A modified L-Scheme to solve nonlinear diffusion problems

MITRA, Koondanibha & POP, Sorin (2019) A modified L-Scheme to solve nonlinear diffusion problems. In: COMPUTERS & MATHEMATICS WITH APPLICATIONS, 77(6), p. 1722-1738.

DOI: 10.1016/j.camwa.2018.09.042
Handle: http://hdl.handle.net/1942/27235

A modified L-Scheme to solve nonlinear diffusion problems

Koondanibha Mitra, Iuliu Sorin Pop

UHasselt Computational Mathematics Preprint
Nr. UP-18-06

Sept. 17, 2018

A modified L-Scheme to solve nonlinear diffusion problems

K. Mitra*[1,2] and I.S. Pop[2,3]

[1] *Eindhoven University of Technology, Department of Mathematics and Computer Science*
[2] *Hasselt University, Faculty of Science*
[3] *University of Bergen, Department of Mathematics*

Abstract

In this work, we propose a linearization technique for solving nonlinear elliptic partial differential equations that are obtained from the time-discretization of a wide variety of nonlinear parabolic problems. The scheme is inspired by the L-scheme, which gives unconditional convergence of the linear iterations. Here we take advantage of the fact that at a particular time step, the initial guess for the iterations can be taken as the solution of the previous time step. First it is shown for quasilinear equations that have linear diffusivity that the scheme always converges, irrespective of the time step size, the spatial discretization and the degeneracy of the associated functions. Moreover, it is shown that the convergence is linear with convergence rate proportional to the time step size. Next, for the general case it is shown that the scheme converges linearly if the time step size is smaller than a certain threshold which does not depend on the mesh size, and the convergence rate is proportional to the square root of the time step size. Finally numerical results are presented that show that the scheme is at least as fast as the modified Picard scheme, faster than the L-scheme and is more stable than the Newton or the Picard scheme.

1 Introduction

In this paper, a linearization technique is considered for the generalized nonlinear advection diffusion equations of the type

$$\partial_t b(u) + \nabla \cdot \vec{F}(\vec{x}, u) = \nabla \cdot [\mathcal{D}(\vec{x}, u)\nabla u] + f(\vec{x}, t, u); \tag{1.1}$$

*Corresponding author: email: k.mitra@tue.nl

completed with suitable boundary and initial conditions. Equation (1.1) appears as mathematical model for many real world applications, like flow through porous media or reactive transport. For the discretization in time, the backward Euler method is often used due to its stability. This changes (1.1) into a sequence of nonlinear elliptic equations. For solving these, linear iterative schemes are required.

A commonly used linearization technique is the Newton scheme (NS). Being quadratically convergent, it is widely used for solving nonlinear equations [2, 15]. However, this quadratic convergence is featured under certain restrictions. In particular, degenerate problems do not fulfill these restrictions [24]. Another drawback of the NS is that it is only locally convergent, meaning that the initial guess for the iterations have to be close enough to the actual solution for the scheme to be convergent [2, 24]. In many cases, this requires extremely small time step sizes limiting the applicability of the NS. For this reason, pre-conditioners, line-search methods and different parametrizations are often used to enhance the robustness of the NS [4, 10].

An alternative to the NS is a modified Picard scheme (PS), proposed in [5]. In [15,16] it is shown that this scheme is quite fast despite having linear convergence. Another linearly converging scheme is the Jäger-Kačur scheme (see [8, 9, 12]). A sufficient condition for convergence was derived in [24] for all the schemes mentioned above. When applied to the Euler implicit discretization of (1.1) one needs to take

$$\tau < C m_b^r h^d, \tag{1.2}$$

to guarantee the convergence of these schemes. Here τ is the time step size, h is the mesh size, d is the spatial dimension of the problem, $m_b \geq 0$ is the lower bound of b' and $C, r > 0$ are constants that depend on the nonlinear functions. For $d \geq 2$ this imposes a severe restriction on the time step size, which can increase the computation time to a great extent. Moreover, in the degenerate case $m = 0$, either a regularized version of the function b has to be used [8, 9, 17] or the initial/boundary data has to be shifted [21] to ensure convergence.

For porous media applications, where all the associated functions are nonlinear and the problem might become degenerate, stability is an important issue. Also, extremely large timescales are involved for such processes and so condition (1.2) cannot always be satisfied. To address this, a fixed point iteration scheme, termed L-scheme or simply LS in the context of this discussion, was proposed in [19,20,23]. The LS is linearly convergent but it has the interesting property of unconditional convergence, meaning that it converges to the time-discrete solution irrespective of the choice of initial guess. This is due to the fact that in the LS, the stabilization terms are estimated globally as opposed to the local estimations used in the NS, the PS or the Jäger-Kačur scheme. However, as shown in [16, 25], this increases the convergence rate of the LS, making it slower when compared to the NS or the PS. This has motivated authors to either use the LS to provide initial guesses for the NS [16] or to apply it in a domain decomposition type approach that boosts the speed of the LS [25].

All of the schemes mentioned above are mostly designed to solve nonlinear elliptic problems and do not use the fact that these are the outcome of the time discretization of a nonlinear parabolic problem. In this context, the solution from the previous time step can be used as initial guess in the iterative process. For solutions that have a good regularity in time, as being the case for parabolic problems, the changes in the solution from one time step to the next one are limited. However the standard schemes do not use this fact and thus their implementation for parabolic and elliptic problems are more or less the same.

The main idea in this work is to exploit the fact that the nonlinear elliptic problems are the result of the time discretization of (1.1). The proposed scheme is a combination of the PS and the LS and it uses local estimations to improve the convergence behaviour of the LS without affecting its stability. After introducing the problem in Section 2 in Section 3 we propose the iterative scheme for simple quasilinear equations, and analyze its convergence. It can be observed that, apart from being unconditionally convergent, the scheme has a convergence rate proportional to the time step size for sufficiently small time steps. This is unlike the LS, where a lower time step increases the convergence rate. Section 4 generalizes these ideas for (1.1). The scheme converges also in the degenerate case, however for small enough time step sizes. For the non-degenerate case, the convergence is linear and the rate is proportional to the square root of the time step size. Finally in Section 5 some numerical experiments are presented. These show that the scheme is more stable than the NS or the PS and converges at least as fast as the PS and sometimes comparable to the NS.

2 The general problem and linearization schemes

Let Ω be a bounded domain in $\mathbb{R}^d$ which has a Lipschitz boundary $\partial\Omega$ and define $Q = \Omega \times [0, T)$ for some $T > 0$. For the rest of our discussion $(\cdot, \cdot)$ and $\left\|\cdot\right\|$ will represent $L^2(\Omega)$ inner product and norm. Any other norm will be presented as $\left\|\cdot\right\|_V$ with V being the corresponding space. The Sobolev space $W^{k,p}(\Omega)$ is the set of functions u defined on Ω such that $D^k u \in L^p(\Omega)$ for the multi-index k, equipped with the norm $\left\|u\right\|_{W^{k,p}(\Omega)} = (\sum_{q\leq k} \int_\Omega |D^q u|^p)^{1/p}$ for $1 \leq p < \infty$ [6]. Further, $H^k(\Omega) = W^{k,2}(\Omega)$ and $H^k_0(\Omega)$ represents the set of elements of $H^k(\Omega)$ which have 0 trace at the boundary $\partial\Omega$ [6].

The Hölder space $\mathcal{C}^{\ell,\delta}(\Omega)$ refers to the space of δ-Hölder continuous functions upto the ℓ^{th} space derivative for the metric $\text{dist}(\vec{x}, \vec{y}) = |\vec{x} - \vec{y}|$, $\vec{x}, \vec{y} \in \mathbb{R}^d$. The associated norms of these spaces are defined in detail in [13]. With these basic definitions stated, we introduce the problem.

2.1 Time-discrete formulation and properties of functions

We consider (1.1) in the space-time domain Q. For simplicity homogeneous Dirichlet boundary condition is assumed on $\partial\Omega \times [0,T]$. Further, let $u_0(\cdot) \in H_0^1(\Omega)$ be the initial condition. We refer to [1] and [18] for results on the existence and uniqueness of weak solutions to (1.1) for this case.

For the time discretization of (1.1) we let $\tau = T/N$, $N \in \mathbb{N}$, be the time step size and $n \in \{1,..,N\}$ represent the time step. Define $t_n = n\tau$. The Euler implicit discretization of (1.1) leads to the sequence of elliptic problems ($n \in \{1,..,N\}$)

$$(\mathcal{P}1)\begin{cases} b(u_n) - b(u_{n-1}) + \tau\nabla\cdot\vec{F}(\vec{x},u_n) = \tau\nabla\cdot[\mathcal{D}(\vec{x},u_n)\nabla u_n] + \tau f(\vec{x},t_n,u_n) \text{ in } \Omega, & (2.1)\\ u_n = 0 \text{ on } \partial\Omega. & (2.2)\end{cases}$$

Below we consider weak solutions to the time discrete problem ($\mathcal{P}1$), defined as:

Definition 0.1. Let $n \in \{1,..,N\}$ and $u_{n-1} \in L^2(\Omega)$ be given. A weak solution to Problem ($\mathcal{P}1$) is a function $u_n \in H_0^1(\Omega)$ satisfying for any test function $\phi \in H_0^1(\Omega)$

$$(b(u_n) - b(u_{n-1}),\phi) + \tau(\mathcal{D}(\vec{x},u_n)\nabla u_n, \nabla\phi) = \tau(\vec{F}(\vec{x},u_n),\nabla\phi) + \tau(f(\vec{x},t_n,u_n),\phi). \quad (2.3)$$

The existence of $u_n \in H_0^1(\Omega)$ for a given $u_{n-1} \in L^2(\Omega)$ is shown in [20]. Note that the sequence $\{u_n\}_{n=1}^N$ can be used to approximate the solution to the original parabolic problem (1.1) (the Rothe method, [11]).

Below we assume the following:

(P. 1) $b : \mathbb{R} \to \mathbb{R}$ is a $\mathcal{C}^2(\mathbb{R})$ function such that $b' \geq m_b$ and $|b'|, |b''| \leq M_b$ for some $m_b, M_b \geq 0$.

(P. 2) $f : Q \times \mathbb{R} \to \mathbb{R}$ is twice differentiable with respect to u and satisfies $T\partial_u f(\vec{x},t,u) \leq b'(u)$ and $|f|, |\partial_u f|, |\partial_{uu} f| \leq M_f$ ($M_f \geq 0$) a.e. in $(\vec{x},t) \in Q$ and $u \in \mathbb{R}$.

(P. 3) $\mathcal{D} : \Omega \times \mathbb{R} \to \mathbb{R}^+$ is a $\mathcal{C}^2(\Omega \times \mathbb{R})$ function. Moreover, there exists $\mathcal{D}_m, \mathcal{D}_M > 0$ such that $\mathcal{D}_m \leq \mathcal{D}$ and $\mathcal{D}, |\partial_u^q \mathcal{D}|, |\partial_{x_j}\mathcal{D}|, |\partial_u\partial_{x_j}\mathcal{D}| \leq \mathcal{D}_M$ for $j \in \{1,..,d\}$, $q \in \{1,2\}$.

(P. 4) $\vec{F} : \Omega \times \mathbb{R} \to \mathbb{R}^d$, with x_j-component denoted by F_j, admits partial derivatives that satisfy $|\partial_u^q F_j|, |\partial_{x_j} F_j|, |\partial_u\partial_{x_j} F_j| \leq M_F$ for $j \in \{1,..,d\}$, $q \in \{1,2\}$ and $M_F \geq 0$.

(P. 5) $u_0 \in H_0^1(\Omega)$. In some cases we also use $u_0 \in W^{2,\infty}(\Omega)$.

In (P. 2) instead of using $\partial_u f \leq 0$ we have used the less restrictive condition $b' \geq T\partial_u f$. At this stage we define the following function which will be used later,

$$z : \Omega \times \mathbb{R} \to \mathbb{R}, \quad z(\vec{x},u) = b(u) - \tau f(\vec{x},t_n,u). \quad (2.4)$$

Observe that actually z depends on the time step size τ and n as well. However the focus here is on constructing an iterative scheme for a fixed time step size τ at a fixed time t_n. Hence, we disregard this dependence. From (P. 1) and (P. 2), $z \in \mathcal{C}^2$ and satisfies $\partial_u z = b' - \tau \partial_u f \geq b'\left(1 - \frac{\tau}{T}\right) \geq m_b\left(1 - \frac{\tau}{T}\right) \geq 0$. So by defining $m = m_b(1 - \frac{\tau}{T})$ we get the inequality

$$\partial_u z \geq m \geq 0 \text{ a.e. for } \vec{x} \in \Omega \text{ and } u \in \mathbb{R}. \tag{2.5}$$

Remark 2.1 (On the properties of the functions)**.** *The situation $m_b = 0$ in (P. 1) gives rise to degeneracy, which will be discussed in detail later. The boundedness of b'' is assumed in our analysis. This is a more severe condition compared to the Lipschitz continuity assumed in [16, 19, 20, 23, 26] and the Hölder continuity assumed in [22]. However, it is true for many porous media flow models, where the form of b is as in* (5.6)*. $\mathcal{D}$ can also be a matrix $\tilde{\mathcal{D}}$, the only constraint being that the condition (P. 3) is satisfied for all the components $\tilde{\mathcal{D}}_{ij}$ and that $\tilde{\mathcal{D}}$ is positive definite. The bounds assumed in (P. 4) guarantee that $\vec{F}$, which corresponds to the 'flux' in physics, is bounded and varies smoothly within Ω.*

Remark 2.2 (Boundary conditions)**.** *For the ease of presentation a zero Dirichlet boundary condition is assumed at the boundary, but the results can be proved for more general boundary conditions, including non-homogeneous Dirichlet, Neumann, Robin or mixed type ones.*

2.2 Standard linearization techniques

For resolving the nonlinear terms in problem ($\mathcal{P}1$), Newton scheme (NS) uses iterations where the values of the nonlinear terms in the current iteration are approximated by Taylor series expansion. For $i \in \mathbb{N}$ let u_n^i stand for the i^{th} iterate at the n^{th} time step and let $\delta u_n^i = u_n^i - u_n^{i-1}$ for $i > 0$. If ϕ_n^i is one of the functions $b, f, z, \mathcal{D}, F_j$ for $u = u_n^i$ and $t = t_n$, then one takes $\phi_n^i \approx \phi_n^{i-1} + \partial_u \phi_n^{i-1} \delta u_n^i$, $\partial_u \phi_n^i$ being the partial derivative of ϕ with respect to u at $u = u_n^i$. With this substitution and given u_n^{i-1}, NS resumes to find δu_n^i that solves

$$(\mathcal{P}_N^{n,i}) \begin{cases} (b'(u_n^{i-1}) - \tau \partial_u f_n^{i-1}) \delta u_n^i - \tau \nabla \cdot \left[\mathcal{D}_n^{i-1} \nabla \delta u_n^i + (\partial_u \mathcal{D}_n^{i-1} \nabla u_n^{i-1} - \partial_u \vec{F}_n^{i-1}) \delta u_n^i \right] \\ = -(b(u_n^{i-1}) - b(u_{n-1})) + \tau \left[\nabla \cdot (\mathcal{D}_n^{i-1} \nabla u_n^{i-1}) - \nabla \cdot \vec{F}_n^{i-1} + f_n^{i-1} \right] \text{ in } \Omega, & (2.6) \\ \delta u_n^i = 0 \text{ on } \partial\Omega. & (2.7) \end{cases}$$

For Problem ($\mathcal{P}_N^{n,i}$), one of the natural initial guesses is $u_n^0 = u_{n-1}$.

