On the optimization of iterative schemes for solving non-linear and/or coupled PDEs

Erlend Storvik

Master thesis in Applied and Computational Mathematics,
Institute of Mathematics,
University of Bergen,
Autumn 2018

Contents

Abstract

In this thesis we study the optimization of iterative schemes as both linearization methods, and as splitting methods for solving non-linear and coupled partial differential equations (PDEs). We consider two equations that are describing processes in porous media; Richards' equation, a possibly degenerate, non-linear and elliptic/parabolic equation that models flow of water in saturated/unsaturated porous media, and Biot's equations, a coupled system of equations that models flow in deformable porous media.

For Richards' equation we compare the numerical properties of several linearization schemes, including the Newton-Raphson method, the modified Picard method and the L-scheme. Additionally, we prove convergence of the linearly and globally convergent L-scheme and discuss theoretically and practically how to choose its stabilization parameter optimally in the sense that convergence is obtained in the least amount of iterations.

The second aim of the thesis is to effectively solve the quasi-static, linear Biot model. We consider the fixed-stress splitting scheme, which is a popular method for iteratively solving Biot's equations. It is well-known that the convergence of the method is strongly dependent on the applied stabilization parameter. We propose a new approach to optimize this parameter, and show theoretically that it does not only depend on the mechanical properties and the coupling coefficient, but also on the fluid's flow properties. The type of analysis presented in this thesis is not restricted to a particular spatial discretization, but we require it to be inf-sup stable. The convergence proof also applies to low-compressible or incompressible fluids, and low-permeable porous media. We perform illustrative numerical examples, including a well-known benchmark problem, Mandel's problem. The results largely agree with the theoretical findings. Furthermore, we show numerically that for conditionally inf-sup stable discretizations, the performance of the fixed-stress splitting scheme behaves in a manner which contradicts the theory provided for inf-sup stable discretizations.

Acknowledgements

First and foremost I would like to thank my supervisors Florin Adrian Radu and Jakub Wiktor Both for their guidance. I have learned incredibly much during the last two years and you have doubtlessly been the greatest contributors. I am especially grateful to you Jakub for always being available to answer my questions and finding bugs in my code. Thank you Florin for suggesting the topics of the thesis and for showing such an interest in my academic life. You are both great mentors.

I would also like to thank the SFB1313 for granting me the "Scholarship Program for Master's Students" allowing me to stay with the CMAT group at the University of Hasselt. The CMAT group at UHasselt, particularly Iuliu Sorin Pop and Carina Bringedal, I thank for hosting me.

Finally, I thank my friends and family, particularly Laila, for being in my life and making me happy (also) when I am not doing mathematics.

Introduction

There are many topics in the field of flow in porous media that are of great societal relevance. Some examples are groundwater simulations, CO_2-storage, life sciences and geothermal energy. Common to all of them is the need to solve partial differential equations. These equations are often non-linear, coupled, time-dependent and possibly degenerate, and therefore require robust numerical methods to solve efficiently. We consider two cases in this thesis; the non-linear, time-dependent and possibly degenerate Richards' equation, and the coupled Biot's equations.

When solving non–linear, time-dependent equations one could apply an explicit temporal discretization to avoid solving a non-linear system at each time step. However, this often requires the time steps to be smaller than what is beneficial. The other way to approach the problem is with an implicit temporal discretization. This requires the application of a non-linear solver. The most popular of these solvers is the Newton-Raphson method, which provides a very fast way to solve the problem, but its local convergence property requires a new bound on the time step size. Moreover, it involves computation of derivatives which might be a costly process. Another alternative is to use a globally convergent fixed-point type solver. One example is the L-scheme, in which one includes a stabilization constant instead of the derivatives in the Newton-Raphson method. While this scheme might converge at a slower speed it has several benefits making it competitive to the Newton-Raphson method.

In this thesis we discuss the theoretical convergence properties of the L-scheme when applied to a special case of two-phase flow; Richards' equation. This equation models flow of water in saturated/unsaturated porous media. In the unsaturated region one assumes that the air moves freely, and therefore its pressure is zero. Hence, the system can be reduced to the single equation describing solely the complementary phase

$$\partial_t(s_w(p)) - \nabla \cdot (\kappa(s_w(p))(\nabla p - \boldsymbol{g})) = f, \tag{1}$$

which was proposed by L.A. Richards in 1931 [1]. The equation and its coefficients are introduced in Section 1.3.6. Already at this point it is worthwhile to notice that the equation contains two nonlinear terms, s_w (saturation) and κ (permeabil-

ity), which when applying an implicit temporal discretization, like implicit Euler, requires the use of a linearization scheme. The most widely used schemes in the literature are the Newton-Raphson method [2, 3], the modified Picard method [4] and the L-scheme [5, 6]. While we focus on the theoretical optimization of the L–scheme, we also compare the performance of the L–scheme with the aforementioned schemes. Moreover, we carry out this comparison with the modified L–scheme [7] and a localized version of the L–scheme. We introduce the localized version of the L-scheme, where we compute its stabilization parameter for each element, in Chapter 2.

