UNIVERSITY OF BERGEN
DEPARTMENT OF MATHEMATICS

Efficient solvers for Richards' equation

Author: Jakob Seierstad Stokke
Supervisors: Florin Adrian Radu & Jakub Wiktor Both

January, 2023

Contents

Introduction **9**

1 Essential theory **11**
1.1 Flow in porous media . . . 11
1.1.1 Porous medium . . . 11
1.1.2 Darcy's law . . . 12
1.1.3 Mass conservation . . . 13
1.1.4 Two-phase flow model . . . 13
1.1.5 Richards' equation . . . 14
1.2 The finite element method . . . 15
1.2.1 Function spaces . . . 15
1.2.2 Variational formulation . . . 18
1.2.3 Existence and uniqueness . . . 19
1.2.4 Galerkin finite element method . . . 21
1.2.5 Mixed finite element method . . . 22
1.2.6 Convergence of FEM . . . 22
1.3 Iterative methods . . . 23
1.3.1 Order of convergence . . . 25
1.4 Linearization methods . . . 27
1.4.1 Newton's method . . . 27
1.4.2 L-scheme linearizations . . . 29
1.4.3 Relaxation strategies for linearization methods . . . 29

2 Solution techniques for Richards' equation **31**
2.1 Temporal discretization . . . 31
2.2 Spatial discretization . . . 32
2.3 Linearizations . . . 32
2.3.1 L-scheme . . . 32
2.3.2 Modified L-scheme . . . 34
2.3.3 Newton's method . . . 35
2.4 Error estimates . . . 38

3 L-scheme/Newton switching **41**
3.1 L-scheme to Newton switching . . . 42
3.2 Newton to L-scheme switching . . . 44
3.3 A-posteriori estimate based adaptive linearization algorithm . . . 46
3.3.1 Computation of equilibrated flux . . . 46
3.3.2 Additional computational considerations . . . 47
3.3.3 Adaptive linearization algorithm . . . 47

4 Numerical results **49**
4.1 Example 1: Strictly unsaturated medium 50
4.1.1 Comparison of convergence properties 51
4.1.2 Switching characteristics 54
4.1.3 Order of convergence 55
4.2 Example 2: Variably saturated medium 57
4.2.1 Comparison of convergence properties 57
4.2.2 Switching characteristics 58
4.3 Example 3: Benchmark problem 59
4.3.1 Comparison of convergence properties 61
4.3.2 Switching characteristics 61
4.3.3 Order of convergence 62
4.4 Conclusions 63

5 Summary **65**

A Convergence proof of L-scheme **71**

B L-adaptivity **73**
B.1 An Adaptive L-scheme algorithm 73
B.2 Numerical results 74

Abstract

In this thesis we study efficient solvers for Richards' equation, a non-linear, degenerate, elliptic-parabolic equation which models flow in saturated/unsaturated porous media. We examine the numerical characteristics of several linearization methods, including the L-scheme, Newton's method and the modified L-scheme. Also Anderson acceleration is used on both the L-scheme and Newton's method. An extension of the linear and global convergence proof of the L-scheme including gravity is shown. In addition to this, the optimal stabilization parameter analysis is extended. For a variant of the Kirchhoff transformed Richards' equation, we show the quadratic convergence of Newton's method. The main result of the thesis is a proposed efficient and robust switching algorithm between the L-scheme and Newton's method, exploiting the unconditional convergence of the L-scheme and the quadratic convergence of Newton's method. We propose an algorithm which adaptively changes between the linearization techniques, based on *a posteriori* estimators. The latter has not previously been done. The performance of the algorithm is tested through realistic examples.

Acknowledgements

First, I would like to thank my supervisors Florin Adrian Radu and Jakub Wiktor Both for their guidance and help. Both of you have taught me a lot and been great mentors in the process of writing this thesis. Florin, I am very grateful for your interest in my mathematical studies from an early stage and for suggesting topics for the thesis. Jakub, I really appreciate your help and advise with both coding and debugging.

I would also like to thank Koondanibha Mitra and Erlend Storvik for insightful discussions. I am also thankful for the inspiration I have received from the porous media group's weekly seminars.

Introduction

Richards' equation is a special case of two-phase flow in porous media, describing the subsurface flow of water in both saturated and unsaturated soils which was first proposed by L.A. Richards in 1931 [52]. It plays a significant role in solving relevant societal problems, e.g., soil erosion, irrigation, and environmental pollution.

This thesis will focus on efficient solvers for Richards' equation, and we will use the pressure head formulation of Richards' equation

$$\partial_t \theta(\psi) - \nabla \cdot (K(\theta(\psi))\nabla(\psi + z)) = f, \tag{0.1}$$

which will be introduced in Section 1.1.5.

Richards' equation is a non-linear, degenerate, elliptic-parabolic equation. This causes challenges when solving the equation numerically, see e.g. the review work of [24]. In general, a solution to Richards' equation lacks regularity [2]. Therefore, it is common to use the backward Euler scheme for the temporal discretization to allow for larger time steps. Also, the benefit of higher order schemes are lost due to the low regularity of solutions. For the spatial discretization we will use Galerkin finite elements as done in [4, 5, 34]. But there are many other alternatives including finite volume schemes [7, 22, 23], multipoint flux approximation [30], or mixed and expanded mixed finite elements [6, 8, 9, 46, 50]. Regardless of the spatial discretization employed, a non-linear finite dimensional problem has to be solved at each time step. The focal point of this thesis will be how to efficiently solve these non-linear problems using iterative linearization techniques.

There are many linearization techniques employed to solve the non-linear problems. Most of the methods are linearly converging fixed-point schemes. One is the Picard method, however it does not perform well for Richards' equation, see e.g. [22]. An alteration was proposed in [16], known as the modified Picard method which performs better. In [45, 53], it was proposed to use a global parameter to stabilize the modified Picard method. This scheme, known as the L-scheme, is more robust and allows for larger time steps. Additionally, the computational time is lower, since no computation of derivatives is needed and the matrices are better conditioned, indicated by numerical results in [34]. Nonetheless, these results also show that the number of iterations needed for convergence is quite high. The convergence rate is also heavily influenced by the global stabilization parameter, see Section 2.3.1. This sensitivity can be decreased by applying Anderson acceleration [3] to the L-scheme. Another scheme is the modified L-scheme [36], which shows similar stability properties as the L-scheme, but has a faster convergence in terms of number of iterations for smaller time steps while still having a linear convergence rate.

The most common technique for solving non-linear problems is Newton's method [9, 46] being quadratically convergent if the initial guess is sufficiently close. However, for degenerate problems, Newton's method may fail to converge. Methods of improving the robustness of Newton's method exists, such as a parametrization switching approach [12]. A different approach is to first compute a few fixed-point iterations, such as using the Picard method [33]

or using the L-scheme [34]. But in these cases the switching has been done heuristically and not by an *a posteriori* indicator.

In this thesis, we explore a hybrid linearization strategy, which adaptively switches between the L-scheme and Newton's method. In this way, one takes advantage of the robustness and global convergence of the L-scheme and the quadratic convergence of Newton's method. The main difficulty of this strategy is deriving reliable estimates to determine the switch between the schemes. *A priori* estimates, such as the one given in Theorem 2.3.3 involves unknowns, including a regularization parameter as we assume a worst case scenario, therefore we seek an *a posteriori* estimate instead. An efficient *a posteriori* estimator for the fully degenerate Richards equation was derived in [37]. Also, in [38] a reliable estimator was derived using a decomposition of the total error into a linearization and discretization component. We are interested in predicting if the linearization component decreases to determine when to switch. In Chapter 3, we derive *a posteriori* switching criteria, from which we propose an adaptive, efficient and reliable switching algorithm for Richard's equation.

Furthermore, we compare the proposed algorithm to the L-scheme, the modified L-scheme and Newton's method through realistic test cases. Also included is the combined L-scheme and Anderson acceleration. Anderson acceleration has also been shown to increase the robustness of iterative methods, consequently we consider a Newton-Anderson algorithm [43]. The comparison considers the number of iterations and the computational time. In addition the order of convergence is studied numerically by concepts in [15].

The main contribution of the thesis is the proposed hybrid linearization algorithm, and it has been submitted for review [54]. The implementation of the algorithm can be found on `https://github.com/MrShuffle/RichardsEquation/releases/tag/v1.0.1`.

Outline

Chapter 1 introduces essential background theory. First we give a short introduction of flow in porous media, where Richards' equation is derived. In Section 1.2 we discuss the finite element method which we will use for the spatial discretization of Richards' equation. Then iterative schemes are introduced, with a particular focus on the order of convergence. Several linearization techniques which is employed in the thesis are discussed in Section 1.4. Also, we introduce Anderson acceleration.

In Chapter 2 we discretize Richards' equation using the backward Euler method in time and continuous Galerkin finite elements in space. We linearize the resulting non-linear problem in three different ways, using the L-scheme, modified L-scheme and Newton's method. Here we extend the previous convergence and optimality analysis of the L-scheme. We also give an error estimate on the solution of the L-scheme. Additionally, by using the Kirchhoff transformation on Richards' equation and introducing a regularization parameter, we prove the quadratic convergence of Newton's method.

In Chapter 3 we derive *a posteriori* estimators to predict the linearization error of the next iterate. Based on these estimators we propose the adaptive solution strategy for Richards' equation.

In Chapter 4 the numerical performance of all schemes is compared in using three realistic examples. They show the robustness and computational efficiency of the adaptive algorithm. We also show the order of convergence for the linearization techniques.

Chapter 1

Essential theory

1.1 Flow in porous media

In the following section we give an overview of some of the fundamentals of flow in porous media. The introduction will primarily focus on Richards' equation and be based upon [48]. We refer to [41] for a more detailed overview of flow in porous media.

1.1.1 Porous medium

A *porous medium* is a material that has pores in it. In this thesis we are particularly interested in media in which the pores are connected, allowing for fluid flow. When only one fluid flows through the pores, we call it single-phase flow; when two fluids flow through the pores, we call it two-phase flow, et cetera. Generally the structure of the porous medium is unknown, such as which part is a pore or solid. Thus one wants to upscale the equations by viewing the medium through a *representative elementary volume* (REV).

The REV is a volume we associate with each point in our domain. Therefore, each point is both in the solid material and the pore space at the same time. Consequently, every property of the porous medium is an average property of the REV. The size of the REV must be large enough to allow for a meaningful average, while still allowing us to differentiate inhomogeneities in the medium. While the REV is small, the average value of for example porosity will tend to oscillate and will dampen as the volume increases and where it flattens out is our desired size of the REV, see Figure 1.1.

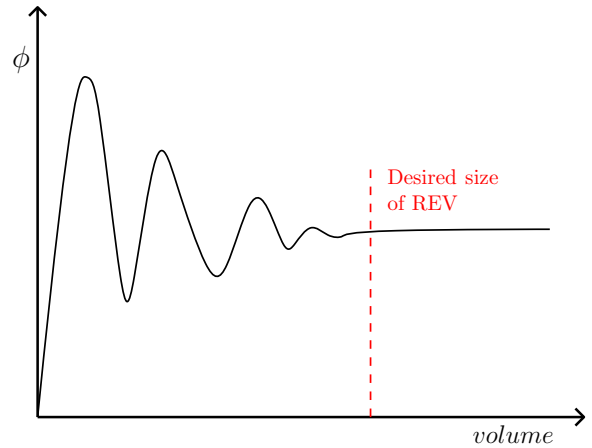


Figure 1.1: Variation of porosity in relation to the size of REV.

An important property of a porous medium is the *porosity* ϕ, which is the volume of the pore space in the REV divided by the volume of the REV

$$\phi = \frac{\text{vol(pores in REV)}}{\text{vol(REV)}}.$$

Another important property is the *saturation* S_α of a fluid phase α, which is the ratio between the volume of a fluid phase α in the REV and the volume of pores in the REV

$$S_\alpha = \frac{\text{vol(fluid phase } \alpha \text{ in REV)}}{\text{vol(pores in REV)}}.$$

From the above we get the definition of the volume of a fluid phase θ_α as the product of the saturation and the porosity

$$\theta_\alpha = S_\alpha \phi.$$

If there are multiple fluids in the porous medium and we assume that the medium is fully saturated, i.e., the pore space is filled with fluids, then

$$\sum_\alpha S_\alpha = 1, \tag{1.1}$$

and as a consequence

$$\sum_\alpha \theta_\alpha = \phi.$$

1.1.2 Darcy's law

Henri Darcy conducted an experiment in 1856 [19], where he measured the flow of water through sand within a tube. He observed that the flow was proportional to the difference in the hydraulic head h between two points and the cross-sectional area A and also inverse proportional to the length of the tube l, i.e.

$$\mathbf{q}_s = -\mathbf{k}\frac{A(h_2 - h_1)}{l}\mathbf{e},$$

where $\mathbf{e}$ is a unit vector describing the direction of the flow and the proportionality coefficient $\mathbf{k}$ is called the hydraulic conductivity. The volumetric flow rate per area is

$$\mathbf{q} = \frac{\mathbf{q}_s}{A},$$

and by letting the length of the tube approach zero results in the differential form of Darcy's law

$$\mathbf{q} = -\mathbf{k}\nabla h. \tag{1.2}$$

In d dimensions $\mathbf{k}$ will be a $\mathbb{R}^{d\times d}$ tensor, which will be symmetric, since at any point in the domain the material will have a direction with maximum hydraulic conductivity and one with minimum. Through dimensional analysis one can derive the following relation

$$\mathbf{k} = \frac{\boldsymbol{\kappa}\rho g}{\mu},$$

where $\boldsymbol{\kappa} \in \mathbb{R}^{d\times d}$ is a property of the porous medium called permeability, g is the gravity, μ is the viscosity and ρ is density of the fluid. Assuming the fluid is incompressible, the hydraulic head in (1.2) can be substituted by the pressure head ψ to obtain the pressure head formulation of Darcy's law

$$\mathbf{q} = -\frac{\boldsymbol{\kappa}\rho g}{\mu}\nabla(\psi + z). \tag{1.3}$$

1.1.3 Mass conservation

A priori the pressure is not known, therefore we do not have enough to describe flow through a porous medium. We can close the system by an important physical principle, namely the conservation of mass. Consider an arbitrary volume Ω and $\boldsymbol{\nu}$ the outward pointing normal vector for a fluid phase. If mass is conserved within Ω, the change of mass flowing through Ω has to be balanced by in- and outflow $\mathbf{Q}$ at the boundary $\partial\Omega$ and mass sources f within Ω,

$$\int_\Omega \partial_t m dV = -\int_{\partial\Omega} \mathbf{Q}\cdot\boldsymbol{\nu} dA + \int_\Omega f dV, \quad \forall\Omega.$$

By using the Gauss' theorem in the boundary flux, we obtain

$$\int_\Omega \partial_t m dV + \int_\Omega \nabla\cdot\mathbf{Q} dV = \int_\Omega f dV, \quad \forall\Omega.$$

Since Ω was arbitrary, we get the local mass balance equation for flow in porous media where $m = \rho\theta$ and the flux as the flow $\mathbf{Q} = \rho\mathbf{q}$,

$$\partial_t(\rho\theta) + \nabla\cdot(\rho\mathbf{q}) = f. \tag{1.4}$$

Combining the mass balance equation and Darcy's law (1.3) with suitable boundary and initial conditions gives a closed model of single phase flow,

$$\begin{cases} \partial_t(\rho\theta) + \nabla\cdot(\rho\mathbf{q}) = f(\mathbf{x},t), & \mathbf{x}\in\Omega, \quad t>0, \\ \mathbf{q} = -\dfrac{\boldsymbol{\kappa}\rho g}{\mu}\nabla(\psi(\mathbf{x},t)+z), & \mathbf{x}\in\Omega, \quad t>0. \end{cases} \tag{1.5}$$

This problem is a linear parabolic equation, however for an incompressible fluid and non-deformable porous medium it becomes an elliptic equation.

1.1.4 Two-phase flow model

We can extend the single-phase flow model by assuming that we have multiple fluids. Since Richards' equation is an example of two phase-flow, we restrict ourselves to two fluids, one wetting fluid w and one non-wetting fluid n.

The saturation of the fluids are governed by the relation (1.1) and we assume that both fluids follow the mass balance principle with no mass transfer between the fluids and Darcy's law. However, in two phase-flow the fluids share the pore space which makes the flow of each fluid more complex and one would expect the hydraulic conductivity to be lower. Therefore one extends Darcy's law by introducing the relative permeability, $\kappa_{r,\alpha}$, which is assumed to be a scalar determined by the saturation, i.e $\kappa_{r,\alpha} = \kappa_{r,\alpha}(S_\alpha)$, where $\alpha = \{n,w\}$. Darcy's law for multiple phases then reads

$$\mathbf{q}_\alpha = -\frac{\kappa_{r,\alpha}(S_\alpha)\boldsymbol{\kappa}_\alpha\rho_\alpha g}{\mu}\nabla(\psi_\alpha + z).$$

Generally the wetting fluid moves easier through narrow pores due to interface forces between the solid and the fluids. This phenomenon can be described by the capillary pressure, p_c, namely the difference between the pressures in the fluids,

$$p_c = p_n - p_w.$$

It has also been observed that the capillary pressure increases when the volume of the wetting fluid decreases, as it will move towards more narrow pores. If the volume of the wetting fluid increases, it will remain in bigger pores causing the capillary pressure to decrease. Therefore, we assume that the capillary pressure is a function of the saturation to the wetting fluid, i.e. $p_c = p_c(S_w)$.

This leads to a consistent two phase-flow model for immiscible fluids

$$\begin{cases} \partial_t(\rho_\alpha \theta_\alpha) + \nabla \cdot (\rho_\alpha \mathbf{q}_\alpha) = f(\mathbf{x},t), & \mathbf{x} \in \Omega, \quad t > 0, \\ \mathbf{q}_\alpha = -\dfrac{\kappa_{r,\alpha}(S_\alpha)\boldsymbol{\kappa}_\alpha \rho_\alpha g}{\mu_\alpha} \nabla(\psi_\alpha(\mathbf{x},t) + z), & \mathbf{x} \in \Omega, \quad t > 0, \\ S_w + S_n = 1, \\ p_c(S_w) = p_n - p_w, \end{cases} \tag{1.6}$$

with appropriate boundary and initial conditions, where the relative permeability and capillary pressure are given functions.

1.1.5 Richards' equation

A special case of two-phase flow is Richards' equation where the fluids are air and water. We assume that air always has a constant pressure, $p_n = 0$, and the density of the water is assumed to be constant, which allows us to simplify the capillary pressure,

$$p_c(S_w) = -p_w = -\psi_w \rho g.$$

In fact the capillary pressure have been shown experimentally to be a decreasing monotone function of saturation, meaning we can invert it

$$S_w = p_c^{-1}(\psi_w \rho g) \Rightarrow \phi\, p_c^{-1}(\psi_w \rho g) = \theta_w(\psi_w).$$

There are different parametrizations of the hydraulic conductivity as a function of the water content based on experiments, e.g. the van Genuchten-Mualem model [25, 39] or the Brooks-Corey model [14]. So the hydraulic conductivity is written as $K(\theta) = \frac{\kappa_{r,w}(\theta_w)\boldsymbol{\kappa}_w \rho_w g}{\mu_w}$.

Then combing mass balance and Darcy's law from the two phase-flow model (1.6), we obtain the pressure head based Richards' equation

$$\partial_t \theta(\psi) - \nabla \cdot (K(\theta(\psi)) \nabla(\psi + z))) = f. \tag{1.7}$$

This is a degenerate elliptic-parabolic equation, as it will degenerate into an elliptic equation if $\partial_t \theta(\psi) = 0$. The degeneracy is depicted in Figure 1.2. The two non-linear terms K and θ will make solving the equation numerically and the analysis quite challenging, as we will see later.

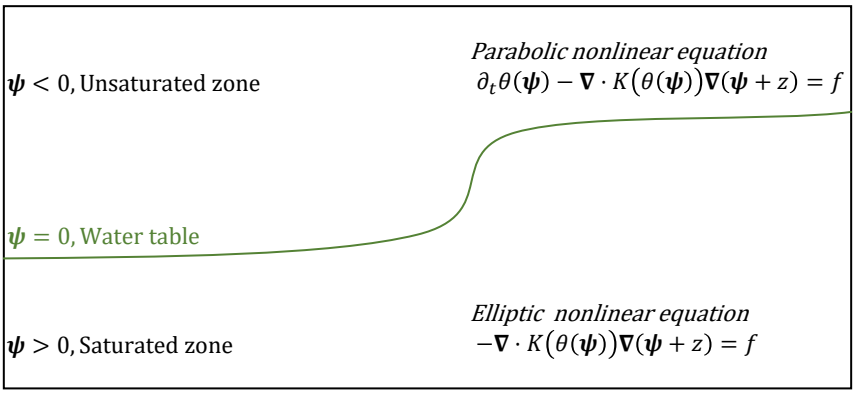


Figure 1.2: Degeneracy of Richards' equation

1.2 The finite element method

Many physical phenomena can be described by partial differential equations (PDEs) which we need to solve. Solving these analytically might be difficult and for non-trivial cases it is impossible to find an exact solution, instead one uses numerical approximation techniques to find an approximate solution. This is done by a finite dimensional discretization in time and space. In time we will employ an implicit discretization, namely the backward Euler method. In space we can consider several different approximation methods e.g., finite volumes, finite differences and finite elements. In this section we will look at the finite element method (FEM), which has a strong mathematical foundation including convergence and stability analysis.

1.2.1 Function spaces

An essential aspect when discussing PDEs and the finite element method is what kind of functions we have and what their properties are. Therefore we give an introduction into Lebesgue and Sobolev spaces, but for a more comprehensive overview see [1].

Definition 1.2.1. For $p \in [1,\infty)$ the L^p-spaces, or Lebesgue spaces are

$$L^p(\Omega) = \left\{ f : \|f\|_{L^p(\Omega)} = \left(\int_\Omega |f|^p dx \right)^{\frac{1}{p}} < \infty \right\},$$

and if $p = \infty$

$$L^\infty(\Omega) = \left\{ f : \|f\|_{L^\infty} = ess\sup_{x\in\Omega} |f| \right\}.$$

Consider functions $f_i \in L^p(\Omega)$ where $i \in \mathbb{N}$, and $f_i(1) = i$. Then $\|f_i\|_{L^p(\Omega)} = 0$ for all i. All these functions can be considered equivalent as the difference in the L^p-norm is zero. Therefore it is common to think of the elements in L^p-spaces as equivalence classes of functions, since the L^p-norm is not "strong" enough to measure point values.

Definition 1.2.2. Let $\{\mathbf{x}_k\}$ be a sequence in a normed linear space, $(U, \|\cdot\|)$. The sequence $\{\mathbf{x}_k\}$ is called a Cauchy sequence if for all $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that $\sup_{i,j \geq N} \|\mathbf{x}_j - \mathbf{x}_i\| \leq \varepsilon, \forall i, j \geq N$.

Definition 1.2.3. If every Cauchy sequence in the space U converges to an element in U, the space is said to be complete. A complete, normed vector space is called a Banach space. A Banach space with an inner product $\langle \cdot, \cdot \rangle$ which induces a norm, $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ is called a Hilbert space.

Theorem 1.2.1 (Riesz-Fischer theorem, [18] Chapter 8)**.** *The L^p-spaces are Banach spaces.*

Definition 1.2.4. Let $\boldsymbol{\alpha}$ be a d-tuple, i.e $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_d)$ and the length $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$. Then

$$D^{\boldsymbol{\alpha}} v := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} v.$$

In the finite element method the variational formulation is defined globally with integrals on Ω, however the derivatives in calculus is defined pointwise. These pointwise values are not needed, but instead derivatives which can be thought of as functions in L^p-spaces are used. Therefore one would prefer a definiton of the derivative in a more general setting suitable to the L^p-spaces. First one needs the set of all locally integrable functions in Ω,

$$L^1_{loc}(\Omega) := \left\{ f : f \in L^1(K), \text{ for any compact } K \subset \Omega \right\}.$$

Note that all continuous functions of Ω is contained in $L^1_{loc}(\Omega)$.

Definition 1.2.5. Let $f \in L^1_{loc}(\Omega)$, then $D^{\boldsymbol{\alpha}} f$ is the weak derivative of f, if there exists a $g \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega g v dx = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega f D^{\boldsymbol{\alpha}} v dx, \quad \forall v \in C_0^\infty(\Omega),$$

where $g = D^{\boldsymbol{\alpha}} f$.

Note that if the classical derivative exists, then the weak derivative exists aswell.

Definition 1.2.6. Let r be a non-negative integer and $p \in [1, \infty)$, the Sobolev norm is

$$\|f\|_{W^{r,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}| \leq r} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}},$$

and in the case $p = \infty$

$$\|f\|_{W^{r,\infty}(\Omega)} := \max_{\boldsymbol{\alpha} \leq r} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}.$$

Then the Sobolev spaces are

$$W^{r,p}(\Omega) = \left\{ f \in L^1_{loc}(\Omega) : \|f\|_{W^{r,p}(\Omega)} < \infty \right\}.$$

Theorem 1.2.2 ([1], chapter 3)**.** *The Sobolev space $W^{r,p}$ is a Banach space.*

A subset of the Sobolev spaces of particular importance is $W^{r,2}$, which will be denoted H^r. In fact, the spaces H^r is associated with an inner product,

$$\langle u, v\rangle_r = \sum_{|\boldsymbol{\alpha}|\leq r} \int_\Omega D^{\boldsymbol{\alpha}} u D^{\boldsymbol{\alpha}} v dx,$$

which induces a norm on H^r,

$$\|u\|_{H^r(\Omega)} = \sqrt{\langle u, u\rangle_r} = \sqrt{\sum_{|\boldsymbol{\alpha}|\leq r} \|D^{\boldsymbol{\alpha}} u\|_{L^2(\Omega)}}.$$

It follows from theorem 1.2.2 that every Sobolev space H^r is a Hilbert space.

Definition 1.2.7. Let $\Omega \subset \mathbb{R}^n$ be bounded and open, then we denote the closure of all indefinitely differentiable functions with compact support in Ω by $H_0^r(\Omega)$.

Often we deal with boundary value problems, as such we must give meaning to how an element of $H^r(\Omega)$ is defined on the boundary of Ω. The issue being that an element in $H^r(\Omega)$ is defined up to a set of points of measure zero, but the boundary is a set of measure zero. As such, the notion of a trace operator T gives meaning to the restriction of an element in $H^r(\Omega)$ to the boundary. We restrict ourselves to only consider $H^1(\Omega)$.