The modified Picard scheme (PS) can be interprated as a simplified version of the NS [5]. Here the Taylor expansion is used only for the function $z = b - \tau f$. For $i \in \mathbb{N}$

and with given u_n^{i-1} one seeks u_n^i solving

$$(\mathcal{P}_P^{n,i}) \begin{cases} \partial_u z_n^{i-1} u_n^i - \tau \nabla \cdot \left[\mathcal{D}_n^{i-1} \nabla u_n^i\right] \\ = \partial_u z_n^{i-1} u_n^{i-1} - (b(u_n^{i-1}) - b(u_{n-1})) + \tau[-\nabla \cdot \vec{F}_n^{i-1} + f_n^{i-1}] \text{ in } \Omega, & (2.8) \\ u_n^i = 0 \text{ on } \partial\Omega. & (2.9) \end{cases}$$

The computational cost per iteration of the PS is lower than for the NS, because the number of gradients to be calculated is less. However, the scheme is linearly convergent and also only conditionally stable. A convergence proof can be found in [24].

The PS takes advantage of the fact that the function z is increasing with respect to u. This insight is taken a step further in the L-scheme (LS). For a $L \geq \max\limits_{u\in\mathbb{R},(\vec{x},t)\in Q}\{\partial_u z\} = \max\limits_{u\in\mathbb{R},(\vec{x},t)\in Q}\{|b'(u) - \tau\partial_u f(\vec{x},t,u)|\}$ and a given u_n^{i-1} the scheme reads (see [19, 20, 23])

$$(\mathcal{P}_L^{n,i}) \begin{cases} L u_n^i - \tau \nabla \cdot \left[\mathcal{D}_n^{i-1} \nabla u_n^i\right] \\ = L u_n^{i-1} - (b(u_n^{i-1}) - b(u_{n-1})) + \tau[-\nabla \cdot \vec{F}_n^{i-1} + f_n^{i-1}] \text{ in } \Omega, & (2.10) \\ u_n^i = 0 \text{ on } \partial\Omega. & (2.11) \end{cases}$$

The difference between the LS and the PS is that, instead of using the factor $\partial_u z_n^{i-1}$, the former uses an upper bound of it, L. This can affect the convergence order [16, 25], but transforms the scheme to a H^1-contraction which converges irrespective of the spatial discretization or the initial guess. Convergence can also be achieved for smaller values of L, e.g. for $L = \frac{1}{2}\max\limits_{u\in\mathbb{R},(\vec{x},t)\in Q}\{\partial_u z\}$ convergence is guranteed (see [16, 25]) although the contraction property can not be ensured in this case. In the next section we show how to improve the speed of the LS while preserving this stability.

Observe that each of the problems, $(\mathcal{P}_N^{n,i})$, $(\mathcal{P}_P^{n,i})$ and $(\mathcal{P}_L^{n,i})$, has unique solution in $H_0^1(\Omega)$ as they are linear and coercive. Also, for the schemes mentioned above, if $u_n^i \to u_n$ in strong sense in $H_0^1(\Omega)$ then u_n is a weak solution of $(\mathcal{P}_1)$ in the sense of Definition 0.1.

For the convergence analysis we let e_n^i denote the error of the iterate at the n^{th} time step,

$$e_n^i = u_n^i - u_n. \tag{2.12}$$

The convergence rate (linear) of the schemes is defined as

$$\alpha = \lim_{i\to\infty} \sup \frac{\left\|e_n^i\right\|_{H^1(\Omega)}}{\left\|e_n^{i-1}\right\|_{H^1(\Omega)}}. \tag{2.13}$$

Notice that $\alpha < 1$ implies convergence. Also worth pointing out is that the H^1-norm has been used to define the convergence rate in (2.13) but the convergence rate will roughly be the same if defined with respect to the L^2 norm [16, 19]. Hence, in Section 5 the L^2-norm has been used to estimate α.

Remark 2.3. *The convergence rate of the LS has the form $\alpha = (L-m)/(L+C\tau) < 1$, for some $C > 0$ and $L > m$ (see [16, 19, 20, 26]). However, for large L, or small τ or m, α approaches 1. This leads to extremely slow convergence of the L-scheme. This issue has been reported for a variety of problems in literature [3, 16, 25].*

The proofs given below make use of the following notations. $[\cdot]_+ = \max(\cdot, 0)$ and $[\cdot]_- = \min(\cdot, 0)$ are the positive and negative cut functions respectively. For any $a, b \in \mathbb{R}$ the interval $\mathcal{I}(a, b)$ is defined as

$$\mathcal{I}(a, b) = \{x : \min(a, b) < x < \max(a, b)\}.$$

We will further use Young's inequality: for $a, b \in \mathbb{R}$ and $\rho > 0$ one has

$$ab \leq \frac{1}{2\rho}a^2 + \frac{\rho}{2}b^2. \tag{2.14}$$

Finally, C_Ω is the constant appearing in the Poincaré inequality.

3 The modified L-scheme

In what follows we discuss a modified form of the L-scheme that preserves its stability property while having an improved convergence rate. The idea is to replace the constant L with a function $L_n^i : \Omega \to \mathbb{R}^+$ defined at each iteration. This leads to the

Modified L-scheme (MS). *Let $n \in \{1, ..., N\}$ and $i \in \mathbb{N}$ be fixed and assume that $u_{n-1}, u_n^{i-1} \in H_0^1(\Omega)$ are given. Find $u_n^i \in H_0^1(\Omega)$ satisfying*

$$\begin{aligned}&(L_n^i u_n^i, \phi) + \tau(\mathcal{D}(\vec{x}, u_n^{i-1})\nabla u_n^i, \nabla\phi)\\ &= (L_n^i u_n^{i-1}, \phi) - (b(u_n^{i-1}) - b(u_{n-1}), \phi) + \tau(\vec{F}(\vec{x}, u_n^{i-1}), \nabla\phi) + \tau(f(\vec{x}, t_n, u_n^{i-1}), \phi),\end{aligned} \tag{3.1}$$

for all $\phi \in H_0^1(\Omega)$, where $L_n^i : \Omega \to \mathbb{R}$ is defined as

$$L_n^i(\vec{x}) = \max\{[b'(u_n^{i-1}(\vec{x})) - \tau\partial_u f(\vec{x}, t_n, u_n^{i-1}(\vec{x})) + \mathfrak{M}\tau], 2\mathfrak{M}\tau\}. \tag{3.2}$$

Here $u_n^0 = u_{n-1}$, and $\mathfrak{M}$ is a positive constant that will be specified later. Observe that $\mathfrak{M} = 0$ corresponds to the PS, and disregarding the $b' - \tau\partial_u f$ term in the definition of L_n^i leads to the LS. In this sense the scheme is a combination of the PS and the LS. Below we show that it inherits the qualities of both schemes. The convergence results and estimates are first obtained for a simple version of $(\mathcal{P}1)$ where $\partial_u\mathcal{D} \equiv 0$ and $\partial_u F_j \equiv 0$ and then for the general problem.

3.1 Quasilinear equations with linear diffusivity and flux

To analyze the scheme we first look at the simpler quasilinear parabolic equation,

$$\partial_t b(u) + \nabla \cdot \vec{F}(\vec{x}) = \nabla \cdot (\mathcal{D}(\vec{x})\nabla u) + f \tag{3.3}$$

which corresponds to $\partial_u \mathcal{D} = 0$ and $\partial_u F_j = 0$ in (1.1). Problems where $\partial_u \mathcal{D} \neq 0$ can also be reduced to this case if $\mathcal{D}$ is separable in the variables, i.e. $\mathcal{D}(\vec{x}, u) = \mathcal{D}_1(\vec{x})\mathcal{D}_2(u)$. By (P. 3), $\mathcal{D}_2 > 0$ and by using the Kirchoff transform $U = \int^u \mathcal{D}_2(\varrho)d\varrho$ (see [1, 19, 27]) one obtains $\nabla \cdot [\mathcal{D}(\vec{x}, u)\nabla u] = \nabla \cdot (\mathcal{D}_1(\vec{x})\nabla U)$. Using U as the primary unknown one gets an equation similar to (3.3). Well-known examples, such as the porous medium equation or the Richards equation, can be reduced to this form.

Recalling the homogeneous boundary conditions, a weak solution to the time discrete version of (3.3) satisfies

$$(b(u_n) - b(u_{n-1}), \phi) + \tau(\mathcal{D}\nabla u_n, \nabla\phi) = \tau(f_n, \phi) + \tau(\vec{F}, \nabla\phi), \tag{3.4}$$

for all $\phi \in H_0^1(\Omega)$. For the subsequent analysis we assume the following:

(A. 1) There exists a $\Lambda > 0$ such that $\|u_n - u_{n-1}\|_{L^\infty(\Omega)} \leq \Lambda\tau$ for all $n \in \mathbb{N}$.

Note that since u_n is, in fact, the time discrete approximation of the solution to the parabolic problem (3.3), Assumption (A. 1) is similar to saying that $\partial_t u \in L^\infty(Q)$. Sufficient conditions for boundedness of $\partial_t u$ in the $L^\infty(\Omega)$-norm can be found in [13, Chapter 5]. Below we present a result that shows the validity of Assumption (A. 1), in this sense.

Proposition 3.1. *For a fixed $n \in \{1, .., N\}$ let u_n solve (3.4) and assume that $m > 0$ and $\nabla \cdot (\mathcal{D}\nabla u_{n-1}) \in L^\infty(\Omega)$. Then $\nabla \cdot (\mathcal{D}\nabla u_n) \in L^\infty(\Omega)$ and a $\Lambda \geq 0$, independent of τ, exists such that $\|u_n - u_{n-1}\|_{L^\infty(\Omega)} \leq \Lambda\tau$.*

The proof is given in Appendix A. Observe that Proposition 3.1 is valid if $u_0 \in W^{2,\infty}(\Omega)$, which extends then to all time steps. Based on Assumption (A. 1) we have

Lemma 3.1. *Assume (A. 1) and that L_n^i satisfies a.e. in Ω*

$$L_n^i \geq \sup\{|b'(\zeta) - \tau\partial_u f(\vec{x}, t_n, \zeta)| : \zeta \in \mathcal{I}(u_n^{i-1}, u_n)\}.$$

Then $\|u_n^i - u_n\|_{L^\infty(\Omega)} \leq \Lambda\tau$ for all $i \in \mathbb{N}$, u_n^i being the solution of (3.1).

Before giving the proof we observe that the result is quite general with respect to the choice of the function L_n^i. However, it will be used for L_n^i either constant or as in (3.2), which corresponds to the LS and the MS.

Proof. The proof is by induction. Assumption (A. 1) guarantees that the assertion holds for $i = 0$. Let the assertion hold for u_n^{i-1}, i.e. $\|u_n^{i-1} - u_n\|_{L^\infty(\Omega)} = \|e_n^{i-1}\|_{L^\infty(\Omega)} \leq \Lambda\tau$. We prove that this implies $\|u_n^i - u_n\|_{L^\infty(\Omega)} = \|e_n^i\|_{L^\infty(\Omega)} \leq \Lambda\tau$.