We also consider the most commonly used mathematical model for flow in deformable porous media, the quasi-static, linear Biot model (see e.g. [8]):

Find $(\mathbf{u}, p)$ such that

$$-\nabla \cdot (2\mu\varepsilon(\boldsymbol{u}) + \lambda\nabla \cdot \boldsymbol{u}I) + \alpha\nabla p = \boldsymbol{f}, \tag{2}$$

$$\frac{\partial}{\partial t}\left(\frac{p}{M} + \alpha\nabla \cdot \boldsymbol{u}\right) - \nabla \cdot (\kappa(\nabla p - \boldsymbol{g}\rho)) = S_f, \tag{3}$$

where (2) models balance of linear momentum and (3) models mass conservation of the fluid. There are two widely used approaches for solving coupled equations: monolithically or by using an iterative splitting algorithm. The former has the advantage of being unconditionally stable, while the latter is much easier to implement, typically building on already available, separate numerical codes for porous media flow and for mechanics. On the other hand, a naive splitting of Biot's equations will lead to an unstable scheme [9]. To overcome this, one adds a stabilization term in either the mechanics equation (the so-called *undrained split scheme* [10]) or in the flow equation (the *fixed-stress splitting scheme* [11]). The splitting methods have very good convergence properties, making them a valuable alternative to monolithic solvers for simulation of the linear Biot model, see e.g. [11, 9, 12, 13]. In Chapter 3 we discuss the fixed-stress splitting scheme, but we remark that a similar analysis can be performed for the undrained split scheme.

The initial derivation of the fixed-stress splitting scheme had a physical motivation [11, 9]: one fixes the (volumetric) stress i.e. imposes

$$K_{dr}\nabla \cdot \mathbf{u}^i - \alpha p^i = K_{dr}\nabla \cdot \mathbf{u}^{i-1} - \alpha p^{i-1}$$

and uses this to replace $\alpha\nabla \cdot \mathbf{u}^i$ in the flow equation. Here K_{dr} is the physical, drained bulk modulus, defined as $K_{dr} = \frac{2\mu}{d} + \lambda$. The resulting stabilization parameter L, from now on called on the *physical* parameter, is $L_{phys} = \dfrac{\alpha^2}{K_{dr}}$. The physical parameter depends on the mechanics and the coupling coefficient. Consequently, L_{phys} was the recommended value for the stabilization parameter, and one assumed that the method is not converging (it is not stable) for $L < L_{phys}$. In

2013, a rigorous mathematical analysis of the fixed-stress splitting scheme was for the first time performed in [12], where the authors show that the scheme is a contraction for any stabilization parameter $L \geq \frac{L_{phys}}{2}$. This analysis was confirmed in [13] for heterogeneous media, using a simpler technique. A natural question arises immediately: is now L_{phys} or $\frac{L_{phys}}{2}$ the optimal stabilization parameter, in the sense that the number of iterations is smallest? The question is relevant, because the number of iterations to achieve convergence can differ considerably depending on the choice of the stabilization parameter [14, 13, 15, 16].

In a recent study [15], the authors considered different numerical settings and looked at the convergence of the fixed-stress splitting scheme. They determined numerically the optimal stabilization parameter for each considered case. This study, together with the previous results presented in [16] and [13] suggest that the optimal parameter is actually a value in the interval $\left[\frac{L_{phys}}{2}, L_{phys}\right]$, depending on the data. In particular, the optimal parameter depends on both the boundary conditions and the flow parameters and is not solely dependent on the mechanics and coupling coefficient.

In this thesis we derive a formula, depending on the mechanical parameters, the coupling coefficient and the flow parameters, for choosing the optimal stabilization parameter for the fixed-stress splitting scheme where the values $\frac{L_{phys}}{2}$ and L_{phys} are obtained as limit situations. We prove first that the fixed-stress splitting scheme converges linearly and then derive a theoretical optimal parameter by minimizing the rate of convergence. The proof techniques in [13] are improved to reach the new results. For this we require the discretization to be inf-sup stable which effectively allows us to control errors in the pressure by those in the stress. A consequence of our theoretical result is that the fixed-stress splitting scheme also converges in the limit case of low-compressible fluids and low-permeable porous media. Finally, we perform numerical computations to test the optimized parameter. In Section 3.5 we find that the numerical results are confirming the theory. In particular, we remark the connection between inf-sup stability and the performance of the fixed-stress splitting scheme: a not inf-sup stable discretization leads to non-monotonic behavior of the splitting scheme with respect to the problems parameters (e.g. the permeability).