Theorem 1.2.3 ([21], Chapter 5)**.** *Let Ω be a bounded Lipschitz domain. Then there exists a bounded linear operator*

$$T : H^1(\Omega) \to L^2(\partial\Omega)$$

such that

$$Tu = u|_{\partial\Omega} \qquad \text{if } u \in H^1(\Omega) \cap C(\bar{\Omega}), \tag{1.8a}$$

$$\|Tu\|_{L^p(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)} \qquad \text{for } u \in H^1(\Omega). \tag{1.8b}$$

Theorem 1.2.4 ([21], Chapter 5)**.** *Let Ω be a bounded Lipschitz domain and $u \in H_0^1(\Omega)$ then*

$$u \in H_0^1(\Omega) \Leftrightarrow Tu = 0 \text{ on } \partial\Omega.$$

Another Hilbert space of importance is $H(\nabla\cdot, \Omega)$, where $f \in H(\nabla\cdot, \Omega)$ implies $f, \nabla \cdot f \in L^2(\Omega)$. An important inequality which is used later is the Poincaré inequality.

Theorem 1.2.5 (Poincaré's inequality)**.** *Let Ω be a bounded open set and $u \in H_0^1(\Omega)$, then there exists a constant C_Ω such that*

$$\|u\|_{L^2(\Omega)} \leq C_\Omega \|\nabla u\|_{L^2(\Omega)}. \tag{1.9}$$

By equipping $H_0^1(\Omega)$ with the inner product of $H^1(\Omega)$, it will be a Hilbert space.

Definition 1.2.8. For any space V, we denote the dual space, i.e. the space of all linear bounded functionals by V' with the dual norm on V' by

$$\|u\|_{V'} = \sup\{\langle u, v\rangle : v \in V, \|u\|_V \leq 1\}.$$

For $H_0^1(\Omega)$, its dual space will be denoted by $H^{-1}(\Omega)$.

1.2.2 Variational formulation

Consider the Poisson equation,

$$\begin{cases} \nabla \cdot (-\nabla u(\mathbf{x})) = f, & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega, \end{cases} \tag{1.10}$$

where $f : \Omega \to \mathbb{R}$, Ω is a bounded and connected domain in $\mathbb{R}^d$ and $\partial\Omega$ is the boundary of Ω.

The first step of the finite element method is to rewrite the PDE in question into its variational formulation. By integrating over the domain Ω and multiplying with a sufficiently regular test function v in a test function space V, in our case $V = H_0^1(\Omega)$, we obtain

$$\int_\Omega \nabla \cdot (-\nabla u)\, v d\mathbf{x} = \int_\Omega f v d\mathbf{x}, \quad \forall v \in H_0^1(\Omega).$$

Using integration by parts one gets,

$$\int_\Omega \nabla u \cdot \nabla v d\mathbf{x} - \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} v = \int_\Omega f v d\mathbf{x}, \quad \forall v \in H_0^1(\Omega),$$

where $\mathbf{n}$ is the outward pointing normal vector of Ω. Since $v \in H_0^1(\Omega)$, the boundary integral will vanish. Also, note that the boundary integral has a meaning due to the trace operator allowing us to define $u, v \in L^2(\Omega)$. The variational formulation of (1.10) then reads; Find u such that

$$\int_\Omega \nabla u \cdot \nabla v d\mathbf{x} = \int_\Omega f v d\mathbf{x}, \quad \forall v \in H_0^1(\Omega). \tag{1.11a}$$

In order for the variational formulation to make sense, the product fv must be integrable and the same goes for $\nabla u \cdot \nabla v$, also the boundary conditions must be satisfied. They are automatically satisfied by our choice of test space $H_0^1(\Omega)$ and the variational formulation will be well defined if also $f \in L^2(\Omega)$. The solution space may be different from the test space, but if they are the same it is called a conformal finite element method, which is what we will use in this thesis. Clearly if u solves (1.10), it is also a solution to (1.11a). However, a solution of (1.11a) needs less smoothness than a solution of (1.10), meaning that it is not necessarily a solution of the original problem.

Another way of writing (1.11a) is as an abstract variational formulation; Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = b(v), \quad \forall v \in H_0^1(\Omega), \tag{1.11b}$$

where $a(\cdot, \cdot)$ is a bilinear form and $b(\cdot)$ is a linear functional. As the integrals define an inner product in H_0^1, this means that in the case for the Poisson equation $a(u, v) = \langle \nabla u, \nabla v \rangle$ and $b(v) = \langle f, v \rangle$.

For non-homogenous Dirichlet boundary conditions, i.e. $u = g$, on $\partial\Omega$, for (1.10), the variational formulation would need the following function space,

$$H_g^1(\Omega) = \left\{ u \in H^1(\Omega) : u = g, \text{ on } \partial\Omega \right\}.$$

The problem with the above space is that when $g \neq 0$ it is not linear. If $w \in H_g^1(\Omega)$, then on the boundary $w + w = 2g$, therefore $w + w \notin H_g^1(\Omega)$. To resolve the problem consider a function

$\hat{g} \in C^\infty$ (or $H^1(\Omega)$) such that $\hat{g}|_{\partial\Omega} = g$, then by defining $\hat{u} = u - \hat{g} \in H_0^1(\Omega)$, one can express the original problem with homogenous boundary conditions.

Neumann boundary conditions are often seen in boundary value problems, as they typically have a strong physical motivation, like free flow across the boundary. The Poisson equation with a Neumann condition can be written as

$$\begin{cases} \nabla \cdot (-\nabla u(\mathbf{x})) = f, & \mathbf{x} \in \Omega, \\ -\nabla u(\mathbf{x}) \cdot \boldsymbol{\nu} = h, & \mathbf{x} \in \Gamma_N \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega \backslash \{\Gamma_N\}, \end{cases} \tag{1.12}$$

where $\boldsymbol{\nu}$ is the outward pointing normal vector. The corresponding variational formulation is; Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v d\mathbf{x} = \int_\Omega f v d\mathbf{x} + \int_{\Gamma_N} h v ds, \quad \forall v \in H_0^1(\Omega). \tag{1.13}$$

In this thesis, we will use an equilibrated flux for the *a posteriori* estimates in the proposed switching algorithm. The flux, $\boldsymbol{\sigma} \in H(\nabla\cdot, \Omega)$, can be found by solving a mixed variational problem stemming from a system of equations. For example, the Poisson equation (1.10), can be written as a system of equations leading to the mixed variational problem; Find $(\boldsymbol{\sigma}, u) \in H(\nabla\cdot, \Omega) \times L^2(\Omega)$ such that

$$\langle \boldsymbol{\sigma}, \boldsymbol{q} \rangle + \langle u, \nabla \cdot \boldsymbol{q} \rangle = 0, \qquad \forall \boldsymbol{q} \in H(\nabla\cdot, \Omega), \tag{1.14a}$$

$$\langle \nabla \cdot \boldsymbol{\sigma}, v \rangle = \langle f, v \rangle, \qquad \forall v \in L^2(\Omega). \tag{1.14b}$$

1.2.3 Existence and uniqueness

Determining whether a solution of a variational problem exists and if it is unique is crucial before attempting to solve the problem. There are many famous existence and uniqueness results, in the following one such result will be presented.

Definition 1.2.9. Let $a(\cdot,\cdot)$ be a bilinear form on a normed vector space V, then

- $a(\cdot,\cdot)$ is continuous if there exists a constant $M > 0$ such that $a(u,v) \leq M\|u\|_V \|v\|_V, \forall u, v \in V$,
- $a(\cdot,\cdot)$ is coercive if there exists a constant $\alpha > 0$ such that $a(u,u) \geq \alpha \|u\|_V^2, \forall u \in V$.

Theorem 1.2.6 (Riesz Representation theorem, [18] Chapter 2)**.** *Let V be a Hilbert space and ϕ a continuous linear functional defined in V. Then any $\phi(u) = \langle u, v \rangle$ is uniquely determined by a $v \in V$.*

Proof. Let ϕ be a continuous linear functional in the Hilbert space V. Let $\ker(\phi)$ the kernel of ϕ, i.e. $\ker(\phi) := \{u \in V : \phi(u) = 0\}$. If $V = \ker(\phi)$ then $\phi(u) = 0$ for all $u \in V$ and $\phi(u) = \langle u, 0 \rangle$. If $V \neq \ker(\phi)$ we observe that $\ker(\phi) \subset V$ is closed since ϕ is continuous. Therefore, by theorem 4 Chapter 2 [18], V can be decomposed into $\ker(\phi)$ and its orthogonal complement $\ker(\phi)^\perp = \{u \in V : \langle u, v \rangle = 0, \forall v \in \ker(\phi)\}$. Thus we can choose a non-zero element $w \in \ker(\phi)^\perp$ such that $\phi(w) = 1$. Note that for all $u \in V$, $u = u - \phi(u)w + \phi(u)w$, where $u - \phi(u)w \in \ker(\phi)$ and $\phi(u)w \in \ker(\phi)^\perp$. The unique element in V is then $\frac{w}{\|w\|^2} \in \ker(\phi)^\perp$ since

$$\left\langle u, \frac{w}{\|w\|^2} \right\rangle = \underbrace{\left\langle u - \phi(u)w, \frac{w}{\|w\|^2} \right\rangle}_{=0} + \left\langle \phi(u)w, \frac{w}{\|w\|^2} \right\rangle = \phi(u) \frac{\langle w, w \rangle}{\|w\|^2} = \phi(u).$$

□

Theorem 1.2.7 (Lax-Milgram, [32] Chapter 3)**.** *Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear, continuous and coercive form and $b(\cdot) : V \to \mathbb{R}$ be a linear and continuous form. Then the variational formulation has a unique solution $u \in V$, such that*

$$a(u,v) = b(v), \quad \forall v \in V. \tag{1.15}$$

Proof. Let the bilinear form be bounded by $M > 0$. For every $u \in V$ we define a linear map $v \mapsto a(u,v)$ for $v \in V$ which is continuous and therefore by the Riesz Representation theorem there exists a unique element $Au \in V'$ such that $Au(v) = a(u,v), \forall v \in V$.

Recall Definition 1.2.8, the linear mapping $A : V \to V'$ is continuous since

$$\|Au\|_{V'} = \sup_{v \in V} \frac{|Au(v)|}{\|v\|_V} = \sup_{v \in V} \frac{|a(u,v)|}{\|v\|_V} \leq M\|u\|_V.$$

By the Riesz Representation theorem we can define a continuous linear function $h : V' \to V$ such that $b(v) = \langle hb, v\rangle, \forall b \in V'$ and $\forall v \in V$. Now solving (1.15) is equivalent to solving $hAu = hb$. Consider the map $T_\lambda v := v - \lambda(hAv - hb)$ for a chosen parameter $\lambda > 0$.

$$\begin{aligned} \|T_\lambda u - T_\lambda v\|_V^2 &= \|T_\lambda w\|_V^2 = \|w - \lambda(hAw)\|_V^2 \\ &= \|w\|_V^2 - 2\lambda\langle hAw, w\rangle + \lambda^2\|hAw\|_V^2 \end{aligned}$$

Due to A being bounded by M and the coercivity of $a(\cdot,\cdot)$, we obtain the following inequalities

$$\begin{aligned} \|hAw\|_V = \|Aw\|_{V'} \leq \|A\|_V\|w\|_V \leq M\|w\|_V, \\ \langle hAw, w\rangle = Aw(w) = a(w,w) \geq \alpha\|w\|_V^2. \end{aligned} \tag{1.16}$$

Then

$$\|T_\lambda u - T_\lambda v\|_V^2 \leq (1 - 2\lambda\alpha + \lambda^2 M^2)\|u - v\|_V^2, \tag{1.17}$$

meaning that if $\lambda \in \left(0, \frac{2\alpha}{M^2}\right)$, T_λ is a contraction and by Banach's fixed-point theorem T_λ has a unique fixed-point u^* which solves $hAu = hb$. □

Remark 1.2.1 (Symmetric bilinear form)**.** *In the case of a symmetric bilinear form $a(\cdot,\cdot)$ which is coercive and continuous, the bilinear form itself becomes a scalar product on the Hilbert space whose induced norm, $\|\cdot\|_a$ is given by $\|u\|_a = \sqrt{a(u,u)}$, often called the energy norm. Then by the coerciveness and continuity of $a(\cdot,\cdot)$, it holds that*

$$\alpha\|u\|^2 \leq a(u,u) = \|u\|_a^2 \leq M\|u\|^2, \quad \forall u \in V,$$

meaning that the norm induced by the Hilbert space and the norm induced by the bilinear form is equivalent. As an immediate consequence, given $b \in V'$ with respect to either norm, the Riesz Representation theorem can be used directly to prove existence and uniqueness of a solution to the variational problem.

By Lax-Milgram one needs to show that a bilinear and linear form is continuous and that the bilinear form is coercive, in order to prove existence and uniqueness for a variational problem. An example is the variational formulation of the Poisson equation (1.11a) where $a(u,v) = \langle \nabla u, \nabla v\rangle$ and $b(v) = \langle f, v\rangle$. Continuity of $b(v)$ and $a(u,v)$ follows from the Cauchy-Schwarz inequality, what remains is showing that $a(\cdot,\cdot)$ is coercive with respect to $\|\cdot\|_{H^1}$. By the Poincaré inequality one obtains,

$$\begin{aligned} a(u,u) &= \int_\Omega \nabla u \cdot \nabla u dx = \frac{1}{2}\|\nabla u\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\nabla u\|_{L^2(\Omega)}^2 \\ &\geq \frac{1}{2C_\Omega}\|u\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\nabla u\|_{L^2(\Omega)}^2 \geq \min\left\{\frac{1}{2}, \frac{1}{2C_\Omega}\right\}\|u\|_{H^1(\Omega)}^2. \end{aligned}$$

Therefore the bilinear form is coercive and the variational formulation of Poisson's equation has a unique solution.

In the case of Neumann boundary conditions (1.12) the boundary integral can be bounded through trace inequalities [32].

1.2.4 Galerkin finite element method

One of the methods for finding an approximate solution of a variational problem like (1.11b) is the Galerkin method, which defines a similar discrete problem over a finite-dimensional subspace, $V_h \subset V$, i.e. find $u_h \in V_h$ such that

$$a(u_h, v_h) = b(v_h), \quad \forall v_h \in V_h. \tag{1.18}$$

Remark 1.2.2 (Well-posedness)**.** *The existence and uniqueness of a solution to the Galerkin formulation also follows from Lax-Milgram. This is due to that every finite dimensional subspace of a Hilbert space is closed and by only defining the inner product to functions in the subspace, it will be a Hilbert space.*

The conformal finite element method is one such Galerkin method which is defined by how the finite-dimensional space, V_h, is constructed. In essence the construction is based upon three features. The first is a subdivision of the domain, Ω, into a finite number of polyhedra, T, which is called a triangulation, $\mathscr{T}_h$, of Ω. The triangulation must happen such that the following properties are satisfied;

($\mathscr{T}_h 1$) Each $T \in \mathscr{T}_h$ is closed and the interior, $\mathring{T} \neq \emptyset$ and connected.

($\mathscr{T}_h 2$) The boundary of each $T \in \mathscr{T}_h$, ∂T is Lipschitz continuous.

($\mathscr{T}_h 3$) $\bar{\Omega} = \cup_{T \in \mathscr{T}_h} T$.

($\mathscr{T}_h 4$) If $T_1, T_2 \in \mathscr{T}_h$ and $T_1 \neq T_2$, then $\mathring{T}_1 \cap \mathring{T}_2 = \emptyset$.

($\mathscr{T}_h 5$) If $T_1, T_2 \in \mathscr{T}_h$, and as will be the case in this thesis are two dimensional triangles and $T_1 \cap T_2 \neq \emptyset$, then the union is either a edge or point of both T_1 and T_2.

Due to ($\mathscr{T}_h 3$) we will restrict ourselves to cases where $\bar{\Omega}$ is a polygon, in order to avoid curved finite elements. We will use continuous Lagrange finite elements, depicted in Figure 1.3 (a).

After obtaining a triangulation, one defines the finite-dimensional spaces being spanned by $v_h \in V_h$ restricted to a polyhedra T, $v_h|_T$, by $P_T = \{v_h|_T : v_h \in V_h\}$. The second feature is then that $P_T, T \in \mathscr{T}_h$ contain polynomials, or functions close enough to polynomials. We will only consider the polynomial space with order 1 polynomials, denoted $\mathscr{P}_1$, because the low regularity of Richards' equation means that one does not necessarily gain a more accurate solution with a higher order polynomial space. Thus, our finite dimensional space will be

$$V_h = \{v_h \in H_0^1(\Omega) | v_{h|T} \in \mathscr{P}_1, T \in \mathscr{T}_h\}.$$

Lastly, is the existence of a canonical basis of V_h, where the support of the basis functions is as small as possible. An example of a basis function which corresponds to each node is $\varphi_i(x_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta.

Now, considering (1.18) let $\{\varphi_i\}_{i=1}^N$ be a basis for V_h, then the solution u_h can be written as a linear combination of the basis functions. Thus we obtain the following problem

$$\sum_{i=1}^{N} a(\varphi_i, \varphi_j)\alpha_i = b(\varphi_j), \quad 1 \leq j \leq N. \tag{1.19a}$$

The problem is now reduced to calculating $a(\varphi_i, \varphi_j)$ and $b(\varphi_j)$ resulting in a linear system of N equations where α_i is the unknown. Let $\mathbf{A}_{i,j} = a(\varphi_j, \varphi_i)$, $\mathbf{b} = b(\varphi_j)$ and $\boldsymbol{\alpha}_i = \alpha_i$ then (1.19a) is equivalent to

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{b}. \tag{1.19b}$$

1.2.5 Mixed finite element method

In order to solve the mixed variational problem (1.14) the finite element spaces should be subspaces of the continuous spaces, i.e. $V_h \subset L^2(\Omega)$ and $\boldsymbol{Q_h} \subset H(\nabla\cdot, \Omega)$. A common paring is the continuous Lagrange elements with first order Raviart-Thomas elements, $\boldsymbol{RT}_1$. $\boldsymbol{RT}_1$ is illustrated in Figure 1.3 (b).

Definition 1.2.10 (First order Raviart-Thomas element)**.** The Raviart-Thomas element of order 1 on S are defined as

$$\boldsymbol{RT}_1(S) = \mathscr{P}_1(S; \mathbb{R}^d) + \boldsymbol{x}\mathscr{P}_1(S).$$

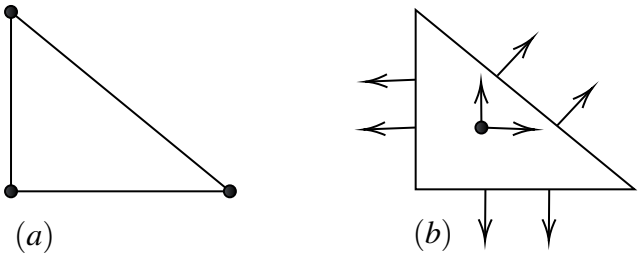


Figure 1.3: Illustration of elements: Continuous Lagrange (a) and Raviart-Thomas (b).

Consider the mixed variational formulation of the Poisson equation (1.14), then by defining $\boldsymbol{Q}_h := \mathrm{RT}_1(\mathscr{T}_h) \cap H(\nabla\cdot, \Omega)$ and $V_h := \{v_h \in H_0^1(\Omega) | v_{h|T} \in \mathscr{P}_1, T \in \mathscr{T}_h\}$, the discrete mixed formulation of the Poisson equation is: Find $(\boldsymbol{\sigma}_h, u_h) \in \boldsymbol{Q}_h \times V_h$ such that

$$\langle \boldsymbol{\sigma}_h, \boldsymbol{q}_h \rangle + \langle u_h, \nabla \cdot \boldsymbol{q}_h \rangle = 0, \qquad \forall \boldsymbol{q}_h \in H(\nabla\cdot, \Omega), \tag{1.20a}$$

$$\langle \nabla \cdot \boldsymbol{\sigma}_h, v_h \rangle = \langle f, v_h \rangle, \qquad \forall v_h \in L^2(\Omega). \tag{1.20b}$$

1.2.6 Convergence of FEM

An essential question is whether the finite element solution u_h is a good approximation of the solution u to the variational formulation.

Lemma 1.2.1 (Céa's lemma, [32] Chapter 2)**.** *Let a bilinear $a(\cdot,\cdot)$ be continuous and coercive, u the solution of Equation* (1.11b) *and u_h be the Galerkin solution of Equation* (1.18)*, then the following error estimate holds*

$$\|u - u_h\| \leq \frac{M}{\alpha} \min_{v_h \in V_h} \{\|u - v_h\|\}. \tag{1.21}$$

Proof. We first note that since both u and u_h solve the variational problem in V_h and $u_h - v_h \in V_h$, we have $a(u-u_h,v) = a(u,v) - a(u_h,v) = b(v) - b(v) = 0$, meaning that the error is orthogonal with respect to the inner product induced by the bilinear form to V_h, referred to as *Galerkin orthogonality*. Since $a(\cdot,\cdot)$ is coercive and continuous and exploiting the Galerkin orthogonality, it follows that

$$\begin{aligned}\alpha\|u-u_h\|^2 \leq & a(u-u_h,u-u_h) = a(u-u_h,v_h-u_h) + a(u-u_h,u-v_h)\\ \leq & a(u-u_h,u-v_h) \leq M\|u-u_h\|\|u-v_h\|.\end{aligned}$$

Thus by dividing by $\|u-u_h\|$ on both sides we obtain

$$\|u-u_h\| \leq \frac{M}{\alpha}\|u-v_h\|,$$

from which (1.21) follows by taking the infimum over $v_h \in V_h$. □

Remark 1.2.3. *Céa's lemma implies that the finite element solution is a quasi-optimal approximation of u in V_h, due to the dependence on ratio of the continuity/coercivity constant. However, it is still the best possible solution in V_h.*

Normally we do not know the solution u which motivates derivation of *a posteriori* estimates. In fact, by using similar ideas as in the proof of the *a priori* estimate in Céa's lemma, the coerciveness of $a(\cdot,\cdot)$ and assuming $u-u_h \in V\backslash\{0\}$ gives

$$\|u-u_h\|_a \leq \frac{a(u-u_h,u-u_h)}{\|u-u_h\|_a} \leq \sup_{v\in V}\frac{a(u-u_h,v)}{\|v\|_a}. \tag{1.22}$$

Note that the last term is the residual of the variational equation, i.e.

$$a(u-u_h,v) = a(u,v) - a(u_h,v) = \langle f,v\rangle - a(u_h,v).$$

Therefore the right hand side of (1.22) can be viewed as a norm of the variational residual. For specific equations one can derive fully computable *a posteriori* upper bounds on $\|u-u_h\|_a$, see e.g. [32] Chapter 4. For Richards' equation the energy norm is non-linear, which means it cannot be directly computed. Therefore we consider an iteration-dependent energy norm (see Section 2.3), which was introduced in [38].

1.3 Iterative methods

In this section an introduction to the theory of iterative schemes for non-linear problems is presented. Different ways of measuring the order of convergence is also introduced.

A non-linear problem can be written in three different ways, let $U \subset \mathbb{R}^d$, $\mathbf{F}: U \to \mathbb{R}^d$ a non-linear function and $\mathbf{b} \in \mathbb{R}^d$,

$$\text{find } \mathbf{x} \in U \text{ such that} \quad \mathbf{F(x)} = \mathbf{b}, \tag{1.23a}$$

$$\text{find } \mathbf{x} \in U \text{ such that} \quad \mathbf{F(x)} = \mathbf{0}, \tag{1.23b}$$

then $\mathbf{x}$ is called a root of (1.23b), or

$$\text{find } \mathbf{x} \in U \text{ such that} \quad \mathbf{F(x)} = \mathbf{x}, \tag{1.23c}$$

which is also known as the fixed-point problem.

In order to solve the non-linear problem one uses an iterative scheme, where one uses a previous approximation to get a better approximation of the solution. For the fixed-point problem (1.23c), one defines the iterative scheme as a sequence

$$\mathbf{F}(\mathbf{x}_{k-1}) = \mathbf{x}_k. \tag{1.24}$$

Typically, the solution is not known *a priori*, naturally one must ask how to determine whether an approximation is close enough to an unknown solution. In order to decide if the scheme converges to a solution of the non-linear equation different fixed-point theorems are typically used. One of the most famous, is the Banach fixed-point theorem. It shows that a solution to the problem exists, and also states that it will be unique.

Definition 1.3.1. Let U and V be two normed spaces and a mapping $\mathbf{F} : U \to V$. If $\mathbf{F}$ satisfies $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}))\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in U, L \in [0,1)$, then $\mathbf{F}$ is called a contraction.

Theorem 1.3.1. *(Banach fixed-point theorem) Let U be a Banach space and let $\boldsymbol{F} : U \to U$ be a contraction on U with contraction constant L. Then there exists only one fixed point $\boldsymbol{x} \in U$ of $\boldsymbol{F}$. Also the fixed-point iteration converges to $\boldsymbol{x}$ for an arbitrary initial value $\boldsymbol{x}_0 \in U$. The following error estimates hold true:*

a) An a posteriori estimate $\|\boldsymbol{x} - \boldsymbol{x}_k\| \leq \dfrac{L}{1-L} \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|$.

b) An a priori estimate $\|\boldsymbol{x} - \boldsymbol{x}_k\| \leq \dfrac{L^k}{1-L} \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|$.

Proof. Let $\{\mathbf{x}_k\}$ be a sequence in a Banach space U and $\mathbf{F} : U \to U$ a contraction on U, then

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_{k-1})\| \leq L\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \leq \cdots L^k \leq \|\mathbf{x}_k - \mathbf{x}_0\|.$$

By applying the triangle inequality we get that for any $k, l \in \mathbb{N}$

$$\begin{aligned}
\|\mathbf{x}_{k+l} - \mathbf{x}_k\| &\leq \|\mathbf{x}_{k+l} - \mathbf{x}_{k+l-1}\| + \|\mathbf{x}_{k+l-1} - \mathbf{x}_{k+l-2}\| + \cdots + \|\mathbf{x}_{k+1} - \mathbf{x}_l\| \\
&\leq (L^{k+l-1} + L^{k+l-2} + \cdots L^k) \|\mathbf{x}_1 - \mathbf{x}_0\| \\
&\leq L^k \sum_{l=0}^{\infty} L^l \|\mathbf{x}_1 - \mathbf{x}_0\| \\
&= \frac{L^k}{1-L} \|\mathbf{x}_1 - \mathbf{x}_0\|,
\end{aligned}$$

which since $L < 1$ means that $\{\mathbf{x}_k\}$ is a Cauchy sequence and therefore it converges. Assuming that $\mathbf{F}$ has two fixed-points, $\mathbf{x}$ and $\mathbf{y}$, we obtain

$$\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{F}(\mathbf{x} - \mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

but since $L < 1$ we must have $\mathbf{x} = \mathbf{y}$ and therefore $\mathbf{x}$ is unique. We also have the inequality

$$\|\mathbf{x} - \mathbf{x}_k\| \leq L\|\mathbf{x} - \mathbf{x}_{k-1}\| \leq L\left(\|\mathbf{x} - \mathbf{x}_k\| + \|\mathbf{x}_k - \mathbf{x}_{k-1}\|\right),$$

from which one can derive the error estimates. □

The Banach fixed-point theorem is a useful tool to simplify checking whether an iterative scheme converges as one only needs to show that it is a contraction. It also gives a way of determining when we are close enough to a solution $\mathbf{x}$ from the *a posteriori* estimate. Since if $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|$ is small, then $\|\mathbf{x} - \mathbf{x}_k\|$ is also small. Therefore one can impose a stopping criterion, $\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \leq \varepsilon$ where ε is a user defined tolerance to decide when the approximation is good enough. We will use a similar stopping criterion (4.1) in a specific norm introduced in Section 2.3.