As $\partial_u \mathcal{D}, \partial_u F_j = 0$, subtracting (3.4) from (3.1), a $\zeta \in \mathcal{I}(u_n^{i-1}, u_n)$ exists such that

$$\begin{aligned}&(L_n^i e_n^i, \phi) + \tau(\mathcal{D}\nabla e_n^i, \nabla\phi) = (L_n^i e_n^{i-1}, \phi) - ((b(u_n^{i-1}) - b(u_n)), \phi) + \tau(f_n^{i-1} - f_n, \phi)\\ &= ((L_n^i - \partial_u z(\zeta))e_n^{i-1}, \phi). \end{aligned} \tag{3.5}$$

Here we used the definition of z from (2.4). With $\phi = [e_n^i - \Lambda\tau]_+$ one gets

$$\begin{aligned}&(L_n^i[e_n^i - \Lambda\tau], [e_n^i - \Lambda\tau]_+) + \Lambda\tau(L_n^i, [e_n^i - \Lambda\tau]_+) + \tau\mathcal{D}_m\big\|\nabla[e_n^i - \Lambda\tau]_+\big\|^2\\ &\leq (L_n^i e_n^i, [e_n^i - \Lambda\tau]_+) + \tau(\mathcal{D}\nabla e_n^i, \nabla[e_n^i - \Lambda\tau]_+) = ((L_n^i - \partial_u z(\zeta))e_n^{i-1}, [e_n^i - \Lambda\tau]_+)\\ &\leq \int_\Omega |L_n^i - \partial_u z(\zeta)||e_n^{i-1}|[e_n^i - \Lambda\tau]_+ \leq \int_\Omega (L_n^i - m)\Lambda\tau[e_n^i - \Lambda\tau]_+ \leq \Lambda\tau(L_n^i, [e_n^i - \Lambda\tau]_+).\end{aligned}$$

In the last estimates we used the inequalities $|e_n^{i-1}| \leq \Lambda\tau$ a.e. and $0 \leq (L_n^i - \partial_u z(\zeta)) \leq L_n^i - m$ for $\zeta \in \mathcal{I}(u_n^{i-1}, u_n)$. Canceling the common terms in both sides gives

$$(L_n^i, [e_n^i - \Lambda\tau]_+^2) + \tau\mathcal{D}_m\big\|\nabla[e_n^i - \Lambda\tau]_+\big\|^2 \leq 0,$$

which implies that $e_n^i < \Lambda\tau$ a.e. in Ω. Similarly taking $\phi = [e_n^i + \Lambda\tau]_-$ one gets that $e_n^i > -\Lambda\tau$ a.e. which concludes the proof. □

For the LS, the condition stated in Lemma 3.1 is satisfied as $L_n^i = L \geq \sup\{|\partial_u z(\vec{x}, \zeta)| : \zeta \in \mathbb{R}\}$. However this leads to an overestimation of L at most points. Below we show that this estimation is improved significantly for the MS, resulting in much better convergence rates.

Theorem 3.1. *Under the Assumptions (A. 1) and(P. 1)-(P. 5) and for* $\mathfrak{M}_0 = \Lambda \max_{u\in\mathbb{R}}\{|b''| + \tau|\partial_{uu} f|\} \geq 0$*, the MS for equation* (3.4) *converges linearly in* $H_0^1(\Omega)$ *for all* $\mathfrak{M} \geq \mathfrak{M}_0$ *and* $\tau > 0$*. More precisely, for* $\mathfrak{M} \geq \mathfrak{M}_0$ *and* $\tau > 0$ *the limit* $u_n = \lim_{i\to\infty} u_n^i$ *exists and is a solution to* (3.4)*, whereas for the convergence rate* α *in* (2.13) *one has* $\alpha < 1$*. Moreover if* $m > 0$ *(the non-degenerate case) then* $\alpha = \mathcal{O}(\tau)$ *for* τ *small enough.*

Proof. By (P. 1) and (P. 2), $\mathfrak{M}_0$ is well defined. Observe that, for any $\zeta \in \mathcal{I}(u_n^{i-1}, u_n)$ there exists a $\zeta_1 \in \mathcal{I}(u_n^{i-1}, \zeta)$ such that

$$\begin{aligned}&|(b'(u_n^{i-1}) - \tau\partial_u f(\vec{x}, t_n, u_n^{i-1})) - (b'(\zeta) - \tau\partial_u f(\vec{x}, t_n, \zeta))| = |\partial_u z(u_n^{i-1}) - \partial_u z(\zeta)|\\ &= |\partial_{uu} z(\zeta_1)||(u_n^{i-1} - \zeta)| \leq \max_{u\in\mathbb{R}}\{|b''| + \tau|\partial_{uu} f|\}\Lambda\tau = \mathfrak{M}_0\tau.\end{aligned}$$

This implies that if $\mathfrak{M} \geq \mathfrak{M}_0$ and $L_n^i = \partial_u z(u_n^{i-1}) + \mathfrak{M}\tau$, then $L_n^i - \partial_u z(\zeta) \geq 0$. Moreover, if $L_n^i = 2\mathfrak{M}\tau$ then $\partial_u z(u_n^{i-1}) \leq \mathfrak{M}\tau$, which means that $\partial_u z(\zeta) \leq \partial_u z(u_n^{i-1}) + \mathfrak{M}_0\tau \leq 2\mathfrak{M}\tau$, giving $L_n^i - \partial_u z(\zeta) \geq 0$. Hence, for $\mathfrak{M} \geq \mathfrak{M}_0$ one has

$$L_n^i - b'(\zeta) + \tau\partial_u f(\vec{x}, t_n, \zeta) \geq 0,$$

which by Lemma 3.1 implies that $\|e_n^i\|_{L^\infty(\Omega)} < \Lambda\tau$ for all $i \in \mathbb{N}$.

Using similar arguments, if $L_n^i = \partial_u z_n^{i-1} + \mathfrak{M}\tau$ and $\mathfrak{M} \geq \mathfrak{M}_0$ then one gets

$$L_n^i - b'(\zeta) + \tau\partial_u f(\vec{x}, t_n, \zeta) = \partial_{uu} z(\zeta_1)(u_n^{i-1} - \zeta) + \mathfrak{M}\tau \leq \Lambda\tau|\partial_{uu} z(\zeta_1)| + \mathfrak{M}\tau \leq 2\mathfrak{M}\tau.$$

If $L_n^i = 2\mathfrak{M}\tau$ then $L_n^i - \partial_u z(\zeta) \leq L_n^i - m \leq 2\mathfrak{M}\tau$. Combining both gives

$$0 \leq L_n^i - b'(\zeta) + \tau\partial_u f(\vec{x}, t_n, \zeta) \leq 2\mathfrak{M}\tau, \tag{3.6}$$

or, more strongly,

$$0 \leq (\mathfrak{M} - \mathfrak{M}_0)\tau \leq L_n^i - b'(\zeta) + \tau\partial_u f(\vec{x}, t_n, \zeta) \leq 2\mathfrak{M}\tau. \tag{3.7}$$

These two inequalities will be used repeatedly in the proofs that follow.

First we prove the convergence of the scheme in $H_0^1(\Omega)$. Taking $\phi = e_n^i$ in (3.5) yields

$$\begin{aligned} &2\mathfrak{M}\tau\left\|e_n^i\right\|^2 + \tau\mathcal{D}_m\left\|\nabla e_n^i\right\|^2 \leq (L_n^i e_n^i, e_n^i) + \tau(\mathcal{D}\nabla e_n^i, \nabla e_n^i) \\ &= ((L_n^i - \partial_u z(\zeta))e_n^{i-1}, e_n^i) \leq 2\mathfrak{M}\tau(|e_n^{i-1}|, |e_n^i|) \leq \mathfrak{M}\tau\left\|e_n^{i-1}\right\|^2 + \mathfrak{M}\tau\left\|e_n^i\right\|^2. \end{aligned} \tag{3.8}$$

Cancelling common terms on both sides and applying the Poincaré inequality one gets

$$\left(\mathfrak{M} + \frac{\mathcal{D}_m}{2}C_\Omega\right)\left\|e_n^i\right\|^2 + \frac{\mathcal{D}_m}{2}\left\|\nabla e_n^i\right\|^2 \leq \mathfrak{M}\left\|e_n^{i-1}\right\|^2,$$

which gives

$$\left\|e_n^i\right\|^2 + \frac{\mathcal{D}_m}{(2\mathfrak{M} + C_\Omega\mathcal{D}_m)}\left\|\nabla e_n^i\right\|^2 \leq \frac{2\mathfrak{M}}{(2\mathfrak{M} + C_\Omega\mathcal{D}_m)}\left(\left\|e_n^{i-1}\right\|^2 + \frac{\mathcal{D}_m}{(2\mathfrak{M} + C_\Omega\mathcal{D}_m)}\left\|\nabla e_n^{i-1}\right\|^2\right).$$

As $\left(\left\|u\right\|^2 + \frac{\mathcal{D}_m}{(2\mathfrak{M}+C_\Omega\mathcal{D}_m)}\left\|\nabla u\right\|^2\right)^{1/2}$ is equivalent to the $H^1(\Omega)$-norm it follows that $\{u_n^i\}_{i\in\mathbb{N}}$ converges in $H^1(\Omega)$. The convergence rate in this equivalent norm is

$$\alpha = \sqrt{2\mathfrak{M}/(2\mathfrak{M} + C_\Omega\mathcal{D}_m)} < 1. \tag{3.9}$$

Observe that the convergence does not depend on the spatial discretization and also holds in the degenerate case when $\partial_u z$ may vanish.

In the non-degenerate case, when $\partial_u z \geq m > 0$, substituting $\phi = e_n^i$ in (3.5) one gets

$$\begin{aligned} &(m + \mathfrak{M}\tau)\left\|e_n^i\right\|^2 + \tau\mathcal{D}_m\left\|\nabla e_n^i\right\|^2 \leq (L_n^i e_n^i, e_n^i) + \mathcal{D}_m\tau\left\|\nabla e_n^i\right\|^2 \\ &= ((L - \partial_u z(\zeta))e_n^{i-1}, e_n^i) \leq 2\mathfrak{M}\tau(|e_n^{i-1}|, |e_n^i|) \leq \frac{2\mathfrak{M}^2\tau^2}{m}\left\|e_n^{i-1}\right\|^2 + \frac{m}{2}\left\|e_n^i\right\|^2. \end{aligned}$$

Canceling the common term in both sides we obtain

$$\left\|e_n^i\right\|^2 + \frac{2\tau\mathcal{D}_m}{(m+2\mathfrak{M}\tau)}\left\|\nabla e_n^i\right\|^2 \leq \frac{4\mathfrak{M}^2\tau^2}{m(m+2\mathfrak{M}\tau)}\left(\left\|e_n^{i-1}\right\|^2 + \frac{2\tau\mathcal{D}_m}{(m+2\mathfrak{M}\tau)}\left\|\nabla e_n^{i-1}\right\|^2\right). \tag{3.10}$$

With τ small enough, one obtains as before, the convergence of u_n^i in $H^1(\Omega)$. The convergence rate is

$$\alpha = \frac{2\mathfrak{M}\tau}{\sqrt{m(m+2\mathfrak{M}\tau)}} < \frac{2\mathfrak{M}\tau}{m}. \tag{3.11}$$

which is less than 1 for $\tau < \frac{m}{2\mathfrak{M}}$. One can also use the inequality $2\mathfrak{M}\tau(|e_n^{i-1}|, |e_n^i|) < \mathfrak{M}\tau(\left\|e_n^{i-1}\right\|^2 + \left\|e_n^i\right\|^2)$ to prove contraction with respect to a different $H^1(\Omega)$-norm with $\alpha = \sqrt{\frac{\mathfrak{M}\tau}{m}}$. Hence, the actual convergence rate is

$$\alpha = \min\left(\frac{2\mathfrak{M}\tau}{\sqrt{m(m+2\mathfrak{M}\tau)}}, \sqrt{\frac{\mathfrak{M}\tau}{m}}, \sqrt{\frac{2\mathfrak{M}}{2\mathfrak{M}+C_\Omega\mathcal{D}_m}}\right). \tag{3.12}$$

□

We conclude this section with a result that shows that for the non-degenerate case even the $L^\infty(\Omega)$ errors, i.e. $\|e_n^i\|_{L^\infty(\Omega)}$, decrease linearly for τ sufficiently small.

Proposition 3.2. *For a fixed $n \in \{1, .., N\}$, let $\{u_n^i\}_{i\in\mathbb{N}}$ be the sequence of functions resulting from the modified L-scheme for* (3.4)*. Assume (A. 1) and (P. 1)-(P. 5). If $m > 0$, then for small enough τ and $\mathfrak{M} \geq \mathfrak{M}_0$,*

$$\left\|u_n^i - u_n\right\|_{L^\infty(\Omega)} < \beta\left\|u_n^{i-1} - u_n\right\|_{L^\infty(\Omega)},$$

for $i \in \mathbb{N}$ and a constant $\beta \in (0, 1)$. Moreover, $\beta = \mathcal{O}(\tau)$.

The proof is given in Appendix A.

Remark 3.1 (Robustness of the MS)**.** *Theorem 3.1 shows that the MS converges irrespective of the spatial discretization and the time step. It converges in the degenerate case $m = 0$ too. Unlike the LS, the convergence rate α is independent of M_b, and scales with τ for small τ. This robustness is extremely helpful for computations, as it is difficult to satisfy condition* (1.2)*, which guarantees the convergence of schemes like NS, PS and Jäger-Kačur, for higher dimensional computations, i.e. $d > 1$. Moreover, the $\alpha = \mathcal{O}(\tau)$ property makes the scheme faster as the time step size is made smaller.*

4 General nonlinear diffusion equation

For the general problem, when $\vec{F}$ and $\mathcal{D}$ are functions of u, convergence can be shown for both the LS and the MS when τ is small enough. For both schemes, the convergence rate does not depend on the spatial discretization. Moreover, for the MS the convergence rate scales with $\sqrt{\tau}$ for small τ values.

For proving the main theorem of this section, we assume

(A. 2) $\|\nabla u_n\|_{L^\infty(\Omega)} \leq \Lambda_1$ for all $n \in \{1, .., N\}$ and some $\Lambda_1 > 0$.

Due to (P. 3), this is equivalent to assuming that the flux is bounded.

Similar to Section 3, the MS works if for all $i \in \mathbb{N}$ one has

$$\|u_n^i - u_n\|_{L^\infty(\Omega)} \leq \Lambda\tau. \tag{4.1}$$

Though non-trivial, this condition is fulfilled under certain assumptions, as follows from

Lemma 4.1. *Let $n \in \{1, .., N\}$ be fixed and Ω be a $\mathcal{C}^2$ domain. Assume (A. 1), $m > 0$ and let $\Lambda_2 > 0$ be such that $\|u_n\|_{W^{2,2q}(\Omega)} \leq \Lambda_2$ for some $q > \frac{d}{2}$, $q \in \mathbb{N}$. Further, assume that a $\Lambda_3 > 0$ exists such that*

$$\|u_n - u_{n-1}\|_{W^{1,2q}(\Omega)} \leq \Lambda_3\tau. \tag{4.2}$$

Let $\{u_n^i\}_{i\in\mathbb{N}}$ be the sequence generated by the MS. Then there exists a $\tilde{\tau} > 0$ such that for all $\tau \leq \tilde{\tau}$ and $i \in \mathbb{N}$, $\|u_n^i - u_n\|_{L^\infty(\Omega)} \leq \Lambda\tau$, $\|u_n^i - u_n\|_{W^{1,2q}(\Omega)} \leq \Lambda_3\tau$ and $u_n^i \in W^{2,2q}(\Omega)$.