Outline

Chapter 1 contains introductory expositions to various topics which will be used throughout the thesis. Specifically, Section 1.1 gives an introduction to the basics of iterative schemes, both in the sense of linearization methods and splitting meth-

ods. We give basic definitions of convergence properties and present the Banach fixed-point theorem. We then introduce the schemes that we use in Chapter 2 and give some information on their stability and rate of convergence. Finally, we discuss two ways to solve coupled equations.

Section 1.2 introduces the finite element methods, which are applied to all spatial discretizations in the later analysis. We give a short introduction to Sobolev spaces and prove the Lax-Milgram theorem for existence and uniqueness for variational problems. Then the Galerkin method is defined and at last we present the conforming finite element method which is later used for the numerical tests.

A brief introduction to the basic equations and language of porous media is provided in Section 1.3. We present the energy/pressure relations, the mass balance equation and Darcy's law of flow in porous media. Most importantly we define the Richards equation and the Biot equations which we consider in Chapter 2 and 3, respectively.

We begin the analysis in Chapter 2. Here, we present a spatial discretization, using conforming finite elements, and a temporal discretization using implicit Euler, of Richards' equation. We then analyze both theoretically and numerically the convergence of the L-scheme applied to Richards' equation. Furthermore, a comparative study of the Newton-Raphson method, the modified Picard method, a locally defined L-scheme, the modified L-scheme and the L-scheme is provided.

In Chapter 3 we analyze the fixed-stress splitting scheme applied to the Biot equations. We first present the discretization, conforming finite elements with P1 elements for the flow equation and P2 elements for the mechanics equation as spatial discretization, and implicit Euler for temporal discretization. The fixed-stress splitting scheme is defined and a convergence proof is provided. We derive a formula for how to optimally choose the stabilization parameter of the splitting scheme. Moreover, we discuss the importance of inf-sup stability of the numerical discretization. Finally, we present a numerical study both testing the theory on the optimality of the stabilization constant and the impact of a stable discretization.

Chapter 1

Basic theory

In this chapter we provide an introduction to the theory that is applied in the next chapters, Chapter 2 and 3. We give a general discussion about iterative schemes as sequences, and then provide more details on linearization schemes and L-scheme type methods for solving coupled equations. In the second section we introduce the Galerkin method and give a brief introduction to Sobolev spaces before defining the conforming finite element method. At last we define the basic nomenclature and equations of flow in porous media, and in particular we define Richards' equation and Biot's equations.

1.1 Iterative solvers for non-linear and/or coupled PDEs

An introduction to the theory of iterative schemes for solving non-linear and/or coupled equations is presented. Some specific linearization schemes and an iterative splitting scheme for solving coupled equation are introduced. The theory is from [17, 18].

1.1.1 Convergence

Quickly explained an iterative scheme is a way of approximating a solution to (in this thesis) either a non-linear or coupled equation. One starts by making a guess of what the solution is, then use that guess to compute an approximation to the solution before applying this approximation to compute a better approximation. This process is called iterating. More precisely, given the fixed-point problem; find x such that $F(x) = x$, we define an iterative scheme as the recursive sequence

$$x_{i+1} = F(x_i) \tag{1.1}$$

where x_0 is user defined and often called the initial guess. The iterative process is continued until one finally reached an approximation which is sufficiently close to the real solution. Of course the question arises; how do we know if we are close to a solution if we do not know the solution? This question is easily addressed in complete spaces. We give the definition in normed spaces here, but they are equivalent in metric spaces.

Definition 1 (Convergence)**.** *Let $\{x_n\}$ be a sequence in a normed space, $(X, \|\cdot\|)$. We say that $\{x_n\}$*

- *is a* **Cauchy sequence** *if for every $\varepsilon > 0$ there exists an $N_\varepsilon \in \mathbb{N}$ such that $\|x_n - x_m\| < \varepsilon$ whenever $n, m \geq N_\varepsilon$.*
- **converges to** *$x \in X$ if for every $\varepsilon > 0$ there exists an $N_\varepsilon \in \mathbb{N}$ such that $\|x_n - x\| < \varepsilon$ whenever $n \geq N_\varepsilon$.*

A sequence $\{x_n\} \subset X$ that converges to $x \in X$ is called **convergent** *in X.*

It is trivial that every convergent sequence is a Cauchy sequence. However, there exist Cauchy sequences that do not converge.

