, the iteration step is introduced through

Problem $P_h^{n,i}$ (The L-scheme).
Let $u_h^{n,i-1} \in W_h$ be given. Find $(u_h^{n,i}, \mathbf{q}_h^{n,i}) \in W_h \times V_h$ s.t. for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$ one has

$$\langle L(u_h^{n,i} - u_h^{n,i-1}) + b(u_h^{n,i-1}), w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,i}, w_h\rangle = \langle b(u_h^{n-1}) + \tau f, w_h\rangle, \tag{5}$$

$$\langle \mathbf{q}_h^{n,i}, \mathbf{v}_h\rangle - \langle u_h^{n,i}, \nabla \cdot \mathbf{v}_h\rangle = 0. \tag{6}$$

As will be seen below, the convergence is obtained without imposing restrictions on the initial guess $u_h^{n,0} \in W_h$, but a natural choice is u_h^{n-1}.

As for Problem P_h^n, the uniqueness of a solution for Problem $P_h^{n,i}$ follows by standard techniques. Specifically, assuming that Problem $P_h^{n,i}$ has two solution pairs $(u_{h,k}^{n,i}, \mathbf{q}_{h,k}^{n,i}) \in W_h \times V_h$ $(k = 1, 2)$ and with $(du_h, \mathbf{dq}_h)$ denoting their difference it holds

$$L\langle du_h, w_h\rangle + \tau\langle \nabla \cdot \mathbf{dq}_h, w_h\rangle = 0,$$
$$\langle \mathbf{q}_h, \mathbf{v}_h\rangle - \langle du_h, \nabla \cdot \mathbf{v}_h\rangle = 0,$$

for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$. Taking in the above $w_h = du_h$, respectively $\mathbf{v}_h = \tau \mathbf{du}_h$, and adding the resulting equations gives

$$L\|du_h\|^2 + \tau\|\mathbf{q}_h\|^2 = 0, \tag{7}$$

which immediately implies uniqueness. Moreover, since Problem $P_h^{n,i}$ is linear and finite dimensional, uniqueness also implies the existence of a solution.

To show the convergence of the scheme we define the errors

$$e_u^{n,i} = u_h^{n,i} - u_h^n, \quad \text{and} \quad e_{\mathbf{q}}^{n,i} = \mathbf{q}_h^{n,i} - \mathbf{q}_h^n,$$

where $(u_h^n, \mathbf{q}_h^n)$ is the solution pair of Problem P_h^n. We use in the next the elementary (in)equalities, holding for any $c, d \geq 0$ and $p, q > 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$

$$c(c-d) = \frac{1}{2}\left(c^2 - d^2 + (c-d)^2\right), \quad \text{and} \quad cd \leq \frac{c^p}{p} + \frac{d^q}{q}. \tag{8}$$

With fixed $\delta > 0$ and $n \in \mathbb{N}$, $n \geq 1$, let $L = \dfrac{1}{\delta}$ and assume $u_h^{n-1} \in W_h$ known. The main result supporting the convergence is

Theorem 4. *Assuming (A1) and $\alpha \in (0,1)$, let $i \in \mathbb{N}$, $i \geq 1$ and $u_h^{n,i-1} \in W_h$ be given. If $(u_h^n, \mathbf{q}_h^n)$ and $(u_h^{n,i}, \mathbf{q}_h^{n,i})$ are the solutions of Problems P_h^n and $P_h^{n,i}$ respectively, there holds*

$$\|e_u^{n,i}\|^2 + \tau\delta R(\delta,\tau)\|e_{\mathbf{q}}^{n,i}\|^2 \leq R(\delta,\tau)\|e_u^{n,i-1}\|^2 + 2C(\alpha)R(\delta,\tau)\delta^{\frac{2}{1-\alpha}}. \tag{9}$$

Here $R(\delta,\tau) = \left(1 + \dfrac{\tau\delta}{C_\Omega^2}\right)^{-1}$, C_Ω being the constant in Lemma 1, and $C(\alpha) = \frac{(1-\alpha)}{2}\big(L_b(2\alpha)^\alpha\big)^{\frac{2}{1-\alpha}}(1+\alpha)^{-\frac{1+\alpha}{1-\alpha}}\sigma(\Omega)$.