Proof. Similar to Lemma 3.1, we give a proof by induction. Assume that $\|e_n^k\|_{L^\infty(\Omega)} \leq \Lambda\tau$, $\|\nabla e_n^k\|_{L^{2q}(\Omega)} \leq \Lambda_3\tau$ and $u_n^k \in W^{2,2q}(\Omega)$ for $k < i$. This is true for $k = 0$ because of Assumption (A. 1) and (4.2). We show that this implies $\|e_n^i\|_{L^\infty(\Omega)} \leq \Lambda\tau$, $\|\nabla e_n^i\|_{L^{2q}(\Omega)} \leq \Lambda_3\tau$ and $u_n^i \in W^{2,2q}(\Omega)$ for small τ. More specifically, we show that there exists a $C > 0$ such that

$$\|e_n^i\| \leq C\tau^2, \quad \|e_n^i\|_{L^\infty(\Omega)} \leq C\tau^{1+\frac{1}{2q}}, \quad \|e_n^i\|_{W^{1,p}(\Omega)} \leq C\tau^{1+\frac{1}{p}}, \quad \|e_n^i\|_{W^{2,2q}(\Omega)} \leq C\tau, \tag{4.3}$$

for all $p \geq 2$ and $\tau > 0$. This proves the lemma for τ small enough.

First, observe that inequality (3.6) holds for L_n^i defined in (3.2). Moreover, the assumption $u_n \in W^{2,2q}(\Omega)$ implies (A. 2), i.e. there exists a $\Lambda_1 > 0$ such that $\|\nabla u_n\|_{L^\infty(\Omega)} \leq \Lambda_1$. This is a direct consequence of Morrey's inequality [6, Chapter 5] and $2q > d$. By the same argument $u_n^{i-1} \in W^{1,\infty}(\Omega) \in H^1(\Omega)$. Hence, by Theorem 5 of [6, Chapter 6] it follows that u_n^i is a classical solution to

$$L_n^i u_n^i - \tau\nabla \cdot (\mathcal{D}_n^{i-1}\nabla u_n^i) = L_n^i u_n^{i-1} - (b_n^{i-1} - b_{n-1}) + \tau\nabla \cdot \vec{F}_n^{i-1} + \tau f_n^{i-1}. \tag{4.4}$$

In particular, u_n^i is essentially bounded and lies in H_0^1 as well.

Subtracting (2.1) from (4.4) and rearranging the terms leads to

$$\begin{aligned} &L_n^i e_n^i - \tau\nabla\cdot(\mathcal{D}_n^{i-1}\nabla e_n^i) \\ &= L_n^i e_n^{i-1} - \partial_u z(\zeta) e_n^{i-1} + \tau\nabla\cdot((\mathcal{D}_n^{i-1}-\mathcal{D}_n)\nabla u_n) - \tau\nabla\cdot(\vec{F}_n^{i-1}-\vec{F}_n), \end{aligned} \tag{4.5}$$

where $\zeta \in \mathcal{I}(u_n^{i-1}, u_n)$. Denoting the terms on the right by I_1, I_2, I_3 we have

$$|I_1| = |(L_n^i - \partial_u z(\zeta))e_n^{i-1}| \le 2\Lambda\mathfrak{M}\tau^2;$$

$$\begin{aligned} |I_2| &= \tau|(\mathcal{D}_n^{i-1}-\mathcal{D}_n)\Delta u_n + (\partial_u\mathcal{D}_n^{i-1}\nabla u_n^{i-1} - \partial_u\mathcal{D}_n\nabla u_n)\cdot\nabla u_n + \sum_{j=1}^{d}(\partial_{x_j}\mathcal{D}_n^{i-1} - \partial_{x_j}\mathcal{D}_n)\partial_{x_j}u_n| \\ &\le \Lambda\mathcal{D}_M\tau^2|\Delta u_n| + \tau|\nabla u_n||\partial_u\mathcal{D}_n^{i-1}\nabla e_n^{i-1} + (\partial_u\mathcal{D}_n^{i-1} - \partial_u\mathcal{D}_n)\nabla u_n| + \tau|\nabla u_n||\sum(\partial_{x_j}\mathcal{D}_n^{i-1} - \partial_{x_j}\mathcal{D}_n)| \\ &\le \Lambda\mathcal{D}_M\tau^2|\Delta u_n| + \tau\mathcal{D}_M\Lambda_1|\nabla e_n^{i-1}| + \mathcal{D}_M\Lambda\Lambda_1^2\tau^2 + d\Lambda_1\mathcal{D}_M\Lambda\tau^2; \end{aligned}$$

$$\begin{aligned} |I_3| &= \tau|\nabla\cdot(\vec{F}_n^{i-1}-\vec{F}_n)| = \tau|(\partial_u\vec{F}_n^{i-1}\cdot\nabla u_n^{i-1} - \partial_u\vec{F}_n\cdot\nabla u_n) + \sum_{j=1}^{d}(\partial_{x_j}F_{j,n}^{i-1} - \partial_{x_j}F_{j,n})| \\ &\le \tau|\partial_u F_n^{i-1}\cdot\nabla e_n^{i-1}| + \tau|(\partial_u\vec{F}_n^{i-1} - \partial_u\vec{F}_n)\cdot\nabla u_n| + \tau|\sum(\partial_{x_j}F_{j,n}^{i-1} - \partial_{x_j}F_{j,n})| \\ &\le \tau M_F|\nabla e_n^{i-1}| + \Lambda M_F\Lambda_1\tau^2 + d\Lambda M_F\tau^2. \end{aligned}$$

Define $S_i := I_1 + I_2 + I_3$. As $\|e_n^{i-1}\|_{W^{1,2q}(\Omega)} \le \Lambda_3\tau$ and $\|\Delta u_n\|_{L^{2q}(\Omega)} \le \Lambda_2$ it follows that a $C_1 > 0$ exists such that

$$\|S_i\|_{L^{2q}(\Omega)} \le C_1\tau^2. \tag{4.6}$$

Now we test (4.5) with the test function $\phi = (e_n^i)^{2q-1} \in H_0^1(\Omega)$. This gives

$$\begin{aligned} &m\|e_n^i\|_{L^{2q}(\Omega)}^{2q} + \tau\mathcal{D}_m(2q-1)\int_\Omega |e_n^i|^{2(q-1)}|\nabla e_n^i|^2 \le (|S_i|, |e_n^i|^{2q-1}) \\ &\le \frac{1}{2qm^{2q-1}}\|S_i\|_{L^{2q}(\Omega)}^{2q} + \frac{(2q-1)m}{2q}\|e_n^i\|_{L^{2q}(\Omega)}^{2q}. \end{aligned}$$

The last inequality follows from repeated application of Young's inequality. This implies

$$\|e_n^i\|_{L^{2q}(\Omega)} \le \frac{1}{m}\|S_i\|_{L^{2q}(\Omega)} \le \frac{C_1}{m}\tau^2.$$

Rewriting (4.5) as

$$\nabla\cdot\left(\mathcal{D}_n^{i-1}\nabla e_n^i\right) = L_n^i\left(\frac{e_n^i}{\tau}\right) - \frac{1}{\tau}S_i, \tag{4.7}$$

we see that the $L^{2q}(\Omega)$-norm of the right hand side is bounded by some constant times τ. As $|\nabla\mathcal{D}_n^{i-1}| \le |\partial_u\mathcal{D}_n^{i-1}\nabla u_n^{i-1}| + \sum_j|\partial_{x_j}\mathcal{D}_n^{i-1}| \le \mathcal{D}_M(|\nabla e_n^{i-1}| + \Lambda_1 + d)$, $|\nabla\mathcal{D}_n^{i-1}| \in L^{2q}(\Omega)$.

Hence, we apply Theorem 15.1 of [14, Chapter 2] to get that $e_n^i \in W^{2,2q}(\Omega) \cap \mathcal{C}^{1,\gamma}(\Omega)$ for $\gamma = 1 - \frac{d}{2q}$. Moreover, there exists a $C > 0$ such that

$$\|e_n^i\|_{W^{2,2q}(\Omega)} \leq C\tau \text{ and } \|\nabla e_n^i\|_{L^\infty(\Omega)} \leq \|e_n^i\|_{\mathcal{C}^{1,\gamma}(\Omega)} \leq C\tau. \tag{4.8}$$

This proves the last statement of (4.3).

Next, we test (4.5) with $\phi = e_n^i$ to get

$$m\|e_n^i\|^2 + \tau\mathcal{D}_m\|\nabla e_n^i\|^2 \leq (S_i, e_n^i) \leq \frac{1}{2m}\|S_i\|^2 + \frac{m}{2}\|e_n^i\|^2. \tag{4.9}$$

As $S_i \in L^{2q}(\Omega) \subseteq L^2(\Omega)$ we get using (4.6) that for some constant $C_2 > 0$

$$\|\nabla e_n^i\|^2 \leq C_2\tau^3. \tag{4.10}$$

This implies that for $p \geq 2$,

$$\|\nabla e_n^i\|_{L^p(\Omega)} = \left(\int_\Omega |\nabla e_n^i|^p\right)^{1/p} \leq \left(\|\nabla e_n^i\|_{L^\infty(\Omega)}^{p-2} \int_\Omega |\nabla e_n^i|^2\right)^{1/p} \leq (C^{p-2}C_2)^{1/p}\tau^{1+\frac{1}{p}}. \tag{4.11}$$

Finally, as $\mathcal{C}^{0,\gamma}(\Omega) \subseteq W^{1,2q}(\Omega)$, see Morrey's inequality (Theorem 5 of [6, Chapter 5]), we get that

$$\|e_n^i\|_{L^\infty(\Omega)} \leq \|e_n^i\|_{\mathcal{C}^{0,\gamma}(\Omega)} \leq C\tau^{1+\frac{1}{2q}}. \tag{4.12}$$

From this we get $\|e_n^i\|_{L^\infty(\Omega)} \leq \Lambda\tau$ if $\tau \leq (\Lambda/C)^{2q}$. Similarly one obtains $\|\nabla e_n^i\|_{L^{2q}(\Omega)} \leq \Lambda_3\tau$ for small τ. □

Before giving the main theorem of this section we give a context that can lead to improved convergence behaviour of the MS. Specifically, we assume

$$|\partial_u\mathcal{D}(\vec{x}, u)| + \sum_{j=1}^{d} |\partial_u F_j(\vec{x}, u)| \leq \Upsilon\partial_u z(\vec{x}, u) \text{ a.e. in } \vec{x} \in \Omega \text{ and } u \in \mathbb{R}. \tag{4.13}$$

Remark 4.1. *The bound* (4.13) *is true, e.g. for the Richards equation (see* (5.1)*, Section 5). In this case the diffusivity and the flux terms are* $\mathcal{D} = k(b(u))$ *and* $\vec{F} = k(b(u))\hat{e}_g$ *(*$\hat{e}_g$ *is a constant unit vector) with* $b : \mathbb{R} \to [0, 1]$ *and* $k \in \mathcal{C}^1([0, 1])$ *giving* $|\partial_u\mathcal{D}|, |\partial_u F_j| \leq \sup_{s\in[0,1]}\{k'(s)\}b'(u)$*. Also the quasilinear system discussed in Section 3 is just a special case when* $\Upsilon = 0$*.*

Theorem 4.1. *For a fixed* $n \in \{1, .., N\}$ *let* $\{u_n^i\}$*,* $i \in \mathbb{N}$ *be the sequence provided by the MS. Assume* (4.1) *holds for* $i \in \mathbb{N}$*. If Assumptions (A. 1)-(A. 2) and (P. 1)-(P. 5) are satisfied then the following holds:*

(a) If inequality (4.13) *is satisfied then there exists a* $\mathfrak{M}_1 > 0$ *and* $\tau^* > 0$ *independent of* m *such that if* $\mathfrak{M} \geq \mathfrak{M}_1$ *and* $\tau \leq \tau^*$ *then* $u_n^i \to u_n$ *in the strong sense in* $H^1(\Omega)$*.*

(b) If $m > 0$ then there exists a $\bar{\tau} > 0$ such that for $\tau < \bar{\tau}$ and $\mathfrak{M} \geq \mathfrak{M}_0$, u_n^i converges linearly to u_n in $H^1(\Omega)$. Moreover $\alpha = \mathcal{O}(\sqrt{\tau})$ for this case.