Example 1. *A simple example of a Cauchy sequence that does not converge is given here. Consider the rationals, $\mathbb{Q}$, and the sequence $x_n = (1 + \frac{1}{n})^n$. This sequence is clearly a Cauchy sequence in $\mathbb{Q}$, but is known to converge to the irrational Euler constant e, hence it is not convergent in $\mathbb{Q}$.*

Spaces with the property that all Cauchy sequences are convergent sequences are very important in computational mathematics. They are called complete spaces.

Definition 2 (Banach space)**.** *A space, X, is called* **complete** *if every Cauchy sequence converges to an element of X. A complete, normed vector space is called a* **Banach** *space.*

Definition 3 (Euclidean space)**.** *We define the Euclidean n-dimensional space by*

$$\mathbb{R}^n := \{\mathbf{x} = (x_1, x_2, ..., x_n) \mid x_i \in \mathbb{R} \text{ for } i = 1, 2, ..., n\}.$$

- *The p-norm on the n-dimensional euclidean space is defined as*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}.$$

The important special cases includes the 1-norm and the 2-norm.

- *In the limit case $p = \infty$ we have the infinity-norm (also called max-norm);*

$$\|\mathbf{x}\|_\infty = \max_{i=1,2,...n} \{x_i\}.$$

Example 2. *The most basic example of known Banach spaces are the Euclidean spaces with the usual Euclidean 2-norm, $(\mathbb{R}^n, \|\cdot\|_2)$.*

The Banach fixed-point theorem which gives a criterion for when a fixed-point iteration converges is now presented.

Definition 4 (Contraction)**.** *Let f be a function between two normed spaces $f : X \to Y$. The function f is called a contraction if there exists a constant $L \in [0, 1)$ such that $\|f(x) - f(y)\|_Y \leq L\|x - y\|_X$ for all $x, y \in X$.*

Theorem 1.1.1 (Banach Fixed-Point Theorem, [17] Chapter 8)**.** *Let X be a Banach space and let $F : X \to X$ be a contraction with contraction constant L. Then F has a unique fixed-point, $F(x^*) = x^*$. Moreover, the sequence starting at some arbitrary $x_0 \in X$ defined as $x_n = F(x_{n-1})$ converges to x^*. The following inequalities hold true and describe the errors of our approximation:*

- *The a priori estimate:* $\|x^* - x_n\| \leq \dfrac{L^n}{1-L}\|x_1 - x_0\|$.

- *The a posteriori estimate:* $\|x^* - x_n\| \leq \dfrac{L}{1-L}\|x_n - x_{n-1}\|$.

Proof. To prove the convergence of the sequence we exploit that we are in a Banach space. Cauchy convergence follows directly from the inequalities

$$\|x_{k+1} - x_k\| = \|F(x_k) - F(x_{k-1})\| \leq L\|x_k - x_{k-1}\| \leq ... \leq L^k\|x_1 - x_0\|.$$

This implies that the sequence is convergent since $L < 1$. Suppose that F has two fixed-points, x^* and x^{**}. The inequality

$$\|x^* - x^{**}\| = \|F(x^*) - F(x^{**})\| \leq L\|x^* - x^{**}\|$$

proves that $x^* = x^{**}$ since $L < 1$. The error-inequalities are both proved in a similar manner through the inequality:

$$\|x^* - x_n\| \leq \|x^* - x_{n+1}\| + \|x_{n+1} - x_n\| \leq L\|x^* - x_n\| + L\|x_n - x_{n-1}\|.$$

□

This is a fundamental theorem, which gives the opportunity to check easily whether an iterative scheme converges. We simply find the function, and check if it is a contraction.

The Banach fixed-point theorem also gives a **stopping criterion** for the iterative scheme (1.1). The a posteriori estimate implies that if $\|x_n - x_{n-1}\|$ is small enough, then $\|x_n - x^*\|$ is also small. In other words, x_n is close to the solution x^*. We stop iterating when the norm of the difference of two consecutive iterations is smaller than some user defined tolerance, ϵ_a, called the absolute tolerance.

- Absolute stopping criterion: $\|x_n - x_{n-1}\| < \epsilon_a$.

On the other hand there are cases where this way of approximating the solution is not the most beneficial, e.g. if the norm of x^*, $\|x^*\|$, is small we would need a tolerance that is correspondingly small to know that we are close to the solution relative to the magnitude of its norm. This gives rise to the definition of a relative stopping criterion where we stop the iteration when $\|x_n - x_{n-1}\| < \epsilon_r \|x_n\|$, for some predetermined relative tolerance, ϵ_r.

- Relative stopping criterion: $\|x_n - x_{n-1}\| < \epsilon_r \|x_n\|$.

A usual approach is to have a combination of the absolute and the relative tolerance as the stopping criterion. We then stop the iteration when $\|x_n - x_{n-1}\| < \epsilon_a + \epsilon_r \|x_n\|$.