Proof. Subtracting (3) and (4) from (5), respectively (6), one gets for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$

$$\langle L(e_u^{n,i} - e_u^{n,i-1}) + b(u_h^{n,i-1}) - b(u_h^n), w_h\rangle + \tau\langle \nabla\cdot e_{\mathbf{q}}^{n,i}, w_h\rangle = 0, \tag{10}$$

$$\langle e_{\mathbf{q}}^{n,i}, \mathbf{v}_h\rangle - \langle e_u^{n,i}, \nabla\cdot\mathbf{v}_h\rangle = 0. \tag{11}$$

By taking $w_h = e_u^{n,i} \in W_h$, respectively $\mathbf{v}_h = \tau e_{\mathbf{q}}^{n,i} \in V_h$, adding the resulting equations and after some algebraic calculations one gets

$$\begin{aligned}\frac{L}{2}\left(\|e_u^{n,i}\|^2 + \|e_u^{n,i} - e_u^{n,i-1}\|^2\right) + \langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i-1}\rangle + \tau\|e_{\mathbf{q}}^{n,i}\|^2 \\ = \frac{L}{2}\|e_u^{n,i-1}\|^2 - \langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i} - e_u^{n,i-1}\rangle.\end{aligned} \tag{12}$$

By (A1), it holds

$$\langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i-1}\rangle \geq L_b^{-\frac{1}{\alpha}}\|b(u_h^{n,i-1}) - b(u_h^n)\|_{\frac{1+\alpha}{\alpha}}^{\frac{1+\alpha}{\alpha}}. \tag{13}$$

Using now the inequality in (8) with $p = \frac{1+\alpha}{\alpha}$, $q = 1+\alpha$, $c = \frac{|b(u_h^{n,i-1}) - b(u_h^n)|}{L_b^{\frac{1}{1+\alpha}}(\frac{2\alpha}{1+\alpha})^{\frac{\alpha}{1+\alpha}}}$ and $d = L_b^{\frac{1}{1+\alpha}}(\frac{2\alpha}{1+\alpha})^{\frac{\alpha}{1+\alpha}}|e_u^{n,i} - e_u^{n,i-1}|$ one gets

$$\begin{aligned}&|\langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i} - e_u^{n,i-1}\rangle| \\ &\leq \frac{1}{2L_b^{\frac{1}{\alpha}}}\|b(u_h^{n,i-1}) - b(u_h^n)\|_{\frac{1+\alpha}{\alpha}}^{\frac{1+\alpha}{\alpha}} + \frac{(2\alpha)^\alpha L_b}{(\alpha+1)^{(\alpha+1)}}\|e_u^{n,i} - e_u^{n,i-1}\|_{1+\alpha}^{1+\alpha}.\end{aligned} \tag{14}$$

From (12), (13) and (14) one obtains

$$\begin{aligned}&\tfrac{L}{2}\left(\|e_u^{n,i}\|^2 + \|e_u^{n,i} - e_u^{n,i-1}\|^2\right) + \tfrac{1}{2}\langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i-1}\rangle + \tau\|e_{\mathbf{q}}^{n,i}\|^2 \\ &\leq \tfrac{L}{2}\|e_u^{n,i-1}\|^2 + \frac{(2\alpha)^\alpha L_b}{(\alpha+1)^{(\alpha+1)}}\|e_u^{n,i} - e_u^{n,i-1}\|_{1+\alpha}^{1+\alpha}.\end{aligned}$$

Using again Young's inequality, but with $p = \frac{2}{1+\alpha}$, $q = \frac{2}{1-\alpha}$, $c = \|e_u^{n,i} - e_u^{n,i-1}\|_{1+\alpha}^{1+\alpha}(\frac{L}{1+\alpha})^{\frac{1+\alpha}{2}}\sigma(\Omega)^{\frac{\alpha-1}{2}}$ and $d = \frac{(2\alpha)^\alpha L_b}{(\alpha+1)^{(\alpha+1)}}(\frac{1+\alpha}{L})^{\frac{1+\alpha}{2}}\sigma(\Omega)^{\frac{1-\alpha}{2}}$ gives