Proof (a). We follow the line of arguments presented in [16] for proving this part. Subtracting (2.3) from (3.1) and taking $\phi = e_n^i$ gives

$$\begin{aligned}&(L_n^i(e_n^i - e_n^{i-1}), e_n^i) + \tau(\mathcal{D}_n^{i-1}\nabla u_n^i - \mathcal{D}_n\nabla u_n, \nabla e_n^i)\\ &= -(z(u_n^{i-1}) - z(u_n), e_n^i) + \tau(\vec{F}_n^{i-1} - \vec{F}_n, \nabla e_n^i).\end{aligned} \tag{4.14}$$

This can be rearranged into the form $T_1 + \tau T_2 + T_3 = T_4 + \tau T_5 + \tau T_6$ where the terms T_j are estimated as:

$$T_1 := (L_n^i(e_n^i - e_n^{i-1}), e_n^i) = \frac{1}{2}\int_\Omega L_n^i|e_n^i|^2 - \frac{1}{2}\int_\Omega L_n^i|e_n^{i-1}|^2 + \frac{1}{2}\int_\Omega L_n^i|e_n^i - e_n^{i-1}|^2;$$

$$T_2 := \int_\Omega \mathcal{D}_n^{i-1}|\nabla e_n^i|^2 \geq \mathcal{D}_m\left\|\nabla e_n^i\right\|^2;$$

$$T_3 := (z(u_n^{i-1}) - z(u_n), e_n^{i-1}) = \int_\Omega \frac{1}{\partial_u z(\zeta_0)}|z(u_n^{i-1}) - z(u_n)|^2;$$

$$T_4 := (z(u_n^{i-1}) - z(u_n), e_n^{i-1} - e_n^i) \leq \frac{1}{2}\int_\Omega \frac{1}{L_n^i}|z(u_n^{i-1}) - z(u_n)|^2 + \frac{1}{2}\int_\Omega L_n^i|e_n^{i-1} - e_n^i|^2;$$

$$T_5 := -((\mathcal{D}_n^{i-1} - \mathcal{D}_n)\nabla u_n, \nabla e_n^i) \leq \frac{1}{\mathcal{D}_m}\|(\mathcal{D}_n^{i-1} - \mathcal{D}_n)\nabla u_n\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2$$

$$\leq \frac{\Lambda_1^2}{\mathcal{D}_m}\left\|\frac{\partial_u\mathcal{D}(\zeta_1)}{\partial_u z(\zeta_1)}(z(u_n^{i-1}) - z(u_n))\right\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2 \leq \frac{(\Upsilon\Lambda_1)^2}{\mathcal{D}_m}\left\|z(u_n^{i-1}) - z(u_n)\right\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2;$$

$$T_6 := (\vec{F}_n^{i-1} - \vec{F}_n, \nabla e_n^i) \leq \frac{1}{\mathcal{D}_m}\|\vec{F}_n^{i-1} - \vec{F}_n\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2$$

$$\leq \frac{1}{\mathcal{D}_m}\sum_{j=1}^d\left\|\frac{\partial_u F_j(\zeta_2)}{\partial_u z(\zeta_2)}(z(u_n^{i-1}) - z(u_n))\right\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2 \leq \frac{\Upsilon^2}{\mathcal{D}_m}\left\|z(u_n^{i-1}) - z(u_n)\right\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2.$$

For the functions ζ_1,$\zeta_2 : \Omega \to \mathbb{R}$ one has $\zeta_{1,2} \in \mathcal{I}(u_n^{i-1}, u_n)$. For T_3, if $z(u_n^{i-1}) = z(u_n)$ then the equality holds for any $\zeta_0 \in \mathbb{R}$ for which $\partial_u z(\zeta_0) \neq 0$. If $z(u_n^{i-1}) \neq z(u_n)$ then by the Mean Value Theorem there exists a $\zeta_0 \in \mathcal{I}(u_n^{i-1}, u_n)$ such that $\partial_u z(\zeta_0) \neq 0$. Hence, there exists a $\zeta_0 : \Omega \to \mathbb{R}$ such that equality for T_3 is satisfied and $\partial_u z(\zeta_0) \neq 0$ a.e. In the following, T_3 will only induce an upper bound for τ which will depend on the lower bound of $\frac{1}{\partial_u z(\zeta_0)}$. Therefore we claim that the choice of ζ_0, discussed above, has no influence on the result.

Putting everything together we get the following inequality

$$\begin{aligned}&\int_\Omega L_n^i|e_n^i|^2 + \tau\mathcal{D}_m\left\|\nabla e_n^i\right\|^2 + \int_\Omega\left(\frac{2}{\partial_u z(\zeta_0)} - \frac{1}{L_n^i} - \frac{2\tau\Upsilon^2}{\mathcal{D}_m}(1 + \Lambda_1^2)\right)|z(u_n^{i-1}) - z(u_n)|^2\\ &\leq \int_\Omega L_n^i|e_n^{i-1}|^2 \leq \int_\Omega L_n^{i-1}|e_n^{i-1}|^2 + \int_\Omega |L_n^i - L_n^{i-1}||e_n^{i-1}|^2.\end{aligned} \tag{4.15}$$

This inequality is useful if all the terms on the left are positive. This is achieved if $\frac{2}{\partial_u z(\zeta_0)} - \frac{1}{L_n^i} - \frac{2\tau\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2) \geq 0$, which will add some restrictions on τ. To see this we define

$$\mathfrak{M}_1 = \max\{4\mathfrak{M}_0, 2\mathfrak{M}_0 + G\}, \tag{4.16}$$

where $\mathfrak{M}_0 = \Lambda \max_{u\in\mathbb{R}}\{|b''| + \tau|\partial_{uu} f|\}$ is defined just as in the proof of Theorem 3.1 and the value of G will be clarified later. Observe that $(\frac{2}{\partial_u z} - \frac{1}{L_n^i}) = \frac{1}{\partial_u z} + \frac{L_n^i - \partial_u z}{L_n^i \partial_u z}$. From (3.7) we get $L_n^i - \partial_u z(\zeta_0) > (\mathfrak{M} - \mathfrak{M}_0)\tau$. Moreover, from (P. 1)-(P. 2), $\partial_u z \leq M_b + TM_f$ and $L_n^i \leq M_b + TM_f + \mathfrak{M}\tau$. This gives

$$\frac{L_n^i - \partial_u z}{L_n^i \partial_u z} \geq \frac{(\mathfrak{M} - \mathfrak{M}_0)\tau}{(M_b + TM_f)(M_b + TM_f + \mathfrak{M}\tau)}. \tag{4.17}$$

To simplify the analysis we note that $\frac{L_n^i - \partial_u z}{L_n^i \partial_u z} - \frac{2\tau\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2) \geq 0$ is a sufficient condition for the positivity of the last term on the left hand side of (4.15) . Inequality (4.17) implies that this is satisfied when

$$\left(1 - \frac{2\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2)(M_b + TM_f)\tau\right)\mathfrak{M} \geq \mathfrak{M}_0 + \frac{2\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2)(M_b + TM_f)^2.$$

Hence by defining

$$\tau^* = \frac{\mathcal{D}_m}{4\Upsilon^2(1+\Lambda_1^2)(M_b + TM_f)} \text{ and } G = \frac{4\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2)(M_b + TM_f)^2,$$

we get that for all $\tau \leq \tau^*$ and $\mathfrak{M} \geq \mathfrak{M}_1$,

$$\frac{L_n^i - \partial_u z}{L_n^i \partial_u z} - \frac{2\tau\Upsilon^2}{\mathcal{D}_m}(1+\Lambda_1^2) \geq 0.$$

Now consider $\tau \leq \tau^*$ and $\mathfrak{M} \geq \mathfrak{M}_1$. Inequality (4.15) is restated as:

$$\int_\Omega L_n^i |e_n^i|^2 + \tau\mathcal{D}_m \left\|\nabla e_n^i\right\|^2 + \int_\Omega \partial_u z(\zeta_0)|e_n^{i-1}|^2 \leq \int_\Omega L_n^{i-1}|e_n^{i-1}|^2 + \int_\Omega |L_n^i - L_n^{i-1}||e_n^{i-1}|^2. \tag{4.18}$$

Observe that Ω can be split into two disjoint sets defined as Ω_1, Ω_2 such that $\Omega_1 = \{\vec{x} \in \Omega : |\partial_u z(\zeta_0)| < \frac{1}{2}\mathfrak{M}\tau\}$ and $\Omega_2 = \{\vec{x} \in \Omega : |\partial_u z(\zeta_0)| \geq \frac{1}{2}\mathfrak{M}\tau\}$. If $\vec{x} \in \Omega_1$, then from $\mathfrak{M} \geq \mathfrak{M}_1 \geq 4\mathfrak{M}_0$ we get

$$\begin{aligned}
\partial_u z(u_n^{i-2}) &\leq \partial_u z(\zeta_0) + \max_{\mathcal{I}(u_n^{i-1}, u_n)}\{\partial_{uu} z\}|\zeta_0 - u_n| + \max_{\mathcal{I}(u_n, u_n^{i-2})}\{\partial_{uu} z\}|u_n - u_n^{i-2}| \\
&\leq \partial_u z(\zeta_0) + \mathfrak{M}_0\tau + \mathfrak{M}_0\tau \leq \frac{1}{2}\mathfrak{M}\tau + \frac{1}{2}\mathfrak{M}\tau
\end{aligned}$$

which implies that $L_n^{i-1} = \max\{\partial_u z(u_n^{i-2}) + \mathfrak{M}\tau, 2\mathfrak{M}\tau\} = 2\mathfrak{M}\tau$. By the same logic, as $|\partial_u z(\zeta_0)| < \frac{1}{2}\mathfrak{M}\tau$, one has $L_n^i = 2\mathfrak{M}\tau$. This means that

$$\int_{\Omega_1} |L_n^i - L_n^{i-1}||e_n^{i-1}|^2 = 0.$$

If $\vec{x} \in \Omega_2$ then $\partial_u z(\zeta_0) \geq \frac{1}{2}\mathfrak{M}\tau \geq 2\mathfrak{M}_0\tau$. There can be four cases. If $\partial_u z(u_n^{i-1}) \leq \mathfrak{M}\tau$ and $\partial_u z(u_n^{i-2}) > \mathfrak{M}\tau$ then $|L_n^i - L_n^{i-1}| = \partial_u z(u_n^{i-2}) - \mathfrak{M}\tau \leq \partial_u z(u_n^{i-2}) - \partial_u z(u_n^{i-1})$. Similarly if $\partial_u z(u_n^{i-1}) > \mathfrak{M}\tau$ and $\partial_u z(u_n^{i-2}) \leq \mathfrak{M}\tau$ then $|L_n^i - L_n^{i-1}| \leq \partial_u z(u_n^{i-1}) - \partial_u z(u_n^{i-2})$. $L_n^i = L_n^{i-1}$ if both $\partial_u z(u_n^{i-2}), \partial_u z(u_n^{i-1}) \leq \mathfrak{M}\tau$ and $|L_n^i - L_n^{i-1}| = |\partial_u z(u_n^{i-1}) - \partial_u z(u_n^{i-2})|$ if both $\partial_u z(u_n^{i-2}), \partial_u z(u_n^{i-1}) > \mathfrak{M}\tau$. Combining everything we can state that

$$|L_n^i - L_n^{i-1}| \leq |\partial_u z(u_n^{i-1}) - \partial_u z(u_n^{i-2})| \leq \max_{\mathcal{I}(u_n^{i-1}, u_n^{i-2})} \{|\partial_{uu} z|\} |u_n^{i-1} - u_n^{i-2}| \leq 2\mathfrak{M}_0\tau.$$

From the above, one obtains

$$\begin{aligned}
&\int_\Omega |L_n^i - L_n^{i-1}||e_n^{i-1}|^2 = \int_{\Omega_2} |L_n^i - L_n^{i-1}||e_n^{i-1}|^2 \leq 2\mathfrak{M}_0\tau \int_{\Omega_2} |e_n^{i-1}|^2 \\
&\leq \int_{\Omega_2} \partial_u z(\zeta_0)|e_n^{i-1}|^2 \leq \int_\Omega \partial_u z(\zeta_0)|e_n^{i-1}|^2.
\end{aligned}$$

Using this result in (4.18) one gets

$$\int_\Omega L_n^i |e_n^i|^2 + \tau\mathcal{D}_m \left\|\nabla e_n^i\right\|^2 \leq \int_\Omega L_n^{i-1}|e_n^{i-1}|^2. \tag{4.19}$$

Taking the sum of (4.19) from $i = 1$ to $i = p$ gives

$$\int_\Omega L_n^p |e_n^p|^2 + \tau\mathcal{D}_m \sum_{i=1}^{p} \left\|\nabla e_n^i\right\|^2 \leq \int_\Omega L_n^0 |e_n^0|^2 \leq \left\|L_n^0\right\|_{L^1(\Omega)} \Lambda^2\tau^2, \tag{4.20}$$

with $L_n^0 = \max\{\partial_u z(\vec{x}, u_{n-1}) + \mathfrak{M}\tau, 2\mathfrak{M}\tau\}$. Similar estimates are given in [8, 17, 21]. In other words, the series $\sum_{i=1}^{\infty} \left\|\nabla e_n^i\right\|^2$ is convergent implying that $\left\|\nabla e_n^i\right\| \to 0$ as $i \to \infty$. This concludes the proof of the first part of Theorem 4.1.