1.3.1 Order of convergence

A central subject in the theory of iterative schemes is how fast the scheme converges. The theory presented is based upon [15]. The notions we introduce hold for multidimensional problems, for simplicity we only consider scalar problems here.

The speed of a converging sequence is normally measured using C-, Q- and R-orders. The C- and Q-orders can be obtained by considering quotient factors

$$Q_p(k) := \frac{|x - x_{k+1}|}{|x - x_k|^p} = \frac{e_{k+1}}{(e_k)^p}, \quad k \geq 0, \quad p \geq 1, \tag{1.25}$$

where one assumes that $x \neq x_k$.

When $p = 1$ we have a few special cases of C-order,

- no C-order, when there does not exist a limit of $\lim_{k\to\infty} Q_1(k)$,
- C-sublinear, if $\lim_{k\to\infty} Q_1(k) = 1$,
- C-linear, if $0 < \lim_{k\to\infty} Q_1(k) < 1$.

All the above cases are the slowest kinds of C-order.

Remark 1.3.1. *If the sequence has an order which is C-linear, then $e_{k+1} < e_k$. Meaning that the errors are strictly monotone. Also, consider the sequence $x_k = \frac{1}{\sqrt{k}}, k$ odd, $x_k = \frac{1}{\sqrt{k-1}}, k$ even for $k \geq 1$. It is clear that $\lim_{k\to\infty} Q_1(k) = 1$, therefore the sequence is C-sublinear. But note that the errors are not monotone, meaning that C-sublinear does not imply monotone errors.*

Definition 1.3.2. A sequence has C-order $p_0 > 1$ if

$$Q_{p_0} := \lim_{k\to\infty} Q_{p_0}(k) \in (0, \infty). \tag{1.26}$$

A problem with the *C*-orders is that there is no explicit expression for p_0 and therefore finding p_0 may be difficult. The Q-orders allow for an easy computation of p_0, some special cases are

- no Q-order, when $\limsup_{k\to\infty} Q_1(k) = \infty$,
- Q-sublinear, if $1 \leq \limsup_{k\to\infty} Q_1(k) < \infty$,
- at least *Q*-linear if $\limsup_{k\to\infty} Q_1(k) < 1$,
- Q-linear, if $0 < \limsup_{k\to\infty} Q_1(k) < 1$.

Definition 1.3.3. A sequence has Q-order $p_0 > 1$ if

$$\lim_{k\to\infty} Q_p(k) = \begin{cases} 0, & p \in [1, p_0), \\ \infty, & p \in (p_0, \infty), \end{cases} \tag{1.27a}$$

or one of the equivalent conditions,

$$\lim_{k\to\infty} Q_L(k) = \lim_{k\to\infty} \frac{\ln(e_{k+1})}{\ln(e_k)} = p_0, \tag{1.27b}$$

$$\lim_{k\to\infty} Q_\Lambda(k) = \lim_{k\to\infty} \frac{\ln\left(\frac{e_{k+2}}{e_{k+1}}\right)}{\ln\left(\frac{e_{k+1}}{e_k}\right)} = p_0. \tag{1.27c}$$

Remark 1.3.2. *Note that the definition of Q-order (1.27a), implies a jump in the convergence profile at $p_0 > 1$. Therefore, in contrast to the definition of C-order, the limit is not required to exist. In fact, a situation might occur when* $\liminf_{k\to\infty} Q_{p_0}(k) = 0$ *and* $\limsup_{k\to\infty} Q_{p_0}(k) = \infty$ *simultaneously.*

A key issue with the Q-orders occurs when a sequence converges, but the rate varies. For example consider the sequence

$$\{x_k\} = \left\{ \frac{1}{4^{\lceil \frac{k}{2} \rceil}} \right\},$$

where $\lceil \cdot \rceil$ is the ceiling function. Clearly the sequence converges to 0, but it has no Q-order. Determining the order of convergence for such a sequence requires considering an averaged quantity and avoiding relating consecutive terms to each other, which is done using the root factors

$$\begin{aligned} R_1(k) :=& |x - x_k|^{\frac{1}{k}} = e_k^{\frac{1}{k}}, \qquad k \geq 1, \\ R_p(k) :=& |x - x_k|^{\frac{1}{p^k}} = e_k^{\frac{1}{p^k}}, \qquad k \geq 0, \quad p > 1. \end{aligned} \tag{1.28}$$

For $p = 1$ we have the following special cases of R-order,

- R-sublinear/no R-order, when $\limsup_{k\to\infty} R_1(k) = 1$,
- at least R-linear if $\limsup_{k\to\infty} R_1(k) < 1$,
- R-linear, if $0 < \limsup_{k\to\infty} R_1(k) < 1$,
- at least R-superlinear if $\limsup_{k\to\infty} R_1(k) = 0$.

Returning to our example above, we see that

$$\lim_{k\to\infty} \left| \frac{1}{4^{\lceil \frac{k}{2} \rceil}} \right|^{\frac{1}{k}} = \frac{1}{2},$$

therefore the sequence has a R-linear order.

Definition 1.3.4. A sequence has R-order $p_0 > 1$ if

$$\lim_{k\to\infty} R_p(k) = \begin{cases} 0, & p \in [1, p_0), \\ 1, & p \in (p_0, \infty), \end{cases} \tag{1.29a}$$

or if following equivalent condition hold

$$\lim_{k\to\infty} R_L(k) = \lim_{k\to\infty} |\ln(e_k)|^{\frac{1}{k}} = p_0. \tag{1.29b}$$

Remark 1.3.3 (Uniqueness of convergence order)**.** *The quotient factors Q_{p_0} goes towards infinity if the denominator is raised to a bigger power than p_0 and by lowering the power goes towards zero. Hence, the C/Q-order p_0 is unique. The uniqueness of R-order p_0 is given provided p_0 exists [42].*

Normally we do not know the solution of a sequence, therefore we will consider computing the order of convergence based upon consecutive iterations. We replace $|x - x_k|$ by $|x_{k+1} - x_k|$ and denote the computational order of convergence by $'$, i.e. C', Q' and R'. If a sequence has

C/Q/R-order p_0, it will also have the corresponding computational order of convergence.[1] The orders are related through the following result.

Theorem 1.3.2 ([15]). *Let $\{x_k\}$ be a convergent sequence, $p_0 > 1$ then*

$$\{C, C'\} \overset{\Rightarrow}{\not\Leftarrow} \{Q, Q'\} \overset{\Rightarrow}{\not\Leftarrow} \{R, R'\}.$$

Remark 1.3.4. *The definition of C', Q', R' are still difficult to compute, as despite not requiring the solution, the order p_0 is still needed. However, Q'_L, Q'_Λ and R'_L are fully computable and of more interest as they give an easy way of approximating p_0.*

The following result gives a simple way of determining the order of convergence.

Theorem 1.3.3 ([46], Lemma 2). *Let $\{e_k\}$ be a sequence of real positive numbers and $p_0 > 1$ such that*

$$e_k \leq \alpha e_{k-1}^{p_0} + \beta e_{k-1}$$

for all $k \geq 1$. Then if

$$\alpha e_0^{p_0 - 1} + \beta < 1,$$

the sequence converges to zero.

1.4 Linearization methods

To solve non-linear equations numerically it is common to use iterative linearization schemes. For the ease of presentation we only consider scalar problems.

1.4.1 Newton's method

The most popular linearization method is Newton's method. It recursively determines approximations of solutions to the non-linear equation utilizing first order Taylor approximations of F. Given an initial guess $x_0 \in \mathbb{R}$, then for $k \geq 1, i \in \mathbb{N}$ the Newton method is; Given an approximation $x_{k-1} \in \mathbb{R}$, find $x_k \in \mathbb{R}$ such that

$$F(x_{k-1}) + F'(x_{k-1})(x_k - x_{k-1}) = 0. \tag{1.30}$$

Definition 1.4.1. If a sequence $\{x_k\}$ converges to x^* for $x_0 \in \hat{U} \subset \mathbb{R}$, where $\hat{U}$ is open, then the sequence is called locally convergent.

Newton's method is known to be locally C-quadratic convergent, for a result in multiple dimensions we refer to [32] Chapter 8. Under certain conditions Newton's method can achieve even faster convergence, see the result below.

Theorem 1.4.1 ([15], theorem 4.8). *Let $x^* \in \mathbb{R}$ be a simple root of f. The Newton method converges locally with C-order $p_0 \geq 2$, where $p_0 \in \mathbb{N}$ if and only if $f''(x^*) = \cdots = f^{(p_0-1)}(x^*) = 0$ and $f^{(p_0)}(x^*) \neq 0$, leading to*

$$Q_{p_0} = \frac{p_0 - 1}{p_0!} \left| \frac{f^{(p_0)}(x^*)}{f'(x^*)} \right|.$$

[1] It is possible to define an equivalent definition using non-linear residuals, but we omit the details and refer to [15]. Theorem 1.3.2 will also hold for the residual based orders with additional assumptions on the mapping (1.23).

Proof. (i) Let $N_f(x) := x - \frac{f(x)}{f'(x)}$ and assume Newtons method converges locally with C-order $p_0 \geq 2$, then

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^{p_0}} = M \in (0,\infty),$$

and we know that

$$N_f'(x^*) = \frac{f(x^*)f^{(2)}(x^*)}{f'(x^*)} = 0. \tag{1.31}$$

Therefore we assume that $N_f^{(j)}(x^*) = 0, \forall\, j = 1,...,l-1$ when $l < p_0$ and we want to show that $N_f^{(l)}(x^*) = 0$. Consider the Taylor expansion of $N_f(x_k)$ around x^*,

$$N_f(x_k) = N_f(x^*) + \frac{N^{(l)}(a_k)}{l!}\,(x_k - x^*)^l,$$

for all $x \in \mathbb{R}$, $a_k \in (x_k, x^*)$. Then

$$x_{k+1} - x^* = N_f(x_k) - N_f(x^*) = \frac{N^{(l)}(a_k)}{l!}(x_k - x^*)^l \Rightarrow \frac{N^{(l)}(a_k)}{l!} = \frac{(x_{k+1} - x^*)(x_k - x^*)^{p_0 - l}}{(x_k - x^*)^{p_0}},$$

and we get

$$\frac{N^{(l)}(x^*)}{l!} = M \lim_{k\to\infty} (x_k - x^*)^{p_0 - l} = \begin{cases} 0, & \text{if } l < p_0, \\ M, & \text{if } l = p_0. \end{cases}$$

Thus, $N_f^{(j)}(x^*) = 0$ for all $j = 1,...,p_0 - 1$ and $N_f^{(p_0)}(x^*) \neq 0$. Consequently, x^* is a root of multiplicity $p_0 - 1$ of N_f', allowing us to rewrite $N_f'(x) = R(x)(x - x^*)^{p_0 - 1}$, where $R(x)$ is a smooth function so that $R(x^*) \neq 0$. Since x^* is a simple root we can write $f(x) = R_2(x)(x - x^*)$ and from (1.31) we obtain

$$f^{(2)}(x^*) = \frac{R(x)f'(x^*)^2(x - x^*)^{p_0 - 2}}{R_2(x)},$$

which implies that $f^{(2)}$ has a root x^* of multiplicity $p_0 - 2$. Meaning $f^{(2)}(x^*) = \cdots = f^{(p_0 - 1)}(x^*) = 0$ and $f^{(p_0)}(x^*) \neq 0$.

(ii) Assume $f''(x^*) = \cdots = f^{(p_0 - 1)}(x^*) = 0$ and $f^{(p_0)}(x^*) \neq 0$. Consider the Taylor expansion of $f(x_k)$ and $f'(x_k)$ around x^*, then it follows that

$$\begin{aligned} x_{k+1} - x^* =& x_k - x^* - \frac{f(x_k)}{f'(x_k)} = \frac{(x_k - x^*)f'(x_k) - f(x_k)}{f'(x_k)} \\ =& \frac{(x_k - x^*)^{p_0}}{f'(x_k)} \left[\frac{f^{(p_0)(b_k)}}{(p_0 - 1)!} - \frac{f^{(p_0)(a_k)}}{p_0!} \right]. \end{aligned}$$

Taking the limit gives

$$Q_{p_0} = \lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^{p_0}} = \lim_{k\to\infty} \frac{\left| \frac{f^{(p_0)(b_k)}}{(p_0-1)!} - \frac{f^{(p_0)(a_k)}}{p_0!} \right|}{|f'(x_k)|} = \frac{p_0 - 1}{p_0!} \left| \frac{f^{(p_0)}(x^*)}{f'(x^*)} \right|.$$

□

It is also possible to accelerate the convergence of Newton's method by simple modifications [35].

1.4.2 L-scheme linearizations

The main issue with Newton's method is that it is only locally convergent. Therefore one wants to modify the method for it to become more robust, but typically this comes at the cost of losing the higher order accuracy. A way to modify Newton's method is to replace the derivative by an approximation, these methods are classified as quasi-Newton methods.

An example of a quasi-Newton method is the L-scheme where the derivative is replaced by a stabilization parameter L. Given an initial guess $x_0 \in \mathbb{R}$, and for $k \geq 1$ $x_{k-1} \in \mathbb{R}$ is known, find $x_k \in \mathbb{R}$ such that

$$F(x_{k-1}) + L(x_k - x_{k-1}) = 0. \tag{1.32}$$

Remark 1.4.1. *Observe that no evaluation of derivatives is required in* (1.32)*, meaning that the L-scheme linearization can be applied to non-smooth problems. As L is constant, it is important to note that constant linearizations are only suitable for non-decreasing Lipschitz continuous non-linearities.*

The choice of L is significant in terms of convergence properties, although one should at most expect only linear convergence. For a given problem one can find an explicit L, e.g., for the Biot's equations see [56] or for Richards' equation see Section 2.3.1, which results in global convergence.

Another quasi-Newton method, is the modified L-scheme [36]. The idea is to exploit the global convergence of the L-scheme while achieving a better convergence rate. Suppose the non-linear problem is defined by two non-linear functions, i.e., $F(x) + G(x) = 0$, where $F'(x) \geq 0$. Then, the modified L-scheme is

$$F(x_{k-1}) + G(x_{k-1}) + L_{k-1}(x_k - x_{k-1}) = 0, \tag{1.33}$$

where

$$L_{k-1} = \max\left\{F'(x_{k-1}) + \tau m, 2\tau m\right\}.$$

The choice of m is of importance, as if $m \geq \sup\{F'(x)\}$ then the scheme would be equivalent to the L-scheme and $m = 0$ would result in a Newton type method. In practice we will choose a m less than $\sup\{F'(x)\}$ to see the full benefit of the linearization scheme.

1.4.3 Relaxation strategies for linearization methods

To increase the robustness and accelerate an iterative solver, relaxation techniques are used. One tries to improve each iteration by introducing a correction step. Common methods include line search for first-order optimality conditions [40] and Anderson acceleration [3]. We will only consider Anderson acceleration, since no assembly of an objective function is required.

Anderson acceleration was first introduced in 1965 by Donald G. Anderson, as a way to accelerate fixed-point iterations with linear convergence and in some cases even cause a non-convergent FPI to converge. The idea is to exploit previous iterates from a fixed-point iteration, like the L-scheme or Newton's method, and combine them to obtain a new iterate.

Given any fixed-point iteration, $\boldsymbol{G}$, Anderson acceleration for multidimensional problems can be written as:

Algorithm 1 Anderson acceleration

Require: $\mathbf{x}_0 \in \mathbb{R}^d$ and $m \geq 1$
  Compute first iterate $\mathbf{x}_1 = \boldsymbol{G}(\mathbf{x}_0)$
  for $k = 1, 2, \ldots$ **do**
    Set depth $m_k = \min\{k, m\}$
    Define residual matrix $\boldsymbol{F}_k = \left[\boldsymbol{f}_{k-m_k}, \ldots, \boldsymbol{f}_k\right]$, where $\boldsymbol{f}_i = \boldsymbol{G}(\mathbf{x}_i) - \mathbf{x}_i$
    Let $\boldsymbol{\alpha} \in \mathbb{R}^{m_k+1}$ s.t $\sum_i \alpha_i = 1$
    Minimize $\|F_k \boldsymbol{\alpha}\|_2$ w.r.t $\boldsymbol{\alpha}$
    Set new iterate $\mathbf{x}_{k+1} = \sum_{i=0}^{m_k} \alpha_i \boldsymbol{G}(\mathbf{x}_i)$

A key issue when implementing AA is that there are many equivalent ways of writing the least-squares problem (i.e. the minimization problem). We choose the same formulation as in [58], namely an unconstrained minimization problem,

$$\min_{\boldsymbol{\gamma}} \|f_k - F_k \boldsymbol{\gamma}\|_2, \tag{1.34}$$

where $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{m_k-1})$, $\alpha_0 = \gamma_0, \alpha_i = \gamma_i - \gamma_{i-1}$ for $1 \leq i \leq m_k - 1$ and $\alpha_{m_k} = 1 - \gamma_{m_k-1}$. The reason for this choice is that the resulting least-squares problem is relatively small and in fact better conditioned than other formulations.

The acceleration of a convergent fixed-point method when using Anderson acceleration is not theoretically guaranteed in general, it may even theoretically diverge when the least-squares problem does not have a unique solution [58]. In the case of contractive fixed-point iterations [57], AA(1) can be shown to be Q-linear without any assumptions on the coefficients. By asserting that the coefficients remain bounded AA(m) have been shown to be locally R-linear. In [20] a theoretical justification is given for why the convergence rate is improved for a contractive fixed-point iteration when close to a fixed-point. Also they show why quadratically converging methods may be slowed down if one applies Anderson acceleration.

A special case with a non-contractive fixed-point iteration is theoretically proven to converge in [11], which means that applying Anderson acceleration to diverging Newton methods might give convergence. In fact, a Newton-Anderson method is proven to converge superlinearly for non-degenerate problems [43] and numerical results indicate superlinear convergence even for degenerate problems.

Chapter 2

Solution techniques for Richards' equation

In this chapter we will look at different numerical solution techniques for Richards' equation (1.7). We will only consider a conforming finite element discretization in space and use a backward Euler discretization in time. The continuous problem we consider is; Find ψ such that

$$\partial_t \theta(\psi) - \nabla \cdot (K(\theta(\psi))\nabla(\psi + z))) = f, \qquad \text{in } \Omega,\ t \in [0,T]. \tag{2.1}$$

For simplicity we assume zero Dirichlet boundary conditions, although the results can be extended to Dirichlet and Neumann conditions in general. In Chapter 4 we consider test cases with both Dirichlet and Neumann boundary conditions. The continuous Galerkin formulation of (2.1) is; Find $\psi \in H_0^1(\Omega)$ such that

$$\langle \partial_t \theta(\psi), v\rangle + \langle K(\theta(\psi))\nabla(\psi + z), \nabla v\rangle = \langle f, v\rangle, \quad \forall v \in H_0^1(\Omega). \tag{2.2}$$

For the existence and uniqueness of solutions to (2.2) we refer to [59].

2.1 Temporal discretization

In general, a solution to Richards' equation lacks regularity, see e.g., [2]. This causes huge challenges when solving the equation numerically. Therefore, it is common to use backward Euler scheme for the discretization in time, see e.g. [50, 51] to allow for larger time steps, and the benefit of higher order schemes are lost due to the low regularity of solutions.

To discretize in time, the interval $[0,T]$ is divided into intervals of time step length $\tau = \frac{T}{N}$, where N is a strictly positive integer, and time steps $t_n = n\tau$ for $n \in \{1,...,N\}$. Denoting $f(t_n,\cdot)$ by f^n subsequently. Then, by applying the backward Euler method in time, the time-discrete Galerkin formulation of Richards' equation reads; given $\psi^{n-1} \in H_0^1(\Omega)$ find $\psi^n \in H_0^1(\Omega)$ such that

$$\frac{1}{\tau}\langle \theta(\psi^n) - \theta(\psi^{n-1}), v\rangle + \langle K(\theta(\psi^n))\nabla(\psi^n + z), \nabla v\rangle = \langle f^n, v\rangle, \tag{2.3}$$

for all $v \in H_0^1(\Omega)$. We assume a homogeneous Dirichlet boundary condition for simplicity, but the results are applicable to Dirichlet and Neumann boundary conditions in general.

2.2 Spatial discretization

For the spatial discretization of Richards' equation we will use linear Galerkin finite elements. Assuming $\Omega \subset \mathbb{R}^d$ is a polygon and the triangulation $\mathscr{T}_h$ is composed of d-simplices where h is the mesh diameter, the Galerkin finite element space is

$$V_h = \{v_h \in H_0^1(\Omega) | v_{h|T} \in \mathscr{P}_1, T \in \mathscr{T}_h\}. \tag{2.4}$$

Applying linear finite element space above, the fully discrete variational formulation of Richards' equation reads; let $\psi_h^{n-1} \in V_h$ be given, then find $\psi_h^n \in V_h$ such that

$$\langle \theta(\psi_h^n) - \theta(\psi_h^{n-1}), v_h \rangle + \tau \langle K(\theta(\psi_h^n)) \nabla(\psi_h^n + z), \nabla v_h \rangle = \tau \langle f^n, v_h \rangle, \tag{2.5}$$

for all $v_h \in V_h$.

2.3 Linearizations

No matter which spatial discretization is used, a non-linear finite dimensional problem has to be solved at each time step. There are various ways of dealing with the two non-linearities K and θ in (2.5). Here we consider three linearization techniques, the L-scheme, the modified L-scheme and Newton's method. Other alternatives include the modifed Picard method [17], the Jäger-Kačůr scheme [28] and in [12] the convergence behaviour of the Newton scheme is improved, especially for degenerate cases, using a parametrization switching approach.

2.3.1 L-scheme

The L-scheme was introduced in [44, 53], the idea is to take advantage of the monotonicity properties of the saturation θ. It can be viewed as a stabilized Picard method. As an initial guess we use $\psi_h^{n-1} = \psi_h^{n,0}$. The L-scheme is; Let $\psi_h^{n-1}, \psi_h^{n,j-1} \in V_h$ and $L > 0$ be given, then find $\psi_h^{n,j}$ such that

$$\begin{aligned} &L\langle(\psi_h^{n,j} - \psi_h^{n,j-1}), v_h\rangle + \tau\langle K(\theta(\psi_h^{n,j-1}))\nabla(\psi_h^{n,j} + z), \nabla v_h\rangle \\ &= \tau\langle f^n, v_h\rangle - \langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^{n-1}), v_h\rangle, \end{aligned} \tag{2.6}$$

for all $v_h \in V_h$. We introduce an energy norm which depends on the iterations for the L-scheme

$$|||\xi|||_{L,\psi_h^{n,j-1}} := \left(\int_\Omega L\xi^2 + |K(\theta(\psi_h^{n,j-1}))^{\frac{1}{2}} \nabla \xi|^2 \right)^{\frac{1}{2}}, \tag{2.7a}$$

and its dual norm

$$|||\zeta|||_{-L,\psi_h^{n,j-1}} := \sup_{\xi \in H_0^1(\Omega), |||\xi|||_{L,\psi_h^{n,j-1}} = 1} \langle \zeta, \xi \rangle. \tag{2.7b}$$

One of the main advantages of the L-scheme is that no computation of derivates is required which allows for usage even to non-smooth problems. There is an increase of robustness when compared to Newton's [34] and the modified Picard method. Another advantage is better condition numbers of the linear system, alongside fewer function evaluations per iteration [34].

To show the convergence of the L-scheme we make the following assumptions.

Assumption 2.3.1. *The saturation θ is Lipschitz continuous with L_θ, monotonically increasing and the derivative is bounded from below by $L_{min} > 0$.*

Assumption 2.3.2. *The permeability K is Lipschitz continuous with L_k and bounded from below by $K_{min} > 0$.*

Assumption 2.3.3. *The solution of (2.5) satisfies $\|\nabla \psi_h^n\|_{L^\infty} \leq M < \infty$.*

The following result is an extension of theorem 2.2.2 in [55] where the gravity term was neglected.

Theorem 2.3.1. *Let Assumptions 2.3.1 to 2.3.3 be true, and let L satisfy*

$$L \geq \frac{L_\theta K_{min}}{2K_{min}(1-\gamma) - \tau L_\theta L_K^2 (1+M)^2}, \tag{2.8}$$

then the L-scheme (2.6) converges C-linearly with rate

$$\sqrt{\frac{L - 2\gamma L_{min}}{L + \frac{\tau K_{min}}{C_\Omega}}}, \tag{2.9}$$

where γ is a constant satisfying

$$0 \leq \gamma < 1 - \frac{\tau L_K^2 (1+M)^2}{2K_{min}}. \tag{2.10}$$

For a proof see Appendix A. A downside of the L-scheme is that it only converges *C*-linearly, compared to Newton's method which can be *C*-quadratic (Theorem 2.3.4). Also, the choice of L will influence the convergence rate (2.9), naturally one would ask how to chose L to achieve the fastest convergence rate. Optimization of the parameter L was extensively studied in [55], we will extend the result regarding optimal choice of L to include the gravity term. However, we will not perform a numerical study with regards to the choice of L, as numerical results in [55] indicate that it may not always be the most optimal choice.

As the idea of choosing an optimal L in the general case follows the same lines as in [55], we only give a brief discussion here on how the choice should be made. We seek to minimize the convergence rate of the L-scheme which depends upon L and γ. Hence, we choose L as small as possible and insert it into the convergence rate (2.9) squared,

$$rate = \frac{L_\theta K_{min} - 4L_{min}K_{min}\gamma(1-\gamma) + 4\gamma L_{min}\tau L_\theta L_K^2 (1+M)^2}{L_\theta K_{min} + \frac{2\tau K_{min}^2 (1-\gamma)}{C_\Omega} - \frac{\tau L_\theta L_K^2 K_{min}(1+M)^2}{C_\Omega}}. \tag{2.11}$$

In order to simplify (2.11) we introduce the following shorthand notation

- $\alpha := L_\theta K_{min}$,
- $\beta := 4K_{min}L_{min}$,
- $\rho := 2L_{min}(1+M)^2 L_K^2 L_\theta \tau$,
- $\lambda := \frac{2K_{min}^2 \tau}{C_\Omega}$,
- $c := \frac{\tau(1+M)^2 L_K^2 L_\theta K_{min}}{C_\Omega}$.