$$\begin{aligned}
&\frac{(2\alpha)^\alpha L_b}{(\alpha+1)^{(\alpha+1)}}\|e_u^{n,i} - e_u^{n,i-1}\|_{1+\alpha}^{1+\alpha} \\
&\quad\le \frac{L}{2}\sigma(\Omega)^{\frac{\alpha-1}{1+\alpha}}\|e_u^{n,i} - e_u^{n,i-1}\|_{1+\alpha}^2 + C(\alpha)L^{\frac{1+\alpha}{\alpha-1}} \\
&\quad\le \frac{L}{2}\|e_u^{n,i} - e_u^{n,i-1}\|^2 + C(\alpha)L^{\frac{1+\alpha}{\alpha-1}},
\end{aligned}$$

where $C(\alpha)$ is defined in the formulation of the theorem. Above we used the inequality $\|f\|_{1+\alpha} \le \sigma(\Omega)^{\frac{1-\alpha}{2(1+\alpha)}}\|f\|_2$, valid for any $f \in L^2(\Omega)$ and $\alpha \in (0,1]$ since Ω is bounded. Now, from the last two estimates it follows that

$$\frac{L}{2}\|e_u^{n,i}\|^2 + \frac{1}{2}\langle b(u_h^{n,i-1}) - b(u_h^n), e_u^{n,i-1}\rangle + \tau\|e_{\mathbf{q}}^{n,i}\|^2 \le \frac{L}{2}\|e_u^{n,i-1}\|^2 + C(\alpha)L^{\frac{1+\alpha}{\alpha-1}}.$$

From (11) and using Lemma 1, a Poincare type inequality $\|e_u^{n,i}\| \le C_\Omega\|e_{\mathbf{q}}^{n,i}\|$ can be obtained. Using this in the above, since $L = 1/\delta$, one obtains (9).

Remark 5. Observe that since $R(\delta,\tau) < 1$ whereas δ has a positive power in the last term on the right of (9), this theorem gives the convergence of the scheme. More precisely, for any chosen tolerance TOL, one can chose δ such that the term $2C(\alpha)\delta^{\frac{2}{1-\alpha}}\frac{R(\delta,\tau)}{1-R(\delta,\tau)} < \frac{1}{2}TOL$. Since this is the sum of the last terms on the right in (9), this can be seen as the total error being accumulated while iterating in one time step. On the other hand, the first term in the right is showing how the error is contracted in one iteration. Thus, choosing $i^* \in N$ large enough s.t. $R(\delta,\tau)^{i^*}\|e_u^{n,0}\|^2 \le \frac{1}{2}TOL$ and applying (9) successively for $i = i^*, i^*-1, \ldots, 1$ one obtains that $\|e_u^{n,i}\|^2 < TOL$. Nevertheless, the convergence rate is worsened with the decrease of δ, as $R(\delta,\tau)$ approaches 1 in this case. From theoretical point of view, this results in an increased number of iterations for obtaining the desired accuracy. This is a rather pessimistic interpretation, as the numerical examples studied in Section 5 indicate that the actual number of iterations is frequently better than what the theorem guarantees.

Remark 6. If b is Lipschitz continuous, the problem reduces to the one studied in [13,21]. In fact, for $\alpha = 1$ the last step in the proof above is superfluous, and the estimate (9) holds with $C(\alpha) = 0$. In this case, the iteration is a contraction, so the convergence is unconditional for any $L \ge L_b$, the Lipschitz constant of b.

Remark 7. Observe that the convergence can be achieved without requiring that the time step size τ is sufficiently small. In fact, when calculating the ratio $\frac{R(\delta,\tau)}{1-R(\delta,\tau)}$ one sees that τ appears in the denominator, so the larger it is, the better the convergence of the iterative scheme. Further, the term

$2C(\alpha)\delta^{\frac{2}{1-\alpha}}\frac{R(\delta,\tau)}{1-R(\delta,\tau)}$ is practically small without taking a too small δ. For example, if $\alpha = 0.5$, the power of δ in this term becomes $\frac{1+\alpha}{1-\alpha} = 3$. Taking $\delta = 0.01$ (hence $L = 100$) gives $\delta^{\frac{1+\alpha}{1-\alpha}} = 10^{-6}$. Also, the number $C(\alpha)$ is small too. In the situation above, if $L_b = 0.5$, and $\sigma(\Omega) = 1$, $C(\alpha) \approx 0.0046$.