(b). To prove (b) we rearrange (4.14) as $T_1' + \tau T_2 = T_3' + \tau T_5 + \tau T_6$, where T_j, $j \in \{2, 5, 6\}$ have been defined before, and T_1', T_3' are

$$\begin{aligned}
T_1' &:= \int_\Omega L_n^i |e_n^i|^2 \geq (m + \mathfrak{M}\tau)\left\|e_n^i\right\|^2; \\
T_3' &:= ((L_n^i - \partial_u z(\zeta))e_n^{i-1}, e_n^i) \leq \mathfrak{M}\tau\left\|e_n^i\right\|^2 + \mathfrak{M}\tau\left\|e_n^{i-1}\right\|^2;
\end{aligned}$$

the last inequality following from (3.6) and Young's inequality. Moreover, a different estimate for T_5, T_6 can be obtained as

$$T_5 \leq \frac{1}{\mathcal{D}_m}\|(\mathcal{D}_n^{i-1} - \mathcal{D}_n)\nabla u_n\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2 \leq \frac{\Lambda_1^2\mathcal{D}_M^2}{\mathcal{D}_m}\|e_n^{i-1}\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2;$$
$$T_6 \leq \frac{1}{\mathcal{D}_m}\|\vec{F}_n^{i-1} - \vec{F}_n\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2 \leq \frac{dM_F^2}{\mathcal{D}_m}\|e_n^{i-1}\|^2 + \frac{\mathcal{D}_m}{4}\|\nabla e_n^i\|^2.$$

Adding in the estimates for all remaining terms, one gets from (4.14),

$$m\left\|e_n^i\right\|^2 + \frac{\tau}{2}\mathcal{D}_m\left\|\nabla e_n^i\right\|^2 \leq \tau\left[\mathfrak{M} + \frac{(dM_F^2 + \mathcal{D}_M^2\Lambda_2^2)}{\mathcal{D}_m}\right]\left\|e_n^{i-1}\right\|^2,$$

which rewrites as

$$\left\|e_n^i\right\|^2 + \frac{\tau\mathcal{D}_m}{2m}\left\|\nabla e_n^i\right\|^2 \leq \frac{\tau}{m}\left[\mathfrak{M} + \frac{(dM_F^2 + \mathcal{D}_M^2\Lambda_1^2)}{\mathcal{D}_m}\right]\left(\left\|e_n^{i-1}\right\|^2 + \frac{\tau\mathcal{D}_m}{2m}\left\|\nabla e_n^{i-1}\right\|^2\right).$$

Taking

$$\bar{\tau} = \frac{m\mathcal{D}_m}{\mathfrak{M}\mathcal{D}_m + (dM_F^2 + \mathcal{D}_M^2\Lambda_1^2)} \text{ and } \alpha = \sqrt{\frac{\tau}{m}\left[\mathfrak{M} + \frac{(dM_F^2 + \mathcal{D}_M^2\Lambda_1^2)}{\mathcal{D}_m}\right]}, \tag{4.21}$$

one observes that the iteration converges in the equivalent $H^1(\Omega)$-norm $\left\|u\right\|_{H^1(\Omega)} = \sqrt{\left\|u\right\|^2 + \frac{\tau\mathcal{D}_m}{2m}\left\|\nabla u\right\|^2}$ and has the convergence rate $\alpha = \mathcal{O}(\sqrt{\tau}) < 1$ if $\tau < \bar{\tau}$. □

5 Numerical results

For the numerical discussions we consider the Richards equation, which is widely used in groundwater modelling. In terms of the capillary pressure p, the non-dimensional Richards equation reads:

$$\partial_t S(p) = \nabla \cdot [k(S(p))(\nabla p - \hat{g})] + f, \tag{5.1}$$

where $\hat{g}$ is the unit vector along the direction of gravity and f is the source term. Richards equation involves nonlinearities in all the terms. The saturation function S is increasing and the permeability function k takes non-negative values. Without entering into the details, as the specific forms will be given later, we mention that in general one has $k(0) = 0$ and $S'(p) \to 0$ as $p \to -\infty$. Further $S'(p) = 0$ for all $p \geq 0$. However, if the flow does not become completely unsaturated, meaning that $S(p) \to 0$ (this being the case when the initial and boundary conditions are taken accordingly), Assumption (P. 3) will be satisfied. Also (4.13) is satisfied as discussed in Remark 4.1.

The theoretical results presented in the previous sections do not depend upon the spatial discretization. Hence, for the numerical results, different methods like finite difference, finite element or finite volume, can be used to discretize (5.1) in space. Here we have used a two point flux approximation finite volume scheme [7] defined on two dimensional triangular unstructured grids. We take

$$\Omega = (0,1) \times (0,1) \text{ and } T = 1, \tag{5.2}$$

and use FVCA8 benchmark meshes of different sizes. For the triangulation $\mathcal{T}$ the mesh size is $h = \sup\{\text{diam}(T) : T \in \mathcal{T}\}$.

With S and k defined as

$$S(p) = \begin{cases} \frac{1}{(1-p)^{\frac{1}{3}}} & \text{if } p < 0 \\ 1 & \text{if } p \geq 0 \end{cases}, \qquad k(S) = S^3, \tag{5.3}$$

the first numerical example is constructed in such a way that

$$\tilde{p}(x,y,t) = 1 - (1+t^2)(1+x^2+y^2), \tag{5.4}$$

is the exact solution (see also [25]). The source term is

$$f(x,y,t) = \frac{2(1-y^2)}{(1+y^2)^2} - \frac{2t}{3\sqrt[3]{(1+t^2)^4(1+y^2)}} - \frac{2x}{(1+t^2)(1+x^2+y^2)^2}. \tag{5.5}$$

The boundary and initial conditions are as in Table 1. $\hat{g}$ points along the positive x-axis.

Table 1: Assumed initial and boundary conditions.

IC	$t = 0$	$p(x,y,0) = \tilde{p}(x,y,0)$	on Ω	
BC	$x = 0:$	$p(0,y,t) = \tilde{p}(0,y,t),$	$x = 1:$	$p(1,y,t) = \tilde{p}(1,y,t),$
	$y = 0:$	$\partial_y p = 0,$	$y = 1:$	$k(S(p))\partial_y p = k(S(\tilde{p}(x,1,t))\partial_y \tilde{p}(x,1,t).$

Figure 1 (left) shows the comparison of the analytical solution $\tilde{p}$ with the numerical solution. The numerical solution is obtained from the MS with $\mathfrak{M} = 1$. The maximum relative error $\|\frac{p-\tilde{p}}{\tilde{p}}\|_{L^\infty(\Omega)}$ is 1.38% and the $L^2(\Omega)$ error is 0.0116 which is of the order of the discretization errors, implying that the computational results are accurate. To ensure correctness, such kind of profile match have been conducted for all the results shown afterwards.

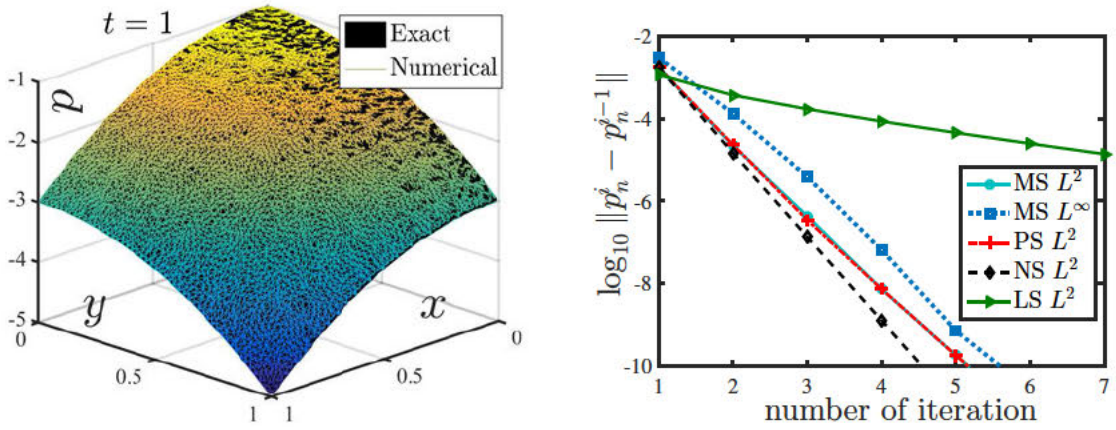

Figure 1: (left) Analytical solution vs. numerical solution for S, k and f given in (5.3), (5.5) and for initial and boundary condition stated in Table 1. The computational parameters are $\mathfrak{M} = 1$, $\tau = 0.001$ and $h = 0.02$. (right) The error decay of the same computation at $t = 0.5$ for different schemes. Both L^2 error and L^∞ error of the MS have been plotted and $L = 0.2$ has been used for the LS.

Figure 1 (right) shows how the errors decrease for the same computation for different schemes. Both L^2 and L^∞ errors of the MS decrease monotonically. The decrease of the L^∞ error shows that Assumption (A. 1) is valid and the linear profile of the L^2 and L^∞ errors shows that the convergence is indeed linear as pointed out in Theorem 4.1 and Proposition 3.2. Observe that the error plot of the PS almost coincides with the error plot of the MS. This is because $\tau = 0.001$ is quite small, and for reasons explained afterwards this is true in general for small values of τ. The NS converges faster than the PS and the MS. The error of the LS decreases linearly but the speed of convergence is considerably less than the other schemes shown. The values of $\mathfrak{M}$ and L used are the optimal values that give fastest convergence in their respective cases. This is explained in detail later.

To illustrate the strength of the MS compared to the other schemes mentioned above, we increase the complexity of the problem by using relations used for real-life simulations of such processes. For this purpose we replace the expressions in (5.3) with the standard van Genuchten relationship: for $\ell > 1$ and $q = 1 - \frac{1}{\ell}$,

$$S(p) = \begin{cases} \frac{1}{(1+(-p)^\ell)^q}, & \text{if } p < 0 \\ 1, & \text{if } p \geq 0 \end{cases}, \qquad k(S) = \sqrt{S}(1 - (1 - S^{1/q})^q)^2. \tag{5.6}$$

In the computations, $\ell = 3$ has been used throughput. Other conditions and definitions remain as in Table 1. The problem is degenerate as inside the domain there are regions where $p > 0$ or $S'(p) = 0$. For the numerical results $\mathfrak{M} = 10$ has been used for the MS and $L = 0.4$ has been used for the LS, unless specified otherwise. These values give optimal convergence for both the schemes. The choice is motivated later.

First a mesh study is conducted where the time step is fixed and the mesh size h is varied from $h = 0.1$ to $h = 0.02$. The results are shown in Figure 2 for two time step sizes $\tau = 0.01$ and $\tau = 0.001$ for a fixed time $t = 0.5$.

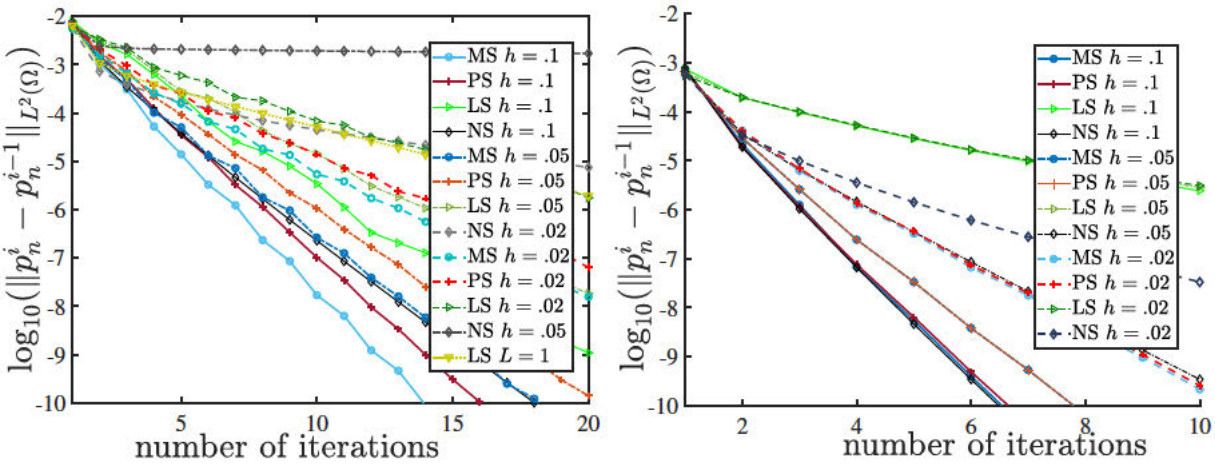

Figure 2: Mesh study of the NS, the PS, the LS and the MS at $t = 0.5$. (left) Results for $\tau = 0.01$; (right) results for $\tau = 0.001$. The mesh sizes used are $h = 0.1, 0.05, 0.02$. The $L = 1$ curve on the left figure corresponds to the error of the LS for $L = 1$ and $h = 0.05$. In all other computations $\mathfrak{M} = 10$ and $L = 0.4$ and the PS corresponds to $\mathfrak{M} = 0$.

It follows from Figure 2 that the MS, the LS and the PS show linear convergence. The MS is faster than the PS for all mesh sizes for $\tau = 0.01$ (Figure 2 (left)). The convergence rate of the LS is higher than for other schemes and in consequence, LS converges slower. In fact for $h = 0.05$ and $h = 0.02$ no further decrease in error is visible after a few iterations in case of the NS. The convergence rates for all the schemes vary with the mesh size. The dependence of the convergence rates of the MS and the LS on mesh size stems from the fact that $\mathfrak{M}$ or L chosen in these cases are not the $\mathfrak{M}_0$ or Lipschitz constant value (which in this case is $L \approx 0.65$) but the values that give the optimal convergence properties. This is also seen in Figure 6. In fact, taking $L = 1$ for the LS results in all the error plots for different mesh sizes being on top of each other. This error characteristic is presented by the curve labeled $L = 1$. Also, it appears that for a fixed time step size, decreasing h makes the convergence slower. This will be explained in detail later.

Figure 2 (right) shows the results for $\tau = 0.001$. The convergence is much faster than the $\tau = 0.01$ case for the MS as it was shown in Theorem 4.1 that $\alpha = \mathcal{O}(\sqrt{\tau})$. The difference between the convergence rates of the PS and of the MS is very small. This is because $\tau << 1$ and $\mathfrak{M} = 10$ so that $\mathfrak{M}\tau << S'(p) \sim \mathcal{O}(1)$ in most of the domain Ω. So difference between the MS and the PS is small. Figure 2 (right) also shows that convergence rate of the LS is stable with respect to the mesh size and is not impacted greatly by the change in time step size. This is probably due to Condition 16 of [16] being satisfied for this value of $\tau = 0.001$. However, the convergence is slower when compared

to the other schemes. As for the NS, for $h = 0.1$ it is marginally faster than the PS and the MS. But for finer meshes, NS becomes slower than the PS or the MS.

The quadratic convergence of the NS is not observed from Figure 2. This might be due to the small time step sizes required for the NS to show quadratic convergence in degenerate cases (see [24] and Table 2). It could also be attributed to the errors of the linear solver as it is known from literature [10, 16, 25] that the stiffness matrices for the NS are relatively ill-conditioned. Indeed, the GMRES solver gives higher residual errors for the NS compared to other methods. The errors introduced in computing the Jacobian could possibly be another reason that causes this deviation.