By differentiating (2.11) and solving for critical values, we obtain one which satisfies (2.10),

$$\gamma_{crit} = \frac{\alpha\beta + \beta\lambda - \beta c + \sqrt{\beta(\alpha^2\beta + \alpha\beta(\lambda - 2\zeta) + \alpha\lambda(\lambda+\rho) + (\lambda - c)(-\beta c + \lambda\rho))}}{\beta\lambda}.$$

Thus, the theoretically optimal choice of γ is,

$$\gamma_{opt} = \begin{cases} \gamma_{crit}, & \text{if } 0 \leq \gamma_{crit} < 1 - \frac{\tau L_K^2 (1+M)^2}{K_{min}}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

This effectively means that the theoretically optimal L is chosen by computing γ_{opt} and inserting into (2.8).

Remark 2.3.1. *In practice determining γ_{opt} is difficult, due to hard to compute constants and may not be the most optimal choice. Therefore in practice we will choose L greater than* $\sup|\theta'(\psi)|/2$ *and lower than* $\sup|\theta'(\psi)|$ *to ensure convergence.*

2.3.2 Modified L-scheme

In [36], a modified L-scheme for Richards' equation is proposed. The idea is to replace the constant L with a function $L^{n,j}$ defined at every iteration. The modified L-scheme is; Let $\psi_h^{n-1}, \psi_h^{n,j-1} \in V_h$ and a function $L^{n,j} : \Omega \to \mathbb{R}^+$ be given, then find $\psi_h^{n,j}$ such that

$$\begin{aligned} &\langle L^{n,j}(\psi_h^{n,j} - \psi_h^{n,j-1}), v_h \rangle + \tau \langle K(\theta(\psi_h^{n,j-1}))\nabla(\psi_h^{n,j} + z), \nabla v_h \rangle \\ &= \tau \langle f^n, v_h \rangle - \langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^{n-1}), v_h \rangle, \end{aligned} \tag{2.13}$$

for all $v_h \in V_h$, where

$$L^{n,j}(\psi_h^{n,j-1}) = \max\left\{ \theta'(\psi_h^{n,j-1}) + \tau m, 2\tau m \right\}, \tag{2.14}$$

and m is chosen with respect to $\mathscr{M}_0$ defined in Theorem 2.3.2.

Remark 2.3.2. *The modified L-scheme can be viewed as a hybrid L-scheme/Picard method as if one disregards $\theta'(\psi_h^{n,j-1})$ in $L^{n,j}$ one would get the L-scheme. Also, if $m = 0$ the modified L-scheme corresponds to the modified Picard method [17].*

Similarly to the L-scheme, we define an iteration-dependent energy norm for the modified L-scheme

$$|||\xi|||_{M,\psi_h^{n,j-1}} := \left(\int_\Omega \left(\theta'(\psi_h^{n,j-1}) + \tau m \right) \xi^2 + \tau |K(\theta(\psi_h^{n,j-1}))^{\frac{1}{2}} \nabla \xi|^2 \right)^{\frac{1}{2}}. \tag{2.15}$$

Under similar assumptions to the L-scheme, but requiring the saturation and permeability to have a bounded second derivative in addition, the following convergence result is obtained.

Theorem 2.3.2 ([36] Theorem 3.1)**.** *For $\mathscr{M}_0 = \max_{\psi \in \mathbb{R}} \{|\theta''|\} \geq 0$, the modified L-scheme (2.13) is C-linearly convergent for all $m \geq \mathscr{M}_0$ and $\tau > 0$. If $m > 0$, i.e. the non-degenerate case, then the convergence rate is $\mathscr{O}(\tau)$ for small enough τ.*

The main advantage of the modified L-scheme is that the convergence rate scales with the time step size, meaning that with small time steps it may even compete with Newton's method. However, generally speaking we are interested in larger time steps. The computational cost is also higher compared to the L-scheme, as the derivative has to be evaluated, but one still does not get the C-quadratic convergence of Newton's method.

2.3.3 Newton's method

The Newton scheme reads; Let $\psi_h^{n-1}, \psi_h^{n,j-1} \in V_h$ be given, then find $\psi_h^{n,j}$ such that

$$\begin{aligned}&\langle \theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}), v_h\rangle + \tau\left\langle (K\circ\theta)'(\psi_h^{n,j-1})\nabla(\psi_h^{n,j-1}+z)(\psi_h^{n,j}-\psi_h^{n,j-1}), \nabla v_h\right\rangle \\ &= \tau\langle f^n, v_h\rangle - \langle \theta(\psi_h^{n,j-1})-\theta(\psi_h^{n-1}), v_h\rangle - \tau\langle K(\theta(\psi_h^{n,j-1}))\nabla(\psi_h^{n,j}+z), \nabla v_h\rangle,\end{aligned} \tag{2.16}$$

for all $v_h \in V_h$. Similarly to the previous linearization schemes considered, we define an iteration-dependent energy norm for Newton's method

$$|||\xi|||_{N,\psi_h^{n,j-1}} := \left(\int_\Omega \theta'(\psi_h^{n,j-1})\,\xi^2 + \tau|K(\theta(\psi_h^{n,j-1}))^{\frac{1}{2}}\nabla\xi|^2\right)^{\frac{1}{2}}. \tag{2.17}$$

A common strategy to analyse Richards' equation is to apply the Kirchhoff transformation [59]

$$\begin{aligned}\mathscr{K} :&\mathbb{R}\to\mathbb{R},\\ &\psi \mapsto \int_0^\psi K(\theta(s))ds,\end{aligned} \tag{2.18}$$

which combines the two non-linearities into one. Since $K(\theta(s))$ is positive, the transformation can be inverted and Richards' equation can be rewritten in terms of $u := \mathscr{K}(\psi)$,

$$\begin{aligned}b(u) &:= \theta(\mathscr{K}^{-1}(u)),\\ k(b(u)) &:= K(\theta(\mathscr{K}^{-1}(u))),\end{aligned}$$

then Richards' equation becomes

$$\partial_t b(u) - \nabla\cdot(\nabla u + k(b(u))\nabla z) = f. \tag{2.19}$$

We make the following assumptions.

Assumption 2.3.4. *$b \in C^1$ is non-decreasing and Lipschitz continuous.*

Assumption 2.3.5. *$k(b(u))$ is continuous and bounded in u and satisfies for all $u_1, u_2 \in \mathbb{R}$,*

$$|k(b(u_2))-k(b(u_1))|^2 \le C_k(b(u_2)-b(u_1))(u_2-u_1).$$

Assumption 2.3.6. *$b(u_0)$ is essentially bounded (by 0 and 1) in Ω and $u_0 \in L^2(\Omega)$.*

Assumption 2.3.7. *$|b'(x)-b'(y)| \le \gamma_1|x-y|$, for all $x, y \in \mathbb{R}$.*

Assumption 2.3.8. *$|(k\circ b)'(x)-(k\circ b)'(y)| \le \gamma_2|x-y|$, for all $x, y \in \mathbb{R}$.*

In order to avoid degeneracy due to b' potentially vanishing we approximate the non-linearity by

$$b_\varepsilon(u) = b(u) + \varepsilon u,$$

where $\varepsilon > 0$ is assumed to be a small regularization parameter. Note that all assumptions for b also applies to b_ε.

Lemma 2.3.1 ([46] Lemma 1)**.** *Let Assumptions 2.3.7 to 2.3.8 be true, then for all $x, y \in \mathbb{R}$ we have*

$$|b_\varepsilon(x)-b_\varepsilon(y)-b_\varepsilon'(y)(x-y)| \le \frac{\gamma_1}{2}|x-y|^2, \tag{2.20a}$$

$$|k(b(x))-k(b(y))-(k\circ b)'(y)(x-y)| \le \frac{\gamma_2}{2}|x-y|^2. \tag{2.20b}$$

For a proof see [47].

The fully discrete variational formulation of Richards' equation is then; Given $u^{n-1} \in V_h$, find $u^n \in V_h$ such that

$$\langle b_\varepsilon(u_h^n) - b_\varepsilon(u_h^{n-1}), v_h\rangle + \tau\langle \nabla u_h^n + k(b(u_h^n))e_z, \nabla v_h\rangle = \tau\langle f^n, v_h\rangle, \quad \forall\, v_h \in V_h. \tag{2.21}$$

The Newton method of (2.21) is; Let $u_h^{n-1}, u_h^{n,j-1} \in V_h$ is given, find $u_h^{n,j} \in V_h$ such that for all $v_h \in V_h$

$$\begin{aligned}&\langle b'_\varepsilon(u_h^{n,j-1})(u_h^{n,j} - u_h^{n,j-1}), v_h\rangle + \tau\langle (k\circ b)'(u^{n,j-1})(u_h^{n,j} - u_h^{n,j-1})\nabla z, \nabla v_h\rangle \\ &= \langle b_\varepsilon(u_h^{n-1}) - b_\varepsilon(u_h^{n,j-1}), v_h\rangle + \tau\langle f^n, v_h\rangle - \tau\langle \nabla u_h^{n,j} + k(b(u_h^{n,j-1}))e_z, \nabla v_h\rangle.\end{aligned} \tag{2.22}$$

By employing similar ideas as in [46] we are able to obtain the following error estimate.

Theorem 2.3.3. *Let Assumptions 2.3.7 to 2.3.8 hold true for small enough τ, then the following estimate holds,*

$$\|e^{n,j}\|^2_{L^2(\Omega)} \le \varepsilon^{-1}\left(\tau\gamma_2^2 + \frac{\gamma_1^2}{2\varepsilon}\right) Ch^{-d}\|e^{n,j-1}\|^4_{L^2(\Omega)}, \tag{2.23}$$

where $C > 0$ does not depend on the discretization parameters.

Proof. By subtracting (2.21) from (2.22) we obtain

$$\begin{aligned}\langle b'_\varepsilon(u_h^{n,j-1})(u_h^{n,j} - u_h^{n,j-1}), v_h\rangle + \tau\langle (k\circ b)'(u^{n,j-1})(u_h^{n,j} - u_h^{n,j-1})\nabla z, \nabla v_h\rangle \\ + \tau\langle \nabla u_h^{n,j} + k(b(u_h^{n,j-1}))\nabla z, \nabla v_h\rangle - \tau\langle \nabla u_h^n + k(b(u_h^n))\nabla z, \nabla v_h\rangle \\ = \langle b_\varepsilon(u_h^n) - b_\varepsilon(u_h^{n,j-1}), v_h\rangle.\end{aligned}$$

Let $v_h = e^{n,j} = u_h^n - u_h^{n,j}$ and note that $e^{n,j} - e^{n,j-1} = u_h^{n,j} - u_h^{n,j-1}$, we then get

$$\begin{aligned}\langle b'_\varepsilon(u_h^{n,j-1})e^{n,j}, e^{n,j}\rangle + \tau\langle (k\circ b)'(u^{n,j-1})e^{n,j}\nabla z, \nabla e^{n,j}\rangle + \tau\langle \nabla u_h^{n,j} - \nabla u_h^n, \nabla e^{n,j}\rangle \\ = \tau\langle (k(b(u_h^n)) - k(b(u_h^{n,j-1})) + k(b)'(u^{n,j-1})e^{n,j-1})\nabla z, \nabla e^{n,j}\rangle \\ + \langle b_\varepsilon(u_h^n) - b_\varepsilon(u_h^{n,j-1}) + b'_\varepsilon(u_h^{n,j-1})e^{n,j-1}, e^{n,j}\rangle.\end{aligned} \tag{2.24}$$

By applying Lemma 2.3.1 on the last term on the right hand side, we obtain the estimate

$$\begin{aligned}&\langle b_\varepsilon(u_h^n) - b_\varepsilon(u_h^{n,j-1}) + b'_\varepsilon(u_h^{n,j-1})e^{n,j-1}, e^{n,j}\rangle \\ &\quad\le \int_\Omega |b_\varepsilon(u_h^n) - b_\varepsilon(u_h^{n,j-1}) + b'_\varepsilon(u_h^{n,j-1})e^{n,j-1}||e^{n,j}|dx \\ &\quad\overset{(2.20a)}{\le} \int_\Omega \frac{\gamma_1}{2}|e^{n,j-1}|^2|e^{n,j}|dx.\end{aligned}$$

Using the inequality $|ab| \le \delta a^2 + \frac{b^2}{4\delta}$ for $a, b \in \mathbb{R}$ and $\delta > 0$ we get

$$\le \int_\Omega \frac{\gamma_1^2}{4\varepsilon}|e^{n,j-1}|^4 + \frac{\varepsilon}{4}|e^{n,j}|^2 dx \le \frac{\gamma_1^2}{4\varepsilon}\|e^{n,j-1}\|^4_{L^4(\Omega)} + \frac{\varepsilon}{4}\|e^{n,j}\|^2_{L^2(\Omega)}. \tag{2.25a}$$

For the first term on the right hand side we apply a similar procedure using Young's inequality

$$\begin{aligned}&\tau\langle (k(b(u_h^n)) - k(b(u_h^{n,j-1})) + (k\circ b)'(u^{n,j-1})e^{n,j-1})\nabla z, \nabla e^{n,j}\rangle \\ &\le \tau\frac{\gamma_2^2}{2}\|e^{n,j-1}\|^4_{L^4(\Omega)} + \frac{\tau}{2}\|\nabla e^{n,j}\|^2_{L^2(\Omega)}.\end{aligned} \tag{2.25b}$$

Furthermore, since $(k\circ b)'(u)$ is bounded, we have

$$\tau\langle(k\circ b)'(u^{n,j-1})(e^{n,j})\nabla z,\nabla e^{n,j}\rangle \leq \tau\int_\Omega |(k\circ b)'(u^{n,j-1})(e^{n,j})||\nabla e^{n,j}| \\ \leq \frac{\tau C_1^2}{2}\|e^{n,j}\|^2_{L^2(\Omega)} + \frac{\tau}{2}\|\nabla e^{n,j}\|^2_{L^2(\Omega)}. \tag{2.25c}$$

Now combining (2.25) on (2.24) we obtain,

$$\langle b'_\varepsilon(u_h^{n,j-1})e^{n,j},e^{n,j}\rangle + \tau\|\nabla e^{n,j}\|^2_{L^2(\Omega)} \leq \tau\frac{\gamma_2^2}{2}\|e^{n,j-1}\|^4_{L^4(\Omega)} + \frac{\tau}{2}\|\nabla e^{n,j}\|^2_{L^2(\Omega)} \\ +\frac{\gamma_1^2}{4\varepsilon}\|e^{n,j-1}\|^4_{L^4(\Omega)} + \frac{\varepsilon}{4}\|e^{n,j}\|^2_{L^2(\Omega)} + \frac{\tau C_1^2}{2}\|e^{n,j}\|^2_{L^2(\Omega)} + \frac{\tau}{2}\|\nabla e^{n,j}\|^2_{L^2(\Omega)}.$$

Also, since $b'_\varepsilon \geq \varepsilon$, and by combing the terms above,

$$\frac{3\varepsilon}{4}\|e^{n,j}\|^2_{L^2(\Omega)} \leq \left(\tau\frac{\gamma_2^2}{2} + \frac{\gamma_1^2}{4\varepsilon}\right)\|e^{n,j-1}\|^4_{L^4(\Omega)} + \tau C_1^2\|e^{n,j}\|^2_{L^2(\Omega)}. \tag{2.26}$$

Using the following inverse estimate for discrete polynomial spaces ([13] p.111)

$$\|e^{n,j-1}\|_{L^4(\Omega)} \leq Ch^{-\frac{d}{4}}\|e^{n,j-1}\|_{L^2(\Omega)}, \tag{2.27}$$

and for $\tau C_1^2 \leq \frac{\varepsilon}{4}$ the estimate becomes

$$\|e^{n,j}\|^2_{L^2(\Omega)} \leq \varepsilon^{-1}\left(\tau\gamma_2^2 + \frac{\gamma_1^2}{2\varepsilon}\right)Ch^{-d}\|e^{n,j-1}\|^4_{L^2(\Omega)}. \tag{2.28}$$

□

Theorem 2.3.4. *Let Assumptions 2.3.4 to 2.3.8 be true, then the Newton method (2.22) is convergent with C-order 2, if* $\tau = \mathcal{O}(\varepsilon^3 h^d)$.

Proof. From [45] Propsition 3.5 we get the stability estimate

$$\|e^0\|^2_{L^2(\Omega)} \leq \frac{C\tau}{\varepsilon}.$$

When $\tau = \mathcal{O}(\varepsilon^3 h^d)$,

$$\varepsilon^{-1}\left(\tau\gamma_2^2 + \frac{\gamma_1^2}{2\varepsilon}\right)Ch^{-d}\frac{\tau}{\varepsilon} < 1, \tag{2.29}$$

then by applying the estimate in Theorem 2.3.3 combined with Theorem 1.3.3, Newton's method converges with C-order 2. □

Remark 2.3.3 (Continuity of the non-linearities)**.** *In the analysis above, the L-scheme and Newton's method was shown to be convergent for Lipschitz continuous non-linearities. Some soil types result in only Hölder continuity which gives unbounded derivatives, thus Newton's method cannot be applied directly. A common approach is to regularize the problem by approximating the non-linearities with ones that are Lipschitz continuous. This would also resolve the choice of L depending on the Lipschitz constant of the saturation. However, convergence for the L-scheme can be obtained without regularization by choosing the stabilization parameter L in a manner which causes the iteration error to become lower than a given threshold, see [10, 49]. The convergence rate will depend upon the Hölder exponent.*

2.4 Error estimates

In this section we give an error estimate in a fixed norm using the iteration-dependent energy norm for the L-scheme. We define the residual $\mathscr{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ of the non-linear problem, and the residual $\mathscr{R}_{\text{lin}}^j : H_0^1(\Omega) \to H^{-1}(\Omega)$ of the j^{th} iterate as

$$\langle \mathscr{R}(u), v\rangle := \langle \theta(u) - \theta(\psi_h^{n-1}), v\rangle + \tau(K(\theta(u))\nabla(u+z), \nabla v) - \tau\langle f^n, v\rangle, \tag{2.30a}$$

$$\begin{aligned}\langle \mathscr{R}_{\text{lin}}^j(u), v\rangle := &\langle L(u - \psi_h^{n,j-1}), v\rangle + \tau(K(\theta(\psi_h^{n,j-1}))\nabla(u+z), \nabla v) \\ &+ \langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^{n-1}) - \tau f^n, v\rangle.\end{aligned} \tag{2.30b}$$

In a similar manner to estimates obtained in [38] we seek a bound on the fixed norm $|||\cdot|||_{-L,\psi_h}$ for a finite element solution to the non-linear problem (2.30a) since the iteration-dependent norm is by definition not fixed. The following estimate follows the same ideas, but with the notable difference that only R-linear convergence is assumed. We make the following assumptions.

Assumption 2.4.1. *Let $\psi, \zeta \in H_0^1(\Omega) \cap L^\infty(\Omega)$ and $v \in H_0^1(\Omega)$ and let K be Lipschitz continuous and bounded from below by $K_{min} > 0$ so that*

$$\langle Lv, v\rangle + \langle K(\theta(\psi))\nabla v, \nabla v\rangle \leq \max\left\{\sup_\Omega \frac{K(\theta(\psi))}{K(\theta(\zeta))}, 1\right\}\left(\langle Lv, v\rangle + \langle K(\theta(\zeta))\nabla v, \nabla v\rangle\right).$$

Assumption 2.4.2. *The L-scheme converges R-linearly in $L^\infty(\Omega)$, i.e.*

$$\left\|\psi_h^{n,j+1} - \psi_h^{n,j}\right\|_{L^\infty(\Omega)}^{\frac{1}{j}} \leq \alpha, \quad \alpha \in (0,1).$$

We assume convergence in $L^\infty(\Omega)$ as the L-scheme (and also the modified L-scheme) for continuous solutions have be shown to converge linearly in $L^\infty(\Omega)$ [36]. We do know that the L-scheme is C-linear, and hence also *R*-linear, as such we make a weaker assumption on the convergence of the L-scheme.

Theorem 2.4.1. *Let Assumptions 2.4.1 to 2.4.2 hold. Let $\psi \in H_0^1(\Omega)$ solve $\langle \mathscr{R}(\psi), v\rangle = 0$ for all $v \in H_0^1(\Omega)$ and $\psi_h \in V_h \subset H_0^\infty(\Omega)$ solve $\langle \mathscr{R}(\psi_h), v_h\rangle = 0$ for all $v_h \in V_h$ where $\mathscr{R}(\cdot)$ is defined in* (2.30a) *then*

$$\frac{|||\mathscr{R}(\psi)|||_{-L,\psi_h^j}}{\left(1+\mathscr{Z}_j\right)^{\frac{1}{2}}} \leq |||\mathscr{R}(\psi)|||_{-L,\psi_h} \leq \left(1+Q_j\right)^{\frac{1}{2}} |||\mathscr{R}(\psi)|||_{-L,\psi_h^j}, \tag{2.31}$$

where

$$Q_j = \left(\frac{1}{1-\alpha}\right)\sup_\Omega\left\{K(\theta(\psi_h^j))\sup_{v\in I^j}\left|\frac{1}{(K\circ\theta)}'(v)\right|\right\},$$

$$\mathscr{Z}_j = \left(\frac{1}{1-\alpha}\right)\sup_\Omega\left\{K(\theta(\psi_h^j))\sup_{v\in I^j}\left|(K\circ\theta)'(v)\right|\right\},$$

and

$$I^j = \left[\psi_h^j - \left(\frac{1}{1-\alpha}\right), \psi_h^j + \left(\frac{1}{1-\alpha}\right)\right].$$

Proof. Assumption 2.4.2 implies that $\psi_h^j \to \psi_h$ when $j \to \infty$, giving

$$\left\| \psi_h - \psi_h^j \right\|_{L^\infty(\Omega)} \leq \sum_{k=j}^{\infty} \left\| \psi_h^{k+1} - \psi_h^k \right\|_{L^\infty(\Omega)} \leq \sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}.$$

Giving

$$\psi_h^j - \left(\frac{1}{1-\alpha} \right) \leq \psi_h \leq \psi_h^j + \left(\frac{1}{1-\alpha} \right), \quad \text{a.e. in } \Omega, \tag{2.32}$$

or equivalently $\psi_h \in I^j$. If $\hat{\varphi} = \arg\max_{\varphi \in H_0^1(\Omega)} \left(|\langle \mathscr{R}(\psi), \varphi \rangle| / |||\varphi|||_{L,\psi_h} \right)$, then

$$\begin{aligned} |||\mathscr{R}(\psi)|||_{-L,\psi_h} &= \frac{|\langle \mathscr{R}(\psi), \hat{\varphi} \rangle|}{|||\hat{\varphi}|||_{L,\psi_h}} = \frac{|\langle \mathscr{R}(\psi), \hat{\varphi} \rangle|}{|||\hat{\varphi}|||_{L,\psi_h^j}} \left(\frac{|||\hat{\varphi}|||_{L,\psi_h^j}}{|||\hat{\varphi}|||_{L,\psi_h}} \right) \\ &\leq |||\mathscr{R}(\psi)|||_{-L,\psi_h^j} \sqrt{\frac{\langle L\hat{\varphi}, \hat{\varphi} \rangle + \langle K(\theta(\psi_h^j))\nabla\hat{\varphi}, \nabla\hat{\varphi} \rangle}{\langle L\hat{\varphi}, \hat{\varphi} \rangle + \langle K(\theta(\psi_h))\nabla\hat{\varphi}, \nabla\hat{\varphi} \rangle}}. \end{aligned}$$

From Assumption 2.4.1 we get

$$|||\mathscr{R}(\psi)|||_{-L,\psi_h} \leq |||\mathscr{R}(\psi)|||_{-L,\psi_h^j} \sqrt{\sup_\Omega \frac{K(\theta(\psi_h^j))}{K(\theta(\psi_h))}}.$$

We will only consider the case when $\sup_\Omega \frac{K(\theta(\psi_h^j))}{K(\theta(\psi_h))} > 1$. The other case immediately results in $|||\mathscr{R}(\psi)|||_{-L,\psi_h} \leq |||\mathscr{R}(\psi)|||_{-L,\psi_h^j}$.

$$\sup_\Omega \frac{K(\theta(\psi_h^j))}{K(\theta(\psi_h))} \leq 1 + \sup_\Omega \left\{ \frac{K(\theta(\psi_h^j)) - K(\theta(\psi_h))}{K(\theta(\psi_h))} \right\} \leq 1 + \sup_\Omega \left\{ K(\theta(\psi_h^j)) \left(\frac{1}{K(\theta(\psi_h))} - \frac{1}{K(\theta(\psi_h^j))} \right) \right\}$$

Let $\psi_h \in I^j$, then

$$\begin{aligned} \sup_\Omega \frac{K(\theta(\psi_h^j))}{K(\theta(\psi_h))} &\leq 1 + \sup_\Omega \left\{ K(\theta(\psi_h^j)) \sup_{v \in I^j} \left| \frac{1}{(K \circ \theta)}'(v) \right| \right\} \left\| \psi_h - \psi_h^j \right\|_{L^\infty(\Omega)} \\ &\leq 1 + \left(\frac{1}{1-\alpha} \right) \sup_\Omega \left\{ K(\theta(\psi_h^j)) \sup_{v \in I^j} \left| \frac{1}{(K \circ \theta)}'(v) \right| \right\} \end{aligned}$$

resulting in the estimate

$$|||\mathscr{R}(\psi)|||_{-L,\psi_h} \leq \left(1 + \left(\frac{1}{1-\alpha} \right) \sup_\Omega \left\{ K(\theta(\psi_h^j)) \sup_{v \in I^j} \left| \frac{1}{(K \circ \theta)}'(\psi) \right| \right\} \right)^{\frac{1}{2}} |||\mathscr{R}(\psi)|||_{-L,\psi_h^j}. \tag{2.33}$$

Similarly, we can derive a lower bound through the inequality,

$$|||\mathscr{R}(\psi)|||_{-L,\psi_h^j} \leq |||\mathscr{R}(\psi)|||_{-L,\psi_h} \sqrt{\sup_\Omega \frac{K(\theta(\psi_h))}{K(\theta(\psi_h^j))}},$$

and if $\psi_h \in I^j$

$$\sup_\Omega \frac{K(\theta(\psi_h^j))}{K(\theta(\psi_h))} \leq 1 + \sup_\Omega \left\{ K(\theta(\psi_h^j)) \sup_{v \in I^j} |(K \circ \theta)'(v)| \right\} \left\| \psi_h - \psi_h^j \right\|_{L^\infty}(\Omega).$$

□

Chapter 3

L-scheme/Newton switching

In this chapter, we develop an adaptive switching algorithm, which utilizes the robustness of the L-scheme and C-quadratic convergence of Newton's method. The switching criteria are based upon *a posteriori* error estimates. The idea behind the adaptive algorithm is to start with the L-scheme and derive an estimator $\eta^j_{L\to N}$ that predicts from the j^{th} and $(j-1)^{\text{th}}$ whether Newton's method will converge using the j^{th} iterate as an initial guess. Another estimator $\eta^j_{N\to L}$ is then derived which predicts the success or failure of Newton's method. If Newton's method is predicted to fail, the algorithm returns to the L-scheme. In addition, we derive an estimator $\eta^j_{L\to L}$ to determine if the L-scheme itself will converge and to tune the value of L accordingly, see Appendix B. A flowchart of the algorithm is illustrated in Figure 3.1. The algorithm is the first combined L-scheme/Newton strategy for Richards' equation based upon robust and reliable switching criteriums and is submitted for publication [54].