- Combined stopping criterion: $\|x_n - x_{n-1}\| < \epsilon_a + \epsilon_r \|x_n\|$.

Remark 1. *Notice that the mean value theorem is a good tool to check whether a real valued function is a contraction or not. It states that for any differentiable function,*

$$f : \mathbb{R} \to \mathbb{R}$$

we have the equality

$$|f(a) - f(b)| = |f'(s)||a - b|$$

for some $s \in [a, b]$. This can be done more general, specifically for vector valued functions and functions of several real or complex variables. If now the derivative of the function f is bounded in absolute value (norm for vectorial functions) by a constant smaller than 1, then the function is a contraction.

It is clear that initial guesses close to the solution make the scheme (1.1) converge in fewer iterations. In fact many methods do not converge at all if the initial guess is too far from the solution, and not all schemes have a contraction property. This gives rise to the notion of a locally convergent method.

Definition 5 (Local convergence)**.** *We say that an iterative scheme* (1.1) *converges locally to x^* if there exists a neighborhood, U, of x^* such that for all initial guesses $x_0 \in U$ it converges to x^*. If U is the entire space then we say that the scheme is globally convergent.*

The last thing we need to define is the notion of how "fast" a method is. This is a concept called the rate of convergence.

Definition 6 (Order of convergence)**.** *Let $\{x_n\}$ be the sequence that arises from an iterative scheme, and suppose that it converges to x^*. Assume that the inequality*

$$\lim_{n\to\infty} \frac{\|x_n - x^*\|}{\|x_{n-1} - x^*\|^k} \leq \mu$$

holds. We say that we have order of convergence k with the special cases

- *linear convergence if $\mu < 1$ and $k = 1$,*
- *sub-linear convergence if $\mu < 1$ and $k \in (0, 1)$,*
- *super-linear convergence if $\mu = 0$ and $k = 1$,*
- *quadratic convergence if $k = 2$.*

We prove now a theorem which helps in determining the order of convergence for different methods for problems in one variable.

Theorem 1.1.2 ([17] Chapter 8)**.** *Let $U \in \mathbb{R}$ be open and $F : U \to U$ be p times continuously differentiable with fixed-point x^*.*

- *For $p = 1$, if $F'(x^*) \neq 0$ and $|F'(x^*)| < 1$*
- *For $p > 1$, if $F'(x^*) = F''(x^*) = ... = F^{(p-1)}(x^*) = 0$ and $F^{(p)}(x^*) \neq 0$*

then the iteration defined by $x_n = F(x_{n-1})$ converges locally to x^* with order of convergence p.

Proof. Consider the Taylor expansion of F around x^*,

$$F(x) = F(x^*) + \frac{F^{(p)}(\xi)}{p!}(x - x^*)^p$$

for all $x \in U$ and some $\xi \in (x, x^*)$.

- For $p = 1$, there exists, from the continuity of F', some interval, I, around x^* such that $|F'(x)| < 1$ for all $x \in I$. Because x^* is a fixed-point of F it follows from the Taylor expansion that we have

$$|x_k - x^*| = |F(x_{k-1}) - F(x^*)| \leq C|x_{k-1} - x^*|,$$

  where $C < 1$. This implies linear convergence.

- For $p > 1$ a similar argument follows from the Taylor expansion,

$$|x_k - x^*| = |F(x_{k-1}) - F(x^*)| \leq C|x_{k-1} - x^*|^p.$$

  This implies order of convergence p.

□

Remark 2. *A way to calculate the order of convergence one experiences numerically is described here. Assume that the scheme has order of convergence k. Then there exists some $\mu \in \mathbb{R}^+$ such that for all $n \geq N$ we have*

$$\|x_n - x^*\| = \mu \|x_{n-1} - x^*\|^k. \tag{1.2}$$

Then also for the previous iterate we have

$$\|x_{n-1} - x^*\| = \mu \|x_{n-2} - x^*\|^k. \tag{1.3}$$

Dividing (1.2) *by* (1.3) *and applying the logarithm to both sides yields the equation*

$$log\left(\frac{\|x_n - x^*\|}{\|x_{n-1} - x^*\|}\right) = k \cdot log\left(\frac{\|x_{n-1} - x^*\|}{\|x_{n-2} - x^*\|}\right).$$

Solving for k gives the very convenient expression

$$k = \frac{log\left(\frac{\|x_n - x^*\|}{\|x_{n-1} - x^*\|}\right)}{log\left(\frac{\|x_{n-1} - x^*\|}{\|x_{n-2} - x^*\|}\right)} \tag{1.4}$$

which applies for all $n \geq 3$.
In practice we rarely know the solution x^* beforehand. An alternative is to precompute a reasonably accurate approximation and use this as x^* when calculating the order.