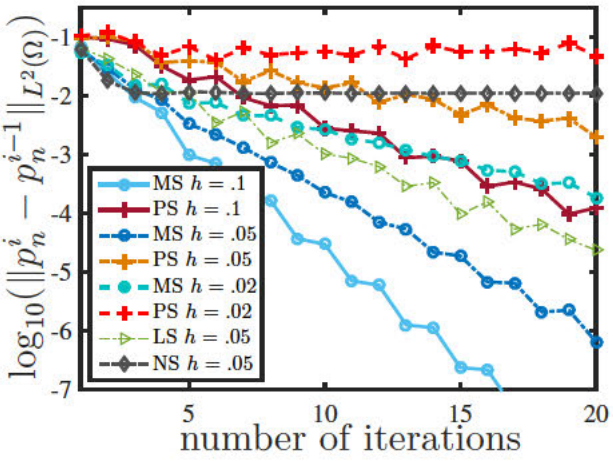

Figure 3: Numerical study of different linearization techniques for $\tau = 0.1$ at $t = 0.5$. The mesh sizes used are $h = 0.1, 0.05, 0.02$. In all the computations $\mathfrak{M} = 10$ and $L = 1$ and the PS corresponds to $\mathfrak{M} = 0$.

One of the advantages of the MS is its stability and for this reason we must look at larger time step sizes. Figure 4 shows the results for $\tau = 0.1$. The PS converges, albeit much slower than the MS, for $h = 0.1$. For $h = 0.05$ the convergence is very slow for the PS and the errors are not decreasing monotonically. In fact for $h = 0.05$ the iterations fail to converge starting from $t = 0.7$. For $h = 0.02$ the PS fails to converge even at $t = 0.5$. Similar behaviour is observed for the NS. For $h = 0.05$ the errors start to increase after a few iterations and the solution diverges at $t = 0.8$. The reason for this behaviour of the PS and the NS is the bound given in (1.2). As the mesh size decreases the time step size has to be reduced in order to guarantee the convergence for the PS and the NS [24]. However this constraint is not there for the MS. Although there is a threshold on the time step size for the MS's convergence, this threshold does not depend on the mesh size and so in all the cases shown in Figure 4, the errors for the MS are decreasing and the convergence is faster when compared to the NS or the PS. This shows that for numerical computations the MS is more stable than the NS or the PS in general.

As the convergence rate of the LS is quite stable with respect to mesh sizes for $L = 1$, the error behaviour of the LS has been plotted only for $h = 0.05$ and $L = 1$ in Figure 4. LS also has an upper bound on the time steps for linear convergence that does not depend upon the spatial discretization [16]. An interesting observation is that for larger time step sizes the convergence of the LS can actually be faster than the convergence of the MS. This is due to the fact that in the computations for $\tau = 0.1$, $\mathfrak{M}\tau \sim L$ and so $\mathfrak{M}\tau + S'(p)$ can be greater than L. So the apparent L value for the MS can be larger than the value of L required to be imposed on the LS. This might make the convergence of the MS slower than of the LS.

Figure 4 shows how the convergence rate α, calculated here as the geometric average of $\left\|p_n^{i+1} - p_n^i\right\| / \left\|p_n^i - p_n^{i-1}\right\|$ over the first 10 iterations, changes with τ. From Theorem 4.1 one expects that for small enough τ the convergence rate should scale with $\sqrt{\tau}$ which implies that the slope of the solid line in Figure 4 should nearly be $\frac{1}{2}$ for small τ values. This is indeed the case. For $\tau \leq .01$ the line is almost parallel to the reference dashed line representing a slope of $\frac{1}{2}$. It is to be noted though that Theorem 4.1 was proved for the non-degenerate case whereas this test case is degenerate.

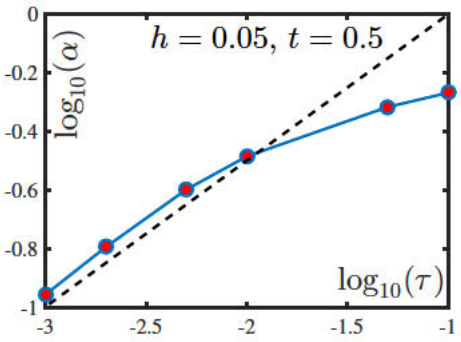

Figure 4: Convergence rate (α) vs. time step size (τ) for $h = 0.05$ at $t = 0.5$. The convergence rates are calculated as the geometric mean of $\left\|p_n^{i+1} - p_n^i\right\| / \left\|p_n^i - p_n^{i-1}\right\|$ over first 10 iterations. The dotted black line shows the slope of 0.5.

Next we study the effect of the choice of $\mathfrak{M}$ on the rate of convergence. One expects from Theorem 4.1 and (4.21) that for small values of $\mathfrak{M}$ the condition $(L_n^i - \partial_u z) > 0$ will not be satisfied whereas large values of $\mathfrak{M}$ would result in a slower convergence. This behaviour is observed precisely if one varies L in the case of the LS [16,25]. From Figure 5 we see that it describes the MS as well. For both time step sizes $\tau = 0.01$ and $\tau = 0.1$ we get that the optimal $\mathfrak{M}$ is close to 10. If one chooses $\mathfrak{M}$ value away from the optimal $\mathfrak{M}$ then α increases. As $\mathfrak{M} \to 0$ the rates tend towards the rate of the PS. In the $\tau = 0.01$ case the effect of an optimal $\mathfrak{M}$ is less pronounced than in the $\tau = 0.1$ case and from Figure 2 (right) one can speculate that it is even less important in the $\tau = 0.001$ case.

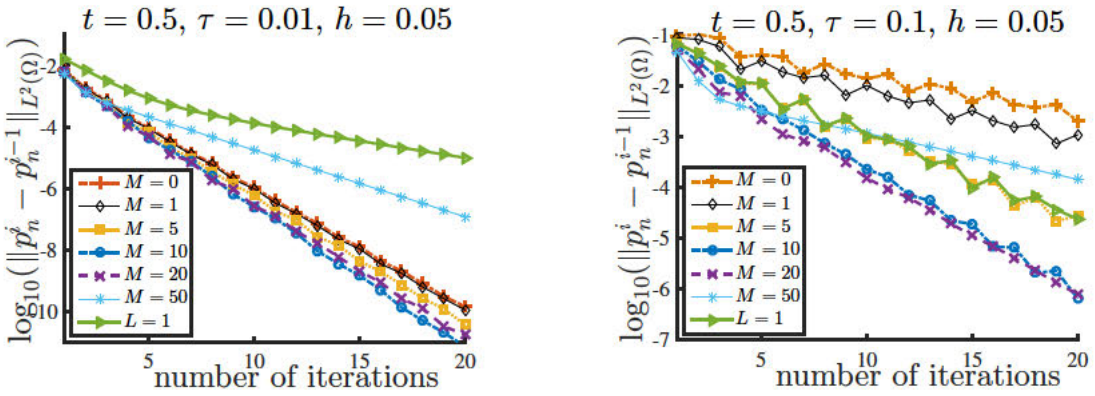

Figure 5: The dependence of the error behaviour on $\mathfrak{M}$ for the MS. (left) Results for $\tau = 0.01$; (right) results for $\tau = 0.1$.

This apriori knowledge can be useful when developing the linearisaion scheme as one can decide what $\mathfrak{M}$ to choose by running the computation for only the first time step with a coarser grid. In our case, we chose $\mathfrak{M} = 10$ for the results presented earlier for this reason.

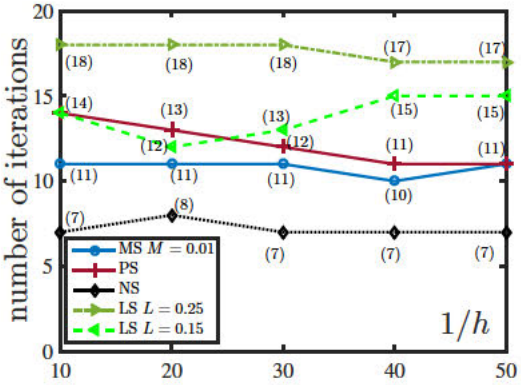

Figure 6: Iterations required by different schemes to have $\|p_n^i - p_n^{i-1}\| < 10^{-5}$, for a fixed $\tau = 1$ and $p_{vad} = -2$. The parameters are taken from [16, Figure 1].

	number of iterations			
τ	1	0.1	0.01	0.001
MS $\mathfrak{M} = 0.01$	18	22	12	7
PS $\mathfrak{M} = 0$	19	22	12	7
NS	-	-	-	7
LS $L = 0.25$	54	50	39	154
LS $L = 0.15$	35	33	26	99

Table 2: Iterations required by different schemes to have $\|p_n^i - p_n^{i-1}\| < 10^{-5}$, for a fixed $h = \frac{1}{40}$ and $p_{vad} = -2$. The parameters are taken from [16, Table 1].

Finally, we present a comparison study with Example 1 of [16]. The functional relationships are taken exactly as stated in [16] but different meshes and linear solvers are used. The initial condition used in this example is

$$p(x, y, 0) = \begin{cases} p_{vad} & \text{in } \Omega \cap \{y < \frac{1}{4}\}, \\ -y + \frac{1}{4} & \text{in } \Omega \cap \{y > \frac{1}{4}\}. \end{cases} \tag{5.7}$$

Here $p_{vad} < 0$ is a constant. The gravity points towards negative y direction and the results are for the first time step, i.e. $n = 1$. As the initial condition is discontinuous,

the NS fails to converge in all cases in our computations. However, the NS converged in most cases if the discontinuity was regularized, e.g. by considering a linear interpolation over a small interval of length 0.1. After this step the results obtained match closely the values given in [16]. In particular, we partly reproduce [16, Figure 1] with $p_{vad} = -2$ and [16, Table 1] with $p_{vad} = -3$ in Figure 6 and Table 2 respectively.

The result shows that the NS is the fastest when it converges but it requires smoother initial conditions and smaller time step sizes to converge. The PS and the MS have comparable stability properties. The MS, as before, is at least as fast as the PS in all cases. The LS is relatively slower and taking a smaller value of L ($L = 0.15$) compared to the Lipschitz constant ($L \approx 0.25$) speeds up the convergence, but makes the convergence rate more susceptible to changes in mesh size. This was seen in Figure 2 (left) too. The convergence rate decreases with the time step size for all the schemes except for the LS. A drastic increase in number of iterations required is seen for the LS at $\tau = 0.001$, see Table 2. This is explained by the fact that the convergence order is $L/(L + C\tau)$, which approaches 1 when τ goes to 0. $\mathfrak{M} = 0.01$ is used for the MS in Figure 6 and Table 2 as it gives the optimal convergence rate.

6 Conclusion

In this paper we propose a linearization scheme for a general class of nonlinear parabolic partial differential equations. We have given a rigorous convergence proof of the scheme and compared its behaviour with that of other linearisation schemes: the Newton scheme (NS), the modified Picard scheme (PS) or the L-scheme (LS), are generally used to solve the sequence of elliptic equations obtained from time-discretization of such problems. The NS and the PS have the drawback that they converge only if the initial guess for the iterations is close enough to the solution of the elliptic problem. For the concerned sequence of elliptic equations, this leads to a severe restriction on time step size. On the other hand, the LS converges irrespective of the initial guess but is much slower than the schemes mentioned above. To resolve theses issues, a combination of the LS and the PS, termed modified L-scheme (MS) in this context, is proposed.

The MS uses local estimates and the solution of the previous time step as the initial guess to improve the convergence behaviour of the LS. This is shown first for quasilinear equations that have linear diffusivity and advection terms. It is proved that the scheme converges linearly irrespective of the spatial discretization and for any time step, even in degenerate cases. Moreover, for small time step sizes, the linear convergence rate is proportional to the time step size, implying that the scheme converges faster as the time step size is made smaller.

Next, this result is generalized to the case when the diffusivity and the advection terms are nonlinear. It is proved that if the time step size is smaller than an upper-bound which is independent of the spatial discretization, the scheme converges even for degenerate

cases. Linear convergence is achieved in the non-degenerate case with a convergence rate that is proportional to the square root of the time step size for sufficiently small time steps.

Finally, numerical results are presented for the Richards equation for all the schemes mentioned above. It is seen that the MS is faster than the PS and the LS. Moreover, the MS is more stable than the NS or the PS in the sense that the MS converges for larger time steps. Numerically it is observed that the MS indeed gives convergence rates proportional to the square root of the time step size. The final numerical results imply that the convergence rate of the MS can be controlled by tuning the parameter $\mathfrak{M}$.

Acknowledgements

K. Mitra is supported by Shell and the Netherlands Organisation for Scientific Research (NWO) through the CSER programme (project 14CSER016) and by the Hasselt University through the project BOF17BL04. The research of I.S. Pop is supported by the Research Foundation-Flanders (FWO) through the Odysseus programme (project G0G1316N). The authors thank the anonymous referees for their thorough review of the manuscript and their enriching comments.

References

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

[2] L. Bergamaschi and M. Putti. Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. *International Journal for Numerical Methods in Engineering*, 45(8):1025–1046, 1999.

[3] M. Borregales, F.A. Radu, K. Kumar, and J.M. Nordbotten. Robust iterative schemes for non-linear poromechanics. *arXiv preprint arXiv:1702.00328*, 2017.

[4] K. Brenner and C. Cances. Improving Newton's method performance by parametrization: The case of the Richards equation. *SIAM Journal on Numerical Analysis*, 55(4):1760–1785, 2017.

[5] M.A Celia, E.T Bouloutas, and R.L. Zarba. General mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research*, 26(7):1483–1496, 1990.

[6] L.C. Evans. *Partial differential equations*. Wiley Online Library, 1988.

[7] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[8] W. Jäger and J. Kačur. Solution of porous medium type systems by linear approximation schemes. *Numerische Mathematik*, 60(1):407–427, 1991.

[9] W. Jäger and J. Kačur. Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 29(5):605–627, 1995.

[10] J.E. Jones and C.S. Woodward. Newton–Krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems. *Advances in Water Resources*, 24(7):763–774, 2001.

[11] J. Kačur. *Method of Rothe in evolution equations*, volume 1192. Springer, 1986.

[12] J. Kačur. Solution to strongly nonlinear parabolic problems by a linear approximation scheme. *IMA Journal of Numerical Analysis*, 19(1):119–145, 1999.

[13] O.A. Ladyzhenskaya, V. Solonnikov, and N. Uraltseva. *Linear and quasilinear parabolic equations of second order*. 1968.