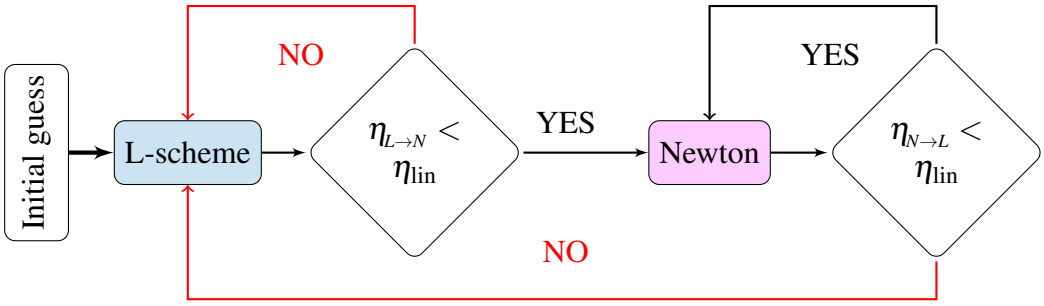


Figure 3.1: Flowchart of Adaptive switching algortihm between L-scheme and Newton's method.

For the L-scheme, it was shown in [38] by Galerkin orthogonality that

$$\underbrace{\left|\!\left|\!\left|\mathscr{R}(\psi_h^{n,j})\right|\!\right|\!\right|^2_{-L,\psi_h^{n,j-1}}}_{\text{total error}} = \underbrace{\left|\!\left|\!\left|\psi_h^{n,j}-\psi_h^{n,i-1}\right|\!\right|\!\right|^2_{L,\psi_h^{n,j-1}}}_{\text{linearization error}} + \underbrace{\left|\!\left|\!\left|\mathscr{R}^j_{\text{lin}}(\psi_h^{n,j})\right|\!\right|\!\right|^2_{-L,\psi_h^{n,j-1}}}_{\substack{\text{discretization error}\\ \text{of the linearization step}}} \tag{3.1}$$

The result holds for the modified L-scheme and the modified Picard method also. This orthogonality relation shows that the embodiment of the linearization and the discretization errors are the terms

$$\eta^j_{\text{lin}} = \left|\!\left|\!\left|\psi_h^{n,j}-\psi_h^{n,j-1}\right|\!\right|\!\right|_{L,\psi_h^{n,j-1}}, \tag{3.2a}$$

$$\eta^j_{\text{disc}} = \left|\!\left|\!\left|\mathscr{R}^j_{\text{lin}}(\psi_h^{n,j})\right|\!\right|\!\right|_{-L,\psi_h^{n,j-1}}. \tag{3.2b}$$

Here we are only interested in the linearization component. In the following $\langle\cdot,\cdot,\rangle$ and $\|\cdot\|$ is the inner product and norm of $L^2(\Omega)$.

3.1 L-scheme to Newton switching

Let us assume that in the $(j+1)^{\text{th}}$-iteration we want to test for switching to the Newton scheme. Let $\hat{\psi}_h^{n,j+1} \in V_h$ be the solution of the Newton scheme (2.16) using $\psi_h^{n,j}$ as the previous iterate, where $\psi_h^{n,j}$ is obtained from the L-scheme. In this section, we will assume the following:

Assumption 3.1.1 (Convection term is not dominant)**.** *For $j \in \mathbb{N}$, we assume that there exists a constant $C_N^j \in [0,2)$ such that*

$$\tau|K(\theta(\psi_h^{n,j}))^{-\frac{1}{2}}(K(\theta))'(\psi_h^{n,j})\nabla(\psi_h^{n,j}+z)|^2 \leq (C_N^j)^2\theta'(\psi_h^{n,j}), \tag{3.3}$$

a.e. in Ω.

Remark 3.1.1. *The above assumption is also required to show the linear problem's coercivity and, thus, the existence of solutions. Observe that since $\psi_h^{n,j}$ is known, the constant C_N^j is fully computable. It additionally satisfies being less than 2 if τ is small and the the numerical flux is bounded. Observe that the assumption holds in the degenerate case when $\theta'(\psi_h^{n,j}) = 0$, since the left-hand side and right hand side both contain the derivative of either the saturation or the permeability, which vanish.*

To cover the degenerate case, we also introduce the concept of an equilibrated flux.

Definition 3.1.1. Let $\Pi_h : L^2(\Omega) \to \mathscr{P}_1(\mathscr{T}_h)$ be the $\mathscr{P}_1$ projection operator, i.e. let $u \in L^2(\Omega)$, then $(\Pi_h u, v_h) = (u, v_h)$ for all $v_h \in \mathscr{P}_1(\mathscr{T}_h)$.

Definition 3.1.2 (Equilibrated flux $\boldsymbol{\sigma}_L^j$ for degenerate regions)**.** For a pre-determined $\varepsilon > 0$, let $\mathscr{T}_{\text{deg}}^{i,\varepsilon} := \{K \in \mathscr{T}_h : \inf \theta'(\psi_h^{n,i}) < \varepsilon \text{ in } K\}$. We define $\boldsymbol{\sigma}_L^j \in \boldsymbol{RT}_1(\mathscr{T}_h) \cap H(\nabla\cdot,\Omega)$ as

$$\nabla\cdot\boldsymbol{\sigma}_L^j = \begin{cases} \frac{1}{\tau}\Pi_h(L(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))) & \text{in } \mathscr{T}_{\text{deg}}^{j,\varepsilon}, \\ 0 & \text{otherwise}. \end{cases} \tag{3.4}$$

For the computation of the equilibrated flux, see Section 3.3.1. Then we have the following result:

Proposition 3.1.1 (L-scheme to Newton switching indicator)**.** *For a given $\psi_h^{n,0}$, $\psi_h^{n-1} \in V_h$, let $\{\psi_h^{n,k}\}_{k=1}^j \subset V_h$ solve* (2.6) *for some $j \in \mathbb{N}$. Let $\tilde{\psi}_h^{n,j+1} \in V_h$ be the solution of* (2.16) *with the previous iterate $\psi_h^{n,j}$. Recall Definition 3.1.2. Let Assumption 3.1.1 be true, then the following estimate holds*

$$\left|\!\left|\!\left| \tilde{\psi}_h^{n,j+1} - \psi_h^{n,j} \right|\!\right|\!\right|_{N,\psi_h^{n,j}} \leq \eta_{L\to N}^j,$$

where

$$\eta_{L\to N}^j := 2/(2-C_N^j)\left([\eta_{L\to N}^{j,\text{source}}]^2 + \tau[\eta_{L\to N,2}^{j,\text{flux}}]^2\right)^{\frac{1}{2}},$$

with

$$\eta_{L\to N}^{j,\text{source}} := \left\| \theta'(\psi_h^{n,j})^{-\frac{1}{2}}(L(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))) \right\|_{\mathscr{T}_h\setminus\mathscr{T}_{\text{deg}}^{j,\varepsilon}},$$

$$\eta_{L\to N}^{j,\text{flux}} := \left\| K(\theta(\psi_h^{n,j}))^{-\frac{1}{2}}\left[(K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1})))\nabla(\psi_h^{n,j}+z) + \boldsymbol{\sigma}_h^j\right] \right\|.$$

Proof. Observe from (2.16) that $\delta\psi_h^{j+1} := \tilde{\psi}_h^{n,j+1} - \psi_h^{n,j} \in V_h$ satisfies

$$\begin{aligned}
&\langle \theta'(\psi_h^{n,j})\delta\psi_h^{j+1}, v_h\rangle + \tau\langle K(\theta(\psi_h^{n,j}))\nabla\delta\psi_h^{j+1}, \nabla v_h\rangle \\
&\quad + \tau\left\langle (K\circ\theta)'(\psi_h^{n,j})\nabla(\psi_h^{n,j}+z)\,\delta\psi_h^{j+1}, \nabla v_h\right\rangle \\
&\quad = \tau\langle f^n, v_h\rangle - \langle \theta(\psi_h^{n,j}) - \theta(\psi_h^{n-1}), v_h\rangle - \tau\langle K(\theta(\psi_h^{n,j}))\nabla\psi_h^{n,j}, \nabla v_h\rangle. \qquad (3.5)
\end{aligned}$$

We choose the following test function $v_h = \delta\psi_h^{j+1}$ in (3.5), then

$$\begin{aligned}
\left|\!\left|\!\left|\delta\psi_h^{j+1}\right|\!\right|\!\right|^2_{N,\psi_h^{n,j}} &\overset{(2.17)}{=} \int_\Omega \left(\theta'(\psi_h^{n,j})|\delta\psi_h^{j+1}|^2 + \tau|K(\theta(\psi_h^{n,j}))^{\frac12}\nabla\delta\psi_h^{j+1}|^2\right) \\
&\overset{(3.5)}{=} \underbrace{-\tau\left\langle (K\circ\theta)'(\psi_h^{n,j})\nabla(\psi_h^{n,j}+z)\,\delta\psi_h^{j+1}, \nabla\delta\psi_h^{j+1}\right\rangle}_{=:T_1} \\
&\quad + \underbrace{\tau\langle f^n, \delta\psi_h^{j+1}\rangle - \langle \theta(\psi_h^{n,j}) - \theta(\psi_h^{n-1}), \delta\psi_h^{j+1}\rangle - \tau\langle K(\theta(\psi_h^{n,j}))\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle}_{=:T_2}.
\end{aligned} \qquad (3.6a)$$

To simplify notation we define $\boldsymbol{\sigma}^j = (K\circ\theta)'(\psi_h^{n,j})\nabla(\psi_h^{n,j}+z)$, and we obtain

$$\begin{aligned}
T_1 &:= -\tau\langle \boldsymbol{\sigma}^j\delta\psi_h^{j+1}, \nabla\delta\psi_h^{j+1}\rangle \\
&\le \left(\tau\int_\Omega |K(\theta(\psi_h^{n,j}))^{-\frac12}\boldsymbol{\sigma}^j|^2(\delta\psi_h^{j+1})^2\right)^{\frac12}\left(\tau\int_\Omega |K(\theta(\psi_h^{n,j}))^{\frac12}\nabla\delta\psi_h^{j+1}|^2\right)^{\frac12} \\
&\overset{(3.3)}{\le} C_N^j\left(\int_\Omega \theta'(\psi_h^{n,j})(\delta\psi_h^{j+1})^2\right)^{\frac12}\left(\tau\int_\Omega |K(\theta(\psi_h^{n,j}))^{\frac12}\nabla\delta\psi_h^{j+1}|^2\right)^{\frac12} \\
&\le \frac{C_N^j}{2}\int_\Omega \left(\theta'(\psi_h^{n,j})|\delta\psi_h^{j+1}|^2 + \tau|K(\theta(\psi_h^{n,j}))^{\frac12}\nabla\delta\psi_h^{j+1}|^2\right) \\
&= \frac{C_N^j}{2}\left|\!\left|\!\left|\delta\psi_h^{j+1}\right|\!\right|\!\right|^2_{N,\psi_h^{n,j}}. \qquad (3.6b)
\end{aligned}$$

Using the divergence theorem we have the relation

$$\begin{aligned}
-\langle \boldsymbol{\sigma}_L^j, \nabla\delta\psi_h^{j+1}\rangle &= \langle \nabla\cdot\boldsymbol{\sigma}_L^j, \delta\psi_h^{j+1}\rangle \\
&\overset{(3.4)}{=} \langle \Pi_h(L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&= \langle L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}}.
\end{aligned}$$

For the last term, using (2.6) and $\delta\psi_h^{j+1} \in V_h$ one has

$$\begin{aligned}
T_2 &:= \tau\langle f^n, \delta\psi_h^{j+1}\rangle - \langle\theta(\psi_h^{n,j}) - \theta(\psi_h^{n-1}), \delta\psi_h^{j+1}\rangle - \tau\langle K(\theta(\psi_h^{n,j}))\nabla\psi_h^j, \nabla\delta\psi_h^{i+1}\rangle \\
&\overset{(2.6)}{=} \langle L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle \\
&\quad - \tau\langle (K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1})))\nabla(\psi_h^{n,j} + z), \nabla\delta\psi_h^{j+1}\rangle \\
&= \langle L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})) - \nabla\cdot\boldsymbol{\sigma}_L^j, \delta\psi_h^{j+1}\rangle \\
&\quad - \tau\langle (K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1})))\nabla(\psi_h^{n,j} + z) + \boldsymbol{\sigma}_L^j, \nabla\delta\psi_h^{j+1}\rangle \\
&= \langle L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_h\setminus\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&\quad - \tau\langle (K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1})))\nabla(\psi_h^{n,j} + z) + \boldsymbol{\sigma}_L^j, \nabla\delta\psi_h^{j+1}\rangle \\
&\overset{(3.4)}{\leq} \langle L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_h\setminus\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&\quad + \tau\eta_{L\to N}^{j,\mathrm{flux}} \|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\| \\
&\leq \eta_{L\to N}^{j,\mathrm{source}} \|\theta'(\psi_h^{n,j})^{\frac{1}{2}}\delta\psi_h^{j+1}\| + \tau\,\eta_{L\to N}^{j,\mathrm{flux}} \|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\|. \qquad (3.6c)
\end{aligned}$$

Combining (3.6) we obtain

$$\frac{2-C_N^j}{2}\left|\!\left|\!\left|\delta\psi_h^{j+1}\right|\!\right|\!\right|_{N,\psi_h^{n,j}}^2 \leq [\eta_{L\to N}^{j,\mathrm{source}}]\ \|\theta'(\psi_h^{n,j})^{\frac{1}{2}}\delta\psi_h^{j+1}\| + \sqrt{\tau}[\eta_{L\to N}^{j,\mathrm{flux}}]\,\sqrt{\tau}\|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\|.$$

By applying the Cauchy-Schwarz inequality we have

$$\begin{aligned}
&[\eta_{L\to N}^{j,\mathrm{source}}]\ \|\theta'(\psi_h^{n,j})^{\frac{1}{2}}\delta\psi_h^{j+1}\| + \sqrt{\tau}[\eta_{L\to N}^{j,\mathrm{flux}}]\,\sqrt{\tau}\|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\| \\
&\leq \left([\eta_{L\to N}^{j,\mathrm{source}}]^2 + \tau[\eta_{L\to N}^{j,\mathrm{flux}}]^2\right)^{\frac{1}{2}} \left(\|\theta'(\psi_h^{n,j})^{\frac{1}{2}}\delta\psi_h^{j+1}\|^2 + \tau\|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\|^2\right)^{\frac{1}{2}},
\end{aligned}$$

which results in

$$\frac{2-C_N^j}{2}\left|\!\left|\!\left|\delta\psi_h^{j+1}\right|\!\right|\!\right|_{N,\psi_h^{n,j}}^2 \leq \left([\eta_{L\to N}^{j,\mathrm{source}}]^2 + \tau[\eta_{L\to N,2}^{j,\mathrm{flux}}]^2\right)^{\frac{1}{2}} \left|\!\left|\!\left|\delta\psi_h^{j+1}\right|\!\right|\!\right|_{N,\psi_h^{n,j}}. \qquad (3.7)$$

□

3.2 Newton to L-scheme switching

Assuming that the L-scheme converges unconditionally, we would only wish to switch back to the L-scheme after using the Newton scheme if its linearization error grew with more iterations. Similarly to the procedure described earlier, we can estimate if this is going to happen in the $(j+1)^{\mathrm{th}}$-step, purely from the iterates up to the j^{th}-step. We introduce another equilibrated flux for this purpose.

Definition 3.2.1 (Equilibrated flux $\boldsymbol{\sigma}_N^j$ for degenerate regions (Newton scheme))**.** We define $\boldsymbol{\sigma}_N^j \in \boldsymbol{RT}_1(\mathscr{T}_h) \cap H(\nabla\cdot,\Omega)$ as

$$\nabla\cdot\boldsymbol{\sigma}_N^j = \begin{cases} \frac{1}{\tau}\Pi_h(\theta'(\psi_h^{n,j-1})(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))) & \text{in } \mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}, \\ 0 & \text{otherwise}. \end{cases} \qquad (3.8)$$

We obtain a result similar to Proposition 3.1.1.

Proposition 3.2.1 (Error control of L-scheme/Newton switching step)**.** *For a given* $\psi_h^{n,0}$, $\psi_h^{n-1} \in V_h$, *let* $\{\psi_h^{n,k}\}_{k=1}^{j+1} \subset V_h$ *solve* (2.16) *for some* $j \in \mathbb{N}$. *Let Assumption 3.1.1 be true, then*

$$\left|\!\left|\!\left| \psi_h^{n,j+1} - \psi_h^{n,j} \right|\!\right|\!\right|_{N,\psi_h^{n,j}} \leq \eta_{N\to L}^j,$$

where

$$\eta_{N\to L}^j := 2/(2-C_N^j)\left([\eta_{N\to L}^{j,\mathrm{source}}]^2 + \tau[\eta_{N\to L}^{j,\mathrm{flux}}]^2\right)^{\frac{1}{2}},$$

with

$$\eta_{N\to L}^{j,\mathrm{source}} := \|\theta'(\psi_h^{n,j})^{-\frac{1}{2}}(\theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})))\|_{\mathscr{T}_h\setminus\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}},$$

$$\eta_{N\to L}^{j,\mathrm{flux}} := \left\| \begin{array}{c} K(\theta(\psi_h^{n,j}))^{-\frac{1}{2}}\boldsymbol{\sigma}_N^j + [K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1}))\nabla(\psi_h^{n,j}+z) \\ -(K\circ\theta)'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1})\nabla(\psi_h^{n,j-1}+z)]K(\theta(\psi_h^{n,j}))^{-\frac{1}{2}} \end{array} \right\|.$$

Proof. Similarly to the proof of Proposition 3.1.1, by inserting the test function $v_h = \delta\psi_h^{j+1} = \psi_h^{n,j+1} - \psi_h^{n,j}$, one has

$$\begin{aligned}
\left|\!\left|\!\left| \delta\psi_h^{j+1} \right|\!\right|\!\right|^2_{N,\psi_h^{n,j}} &\overset{(2.17)}{=} \int_\Omega \left(\theta'(\psi_h^{n,j})|\delta\psi_h^{j+1}|^2 + \tau|K(\theta(\psi_h^{n,j}))^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}|^2\right) \\
&\overset{(3.5)}{=} \underbrace{-\tau\left\langle (K\circ\theta)'(\psi_h^{n,j})\nabla(\psi_h^{n,j}+z)\,\delta\psi_h^{j+1}, \nabla\delta\psi_h^{j+1}\right\rangle}_{=:T_1} \\
&\quad + \underbrace{\tau\langle f^n, \delta\psi_h^{j+1}\rangle - \langle\theta(\psi_h^{n,j}) - \theta(\psi_h^{n-1}), \delta\psi_h^{j+1}\rangle - \tau\langle K(\theta(\psi_h^{n,j}))\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle}_{=:T_2}.
\end{aligned} \tag{3.9a}$$

T_1 is estimated the same way as in (3.6b),

$$T_1 \leq \frac{C_N^j}{2}\left|\!\left|\!\left| \delta\psi_h^{j+1} \right|\!\right|\!\right|_{N,\psi_h^{n,j}}. \tag{3.9b}$$

For the last term, using (2.16), and $\delta\psi_h^{j+1} \in V_h$ one has

$$\begin{aligned}
T_2 &:= \tau\langle f^n, \delta\psi_h^{j+1}\rangle - \langle\theta(\psi_h^{n,j}) - \theta(\psi_h^{n-1}), \delta\psi_h^{j+1}\rangle - \tau\langle K(\theta(\psi_h^{n,j}))\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle \\
&= \tau\langle f^n, \delta\psi_h^{j+1}\rangle - \langle\theta(\psi_h^{n,j-1}) - \theta(\psi_h^{n-1}), \delta\psi_h^{j+1}\rangle - \tau\langle K(\theta(\psi_h^{n,j-1}))\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle \\
&\quad - \langle\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}), \delta\psi_h^{j+1}\rangle - \tau\langle[K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1}))]\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle \\
&\overset{(2.16)}{=} \langle\theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle \\
&\quad + \tau\langle(K\circ\theta)'(\psi_h^{n,j-1})\nabla(\psi_h^{n,j-1}+z)(\psi_h^{n,j}-\psi_h^{n,j-1}), \nabla\delta\psi_h^{j+1}\rangle \\
&\quad - \tau\langle[K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1}))]\nabla(\psi_h^{n,j}+z), \nabla\delta\psi_h^{j+1}\rangle
\end{aligned}$$

Using the divergence theorem and $\delta\psi_h^{j+1} \in V_h$ for the reduction

$$\begin{aligned}
-\langle\boldsymbol{\sigma}_N^j, \nabla\delta\psi_h^{j+1}\rangle &= \langle\nabla\cdot\boldsymbol{\sigma}_N^j, \delta\psi_h^{j+1}\rangle \\
&\overset{(3.8)}{=} \langle\Pi_h(\theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&= \langle(\theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1}))), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}},
\end{aligned}$$

we get

$$\begin{aligned}
T_2 &= \langle \theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j})-\theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_h\setminus\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&\quad + \tau\langle (K\circ\theta)'(\psi_h^{n,j-1})\nabla(\psi_h^{n,j-1}+z)(\psi_h^{n,j}-\psi_h^{n,j-1}), \nabla\delta\psi_h^{j+1}\rangle \\
&\quad - \tau\langle [K(\theta(\psi_h^{n,j}))-K(\theta(\psi_h^{n,j-1}))]\nabla(\psi_h^{n,j}+z) - \boldsymbol{\sigma}_N^j, \nabla\delta\psi_h^{j+1}\rangle \\
&\leq \langle \theta'(\psi_h^{n,j-1})(\psi_h^{n,j}-\psi_h^{n,j-1}) - (\theta(\psi_h^{n,j})-\theta(\psi_h^{n,j-1})), \delta\psi_h^{j+1}\rangle_{\mathscr{T}_h\setminus\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}} \\
&\quad + \tau\eta_{N\to L}^{j,\mathrm{flux}}\|K(\theta(\psi_h^{n,j}))^{\frac{1}{2}}\delta\psi_h^{n,j+1}\| \\
&\leq [\eta_{N\to L}^{j,\mathrm{source}}]\cdot\|\theta'(\psi_h^{n,j})^{\frac{1}{2}}\delta\psi_h^{j+1}\| + \sqrt{\tau}\,[\eta_{N\to L}^{j,\mathrm{flux}}]\cdot\sqrt{\tau}\|K(\psi_h^{n,j})^{\frac{1}{2}}\nabla\delta\psi_h^{j+1}\| \qquad (3.9c)
\end{aligned}$$

By combing (3.9) and applying the Cauchy-Schwarz inequality with the definition of $\eta_{N\to L}$ one obtains the estimate. □

Remark 3.2.1 (Effectivity of the estimators $\eta_{L\to N}^j$ and $\eta_{N\to L}^j$)**.** *The estimators $\eta_{L\to N}^j$ and $\eta_{N\to L}^j$ predict the linearization error $\eta_{\mathrm{lin}}^{j+1}$ of the $(j+1)^{\mathrm{th}}$ iteration if done using the Newton scheme* (2.16)*. When the iterations are performed using Newton's method, the sharpness of the estimates can be measured using the* ***effectivity index****, i.e., if $(j+1)^{\mathrm{th}}$ iteration is Newton then*

$$\textit{(Eff. Ind.)}_j := \begin{cases} \eta_{L\to N}^j/\eta_{\mathrm{lin}}^{j+1} & \textit{if } j^{\mathrm{th}} \textit{ iteration is L-scheme}, \\ \eta_{N\to L}^j/\eta_{\mathrm{lin}}^{j+1} & \textit{if } j^{\mathrm{th}} \textit{ iteration is Newton}. \end{cases} \qquad (3.10)$$

It is always greater than one because of Propositions 3.1.1 and 3.2.1, and an effectivity index close to one indicates a precise estimate. With the exception of (3.6b)*, where the term T_1 is bounded above using the global approximation in Assumption 3.1.1, the estimators are expected to be quite accurate because the Cauchy-Schwarz inequality is primarily used to derive them. This expected sharpness is demonstrated by the numerical experiments in Chapter 4, see in particular Figures 4.6 and 4.11.*

3.3 A-posteriori estimate based adaptive linearization algorithm

After some considerations, we propose the following switching algorithm based on the above estimates.

3.3.1 Computation of equilibrated flux

Recalling Definitions 3.1.2 and 3.2.1, we propose a simple algorithm to compute an equilibrated flux $\boldsymbol{\sigma}_h \in \mathrm{RT}_1(\mathscr{T}_h)\cap H(\nabla\cdot,\Omega)$ satisfying $\nabla\cdot\boldsymbol{\sigma}_h = \Pi_h f$ in $\mathscr{T}_{\mathrm{deg}}^{j,\varepsilon}$, and $\nabla\cdot\boldsymbol{\sigma}_h = 0$ otherwise, where $f\in L^2(\Omega)$. Defining $\boldsymbol{Q}_h := \mathrm{RT}_1(\mathscr{T}_h)\cap H(\nabla\cdot,\Omega)$ and $V_h := \{v_h\in H_0^1(\Omega) | v_{h|T}\in \mathscr{P}_1, T\in\mathscr{T}_h\}$, we seek a pair $(\boldsymbol{\sigma}_h, r_h)\in \boldsymbol{Q}_h\times V_h$ that satisfies the mixed finite element problem,

$$\langle K(1)^{-1}\boldsymbol{\sigma}_h, \boldsymbol{q}_h\rangle - \langle r_h, \nabla\cdot\boldsymbol{q}_h\rangle = 0, \qquad \forall \boldsymbol{q}_h\in\boldsymbol{Q}_h, \qquad (3.11a)$$

$$\langle \nabla\cdot\boldsymbol{\sigma}_h, v_h\rangle = \langle f, v_h\rangle, \qquad \forall v_h\in V_h. \qquad (3.11b)$$

The advantage of this flux is that it minimizes $\|K(1)^{-\frac{1}{2}}\boldsymbol{\sigma}_h\|$ which appears in the estimates in Propositions 3.1.1 and 3.2.1.