4 Iterative schemes based on regularisation

As follows from the above, the iterations introduced through Problem $P_h^{n,i}$ converge also for the case of a Hölder continuous b and do not involve computing any derivatives. However, the iterations only converge linearly. A natural question appears: what is the performance of the new scheme in comparison with the Newton or the L-scheme, but applied for the regularised problems. To study this aspect we first present below these schemes and discuss their convergence.

For simplicity we consider the function $b : \mathbb{R} \to \mathbb{R}$, $b(u) = (\max\{u, 0\})^\alpha$. Observe that b is not Lipschitz for arguments approaching 0 from above. For regularising it we let $\varepsilon > 0$ and consider the function $b_\varepsilon : \mathbb{R} \to \mathbb{R}$,

$$b_\varepsilon(u) = \alpha\varepsilon^{\alpha-1}u + (1-\alpha)\epsilon^\alpha,$$

if $u \in (0, \varepsilon)$, whereas $b_\varepsilon(u) = b(u)$ everywhere else. Clearly, $b_\varepsilon(\cdot)$ is non-decreasing, and both $b_\varepsilon(\cdot), b'_\varepsilon(\cdot)$ are Lipschitz continuous with the Lipschitz constants $L_{b_\varepsilon} = \alpha\varepsilon^{\alpha-1}$, respectively $L_{b'_\varepsilon} = \alpha(1-\alpha)\varepsilon^{\alpha-2}$. Moreover, it holds $0 \le b(x) - b_\varepsilon(x) \le (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}\varepsilon^\alpha$.

As before, for given $\varepsilon > 0$ and $u_{h,\varepsilon}^{n-1} \in W_h$ (observe the dependency of the solution on ε), and with $i \in \mathbb{N}$, $i > 0$ being the iteration index, the Newton iterations for Problem P_h^n are defined through

Problem $NEWTON_h^{n,i}$.
Let $u_{h,\varepsilon}^{n,i-1} \in W_h$ be given. Find $(u_{h,\varepsilon}^{n,i}, \mathbf{q}_{h,\varepsilon}^{n,i}) \in W_h \times V_h$ s. t. for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$

$$\langle b_\varepsilon(u_{h,\varepsilon}^{n,i-1}) + b'_\varepsilon(u_{h,\varepsilon}^{n,i-1})(u_{h,\varepsilon}^{n,i} - u_{h,\varepsilon}^{n,i-1}), w_h\rangle + \tau\langle\nabla\cdot\mathbf{q}_{h,\varepsilon}^{n,i}, w_h\rangle = \langle b_\varepsilon(u_{h,\varepsilon}^{n-1}) + \tau f, w_h\rangle, \quad (15)$$

$$\langle\mathbf{q}_{h,\varepsilon}^{n,i}, \mathbf{v}_h\rangle - \langle u_{h,\varepsilon}^{n,i}, \nabla\cdot\mathbf{v}_h\rangle = 0. \quad (16)$$

*Remark 8 (***Regularised L-scheme).** A L-scheme for the regularised problem is obtained by replacing $b'_\varepsilon(u_{h,\varepsilon}^{n,i-1})$ with $L \ge 0$ in (15). The resulting scheme is convergent for $L \ge L_{b_\varepsilon}/2$, as proved in [13,19,20]. Moreover, the convergence holds in H^1 and for any initial guess, under very mild restrictions on the time step, but it is only linear. It is worth emphasising on the difference between the L-scheme in Section 3, designed for Hölder continuous non-linearities, and the L-scheme for the regularised problems. In the former case the errors at each iteration step consist of two components, one

that is contracted, and another that accumulates. The choice of the L parameter is driven by these two: first, the accumulated errors should remain below a threshold $\frac{1}{2}TOL$, and second the contracted ones reduce to the same threshold. For the latter the problem is regularised so that the non-linearities become Lipschitz continuous, and then the L parameter is taken as L_{b_ε}.