We will split the remainder of this section on iterative solvers into two subsections. The first one concerns non-linear equations, and the second concerns coupled problems.

1.1.2 Non-linear equations

Approximating solutions to non-linear problems has a long history, with possibly the most famous method being due to Newton. However, before going into details about the specific methods we stress that there are several ways of stating a non-linear problem. The first is to pose the problem as finding a root of a function, $f(x) = 0$. The second is to find a fixed-point, $f(x) = x$. Both equations are in practice equivalent and any non-linear problem can be stated in either way.

The Newton-Raphson method

The Newton-Raphson method is the only method presented here of quadratic order of convergence. However, there are some properties of this method that are not good. It is only locally convergent, and the method requires the computation of derivatives, which can be a costly process. Suppose that we want to solve the equation, $\boldsymbol{f}(\mathbf{x}) = \mathbf{0}$ for some $\boldsymbol{f} : \mathbb{R}^k \to \mathbb{R}^k$. Then the Newton-Raphson method reads

$$\mathbf{x}_n = \mathbf{x}_{n-1} - \mathbf{D}_{\boldsymbol{f}}^{-1}(\mathbf{x}_{n-1})\boldsymbol{f}(\mathbf{x}_{n-1}). \tag{1.5}$$

Here, $\mathbf{D}_{\boldsymbol{f}}$ is the Jacobian of $\boldsymbol{f}$, and the method is only well-posed if $\boldsymbol{f}$ is differentiable and $\mathbf{D}_{\boldsymbol{f}}$ is invertible in a neighborhood of $\mathbf{x}^*$. It is easy to see that the formula corresponds exactly to linearizing $\boldsymbol{f}$ about $(\mathbf{x}_{n-1}, \boldsymbol{f}(\mathbf{x}_{n-1}))$ and denoting by $\mathbf{x}_n$ the root of this linearization.

We remark that the most computationally efficient way to implement the Newton-Raphson method (1.5) will be through solving the linear system

$$\mathbf{D}_{\boldsymbol{f}}(\mathbf{x}_{n-1})\boldsymbol{\delta}_n = -\boldsymbol{f}(\mathbf{x}_{n-1})$$

with respect to the increment, $\boldsymbol{\delta}_n$, and then defining $\mathbf{x}_n := \boldsymbol{\delta}_n + \mathbf{x}_{n-1}$. It follows from Theorem 1.1.2 that the Newton-Raphson method is of quadratic order of convergence for functions in $C^3(\mathbb{R})$.

Definition 7 (Space of continuous and differentiable functions)**.** *Let Ω be an open and connected subset of $\mathbb{R}^n$. We define the space of continuous function on Ω as*

$$C(\Omega) := \{f : \Omega \to \mathbb{R} \mid f \quad \textit{is continuous}\}.$$

We can further define the space of k times differentiable functions on Ω as

$$C^k(\Omega) := \{f : \Omega \to \mathbb{R} \mid f \quad \textit{has continuous derivatives of order } k\}.$$

Corollary 1.1.3. *Let $f \in C^3(\mathbb{R})$ and suppose that $f(x^*) = 0$, $f'(x^*) \neq 0$. The Newton-Raphson method converges locally quadratically to x^*.*

Proof. Define $F(x) := x - \frac{f(x)}{f'(x)}$ as the fixed-point function corresponding to the Newton-Raphson method. By Theorem 1.1.2 it suffices to show that $F'(x^*) = 0$, and that F'' exists. There exists, from the continuity of f', an interval, I, around x^* such that $f'(x) \neq 0$ for all $x \in I$. It follows then by the calculation

$$F'(x^*) = 1 - \frac{f'(x^*)f'(x^*) - f(x^*)f''(x^*)}{f'(x^*)^2} = \frac{f(x^*)f''(x^*)}{f'(x^*)}$$

that $F'(x^*) = 0$, and because $f \in C^3(\mathbb{R})$ we know that F'' exists and is continuous. □

Remark 3. *Suppose that we want to solve the problem $f(x) = 0$, where $f(x) = g(x)+h(x)$. An alternative way to apply a Newton-Raphson-like method here could be to only include the derivative of either g or h, e.g.*

$$g'(x_{n-1})\delta_n = -g(x_{n-1}) - h(x_{n-1}).$$

This might be an advantage in some cases if the computational cost of computing the derivative of a part of the terms is costly, if $h \notin C^1$, or in the multidimensional case it might be of an advantage for the symmetry of the problem. However, due to an inexact derivative the the quadratic convergence is lost. One prominent example is the modified Picard method as a linearization of the Richards equation. For this the non-linear permeability is not linearized but the saturation is, see Chapter 2.