3.3.2 Additional computational considerations

To speed up the computations of the switching criteria, we make a few more reductions

- **[Equilibrated flux]** If the saturated domain is much smaller than the unsaturated domain, then we take $\boldsymbol{\sigma}_L^j = \boldsymbol{\sigma}_N^j = 0$.
- **[Switching condition]** The condition $\eta_{L\to N}^j \leq \eta_{\text{lin}}^j$ may only happen after many iterations. As a result, we will use the criteria $\eta_{L\to N}^j < C_{\text{tol}}\eta_{\text{lin}}^j$ for a constant $C_{\text{tol}} > 1$ to speed up the switch between L-scheme and Newton's method.

3.3.3 Adaptive linearization algorithm

We propose the following adaptive algorithm:

Algorithm 2 L-scheme/Newton a-posteriori switching

Require: $\boldsymbol{\psi}^{n,0} \in L^2(\Omega)$ as initial guess.
Ensure: Scheme=L-scheme, $C_{\text{tol}} = 1.5$
  for i=1,2,.. **do**
    if Scheme=L-scheme **then**
      Compute iterate using L-scheme, i.e., (2.6)
      if $C_N^i \geq 2$ **then continue**.
      else if $\eta_{L\to N}^i \leq C_{\text{tol}}\eta_{\text{lin}}^i$ **then**
        Set Scheme=Newton
    else
      Compute iterate using Newton, i.e., (2.16)
      if $\eta_{N\to L}^i > \eta_{\text{lin}}^i$ **then**
        Set Scheme=L-scheme

Remark 3.3.1 (Computational cost of the estimators)**.** *In the non-degenerate case, the switching indicators $\eta_{L\to N}^j$ and $\eta_{N\to L}^j$ can be directly computed from the iterates $\psi_h^{n,j}$ and $\psi_h^{n,j-1}$ by setting $\boldsymbol{\sigma}_L^j = \boldsymbol{\sigma}_N^j = \mathbf{0}$, see Propositions 3.1.1 and 3.2.1. As a result, compared to the cost of assembly and solution of a linear system, the cost of computing the estimators is minimal. The L/N scheme generally performs similarly or better than the Newton scheme time-wise for the cases considered in this thesis, since the L-scheme iterations are less expensive than the Newton iterations. This is evident from the numerical experiments, e.g. see Figure 4.4. In the degenerate case, global computation are required for computing $\boldsymbol{\sigma}_L^j$ and $\boldsymbol{\sigma}_N^j$ if they are used. The computation of these equilibrated fluxes can be made relatively inexpensive by precomputing the associated stiffness matrices which are constant. By only evaluating the estimators on a subset of iterations, the computational cost can be further reduced. For the sake of simplicity, we choose not to pursue this choice.*

Remark 3.3.2 (L-scheme adaptivity)**.** *To help accelerate the convergence of the L-scheme, we additionally suggest an algorithm in Appendix B for adaptively choosing L. This can be used directly in conjunction with Algorithm 2 to hasten the convergence of the composite scheme. For the sake of presentational simplicity, we have chosen not to combine these schemes.*

Remark 3.3.3 (Generality of the results)**.** *Although the analysis above focuses on the switching between the L-scheme and the Newton method, the same techniques can be extended to cover*

switching between the modified L-scheme or modified Picard method and Newton. Also, using Anderson acceleration on only the iterates from linearly converging methods is a possibility, but we choose not to pursue this. Furthermore, the L-adaptive strategy in Appendix B can be extended to the modified L-scheme to adaptively select the parameter $m > 0$.

Chapter 4

Numerical results

In this chapter, we demonstrate the effectiveness and robustness of the proposed hybrid L/N-scheme and compare it to the linearization schemes mentioned above, including Anderson acceleration of the L-scheme and Newton's method. Since the L-scheme's convergence depends heavily on a tuning parameter, we choose two different values, L_1 and L_2 in the performance comparison. Here, L_1 is a quasi-optimal choice of tuning parameter and will be defined for each specific subproblem, see Table 4.1, and $L_2 = \sup\{\theta'(\psi)\}$. The quasi-optimal choice L_1 is always chosen for the L-scheme iterations for the L/N-scheme. For the modified L-scheme we always choose the stabilization parameter as $m = \sup\{|\theta''|\}$.

We examine the number of iterations, order of convergence and computational time of the schemes. Computational time refers to the time required for assembly of linear systems, linear solvers, computation of the switching indicators and the iteration-dependent energy norm used as a stopping criterion. All experiments have been performed on an Acer Swift 3, with an Intel core i7-1165G7-processor. A direct solver is used to solve all linear systems. The corresponding iteration-dependent energy-norm for the pressure head is used as a stopping criterium,

$$\left|\!\left|\!\left| \psi_h^{n,j} - \psi_h^{n,j-1} \right|\!\right|\!\right|_{\mathfrak{L},\psi^{j-1}} \leq 10^{-7}, \tag{4.1}$$

with $\mathfrak{L} \in \{L, N, M\}$.

For the numerical experiments, three different examples are considered:

- Example 1: The first example is a strictly unsaturated medium, where the flow is always partially saturated. It is taken from [27], but we disregard the surfactant transport.

- Example 2: The second example considers a variably saturated medium with extraction/injection in the unsaturated zone and can be found in [34].

- Example 3: The last example is a known benchmark problem that is studied in [26, 31, 34]. It models the recharge of a groundwater reservoir from a drainage trench using a time-dependent Dirichlet boundary condition.

In all examples the parametrization of saturation and permeability will be the van Genucthen-

Mualem model

$$\theta(\psi) = \begin{cases} \theta_R + (\theta_S - \theta_R)\left[\frac{1}{1+(-\alpha\psi)^n}\right]^{\frac{n-1}{n}}, & \psi \leq 0, \\ \theta_S, & \psi > 0, \end{cases}$$
$$K(\theta_e(\psi)) = \begin{cases} K_s\theta_e(\psi)^{\frac{1}{2}}\left[1-\left(1-\theta_e(\psi)^{\frac{n}{n-1}}\right)^{\frac{n-1}{n}}\right]^2, & \psi \leq 0, \\ K_s, & \psi > 0, \end{cases} \tag{4.2}$$

where

$$\theta_e(\psi) = \frac{\theta(\psi) - \theta_R}{\theta_S - \theta_R},$$

and θ_S and θ_R is the water volume and the residual water content respectively. Also, K_s is the hydraulic conductivity of the fully saturated porous medium and α and n are soil characteristics.

Parameters	Example 1	Example 2	Example 3
van Genuchten-Mualem			
θ_R	0.026	0.026	0.131
θ_S	0.42	0.42	0.396
K_S	0.12	0.12	$4.96 \cdot 10^{-2}$
α	0.551	0.95	0.423
n	2.9	2.9	2.06
L-scheme			
L_1	0.1	0.15	$3.501 \cdot 10^{-3}$
$L_2(L_\theta)$	0.136	0.2341	$4.501 \cdot 10^{-3}$
Modified L-scheme			
m	0.14125	0.419	0.0447

Table 4.1: Parameter values for all test cases.

Remark 4.0.1. *The computational times of the modified L-scheme are not included, as the implementation has not been done with regards to speed. However, we note that the modified L-scheme requires evaluation of the derivative thus causing the assembly to be slower than the L-scheme, but should be computationally faster than Newton's method per iteration. The computational time of Newton Anderson acceleration will also be omitted.*

The finite element implementation is done in Python and uses the simulation toolbox PorePy [29] for grid management. It is available at `https://github.com/MrShuffle/RichardsEquation/releases/tag/v1.0.1`.

4.1 Example 1: Strictly unsaturated medium

The parameters for this example are given in Table 4.1 Example 1. We consider a strictly unsaturated medium, where the domain is given by $\Omega = \Omega_1 \cup \Omega_2$, where $\Omega_1 = [0,1] \times [0,1/4]$

and $\Omega_2 = [0,1] \times (1/4,1]$. The initial pressure head profile is

$$\psi^0(x,z) = \begin{cases} -z - 1/4, & (x,z) \in \Omega_1, \\ -4, & (x,z) \in \Omega_2, \end{cases}$$

where x represents the positional variable in the horizontal direction and z in the vertical direction. At the top boundary a constant Dirichlet condition is used, equal to the initial value at all times. For the rest of the boundary, no-flow boundary conditions are used. We choose the following source term

$$f(x,z) = \begin{cases} 0, & (x,z) \in \Omega_1, \\ 0.06 \cos\left(\frac{4}{3}\pi(z)\right) \sin(x), & (x,z) \in \Omega_2. \end{cases}$$

In this example the solutions will be computed over different time intervals, $[0,T]$, where $T = \tau$. The solution at $T = 1$ is portrayed in Figure 4.1.

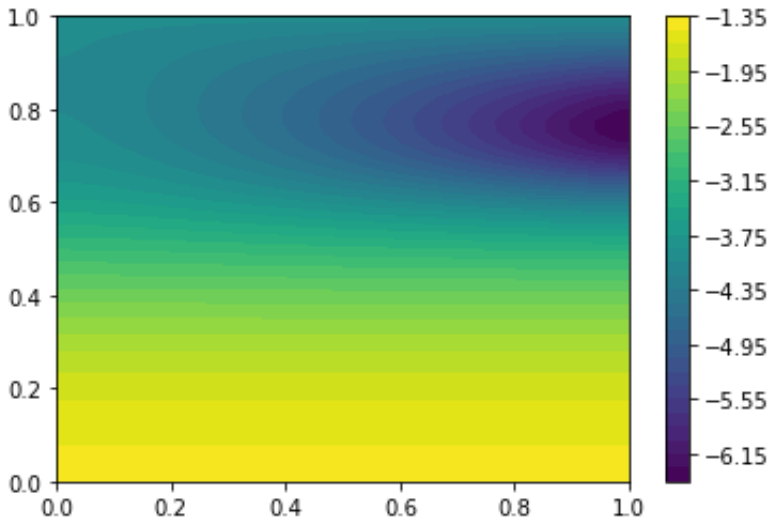


Figure 4.1: Strictly unsaturated medium (Example 1): Pressure head profile at $T = 1$.

4.1.1 Comparison of convergence properties

The performance results for Example 1 are shown in terms of the number of iterations necessary for various mesh sizes in Figure 4.2. As expected the modified L-scheme and L-scheme are robust and converge for all mesh sizes, with the modified L-scheme using the fewest iterations of the two. Anderson acceleration of the L-scheme (using L_1) converges with fewer iterations, which shows the effective acceleration of a linear contractive fixed point iteration. Newton's method only converges for sufficiently coarse meshes. However, when it converges it uses fewer iterations than the linearly convergent and accelerated schemes. Although the Newton Anderson acceleration uses more iterations than Newton's method, it interestingly slightly increases the robustness. Finally, the L/N-scheme always uses the fewest number of iterations, being identical with Newton's method.

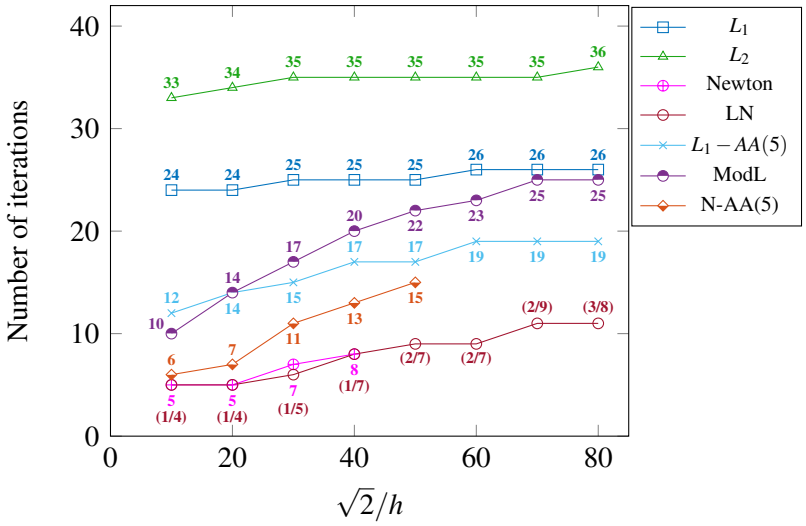


Figure 4.2: Strictly unsaturated medium (Example 1): Number of iterations required for fixed time step $\tau = 0.01$ and multiple mesh sizes. The numbers in the red parentheses correspond to (number of L-scheme iterations/number of Newton iterations).

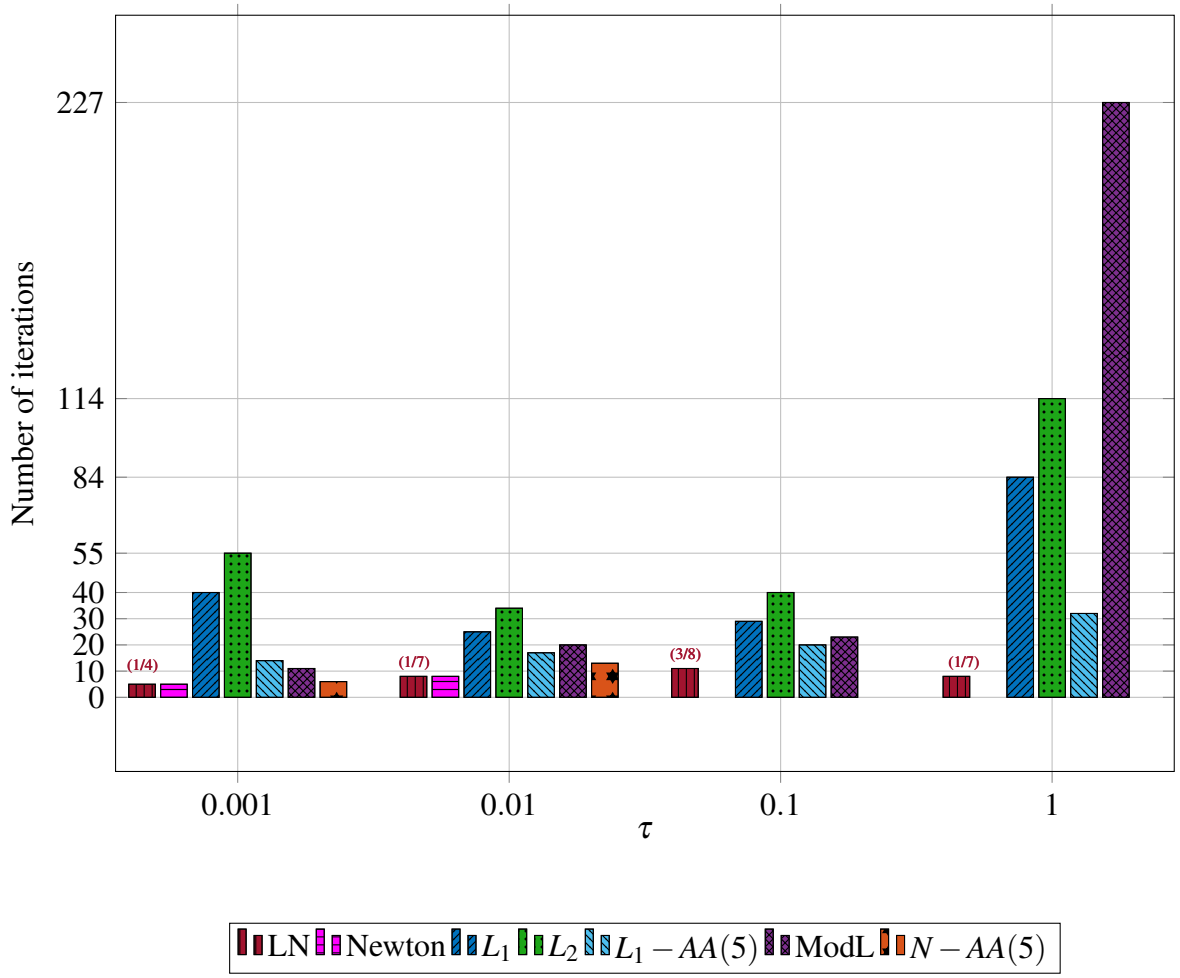


Figure 4.3: Strictly unsaturated medium (Example 1): Total number of iterations for all numerical schemes for different time step lengths and fixed $h = \sqrt{2}/40$. Missing plot means no convergence.

For different time step sizes the total number of iterations required is given in Figure 4.3. The L-schemes and modified L-scheme converge in every scenario. For smaller time steps the modified L-scheme is faster than the L-schemes as the convergence rate scales with the time step size. Anderson acceleration of the L-scheme uses almost as many iterations as the modified L-scheme due to a smaller time step size meaning the distance to a fixed-point is very small initially. For $\tau = 1$ the worst performing L-scheme uses almost half as many iterations as the modified L-scheme. Note that no optimization of m is done here. There is a huge benefit of applying Anderson acceleration to the L-scheme, as the total number of iterations is reduced by more than half. Newton's method and the accelerated Newton converge only for time step sizes smaller than or equal to $\tau = 0.01$. The hybrid method uses the fewest number of iterations for all time step sizes, equal to Newton when converging. For larger time steps the hybrid method performs similarly to what would be expected of Newton's method.

The performance of all schemes, except the modified L-scheme and the Anderson accelerated Newton's method, with regards to computational time for varying mesh sizes is displayed in Figure 4.4. Most significantly, the L/N-scheme performs almost equal to Newton's method when it converges. In addition to this, the hybrid method maintains the same performance in cases where Newton's method fails to converge. The computational time of the L-schemes is consistent with the number of iterations reported with the stabilizing parameter L_1 being the fastest. Applying Anderson acceleration to the L-scheme (using L_1) leads to a 30% reduction in computational time for the finest mesh considered, but still being approximately 168% more computationally expensive than the L/N-scheme.

For fixed mesh size $h = \sqrt{2}/40$ and variable time step size similar observations are made for the CPU time, see Figure 4.5. The hybrid method performs best in all scenarios, being considerably faster than the L-schemes. There is a major reduction in time when applying Anderson acceleration to the fastest L-scheme. In particular, the CPU time is reduced by more than half for $\tau = 1$ and $\tau = 0.001$.

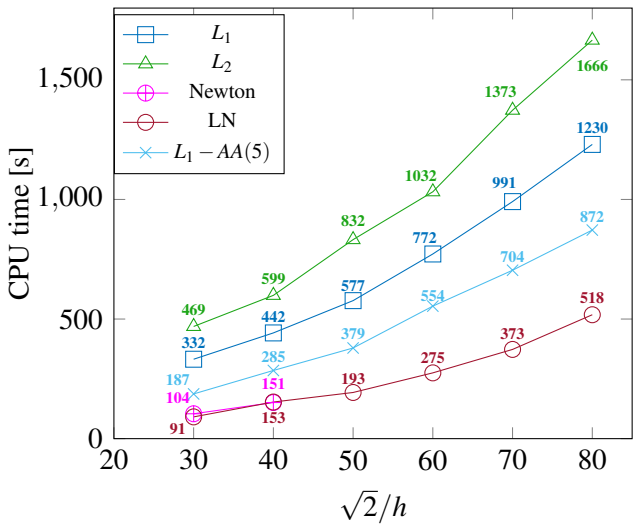


Figure 4.4: Strictly unsaturated medium (Example 1): Computational times for varying mesh sizes and fixed time step size $\tau = 0.01$.

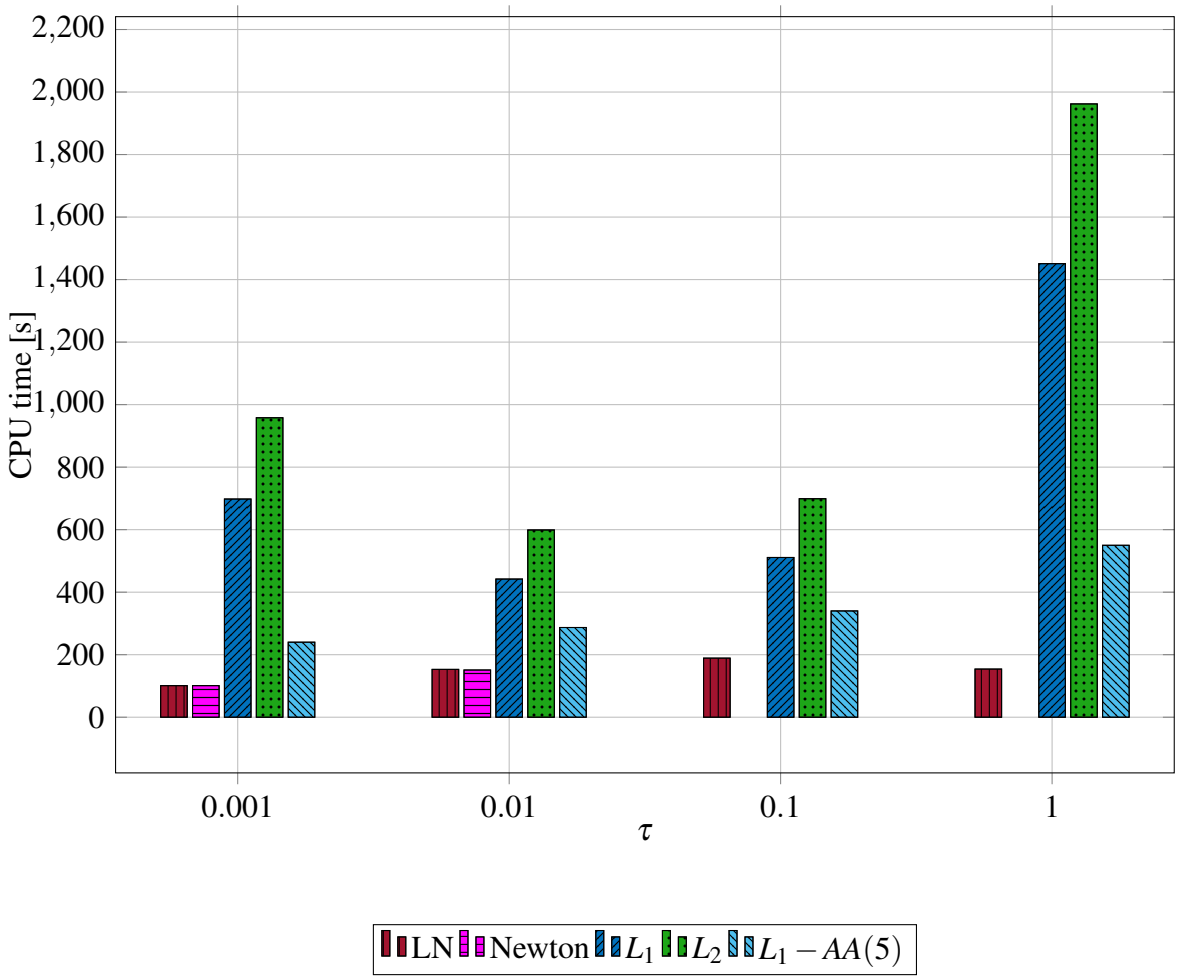
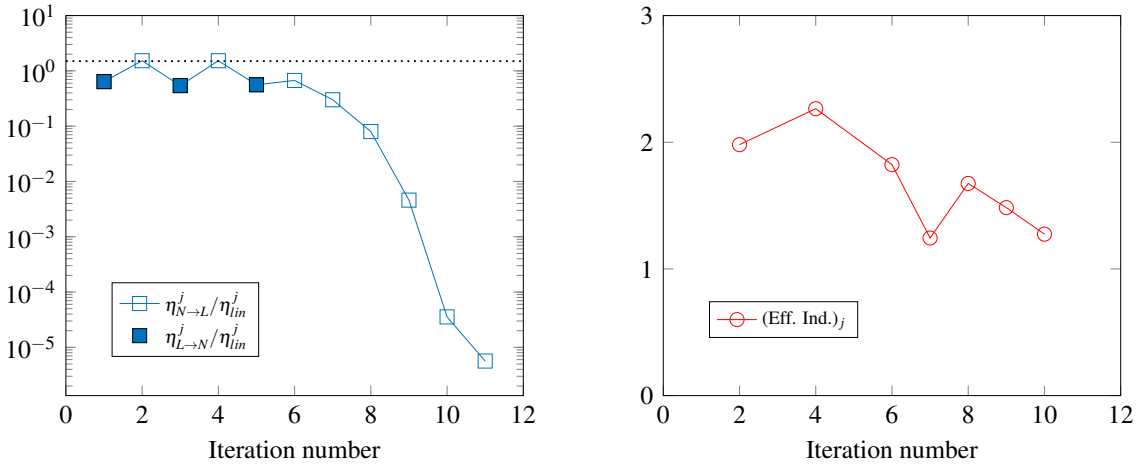


Figure 4.5: Strictly unsaturated medium (Example 1): Computational time for different time step lengths and fixed $h = \sqrt{2}/40$.

4.1.2 Switching characteristics

Finally, a closer look is given to the dynamic switch between Newton's method and the L-scheme. In Figure 4.6, the swithing indicators at each iteration are displayed for fixed mesh and time step size. The example particularly highlights the ability of the L/N-scheme to switch back and forth between both linearizations. Furthermore, the final number of L-scheme iterations is kept to a bare minimum. The effectivity indexes introduced in (3.10) and discussed in Remark 3.2.1 are also plotted. The effectivity index is greater than 1 in all cases, which validates Propositions 3.1.1 and 3.2.1 and it stays between 1.27 to 2.3, implying that the estimators $\eta^i_{L\to N}$ and $\eta^i_{N\to L}$ are sharp.



(a) Switching indicators.

(b) Efficency index.

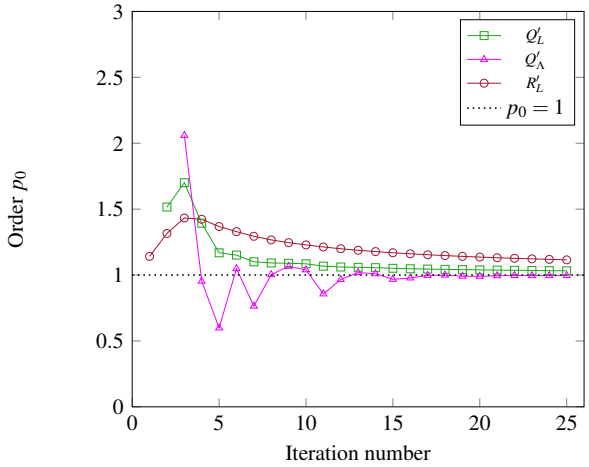
Figure 4.6: Strictly unsaturated medium (Example 1): Evolution of switching indicators for L/N-scheme for fixed $h = 80$ and $\tau = 0.01$. The dashed line is $C_{tol} = 1.5$, the switching criterion from L-scheme to Newton's method. The L/N-scheme oscillates between the linearization strategies, but eventually recovers. The effectivity indices (3.10) of the Newton iterations, introduced in Remark 3.2.1, are also plotted.