*Remark 9 (***Convergence of the regularised Newton scheme).** Two issues concerning the convergence appear in this case. First, the solution u_ε of the regularised problem should not be too far from u, the solution to the original problem. This means that ε should be sufficiently small. On the other hand, the advantage of the Newton scheme is its quadratic convergence. Guaranteeing it requires typically a small τ because the scheme is only locally convergent, so the initial guess of the iteration should not be too far from the solution and the choice at hand is the solution at the previous time step. However, τ and ε are correlated. So satisfying both requirements might be quite challenging, if not impossible in certain computations. If one assumes additionally that $b' \geq m > 0$, which rules out the *fast diffusion* case, the sufficient condition for convergence is to choose $\tau = O(\varepsilon^a h^{d/2})$, with a depending on the Hölder exponent (see [20]). In the case $b' \geq 0$, one can further perturb b so that b'_ε is bounded away from 0, e.g. by taking $b^{new}_\varepsilon(u) = \varepsilon u + b_\varepsilon(u)$ with $b_\varepsilon(u)$ given before. Then the convergence is guaranteed for similar constraints, possibly with a different exponent a.

To summarise, the convergence of Newton's scheme depends on the choice of the discretisation and regularisation parameters. Fixing two parameters, e.g. h and ε, only a sufficiently small τ will guarantee the convergence. Alternatively, for fixed τ and ε, the mesh size can not be too small, and if the Newton scheme diverges, refining the mesh will not help. In other words, to achieve a certain accuracy, e.g. by letting $\varepsilon \searrow 0$, the convergence condition for the Newton scheme might become very restrictive.

5 Numerical examples

In this section we provide numerical examples to illustrate the performance of the scheme. We use the example mentioned in Section 4, $b(u) = \max\{u, 0\}^\alpha$, for $\alpha = 0.5$. The domain is the square $\Omega = (0,1)\times(0,1)$, and the time interval is $t \in (0.0, 0.5]$. To evaluate the convergence we choose the source term, the boundary conditions and the initial condition such that the exact solution is

$$u(t,x,y) = -\frac{1}{2} + 16\, x(1-x)\, y(1-y)(t+0.5). \tag{17}$$

For the discretisation we consider a 32×32 mesh with different time step sizes $\tau \in \{0.05, 0.025, 0.0125\}$, resulting in 10, 20, respectively 40 time steps. To differentiate between the errors brought by the discretisation itself and those related to the iterative solver, we first compute a very accurate

approximation of the non-linear, fully discrete systems. Specifically, with Δu^i and $\Delta \mathbf{q}^i$ denoting the difference between two iterates, the reference solution is the iteration satisfying

$$\|\Delta u^i\|_{L^2(\Omega)} + \|\Delta \mathbf{q}^i\|_{L^2(\Omega)} < 10^{-8}, \text{ and } \frac{\|\Delta u^i\|_{L^2(\Omega)}}{\|u^i\|_{L^2(\Omega)}} + \frac{\|\Delta \mathbf{q}^i\|_{L^2(\Omega)}}{\|\mathbf{q}^i\|_{L^2(\Omega)}} < 10^{-8}.$$

This solution, called below u_h, was computed with the L-type scheme in Section 3 to avoid additional regularisation errors. Having obtained u_h we proceed by testing the three schemes discussed here, the L-scheme in the framework discussed in Section 3 (called HL), and the two (Newton and L) in Section 4, involving a regularisation step.

In agreement with the result stated in Theorem 4 we choose an admissible tolerance TOL to be used as stopping criterion for the different iteration schemes. Specifically, $u_h^\star$ is accepted as numerical solution if it satisfies $\|u_h^\star - u_h\|_{L^2(\Omega)} < TOL$ where u_h is the (accurate) solution from above.

We consider different tolerances, namely $TOL \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. For the regularisation based schemes, the problem is first regularised by taking $\varepsilon \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. For the L-scheme we take $L = \varepsilon^{\alpha-1}$, the Lipschitz constant of b_ε. For the HL-scheme we take $L = \frac{1}{\delta}$ where δ is such that the condition in Remark 5 on the accumulated error is met.

Table 1 presents the total number of Newton iterations and the corresponding, average number of iterations per time step for given different tolerances TOL, regularisation parameters ε and time step sizes τ. Observe that the parameters TOL and ε should be correlated to avoid that the regularisation error becomes dominating. In other words, a smaller TOL requires a smaller ε for obtaining the convergence. In the same spirit, a smaller τ requires smaller TOL and ε. For $\tau = 0.0125$, it becomes almost impossible to obtain solutions within the required accuracy by using the Newton scheme, as ε has to be very small and then the condition number of the Jacobian becomes very high. This is evidenced by the appearance of cases where the Newton scheme did not converge, which are mentioned as `nc`. In summary, the Newton scheme fails to converge if either the regularisation parameter ε is too large for the chosen tolerance TOL, or if ε is too low, which makes the problem very badly conditioned. Clearly, if convergent, the Newton scheme requires the least number of iterations among all schemes.