The L-scheme

The main scheme that we analyze in this thesis is the L-scheme [5, 6]. It is a quasi-Newton method, with the benefit that it requires no computation of derivatives. For some problems the L-scheme is also globally convergent. The drawback is that the rate of convergence is only linear.

Suppose that we want to solve the equation $\boldsymbol{f}(\mathbf{x}) = \mathbf{0}$. Then the most naive quasi-Newton iterative scheme would be given by the formula

$$\boldsymbol{\delta}_n = \mathbf{x}_n - \mathbf{x}_{n-1} = -\boldsymbol{f}(\mathbf{x}_{n-1})$$

which corresponds to setting $\mathbf{D}_{\boldsymbol{f}}(\mathbf{x}_{n-1}) = \mathbf{I}$. The L-scheme will be a relaxation of this method and reads

$$L\boldsymbol{\delta}_n = -\boldsymbol{f}(\mathbf{x}_{n-1})$$

where L is a parameter to be chosen. Chapter 2 considers the optimal way to choose this parameter L for Richards' equation.

Similar to the Newton-Raphson we can apply Theorem 1.1.2 to show that the L-scheme converges linearly for real functions in one variable.

Corollary 1.1.4. *Suppose that $f \in C^1(\mathbb{R})$ has a unique root, x^*, and bounded, positive derivatives. If L is chosen greater than*

$$L_f = \sup_{x \in \mathbb{R}} \{f'(x)\}$$

the L-scheme converges linearly to x^*.

Proof. Define $F(x) := x - \frac{f(x)}{L}$ as the fixed-point function corresponding to the L-scheme. Now by Theorem 1.1.2 we have that since

$$0 < F'(x) = 1 - \frac{f'(x)}{L} < 1$$

the L-scheme converges linearly to x^*. □

Modified L-scheme

Another approach to solve non-linear equations is to use what we call the modified L-scheme, due to [7]. The idea here is to have a method that will be of faster convergence than the L-scheme, but globally convergent in contrast to the Newton-Raphson. Suppose we are solving the problem $g(x) + h(x) = 0$, where $g'(x) \geq 0$. We then define the modified L-scheme as

$$M(x_{n-1})\delta_n = -g(x_{n-1}) - h(x_{n-1})$$

where

$$M(x_n) = max\{[g'(x_n) + m], 2m\}.$$

The constant m is problem dependent and will be specified where it is applied. Recognize that if $m = 0$ we are in a Newton-Raphson-type scheme. If however $m \geq max\{g'(x)\}$ the method is equivalent to the L-scheme. By this reasoning we see that m should be chosen smaller than $max\{g'(x)\}$ to see the full potential of the scheme.

1.1.3 L-scheme for coupled problems

Here we introduce two L-scheme-type methods for coupled problems, a monolithic scheme, and a splitting scheme. They are methods for solving coupled equations where the coupled terms appear in non-linearities. Suppose that we want to solve the system of equations

$$\begin{cases} \boldsymbol{F}(\mathbf{x}) + \boldsymbol{G}(\mathbf{y}) = \mathbf{0} \\ \boldsymbol{H}(\mathbf{x}) + \boldsymbol{K}(\mathbf{y}) = \mathbf{0}, \end{cases} \tag{1.6}$$

where

$$\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{K} : \mathbb{R}^n \to \mathbb{R}^n.$$

Let $L_1 > 0$ and $L_2 > 0$ be two positive constants. The iteration step in the monolithic scheme is defined by the following; given $(\mathbf{x}_{n-1}, \mathbf{y}_{n-1})$ solve

$$\begin{cases} \boldsymbol{F}(\mathbf{x}_{n-1}) + L_1(\mathbf{x}_n - \mathbf{x}_{n-1}) + \boldsymbol{G}(\mathbf{y}_{n-1}) = \mathbf{0} \\ \boldsymbol{H}(\mathbf{x}_{n-1}) + \boldsymbol{K}(\mathbf{y}_{n-1}) + L_2(\mathbf{y}_n - \mathbf{y}_{n-1}) = \mathbf{0}. \end{cases} \tag{1.7}$$

We start the iteration with some initial guess $(\mathbf{x}_0, \mathbf{y}_0)$ and then iterate until some stopping criteria is reached.