4.1.3 Order of convergence

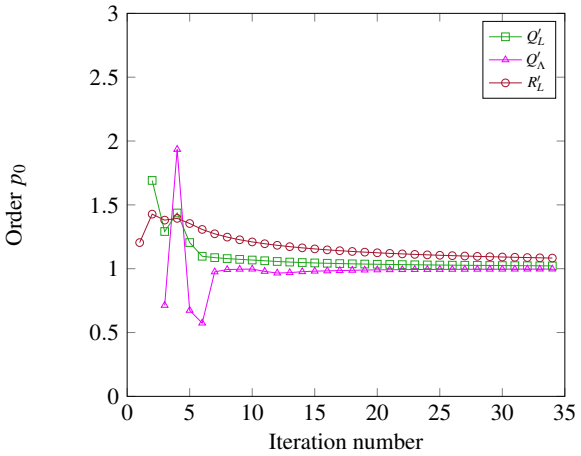
The convergence order for Example 1 of all linearization schemes, except the hybrid method, for fixed mesh size is depicted in Figure 4.7. The L/N-scheme is not included due to the sequence of iterates generated stemming from linearizations with different convergence orders. Although the different convergence orders are defined at a limit, it will always be limited by machine precision and therefore base the convergence order from a finite number of iterations.

As expected, the L-schemes converge linearly with all definitions of convergence order, but Q'_Λ is unstable in the first iterations. For the modified L-scheme, the limit of Q'_L and R'_L approaches 1 indicating linear convergence, but Q'_Λ oscillates between values above and below 1. Newton's method converges in very few iterations, but starts to increase towards 2 at the final iterations.

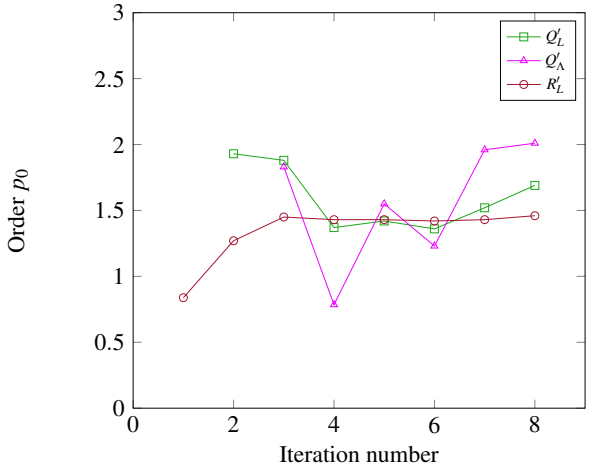
The convergence order of the L-scheme-Anderson acceleration approaches 1, indicating a linear convergence. However, at specific iterations the convergence order is much higher, as R'_L, Q'_L both make a small jump. Q'_λ also makes a jump. This is due to large changes in the iterates, and that the residual becomes significantly smaller at the same time. The same behaviour is also exhibited by Newton-Anderson acceleration which converges superlinearly/sub-quadratic.



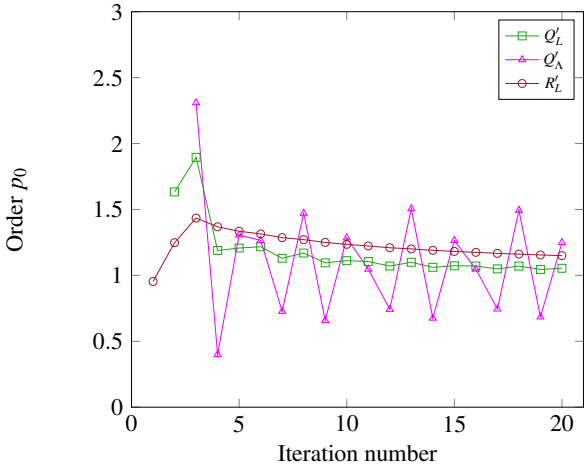
(a) Order of convergence L-scheme.



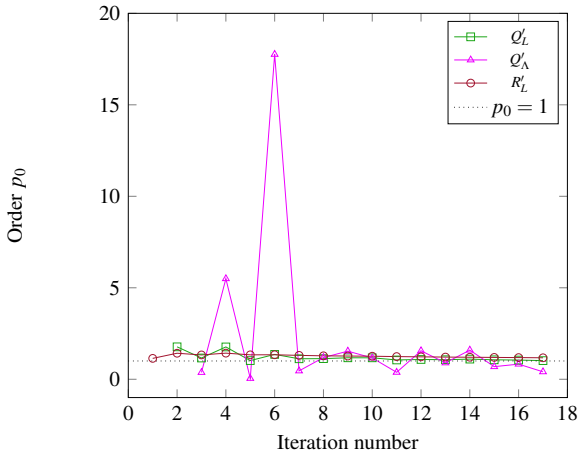
(b) Order of convergence L-scheme.



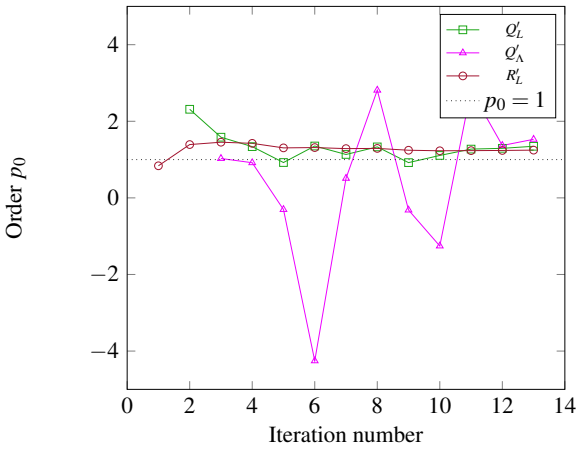
(c) Order of convergence Newton's method.



(d) Order of convergence Modified L-scheme.



(e) Order of convergence L-scheme-Anderson acceleration with depth 5.



(f) Order of convergence Newton-Anderson acceleration with depth 5.

Figure 4.7: Strictly unsaturated medium (Example 1): Computational order of convergence at each iteration for multiple numerical methods for fixed $h = \sqrt{2}/40$ and $\tau = 0.01$.

4.2 Example 2: Variably saturated medium

The parameters for this example are listed in Table 4.1 Example 2. The domain is divided into two parts, one a vadoze zone Ω_{vad} and the other the region below the water table, i.e., the groundwater zone Ω_{gw}. Let $\Omega = \Omega_{gw} \cup \Omega_{vad}$, where $\Omega_{gw} = [0,1] \times [0,1/4)$ and $\Omega_{vad} = [0,1] \times [1/4,1]$. We choose the pressure head to initially be given by

$$\psi^0(x,z) = \begin{cases} -z + 1/4, & (x,z) \in \Omega_{gw}, \\ -3, & (x,z) \in \Omega_{vad}, \end{cases}$$

where x represents the positional variable in the horizontal direction and z in the vertical direction. A constant Dirichlet condition is used on the surface, being equal to the initial condition at all times. For the rest of the boundary, no-flow boundary conditions are used. We use the following source term

$$f(x,z) = \begin{cases} 0 & (x,z) \in \Omega_{gw} \\ 0.006 \cos\left(\frac{4}{3}\pi(z-1)\right) \sin(2\pi x) & (x,z) \in \Omega_{vad}. \end{cases}$$

The solution is computed over the time interval $t \in [0, 0.01]$ and we only take one time step.

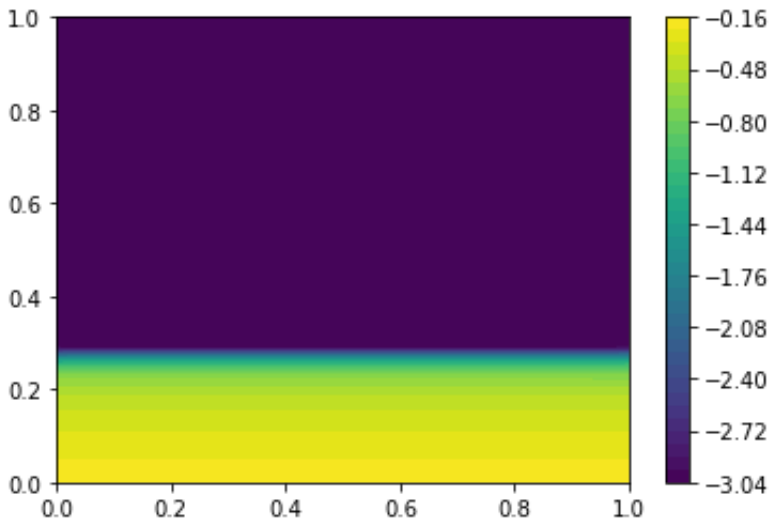


Figure 4.8: Variably saturated medium (Example 2): Pressure head profile at $T = 0.01$.

4.2.1 Comparison of convergence properties

The number of iterations for Example 2 for various mesh sizes and a fixed time step for all linearization schemes is shown in Figure 4.9. Once more, in every instance, the L-schemes and modified L-scheme converge with the modified L-scheme using less. In this case, Newton's method does not converge for any mesh size. Interestingly, there is a clear increase in robustness of Newton's method when applying Anderson acceleration, although it does not converge on the finest meshes considered. In fact, for the coarsest meshes considered, it uses the fewest iterations being equal to the hybrid method in some cases. The L/N-scheme requires the fewest number of iterations in every case but one.

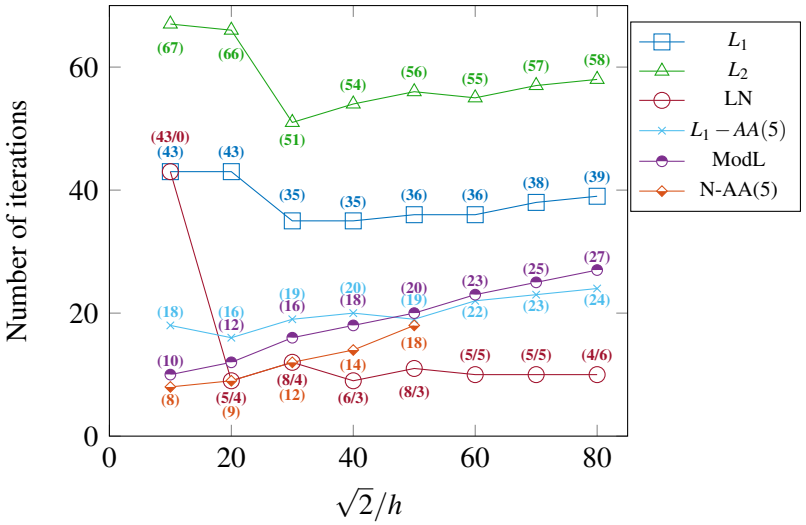


Figure 4.9: Variably saturated medium (Example 2): Total number of iterations for fixed $\tau = 0.01$ and varying mesh size. The numbers in the red parentheses correspond to (number of L-scheme iterations/number of Newton iterations).

The linearization schemes' CPU time performance is compared in Figure 4.10. The L-scheme using L_1 is less expensive than the other L-scheme, and both use computational time consistent with the number of iterations. However, the L-scheme (using L_1) requires approx. 164% of the computational time of the Anderson accelerated L-scheme. The L/N-scheme is the fastest in every scenario, requiring less than half the computational time of AA.

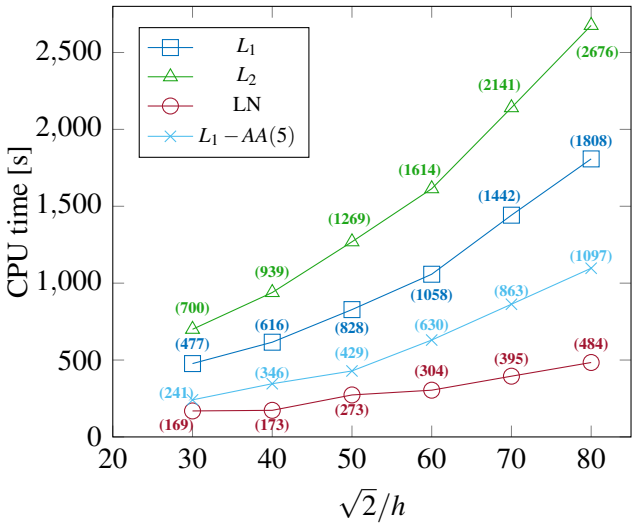
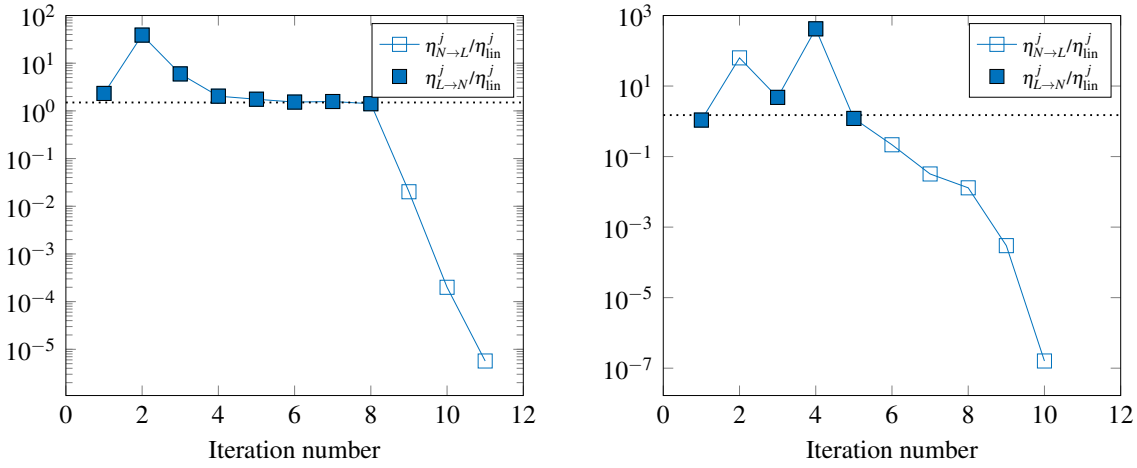


Figure 4.10: Variably saturated medium (Example 2): Total number of iterations for fixed $\tau = 0.01$ and varying mesh size.

4.2.2 Switching characteristics

At last a more thorough examination of the dynamic switch between Newton's method and the L-scheme is conducted. In Figure 4.11, the evolution of the switching indicators is shown for fixed time step and two different mesh sizes. For both mesh sizes the initial $\eta_{L\to N}$ is highly dependent on the initial data, for the finest mesh the switch happens immediately but it recovers.

Note that for the first iteration for $h = \sqrt{2}/80$ $\eta^1_{L\to N}$ is greater than 1. The example indicates a slight mesh dependence of the switch from L-scheme to Newton's method as the number of L-scheme iterations needed before the switch varies with the mesh size, see Figure 4.9.



(a) $h = \sqrt{2}/50$ and $\tau = 0.01$.

(b) $h = \sqrt{2}/80$ and $\tau = 0.01$.

Figure 4.11: Variably saturated medium (Example 2): Evolution of switching indicators for L/N-scheme for two mesh sizes and $\tau = 0.01$. The dashed line is $C_{tol} = 1.5$, the switching criterion from L-scheme to Newton's method.

Remark 4.2.1. *Note that no switch happens on the coarsest mesh. This motivates the use of applying Anderson acceleration only to the L-scheme iterates of the hybrid algorithm. Although no experimentation has been done in this regard, it is worth noting that in this specific case applying AA causes a switch after 5 accelerated L-scheme iterates converging with a total of 8 iterations.*

4.3 Example 3: Benchmark problem

The parameters for this example are found in Table 4.1 Example 3. The van Genucthen-Mualem parameters for this example represents silt loam [25]. Here an additional stablizing parameter is used, $L_3 = 1.501 \cdot 10^{-3}$, which does not satisfy (2.8). We consider a known benchmark problem, also considered in [34]. This example models the recharge of a groundwater reservoir from a drainage trench. The domain $\Omega \subset \mathbb{R}^2$ represents a vertical section of the subsurface. The drainage trench is simulated by a time dependent Dirichlet boundary condition on parts of the upper boundary. A constant Dirichlet condition is also used on the lower right side of the domain. For the remaining parts of the boundary, no-flow conditions are employed. The geometry of Ω is given by

$$\begin{aligned} \Omega &= [0,2] \times [0,3], \\ \Gamma_{D_1} &= [0,1] \times (3), \\ \Gamma_{D_2} &= (2) \times [0,1], \\ \Gamma_N &= \Omega \backslash \{\Gamma_{D_1} \cup \Gamma_{D_2}\}, \end{aligned}$$

and the initial and boundary conditions are

$$\begin{aligned}
&\psi^0 = 1 - z \\
&\psi = \begin{cases} -2 + 35.2t, & \text{if } t \leq \frac{1}{16}, \quad \text{on } \Gamma_{D_1}, \\ 0.2, & \text{if } t > \frac{1}{16}, \quad \text{on } \Gamma_{D_1}, \\ 1 - z, & \text{on } \Gamma_{D_2}, \end{cases} \\
&-K(\theta(\psi))\nabla(\psi + z) \cdot \boldsymbol{\nu} = 0, \quad \text{on } \Gamma_N,
\end{aligned}$$

where $\boldsymbol{\nu}$ is the outward pointing normal vector. We take 9 time steps with time step size $\tau = 1/48$ where the time unit is in days. The solution is computed on a regular mesh consisting of 2501 nodes and the final pressure head profile is depicted in Figure 4.12.

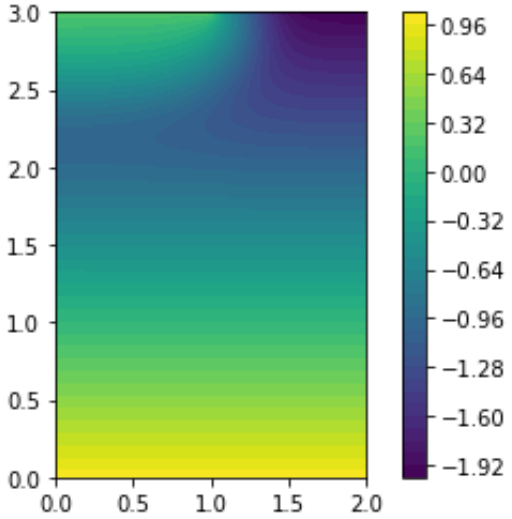


Figure 4.12: Benchmark problem (Example 3): Pressure head profile at final time 4.5 hours.

	No. Itr	CPU time
L_1	274	6136
L_2	330	7356
L_3	–	–
Newton	39	980
L/N	(10/30)	1021
$L_1 - AA$	105	2324
$L_3 - AA$	132	2934
ModL	90	
N-AA	44	

Table 4.2: Benchmark problem (Example 3): Comparison of number of iterations and computational time for 2501 nodes, $\tau = 1/48$ and final time $T = 3/16$. The numbers for L/N-scheme in parenthesis correspond to (number of L-scheme iterations/number of Newton iterations).

4.3.1 Comparison of convergence properties

The performance results for example 3 are shown in Table 4.2. All schemes converge for this example, except for the L-scheme which uses L_3. The Newton's method requires the least amount of iterations, one less than the hybrid method. However, both use substantially fewer iterations than the linearly converging schemes. The modified L-scheme uses fewer iterations than the L-schemes, visualizing the benefit of the convergence rate scaling with the time step. But the L-scheme Anderson acceleration (using L_1) is comparable to the modified L-scheme in terms of number of iterations. The Newton Anderson acceleration slows down the convergence of Newton's method, although not as much when compared to the prior examples. Applying Anderson acceleration to the non contractive L-scheme (using L_3) causes the scheme to converge. In addition to converging, it also uses far less iterations than the L-schemes which converges.

The computational cost of the L-schemes is consistent with the expense per iteration, being significantly slower than Newton and the hybrid method. Furthermore, both Anderson accelerated L-schemes use less than half the computational time as the fastest L-scheme. More importantly, in terms of computational time, the L/N-scheme performs similarly to Newton's method. It is slightly slower as a result of the additional iteration.

4.3.2 Switching characteristics

The dynamic switch is thoroughly examined in this section. As seen in the number of iterations, except for one time step, only one L-scheme iteration is required per time step. This indicates a successful dynamic switch for almost all time steps. In Figure 4.13 the efficiency indices for the Newton iterates are depicted. For all iterations in a given time step the index is above 1, which further validates Propositions 3.1.1 and 3.2.1. In the first time steps the index remains below 2, but as the saturated region grows, the sharpness of the estimate becomes slightly worse.

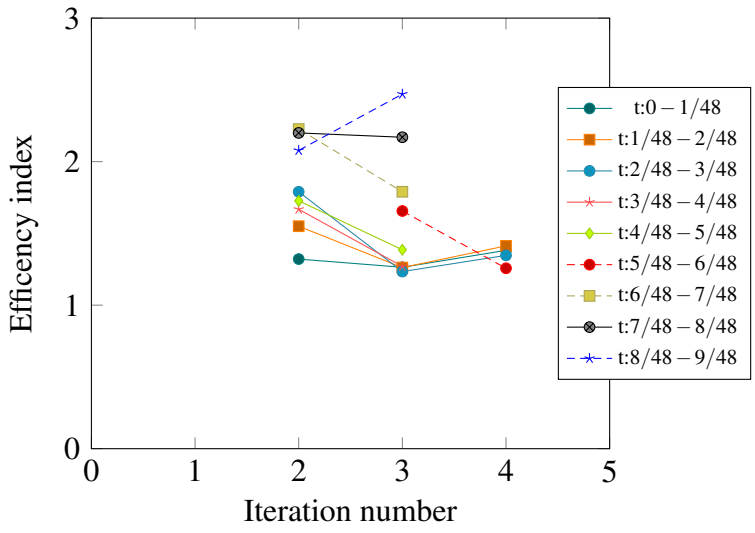
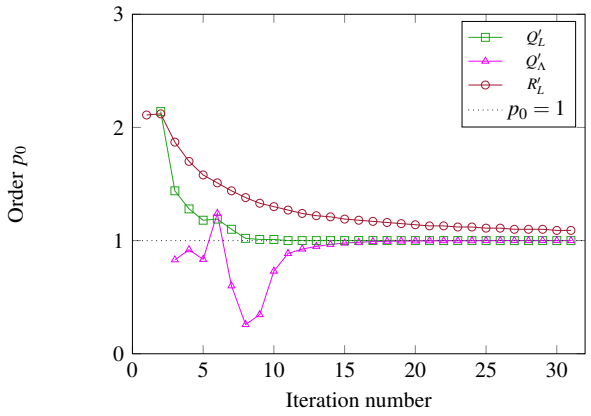
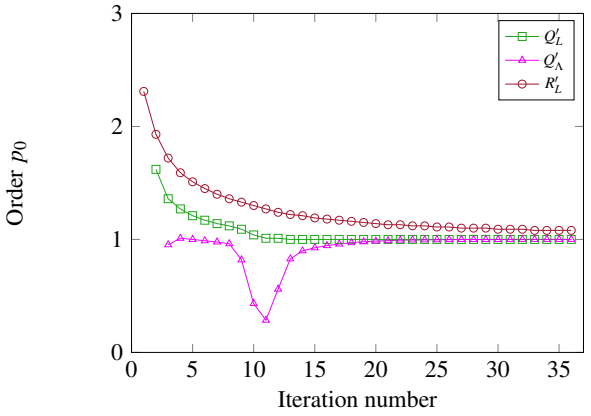


Figure 4.13: Benchmark problem (Example 3): Efficiency indexes of the Newton iterations at every time step. L-scheme iterations are omitted.

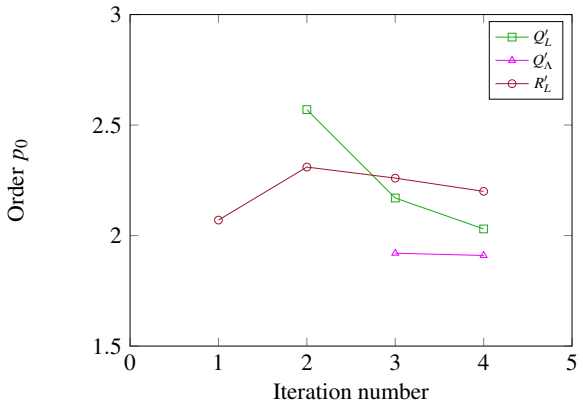
4.3.3 Order of convergence



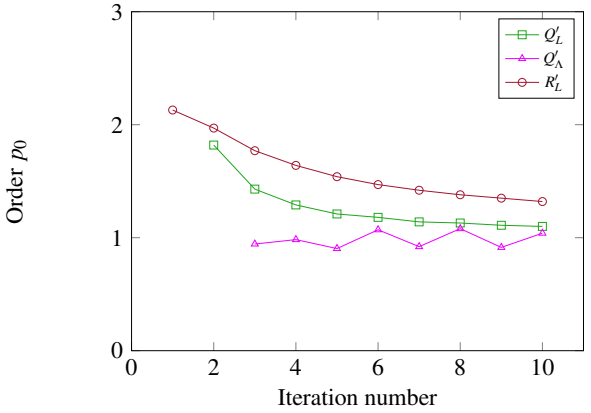
(a) Order of convergence L_1-scheme.



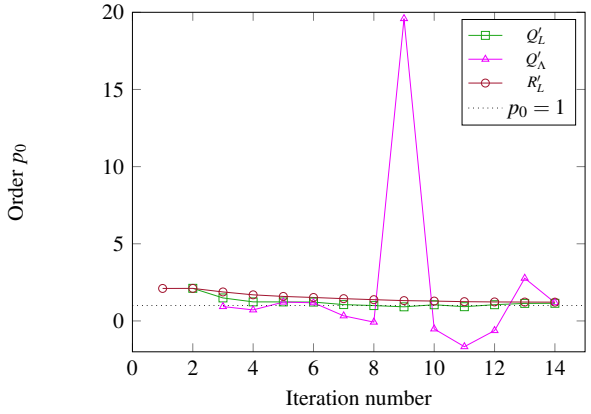
(b) Order of convergence L_2-scheme.