Similar experiments have been performed for the standard L-scheme, applied after regularising the problem. Recalling b_ε is Lipschitz, we set $L = L_{b_\varepsilon} = \alpha\varepsilon^{\alpha-1}$. The actual values are given in Table 2. Table 3 presents the convergence results. As for the Newton scheme, one needs to correlate the parameters TOL, ε, and τ. To ensure convergence, if TOL is small ε should be small enough, otherwise the regularisation error will dominate and the convergence criterion will not be met. This is the reason why the L scheme, though unconditionally convergent in theory, is marked as not convergent for the case $\varepsilon = 10^{-3}$, if $TOL = 10^{-4}$ or 10^{-5}. Also, observe that $L = L_{b_\varepsilon}$ blows up with $\varepsilon \searrow 0$, while the convergence rate approaches 1 if L is large, or τ

		N-iterations per time step		
TOL	ε	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.0125$
1e−3	1e−3	1.7	1.2	1.2
1e−3	1e−4	1.6	1.3	nc
1e−3	1e−5	1.6	1.3	nc
1e−4	1e−3	2.2	2.1	nc
1e−4	1e−4	2.3	2.4	nc
1e−4	1e−5	2.3	2.3	nc
1e−5	1e−3	nc	nc	nc
1e−5	1e−4	3.1	3.0	nc
1e−5	1e−5	3.1	3.2	nc

Table 1. Results for the Newton scheme. The scheme does not converge (nc) for the smallest time step and if ε is not in agreement with TOL.

ε	1e−3	1e−4	1e−5
L	16	50	159

Table 2. L values for the standard L-scheme, obtained for different values of ε.

is small (see [19]). Therefore if ε and τ are small, combined with the finite precision arithmetic may lead to the divergence of the L-scheme.

This also explains why the number of L-scheme iterations increases drastically with the decrease of the regularisation parameter. Compared to the Newton scheme, the number of L-iterations is much larger. On the other hand, the L-scheme is more robust than the Newton scheme, allowing to compute the solution for small time steps τ or for small regularisation parameters ε.

		L-iterations per time step		
TOL	ε	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.0125$
1e−3	1e−3	30.5	38.9	48.4
1e−3	1e−4	96.9	124.6	155.2
1e−3	1e−5	305.8	394.6	492.8
1e−4	1e−3	47.9	nc	nc
1e−4	1e−4	150.5	202.9	273
1e−4	1e−5	475.1	643.7	870.7
1e−5	1e−3	nc	nc	nc
1e−5	1e−4	204.5	281.5	nc
1e−5	1e−5	645.9	889.1	1247.9

Table 3. Results for the standard L-scheme. The scheme does not converge (nc) if ε is not in agreement with TOL.

Finally we draw our attention to the HL-scheme, where the parameter L is chosen as mentioned in Remark 5, depending on TOL. Since the domain is the unit square one has $C_\Omega = \sigma(\Omega) = 1$ and thus $R(\delta, \tau) = (1 + \tau\delta)^{-1}$.

For $\alpha = 0.5$, to reduce the accumulated errors below $\frac{1}{2}TOL$ one needs to take $\delta < \frac{3}{2}(\tau TOL)^{\frac{1}{3}}$, while $L = \frac{1}{\delta}$. The corresponding values are given in Table 4. Observe that the values of L in this case are similar to the ones for the standard L scheme, except for the smallest tolerance. Also, the L values increase for smaller TOL and smaller time steps τ, which was not the case for the standard L scheme.

	L-parameters for the HL-scheme		
TOL	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.0125$
1e−3	19	23	29
1e−4	40	50	62
1e−5	84	106	134

Table 4. The L parameters for the HL-scheme, computed for different values of TOL and τ. The total iteration error is guaranteed below TOL (see Remark 5).