[14] O.A. Ladyzhenskaya, N.N. Uraltseva, and L. Ehrenpreis. *Linear and quasilinear elliptic equations*. Academic Press, 1968.

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

[16] F. List and F.A. Radu. A study on iterative methods for solving Richards' equation. *Computational Geosciences*, 20(2):341–353, 2016.

[17] E. Magenes, R.H. Nochetto, and C. Verdi. Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 21(4):655–678, 1987.

[18] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic–parabolic equations. *Journal of Differential Equations*, 131(1):20–38, 1996.

[19] I.S. Pop, F.A. Radu, and P. Knabner. Mixed finite elements for the Richards equation: linearization procedure. *Journal of Computational and Applied Mathematics*, 168(1):365–373, 2004.

[20] I.S. Pop and W.A. Yong. On the existence and uniqueness of a solution for an elliptic problem,. *Studia Universitatis Babeş-Bolyai Mathematica*, 45:97–107, 2000.

[21] I.S. Pop and W.A. Yong. A numerical approach to degenerate parabolic equations. *Numerische Mathematik*, 92(2):357–381, Aug 2002.

[22] F.A. Radu, K. Kumar, J.M. Nordbotten, and I.S. Pop. A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities. *IMA Journal of Numerical Analysis*, 38(2):884–920, 2018.

[23] 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. *Journal of Computational and Applied Mathematics*, 289:134 – 141, 2015.

[24] F.A. Radu, I.S. Pop, and P. Knabner. Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations. In *Numerical Mathematics and Advanced Applications*, pages 1192–1200. Springer, 2006.

[25] D. Seus, K. Mitra, I.S. Pop, F.A. Radu, and C. Rohde. A linear domain decomposition method for partially saturated flow in porous media. *Computer Methods in Applied Mechanics and Engineering*, 333:331 – 355, 2018.

[26] M. Slodicka. A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media. *SIAM Journal on Scientific Computing*, 23(5):1593–1614, 2002.

[27] J. Wang and C.V. Pao. Finite difference reaction–diffusion equations with nonlinear diffusion coefficients. *Numerische Mathematik*, 85(3):485–502, 2000.

A Proof of Propositions 3.1 and 3.2

Proof of Proposition 3.1. We consider the L-scheme for (3.4). Let u_n^i be the solution of the i^{th} iteration of the L-scheme (2.10) initiated by $u_n^0 = u_{n-1}$. We use the properties of the sequence $\{u_n^i\}_{i\in\mathbb{N}}$ to prove Proposition 3.1. From (2.10) and (3.1), u_n^i satisfies

$$(Lu_n^i,\phi)+\tau(\mathcal{D}\nabla u_n^i,\nabla\phi) = (Lu_n^{i-1},\phi)-(b(u_n^{i-1})-b(u_{n-1}),\phi)+\tau(\vec{F},\nabla\phi)+\tau(f_n^{i-1},\phi). \quad \text{(a.1)}$$

Let ρ^i denote the difference between consecutive iterates, i.e.

$$\rho^i = u_n^{i+1} - u_n^i.$$

Observe that for $i = 0$, ρ^0 solves the problem

$$L(\rho^0,\phi) + \tau(\mathcal{D}\nabla\rho^0,\nabla\phi) = \tau[(f(\vec{x},t_n,u_{n-1}),\phi) + (\vec{F},\nabla\phi) - (\mathcal{D}\nabla u_{n-1},\nabla\phi)].$$

As $\nabla\cdot(\mathcal{D}\nabla u_{n-1}) \in L^\infty(\Omega)$ is assumed in Proposition 3.1, ρ^0 satisfies the equation

$$L\rho^0 - \tau\nabla\cdot(\mathcal{D}\nabla\rho^0) = \tau[-\nabla\cdot\vec{F} + f(\vec{x},t_n,u_{n-1}) + \nabla\cdot(\mathcal{D}\nabla u_{n-1})],$$

in the classical sense (see [6][Section 6.3]). As the term $f(\vec{x}, t_n, u_{n-1}) - \nabla \cdot \vec{F} + \nabla \cdot (\mathcal{D}\nabla u_{n-1})$ is bounded, the maximum principle applies [6][Section 6.4] and therefore a $\mathfrak{C}_0 > 0$, independent of τ, exists such that

$$\|\rho^0\|_{L^\infty(\Omega)} = \mathfrak{C}_0\tau.$$

We claim that there exists a $\beta \in (0,1)$ such that

$$\|\rho^k\|_{L^\infty} \leq \beta \|\rho^{k-1}\|_{L^\infty}. \tag{a.2}$$

The value of β will be specified later. The proof of (a.2) is similar to the one for Lemma 3.1.

Subtracting (a.1) for i from the same equation written for $(i+1)^{\text{th}}$, we get an expression similar to (3.5), i.e.

$$L(\rho^i, \phi) + \tau(\mathcal{D}\nabla\rho^i, \nabla\phi) = ((L - \partial_u z(\zeta))\rho^{i-1}, \phi),$$

where $\zeta : \Omega \to \mathbb{R}$ is a function satisfying $\zeta \in \mathcal{I}(u_n^i, u_n^{i-1})$. Let $\mathfrak{C}_{i-1} = \|\rho^{i-1}\|_{L^\infty(\Omega)}/\tau$. Taking $\phi = [\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+$ as test function in the above and using the bounds for $\mathcal{D}$ gets

$$L(\rho^i, [\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+) + \tau\mathcal{D}_m\left\|\nabla[\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+\right\|^2 \leq \int_\Omega (L - \partial_u z(\zeta))\rho^{i-1}[\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+,$$

which implies that

$$L\|[\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+\|^2 + \tau\mathcal{D}_m\left\|\nabla[\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+\right\|^2 \leq \int_\Omega ((L - \partial_u z(\zeta))\rho^{i-1} - L\beta\mathfrak{C}_{i-1}\tau)[\rho^i - \beta\mathfrak{C}_{i-1}\tau]_+.$$

With this, we analyze the sign of the expression $(L - \partial_u z(\zeta))\rho^{i-1} - L\beta\mathfrak{C}_{i-1}\tau$. If $\rho^{i-1} \leq 0$ then this is clearly negative. If $\rho^{i-1} > 0$, since $\|\rho^{i-1}\|_{L^\infty(\Omega)} = \mathfrak{C}_{i-1}\tau$ and $0 \leq L - \partial_u z(\zeta) \leq L - m$, see (2.5), by taking

$$\beta = (L - m)/L < 1,$$

one obtains $(L - \partial_u z(\zeta))\rho^{i-1} - L\beta\mathfrak{C}_{i-1}\tau \leq [(L - m) - L\beta]\mathfrak{C}_{i-1}\tau = 0$. This shows that $\rho^i \leq \beta\mathfrak{C}_{i-1}\tau$ a.e. The inequality $\rho^i \geq -\beta\mathfrak{C}_{i-1}\tau$ follows in a similar fashion.

Observe that $\sum_{i=0}^\infty \|\rho^i\|_{L^\infty(\Omega)} \leq \mathfrak{C}_0\tau\sum_{i=0}^\infty \beta^i = \mathfrak{C}_0\tau/(1-\beta)$ and as $\rho^i = u_n^{i+1} - u_n^i$, it implies that $u_n^i \to u_n$ in the $L^\infty(\Omega)$-norm as $i \to \infty$. Note that, u_n^i also converges to u_n in the $H^1(\Omega)$ norm as shown in [19] and due to the uniqueness of the weak solutions of (3.4) (see [18]), u_n is the unique solution of (3.4). Finally, defining $\Lambda = \mathfrak{C}_0/(1-\beta)$ we get

$$\|u_n - u_{n-1}\|_{L^\infty(\Omega)} \leq \sum_{i=0}^\infty \|\rho^i\|_{L^\infty(\Omega)} \leq \Lambda\tau.$$

Rewriting (3.4) as

$$(\nabla \cdot (\mathcal{D}\nabla u_n), \phi) = \left(\frac{1}{\tau}(b(u_n) - b(u_{n-1})) + \nabla \cdot F - f_n, \phi\right),$$

we see right away that $\nabla \cdot (\mathcal{D}\nabla u_n) = \frac{1}{\tau}(b(u_n) - b(u_{n-1})) - f_n + \nabla \cdot F$ a.e. and as the terms on the right hand side are bounded in the $L^\infty(\Omega)$-norm, so is $\nabla \cdot (\mathcal{D}\nabla u_n)$. □

Proof of Proposition 3.2. The proof is almost identical to the proof of Proposition 3.1. We subtract (3.4) from (3.5) and follow the steps of the proof of Proposition 3.1 to obtain that $\left\|u_n^i - u_n\right\|_{L^\infty(\Omega)} \leq \beta \left\|u_n^{i-1} - u_n\right\|_{L^\infty(\Omega)}$ if

$$\beta \geq \max \left\{ \frac{L_n^i - \partial_u z(\zeta)}{L_n^i} \right\}, \quad \zeta \in \mathcal{I}(u_n, u_n^{i-1}).$$

Observe that $L_n^i \geq m + \mathfrak{M}\tau$. Moreover, from (3.6) we obtain, $L_n^i - \partial_u z(\zeta) \leq 2\mathfrak{M}\tau$ for $\mathfrak{M} \geq \mathfrak{M}_0$. Hence,

$$\frac{L_n^i - \partial_u z(\zeta)}{L_n^i} \leq \frac{2\mathfrak{M}\tau}{m + \mathfrak{M}\tau} \leq \frac{2\mathfrak{M}\tau}{m}. \tag{a.3}$$

By defining,

$$\beta = \frac{2\mathfrak{M}\tau}{m}, \tag{a.4}$$

we observe that for $\tau < \frac{m}{2\mathfrak{M}}$, $\beta < 1$ and $\beta = \mathcal{O}(\tau)$.

□

UHasselt Computational Mathematics Preprint Series

2018

UP-18-06 *Koondanibha Mitra, Iuliu Sorin Pop*, **A modified L-Scheme to solve nonlinear diffusion problems**, 2018

UP-18-05 *David Landa-Marban, Na Liu, Iuliu Sorin Pop, Kundan Kumar, Per Pettersson, Gunhild Bodtker, Tormod Skauge, Florin A. Radu*, **A pore-scale model for permeable biofilm: numerical simulations and laboratory experiments**, 2018

UP-18-04 *Florian List, Kundan Kumar, Iuliu Sorin Pop and Florin A. Radu*, **Rigorous upscaling of unsaturated flow in fractured porous media**, 2018

UP-18-03 *Koondanibha Mitra, Hans van Duijn*, **Wetting fronts in unsaturated porous media: the combined case of hysteresis and dynamic capillary**, 2018

UP-18-02 *Xiulei Cao, Koondanibha Mitra*, **Error estimates for a mixed finite element discretization of a two-phase porous media flow model with dynamic capillarity**, 2018

UP-18-01 *Klaus Kaiser, Jonas Zeifang, Jochen Schütz, Andrea Beck and Claus-Dieter Munz*, **Comparison of different splitting techniques for the isentropic Euler equations**, 2018

2017

UP-17-12 *Carina Bringedal, Tor Eldevik, Øystein Skagseth and Michael A. Spall*, **Structure and forcing of observed exchanges across the Greenland-Scotland Ridge**, 2017

UP-17-11 *Jakub Wiktor Both, Kundan Kumar, Jan Martin Nordbotten, Iuliu Sorin Pop and Florin Adrian Radu*, **Linear iterative schemes for doubly degenerate parabolic equations**, 2017

UP-17-10 *Carina Bringedal and Kundan Kumar*, **Effective behavior near clogging in upscaled equations for non-isothermal reactive porous media flow**, 2017

UP-17-09 *Alexander Jaust, Balthasar Reuter, Vadym Aizinger, Jochen Schütz and Peter Knabner*, **FESTUNG: A MATLAB / GNU Octave toolbox for the discontinuous Galerkin method. Part III: Hybridized discontinuous Galerkin (HDG) formulation**, 2017

UP-17-08 *David Seus, Koondanibha Mitra, Iuliu Sorin Pop, Florin Adrian Radu and Christian Rohde*, **A linear domain decomposition method for partially saturated flow in porous media**, 2017

UP-17-07 *Klaus Kaiser and Jochen Schütz*, **Asymptotic Error Analysis of an IMEX Runge-Kutta method**, 2017

UP-17-06 *Hans van Duijn, Koondanibha Mitra and Iuliu Sorin Pop*, **Travelling wave solutions for the Richards equation incorporating non-equilibrium effects in the capillarity pressure**, 2017

UP-17-05 *Hans van Duijn and Koondanibha Mitra*, **Hysteresis and Horizontal Redistribution in Porous Media**, 2017

UP-17-04 *Jonas Zeifang, Klaus Kaiser, Andrea Beck, Jochen Schütz and Claus-Dieter Munz*, **Efficient high-order discontinuous Galerkin computations of low Mach number flows**, 2017

UP-17-03 *Maikel Bosschaert, Sebastiaan Janssens and Yuri Kuznetsov*, **Switching to nonhyperbolic cycles from codim-2 bifurcations of equilibria in DDEs**, 2017

UP-17-02 *Jochen Schütz, David C. Seal and Alexander Jaust*, **Implicit multiderivative collocation solvers for linear partial differential equations with discontinuous Galerkin spatial discretizations**, 2017

UP-17-01 *Alexander Jaust and Jochen Schütz*, **General linear methods for time-dependent PDEs**, 2017

2016

UP-16-06 *Klaus Kaiser and Jochen Schütz*, **A high-order method for weakly compressible flows**, 2016

UP-16-05 *Stefan Karpinski, Iuliu Sorin Pop, Florin A. Radu*, **A hierarchical scale separation approach for the hybridized discontinuous Galerkin method**, 2016

UP-16-04 *Florin A. Radu, Kundan Kumar, Jan Martin Nordbotten, Iuliu Sorin Pop*, **Analysis of a linearization scheme for an interior penalty discontinuous Galerkin method for two phase flow in porous media with dynamic capillarity effects** , 2016

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

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

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