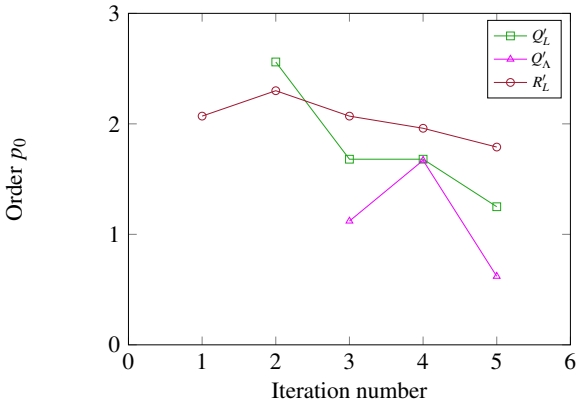
(c) Order of convergence Newton's method.



(d) Order of convergence Modified L-scheme.



(e) Order of convergence L-scheme-Anderson acceleration with depth 5.



(f) Order of convergence Newton-Anderson acceleration with depth 5.

Figure 4.14: Benchmark problem: Computational order of convergence at each iteration for multiple numerical methods from $t = 4/48$ to $t = 5/48$ and 2501 nodes.

The convergence order for Example 3 of all linearization schemes for fixed mesh size is depicted in Figure 4.14. The L-schemes and modified L-scheme converge linearly while Newton's method converges quadratically as expected. The Anderson acceleration of L-scheme and Newton's method both exhibit super-linear and sub-quadratic convergence respectively. For the acceleration of the L-scheme the convergence order Q'_Λ is negative at some iterations.

This shows why relating consecutive iterations causes less accurate estimates for the order of convergence.

4.4 Conclusions

In this chapter we tested the newly proposed switching algorithm on realistic examples and compared it to the L-scheme, the modified L-scheme, Newton's method and the Anderson accelerated L-scheme and Newton's method. The *a posteriori* estimators determining the switch between the L-scheme and Newton's method are reliable and efficient for the examples considered in the thesis. Therefore the dynamic switch between the linearization schemes are quite successful. In almost all cases the adaptive algorithm uses the fewest number of iterations and is computationally faster than the other schemes. Consequently the proposed algorithm is a good alternative solution strategy for Richards' equation, as it is both robust and quadratically convergent after the switch to Newton's method.

We observed that Anderson acceleration for the L-scheme decreased the number of iterations and increased for Newton's method. For Newton's method, the robustness was also increased. In addition, for a non-contractive L-scheme the use of Anderson acceleration caused convergence.

Chapter 5

Summary

In this thesis we have sought to solve Richards' equation. In time we applied a backward Euler discretization and in space we used a continuous Galerkin finite element discretization. The focal point was how to efficiently solve the resulting non-linear problem. We considered several linearization techniques, the L-scheme, the modified L-scheme and Newton's method and also Anderson acceleration applied to the L-scheme and Newton's method. For the L-scheme we gave a convergence proof and extended the previously existing optimality analysis to include the gravity term. In addition, we also gave an error estimate on the solution of the L-scheme. For Newton's method applied to a variant of Richards' equation after Kirchhoff transformation, we also proved the quadratic convergence if the initial guess is sufficiently close.

We proposed an adaptive algorithm between the L-scheme and Newton's method. This way we utilized the quadratic convergence of Newton's method when it converges and the robustness of the L-scheme. In order to determine when to switch between the two schemes, we derived reliable and efficient *a posteriori* indicators which predict the linearization error of the subsequent iteration. The algorithm always starts using the L-scheme, and at every iteration checks to see if the linearization error is predicted to decrease by switching to Newton's method. If this is the case, then Newton's method is used, otherwise the L-scheme is used for the next iteration. Hence, the adaptive scheme is now robust and quadratically convergent after switching to Newton's method.

The proposed algorithm is assessed on realistic examples. They demonstrate that the algorithm is as robust as the L-scheme and converges even when Newton's method fails. Furthermore, when Newton converges, the hybrid scheme takes roughly the same number of iterations and computational time as Newton's method while being significantly faster than other linearization and acceleration techniques.

Bibliography

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces* (Pure and applied mathematics). 2003, vol. 140, ISBN: 9780120441433.

[2] H. Alt, S. Luckhaus, and A. Visintin, "On nonstationary flow through porous media," *Annali di Matematica Pura ed Applicata*, vol. 136, no. 1, pp. 303–316, 1984.

[3] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *Journal of the ACM (JACM)*, vol. 12, no. 4, pp. 547–560, 1965.

[4] T. Arbogast, M. Obeyesekere, and M. F. Wheeler, "Numerical Methods for the Simulation of Flow in Root-Soil Systems," *SIAM journal on numerical analysis*, vol. 30, no. 6, pp. 1677–1702, 1993, ISSN: 0036-1429.

[5] T. Arbogast, "An error analysis for Galerkin approximations to an equation of mixed elliptic-parabolic type," *Technical Report TR90-33, Department of Computational and Applied Mathematics, Rice University, Houston, TX*, 1990.

[6] T. Arbogast, M. F. Wheeler, and N. Y. Zhang, "A non-linear mixed finite element method for a degenerate parabolic equation arising in flow in porous media," *SIAM J. Numer. Anal. 33*, pp. 1669–1687, 1996.

[7] S. Bassetto, C. Cancès, G. Enchéry, and Q.-H. Tran, "On several numerical strategies to solve Richards' equation in heterogeneous media with finite volumes," *Computational geosciences*, vol. 26, no. 5, pp. 1297–1322, 2022, ISSN: 1420-0597.

[8] M. Bause and P. Knabner, "Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods," *Adv. Water Resources 27*, pp. 565–581, 2004.

[9] 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*, vol. 45, no. 8, pp. 1025–1046, 1999, ISSN: 0029-5981.

[10] J. W. Both, K. Kumar, J. M. Nordbotten, I. S. Pop, and F. A. Radu, "Iterative Linearisation Schemes for Doubly Degenerate Parabolic Equations," in *Numerical Mathematics and Advanced Applications ENUMATH 2017*, ser. Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2019, pp. 49–63, ISBN: 9783319964140.

[11] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, "Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media," *Computers and mathematics with applications (1987)*, vol. 77, no. 6, pp. 1479–1502, 2019, ISSN: 0898-1221.

[12] K. Brenner and C. Cances, "Improving Newton's method performance by parametrization: the case of the Richards equation," *SIAM Journal on Numerical Analysis*, vol. 55, no. 4, pp. 1760–1785, 2017.

[13] S. C. Brenner, *The mathematical theory of finite element methods*, 2002.

[14] R. Brooks and A. Corey, "Properties of porous media affecting fluid flow," *Journal of the Irrigation and Drainage Division*, vol. 92, no. 2, pp. 61–90, 1966.

[15] E. Cătinaş, "How Many Steps Still Left to x?" eng, *SIAM review*, vol. 63, no. 3, pp. 585–624, 2021, ISSN: 0036-1445.

[16] M. Celia, E. Bouloutas, and R. Zarba, "General mass-conservative numerical solution for the unsaturated flow equation," *Water Resources Research*, vol. 26, no. 7, pp. 1483–1496, 1990.

[17] M. A. Celia, E. T. Bouloutas, and R. L. Zarba, "A general mass-conservative numerical solution for the unsaturated flow equation," *Water resources research*, vol. 26, no. 7, pp. 1483–1496, 1990, ISSN: 0043-1397.

[18] W. Cheney, *Analysis for applied mathematics* (Graduate texts in mathematics). Springer, 2001, vol. 208, ISBN: 0387952799.

[19] H. Darcy, "Les Fontaines Publiques de la Ville de Dijon," 1856.

[20] C. Evans, S. Pollock, L. G. Rebholz, and M. Xiao, "A Proof That Anderson Acceleration Improves the Convergence Rate in Linearly Converging Fixed-Point Methods (But Not in Those Converging Quadratically)," *SIAM journal on numerical analysis*, vol. 58, no. 1, pp. 788–810, 2020, ISSN: 0036-1429.

[21] L. C. Evans, *Partial differential equations*, 2010.

[22] R. Eymard, M. Gutnic, and D. Hilhorst, "The finite volume method for Richards equation," *Computational geosciences*, vol. 3, no. 3-4, pp. 259–294, 1999, ISSN: 1420-0597.

[23] R. Eymard, D. Hilhorst, and M. Vohralik, "A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems," *Numerische Mathematik*, vol. 105, no. 1, pp. 73–131, 2006, ISSN: 0029-599X.

[24] M. W. Farthing and F. L. Ogden, "Numerical Solution of Richards' Equation: A Review of Advances and Challenges," *Soil Science Society of America Journal*, vol. 81, no. 6, pp. 1257–1269, 2017.

[25] M. T. van Genuchten, "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils," *Soil Science Society of America journal*, vol. 44, no. 5, pp. 892–898, 1980, ISSN: 0361-5995.

[26] R. Haverkamp, M. Vauclin, J. Touma, P. J. Wierenga, and G. Vachaud, "A Comparison of Numerical Simulation Models For One-Dimensional Infiltration," *Soil Science Society of America Journal*, vol. 41, no. 2, pp. 285–294, 1977.

[27] D. Illiano, I. S. Pop, and F. A. Radu, "Iterative schemes for surfactant transport in porous media," *Computational geosciences*, vol. 25, no. 2, pp. 805–822, 2021, ISSN: 1420-0597.

[28] W. Jäger and J. Kačur, "Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 29, no. 5, pp. 605–627, 1995.

[29] E. Keilegavlen, R. Berge, A. Fumagalli, M. Starnoni, I. Stefansson, J. Varela, and I. Berre, "Porepy: An open-source software for simulation of multiphysics processes in fractured porous media," *Computational geosciences*, vol. 25, no. 1, pp. 243–265, 2021, ISSN: 1420-0597.

[30] R. A. Klausen, F. A. Radu, and G. T. Eigestad, "Convergence of MPFA on triangulations and for Richards' equation," *Int. J. for Numer. Meth. Fluids 58*, pp. 1327–1351, 2008.

[31] P. Knabner, "Finite element simulation of saturated-unsaturated flow through porous media," in *Large Scale Scientific Computing*, P. Deuflhard and B. Engquist, Eds. Birkhäuser Boston, 1987, pp. 83–93, ISBN: 978-1-4684-6754-3.

[32] P. Knabner and L. Angerman, *Numerical Methods for Elliptic and Parabolic Partial Differential Equations* (Texts in Applied Mathematics). Springer, 2003, vol. 44, ISBN: 038795449X.

[33] F. Lehmann and P. Ackerer, "Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media," *Transport in Porous Media*, vol. 31, no. 3, pp. 275–292, 1998.

[34] F. List and F. A. Radu, "A study on iterative methods for solving Richards' equation," *Computational geosciences*, vol. 20, no. 2, pp. 341–353, 2016, ISSN: 1420-0597.

[35] T. J. McDougall and S. J. Wotherspoon, "A simple modification of newton's method to achieve convergence of order 1+root 2," *Applied mathematics letters*, vol. 29, pp. 20–25, 2014, ISSN: 0893-9659.

[36] K. Mitra and I. Pop, "A modified L-scheme to solve nonlinear diffusion problems," *Computers and mathematics with applications (1987)*, vol. 77, no. 6, pp. 1722–1738, 2019, ISSN: 0898-1221.

[37] K. Mitra and M. Vohralík, "A posteriori error estimates for the Richards equation," working paper or preprint, Aug. 2021. [Online]. Available: `https://hal.inria.fr/hal-03328944`.

[38] K. Mitra and M. Vohralík, "Reliable, efficient, and robust a posteriori estimates for nonlinear elliptic problems: An orthogonal decomposition result based on iterative linearization," *In Preparation (To be submitted in Jan 2023)*,

[39] Y. Mualem, "New model for predicting the hydraulic conductivity of unsaturated porous media," *Water resources research*, vol. 12, no. 3, pp. 513–522, 1976, ISSN: 0043-1397.

[40] J. Nocedal, *Numerical Optimization*, 2006.

[41] J. M. Nordbotten and M. A. Celia, *Geological storage of CO2: modeling approaches for large-scale simulation*. Wiley, 2012, ISBN: 111813706X.

[42] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000.

[43] S. Pollock and H. Schwartz, "Benchmarking results for the Newton–Anderson method," *Results in Applied Mathematics*, vol. 8, p. 100 095, 2020, ISSN: 2590-0374.

[44] I. S. Pop, F. A. Radu, and P. Knabner, "Mixed finite elements for the Richards' equation: linearization procedure," *Journal of computational and applied mathematics*, vol. 168, no. 1-2, pp. 365–373, 2004, ISSN: 0377-0427.

[45] F. A. Radu, I. S. Pop, and P. Knabner, "Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation," *SIAM journal on numerical analysis*, vol. 42, no. 4, pp. 1452–1478, 2004, ISSN: 0036-1429.

[46] F. A. Radu, I. S. Pop, and P. Knabner, "On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation," *Numerical Mathematics and Advanced Applications*, vol. 42, pp. 1194–1200, 2006.

[47] F. A. Radu, *Mixed finite element discretization of Richards' equation: error analysis and application to realistic infiltration problems*. Naturwissenschaftliche Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg, 2004.

[48] F. A. Radu, *MAT254 Flow in Porous Media*, `https://mitt.uib.no/courses/29788/files/folder/LectureNotes`, Accessed online 08-August-2021.

[49] 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*, vol. 38, no. 2, pp. 884–920, 2018, ISSN: 0272-4979.

[50] F. A. Radu and W. Wang, "Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media," *Nonlinear analysis.*, vol. 15, pp. 266–275, 2014, ISSN: 1468-1218.

[51] F. A. Radu, I. S. Pop, and P. Knabner, "Error estimates for a mixed finite element discretization of some degenerate parabolic equations," *Numer. Math. 109*, pp. 285–311, 2008.

[52] L. A. Richards, "Capillary conduction of liquids through porous mediums," *Physics*, vol. 1, no. 5, pp. 318–333, 1931.

[53] 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*, vol. 23, no. 5, pp. 1593–1614, 2002, ISSN: 1064-8275.

[54] J. S. Stokke, K. Mitra, E. Storvik, J. W. Both, and F. A. Radu, "An adaptive solution strategy for Richards' equation," 2023. arXiv: 2301.02055.

[55] E. Storvik, *On the optimization of iterative schemes for solving non-linear and/or coupled PDEs*, 2018.

[56] E. Storvik, J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, "On the optimization of the fixed-stress splitting for Biot's equations," *International Journal for Numerical Methods in Engineering*, vol. 120, no. 2, pp. 179–194, 2019.

[57] A. Toth and C. T. Kelley, "Convergence analysis for anderson acceleration," *SIAM journal on numerical analysis*, vol. 53, no. 2, pp. 805–819, 2015, ISSN: 0036-1429.

[58] H. F. Walker and P. Ni, "Anderson acceleration for fixed-point iterations," *SIAM journal on numerical analysis*, vol. 49, no. 3/4, pp. 1715–1735, 2011, ISSN: 0036-1429.

[59] H. Wilhelm Alt and S. Luckhaus, "Quasilinear elliptic-parabolic differential equations," *Mathematische Zeitschrift*, vol. 183, no. 3, pp. 311–341, 1983, ISSN: 0025-5874.

Appendix A

Convergence proof of L-scheme

Proof of Theorem 2.3.1. In the following $\|\cdot\|_{L^2(\Omega)}$ is denoted by $\|\cdot\|$ to simplify notation. First, by subtracting (2.5) from (2.6) and choosing $v_h = e^{n,j} = \psi_h^{n,j} - \psi_h^n$ we obtain

$$\begin{aligned}
&\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j}\rangle + L\langle (\psi_h^{n,j} - \psi_h^{n,j-1}), e^{n,j}\rangle \\
&\quad + \tau\langle K(\theta(\psi_h^{n,j-1}))\nabla\psi_h^{n,j} - K(\theta(\psi_h^n))\nabla\psi_h^n, \nabla e^{n,j}\rangle \\
&\quad + \tau\langle \Big(K(\theta(\psi_h^{n,j-1})) - K(\theta(\psi_h^n))\Big)\nabla z, \nabla e^{n,j}\rangle = 0
\end{aligned}$$

We split the saturation term and let some γ satisfy (2.10) such that

$$\begin{aligned}
&\gamma\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j-1}\rangle + (1-\gamma)\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j-1}\rangle \\
&\quad \langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j} - e^{n,j-1}\rangle + L\langle (e^{n,j} - e^{n,j-1}), e^{n,j}\rangle \\
&\quad + \tau\langle K(\theta(\psi_h^{n,j-1}))\nabla\psi_h^{n,j} - K(\theta(\psi_h^n))\nabla\psi_h^n, \nabla e^{n,j}\rangle \\
&\quad + \tau\langle \Big(K(\theta(\psi_h^{n,j-1})) - K(\theta(\psi_h^n))\Big)\nabla z, \nabla e^{n,j}\rangle = 0.
\end{aligned} \tag{A.1}$$

We have the algebraic relation

$$L\langle (e^{n,j} - e^{n,j-1}), e^{n,j}\rangle = \frac{L}{2}\|e^{n,j}\|^2 + \frac{L}{2}\|e^{n,j} - e^{n,j-1}\|^2 - \frac{L}{2}\|e^{n,j-1}\|^2. \tag{A.2a}$$

Furthermore, by Assumption 2.3.1 θ is Lipschitz continuous and monotonically increasing and the derivative is bounded from below by L_{min}, we obtain the following estimates

$$(1-\gamma)\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j-1}\rangle \geq \frac{1-\gamma}{L_\theta}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2, \tag{A.2b}$$

$$\gamma\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j-1}\rangle \geq \gamma L_{min}\|e^{n,j-1}\|^2. \tag{A.2c}$$

Inserting the relations (A.2) into (A.1), we get the inequality

$$\begin{aligned}
&\gamma L_{min}\|e^{n,j-1}\|^2 + \frac{1-\gamma}{L_\theta}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2 + \Big\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j} - e^{n,j-1}\Big\rangle \\
&+ \frac{L}{2}\|e^{n,j}\|^2 + \frac{L}{2}\|e^{n,j} - e^{n,j-1}\|^2 + \tau\Big\langle K(\theta(\psi_h^{n,j-1}))\nabla e^{n,j}, \nabla e^{n,j}\Big\rangle \\
&+ \tau\Big\langle \Big(K(\theta(\psi_h^{n,j-1})) - K(\theta(\psi_h^n))\Big)\Big(\nabla\psi_h^{n,j} + \nabla z\Big), \nabla e^{n,j}\Big\rangle \leq \frac{L}{2}\|e^{n,j-1}\|^2
\end{aligned} \tag{A.3}$$

Using Young's inequality

$$\begin{aligned}
&\langle \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n), e^{n,j} - e^{n,j-1} \rangle \\
\leq &\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\| \|e^{n,j} - e^{n,j-1}\| \\
\leq &\frac{1}{2L}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2 + \frac{L}{2}\|e^{n,j} - e^{n,j-1}\|^2
\end{aligned} \tag{A.4a}$$

By Assumptions 2.3.2 to 2.3.3, the permeability is Lipschitz continuous with L_K and the finite element solution satisfies $\|\nabla \psi_h^n\|_{L^\infty} \leq M < \infty$, and applying Young's inequality again we get

$$\begin{aligned}
&\tau \langle \left(K(\theta(\psi_h^{n,j-1})) - K(\theta(\psi_h^n)) \right) \left(\nabla \psi_h^{n,j} + \nabla z \right), \nabla e^{n,j} \rangle \\
&\leq \tau \left\| \left(K(\theta(\psi_h^{n,j-1})) - K(\theta(\psi_h^n)) \right) \left(\nabla \psi_h^{n,j} + \nabla z \right) \right\| \left\| \nabla e^{n,j} \right\| \\
&\leq \tau L_K (M+1)^2 \left\| \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n) \right\| \left\| \nabla e^{n,j} \right\| \\
&\leq \frac{\tau L_K^2 (M+1)^2}{2K_{min}} \left\| \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n) \right\|^2 + \frac{\tau K_{min}}{2} \left\| \nabla e^{n,j} \right\|^2
\end{aligned} \tag{A.4b}$$

Inserting the inequalities (A.4) into (A.3) and by Assumption 2.3.2 there exists a bound from below of the permeability, K_{min}, we are able to obtain

$$\begin{aligned}
\gamma L_{min} \|e^{n,j-1}\|^2 + \frac{1-\gamma}{L_\theta}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2 + \frac{L}{2}\|e^{n,j}\| + \frac{L}{2}\|e^{n,j} - e^{n,j-1}\|^2 \\
+ \tau K_{min} \|\nabla e^{n,j}\|^2 \leq \frac{L}{2}\|e^{n,j-1}\|^2 + \frac{\tau L_K^2 (M+1)^2}{2K_{min}}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2 + \frac{\tau K_{min}}{2}\|\nabla e^{n,j}\|^2 \\
+ \frac{1}{2L}\|\theta(\psi_h^{n,j-1}) - \theta(\psi_h^n)\|^2 + \frac{L}{2}\|e^{n,j} - e^{n,j-1}\|^2.
\end{aligned} \tag{A.5}$$

Combining and rearranging terms the terms in (A.5) and applying Poincaré's inequality,

$$\begin{aligned}
&\left(\frac{1-\gamma}{L_\theta} - \frac{\tau L_K^2 (1+M)^2}{2K_{min}} - \frac{1}{2L} \right) \left\| \theta(\psi_h^{n,j-1}) - \theta(\psi_h^n) \right\|^2 \\
&+ \frac{L}{2} \left\| e^{n,j} \right\|^2 + \frac{\tau K_{min}}{2C_\Omega} \left\| e^{n,j} \right\|^2 \leq \left(\frac{L}{2} - \gamma L_{min} \right) \left\| e^{n,j-1} \right\|^2
\end{aligned} \tag{A.6}$$

Since L satisfies (2.8) we have

$$\frac{1-\gamma}{L_\theta} - \frac{\tau L_K^2 (1+M)^2}{2K_{min}} - \frac{1}{2L} \geq 0, \tag{A.7}$$

and obtain the error estimate

$$\left\| e^{n,j} \right\|^2 \leq \frac{L - 2\gamma L_{min}}{L + \frac{\tau K_{min}}{C_\Omega}} \left\| e^{n,j-1} \right\|^2. \tag{A.8}$$

□

Appendix B

L-adaptivity

As shown in Theorem 2.3.1, the L-scheme converges unconditionally if L satisfies (2.8). However, numerical results in [34] suggest that the optimal rate of convergence of the L-scheme is obtained for a considerably smaller L although convergence cannot always be guaranteed for such values. Hence, to speed up the computations, it is possible to start the iterations with a smaller value of L and then use a posteriori estimates to decide if L is to be increased or not. Analogous to Propositions 3.1.1 and 3.2.1 the result we are going to use for this purpose is:

Proposition B.0.1 (Error control of L-scheme)**.** *For a given $\psi_h^{n,0}$, $\psi_h^{n-1} \in V_h$, let $\{\psi_h^{n,k}\}_{k=1}^{j+1} \subset V_h$ solve* (2.6) *for some $j \in \mathbb{N}$. Then, one has*

$$\left|\left|\left| \psi_h^{n,j+1} - \psi_h^{n,j} \right|\right|\right|_{L,\psi_h^{n,j}} \leq \eta_{L\to L}^j,$$

where

$$\eta_{L\to L}^j := \left([\eta_{L\to L,1}^j]^2 + \tau [\eta_{L\to L,2}^j]^2 \right)^{\frac{1}{2}}$$

with

$$\eta_{L\to L,1}^j := \|L^{-\frac{1}{2}}(L(\psi_h^{n,j} - \psi_h^{n,j-1}) - (\theta(\psi_h^{n,j}) - \theta(\psi_h^{n,j-1})))\|,$$
$$\eta_{L\to L,2}^j := \left\| (K(\theta(\psi_h^{n,j})) - K(\theta(\psi_h^{n,j-1})))K(\theta(\psi_h^{n,j}))^{-\frac{1}{2}} \nabla(\psi_h^{n,j} + z) \right\|.$$

The detailed proof is omitted. Observe that neither Assumption 3.1.1 nor any specific handling of the degenerate domains is necessary for the estimate shown above.

B.1 An Adaptive L-scheme algorithm

Based on Proposition B.0.1, we propose an algorithm that selects optimal L-values adaptively.

Algorithm 3 The L-adaptive scheme

Require: $\boldsymbol{\psi}^{n,0} \in L^2(\Omega)$ as initial guess, $L_M := \sup_{\psi\in\mathbb{R}} \theta'(\psi)$, and $L_m := L_M/8$
Ensure: $C_{L\to L} = \sqrt{2}$, $L = L_m$
  for i=1,2,.. **do**
    Compute iterate using L-scheme, i.e., (2.6)
    if $\eta_{L\to L}^i > 1$ **then**
      Replace $L_m = L$, $L = \min(C_{L\to L}L, L_M)$, and **continue**.
    else if $\eta_{L\to L}^j > 0.8$ for $j \in \{i, i-1, i-2\}$ **then**
      Replace $L = \max(0.9L, 1.1L_m)$ and **continue**.

B.2 Numerical results

In Figure B.1, we show that the L-adaptive scheme outperforms a fixed L-approach. Because of the large time step size, $L_\theta/2$ is too small for convergence in this case. The number of iterations is decreased by 20 when compared to a fixed L_1 with the same mesh size and time step size, see Figure 4.3. For smaller time steps, the numerical results show that Algorithm 3 requires roughly the same number of iterations as a fixed and optimized $L = L_1$ less than L_θ. The advantage of such an adaptive technique is that no L optimization study is required prior to the simulation. However, because the L-adaptive strategy does not significantly improve the behavior of the L-scheme over the optimized $L = L_1$, we chose not to include it in Algorithm 2. Finally, in all examples, $\eta_{L\to L}$ was only observed to become greater than one for non-convergent parameters L and converges to a limit less than one when the L-scheme converged.

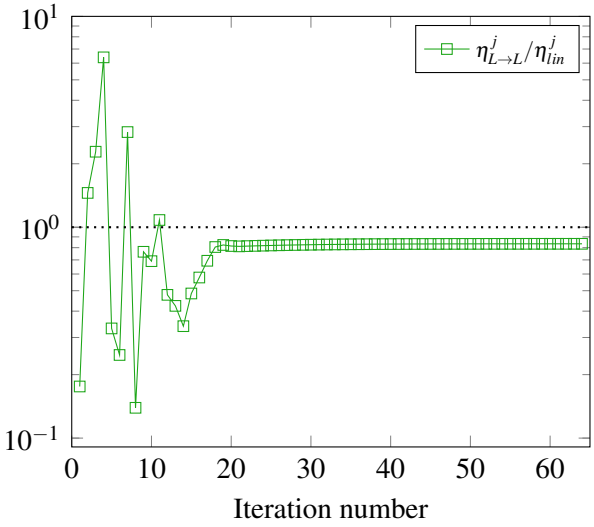


Figure B.1: Strictly unsaturated medium (Example 1): L-scheme with L-adaptivity and initial stabilization parameter $L_0 = L_2/8$, $h = \sqrt{2}/40$ and $\tau = 1$.