The convergence results are given in Table 5. Since the L parameters have similar values for both L-type schemes, the number of iterations in both schemes is comparable whenever the standard L-scheme converges. However, for the HL-scheme, L can be chosen automatically, based on the required tolerance TOL and on the time step size τ. This leads to faster convergence rates, based on the theoretically results. Nevertheless, decreasing the tolerance TOL implies an increasing L, which deteriorates the convergence rate. However, the HL-scheme converged for all combinations of parameters.

	HL-iterations per time step		
TOL	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.0125$
1e−3	37.0	57.2	89.5
1e−4	120.4	202.5	338.2
1e−5	343.3	596.2	1057.4

Table 5. Results for the standard HL-scheme. The scheme converges for all values of TOL and all time steps τ.

When comparing the three schemes, it becomes clear that the Newton scheme requires the least number of iterations whenever it converges. On the other hand, the Newton scheme was the one which did not converge in the most of the cases considered here, so it is least robust. Also, the convergence criterion is not always met for the standard L-scheme due to regularisation. Both schemes require a regularisation step. Instead, no regularisation is needed for the HL-scheme. Clearly, it requires more iterations than the Newton scheme, but generally less than the standard L-scheme. Most important,

it displayed a robust behaviour, as it converged in all experiments. In fact, this convergence can be achieved for any tolerance TOL and time step τ.

It is worth mentioning that, next to the number of iterations, the total execution time is influenced by two factors: the time for solving the linear systems at each iteration, and the time for assembling the discretisation matrices. Among all three schemes, the Newton scheme is closest to generate ill conditioned matrices, if not singular. Therefore the linear solvers are more expensive than in the case of the L-type schemes. Moreover, the linear system needs to be reassembled completely every iteration, as the Jacobian depends on the current iteration, and involves many function evaluations. The L-type schemes behave better in this respect. For the example presented above, the emerging linear systems involve the discrete Laplacian and the discretisation of the identity operator multiplied by L. This not only generates better conditioned matrices, but these matrices remain unchanged for every iteration. In this case, a solver based on the LU-decomposition is an effective approach, as this decomposition needs to be performed only once.

6 Conclusion

We discuss iterative schemes for solving the fully discrete non-linear systems obtained by a backward Euler - lowest order Raviart-Thomas mixed finite element discretisation of a class of doubly degenerate parabolic problems. Appearing as models of practical relevance, the non-linear function involved in the model must be increasing and Hölder continuous, but may remain constant over intervals. In consequence, two kinds of degeneracy are allowed, slow and fast diffusion. This leads to fully discrete systems that have singular Jacobians, which brings difficulties in finding robust iterative solvers.

We present here an approach inspired by the L-scheme, which is suited for the case of Hölder continuous non-linearities. To apply the Newton scheme or the standard L-scheme in such a case, one needs to regularise first the problem, i.e. to approximate the non-linearity by a Lipschitz continuous one. This step is associated with additional errors. If highly accurate approximations of the exact, fully discrete solutions are needed, the regularisation step may be the cause of the fact that the convergence is very slow, if not impossible. The scheme discussed here makes no use of any regularisation. Instead, the parameter L is chosen not as the Lipschitz constant of the non-linearity, but in such a way that the error has a guaranteed decay below any chosen tolerance. We provide a rigorous proof for this decay, which also gives a practical way to choose the parameter L.

We present numerical experiments where we compare the behaviour of the three schemes: Newton, standard L, and the L-variant proposed here. As resulting from these experiments, the Newton scheme requires the least number of iterations, but is also the least robust of all as there were the most cases where it did not converge. The standard L-scheme is more robust, at

the expense of a high number of iterations. Also, convergence could not be achieved in all cases, in particular if the regularisation parameter is not in agreement with the required tolerance. The new scheme is improving these aspects: it shows convergence for any required tolerance, and any choice of the time step size. Nevertheless, an optimisation of the choice of L and possibly in combination with an optimal linear solver can make the proposed scheme an effective alternative to the traditional ones.

Acknowledgement

The research is partially supported by the Norwegian Research Council (NFR) through the NFR-DAAD grant 255715, the VISTA project AdaSim 6367 and the project Toppforsk 250223, Lab2Field 811716, by Statoil through the Akademia Grant and by the Research Foundation-Flanders (FWO) through the Odysseus programme (project G0G1316N).

References

1. Alt, H.W., Luckhaus, S., Quasilinear elliptic-parabolic differential equations, Math. Z. **183** (1983), 311–341.
2. Brezzi, F., Fortin, M., Mixed and Hybrid Finite Element Methods, Springer-Verlag, New York, 1991.
3. Arbogast, T., The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow, J. Nonlinear Anal. Theory Methods Appl. **19** (1992), 1009-1031.
4. Bergamashi, N., Putti, M., Mixed finite elements and Newton-type linearizations for the solution of Richards' equation, Internat. J. Numer. Meth. Engrg. **45** (1999), 1025-1046.
5. Brenner, K., Cances, C., Improving Newton's method performance by parametrization: the case of the Richards equation, SIAM J. Numer. Anal. 55, 1760–1785, 2017.
6. Celia, M., Bouloutas, E., Zarba, R., A general mass-conservative numerical solution for the unsaturated flow equation, Water Resour. Res. **26** (1990), 1483–1496.
7. Chen, Z., Degenerate two-phase incompressible flow. Existence, uniqueness and regularity of a weak solution, J. Differential Equations **171** (2001), 203-232.
8. Cherfils, L., Choquet, C., Diedhiou, M.M., Numerical validation of an upscaled sharp-diffuse interface model for stratified miscible flows, Math. Comput. Simulation **137** (2017), 246-265.
9. Douglas Jr., J., Roberts, J., Global estimates for mixed methods for second order elliptic problems, Math. Comp. **45** (1985), 39–52.
10. Farthing M.W., Ogden, F.L., Numerical solution of Richards equation: a review of advances and challenges, Soil Sci. Soc. Am. J. (2017), doi:10.2136/sssaj2017.02.0058

11. Jäger, W., Kačur, J., Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes, Math. Model. Num. Anal. **29** (1995), 605–627.
12. Lehmann, F., Ackerer, Ph., Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media, Transp. Porous Med. **31** (1998), 275–292.
13. List, F., Radu, F.A., A study on iterative methods for Richards' equation, Comput. Geosci. **20** (2016), 341–353.
14. Nochetto, R.H., Verdi, C., Approximation of degenerate parabolic problems using numerical integration, SIAM J. Numer. Anal. **25** (1988), 784–814.
15. Nordbotten, J.M., Celia, M.A., Geological Storage of CO2. Modeling Approaches for Large-Scale Simulation, John Wiley and Sons, Hokoben, New Jersey, 2012.
16. Otto, F., L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations, J. Differential Equations. **131** (1996), 20-38.
17. Paniconi, C., Putti, M., A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems, Water Resour. Res., **30** (1994), 3357–3374.
18. Park, E.J., Mixed finite elements for non-linear second-order elliptic problems, SIAM J. Numer. Anal. **32** (1995), 865–885.
19. Pop, I.S., Radu, F.A., Knabner, P., Mixed finite elements for the Richards' equations: linearization procedure, J. Comput. Appl. Math. **168** (2004), 365–373.
20. Radu, F.A., Pop, I.S., Knabner, P., *On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation*, Numerical Mathematics and Advanced Applications (A. Bermudez de Castro, D. Gomez, P. Quintela, P. Salgado, eds.), Springer Verlag, 2006, 1192–1200.
21. Radu, F.A., Pop, I.S., Knabner, P., Error estimates for a mixed finite element discretization of some degenerate parabolic equations, Numer. Math. **109** (2008), 285–311.
22. Radu, F.A., Kumar, K., Nordbotten, J.M., Pop, I.S., A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media, arHiv: 1512.08387 (2015).
23. Radu, F.A., Kumar, K., Nordbotten, J.M., Pop, I.S., A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities, IMA J. Numer. Anal. (2018), doi:10.1093/imanum/drx032.
24. Radu, F.A., Nordbotten, J.M., Pop, I.S., Kumar, K., A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media, J. Comput. Appl. Math. **289** (2015), 134–141.
25. Slodicka, M., 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), 1593–1614.

26. Temam, R., Navier-Stokes Equations: Theory and Numerical Analysis, AMS Chelsea Publishing, Providence, RI, 2001.
27. Yong, W.A., Pop, I.S., A numerical approach to porous medium equations, Preprint 95-50 (SFB 359), IWR, University of Heidelberg, 1996.