The splitting scheme is defined in a similar way. Given $(\mathbf{x}_{n-1}, \mathbf{y}_{n-1})$ solve

$$\begin{cases} (i)\ \boldsymbol{F}(\mathbf{x}_{n-1}) + L_1(\mathbf{x}_n - \mathbf{x}_{n-1}) + \boldsymbol{G}(\mathbf{y}_{n-1}) = \mathbf{0} \\ (ii)\ \boldsymbol{H}(\mathbf{x}_n) + \boldsymbol{K}(\mathbf{y}_{n-1}) + L_2(\mathbf{y}_n - \mathbf{y}_{n-1}) = \mathbf{0}. \end{cases} \tag{1.8}$$

Algorithmically this scheme is different to the monolithic one. In the splitting scheme (1.8) we start with an initial guess $(\mathbf{x}_0, \mathbf{y}_0)$, and solve first equation (1.8)(i) for $\mathbf{x}_1$, then solve (1.8)(ii) for $\mathbf{y}_1$ using $\mathbf{x}_1$ in $\boldsymbol{H}$ that we already computed. We continue this process of iterating between the two equations, until some stopping criteria is satisfied.

Both of these schemes are closely related to the L-scheme of the previous section, Section 1.1.2. The two parameters, L_1 and L_2 should be chosen with respect to the non-linearities and in some cases it might be okay to have spatially dependent parameters, and not only constants. Newton-Raphson-like choices might then give a higher order of converge, e.g. $L_1(\mathbf{x}_{n-1}) = \mathbf{D}_{\boldsymbol{F}}(\mathbf{x}_{n-1})$ and $L_2(\mathbf{y}_{n-1}) = \mathbf{D}_{\boldsymbol{K}}(\mathbf{y}_{n-1})$.

In some problems $\boldsymbol{F}$ and/or $\boldsymbol{K}$ are linear terms. Then one simply evaluates them in $\mathbf{x}_n$ and/or $\mathbf{y}_n$ instead of $\mathbf{x}_{n-1}$ and/or $\mathbf{y}_{n-1}$ to enhance the performance of the scheme.

Remark 4 (Convergence)**.** *To check whether such a method converges linearly we define the error ratios*

$$\alpha_x = \lim_{n\to\infty} \frac{\|\mathbf{e}_x^n\|}{\|\mathbf{e}_x^{n-1}\|} \quad \textit{and} \quad \alpha_y = \lim_{n\to\infty} \frac{\|\mathbf{e}_y^n\|}{\|\mathbf{e}_y^{n-1}\|},$$

where $\mathbf{e}_x^n = \mathbf{x}^n - \mathbf{x}^{n-1}$ *and* $\mathbf{e}_y^n = \mathbf{y}^n - \mathbf{y}^{n-1}$ *and conclude through the Banach fixed-point theorem that the method converges if* $\alpha_x < 1$ *and* $\alpha_y < 1$ *and the method is a contraction.*

1.2 The finite element method – FEM

Many physical problems require partial differential equations (PDEs) to be solved. As these are seldom suitable to solve analytically, one needs numerical approximations. There are several ways to approximate solutions to PDEs, e.g. finite differences, finite volumes and the finite element method. In this section we consider the finite element method, or FEM, which is one of the most popular techniques.

1.2.1 Variational problems

The finite element approach for solving PDEs starts with rewriting our PDE to a variational or weak formulation.

Consider the Poisson equation

$$\begin{cases} -\Delta u = f \ in \ \Omega \\ u = 0 \ on \ \partial\Omega, \end{cases} \tag{1.9}$$

where $\Delta = \nabla^2$ is the Laplace operator, Ω is a connected and bounded domain in $\mathbb{R}^n$, $\partial\Omega$ is the boundary of Ω and $f \in C(\Omega)$ (see Definition 7). Multiply the differential equation by a test function from the test space

$$V = \{v \in C^1(\Omega) : v = 0 \ on \ \partial\Omega\} \tag{1.10}$$

and integrate over the domain and the problem becomes; find u such that

$$-\int_\Omega v\Delta u dx = \int_\Omega v f dx$$

for all $v \in V$. Gauss' theorem (integration by parts) reduces the problem to; find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v dx - \int_{\partial\Omega} v\nabla_{\mathbf{n}} u d\sigma = \int_\Omega f v dx$$

for all $v \in V$, where $\mathbf{n}$ is the outwards pointing normal vector of Ω. Since $v = 0$ on $\partial\Omega$ the problem is to find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx \tag{1.11}$$

for all $v \in V$. It is trivial that if u solves (1.9) then u also solves (1.11). However, if u solves (1.11) it might only be weakly differentiable and needs certainly not be two times differentiable. Hence, a function solving (1.11) might not solve (1.9). This is what we call a variational formulation of the problem.

- If the boundary conditions are not homogeneous, but still continuous and of Dirichlet type, $u = u_D$ on $\partial\Omega$, the boundary term does not vanish like in (1.11). The common approach then is to rewrite the problem

$$\begin{cases} -\Delta u = f \ in \ \Omega \\ u = u_D \ on \ \partial\Omega \end{cases}$$

as

$$\begin{cases} -\Delta w = f \ in \ \Omega \\ w = 0 \ on \ \partial\Omega \end{cases}$$

where $w = u - \tilde{u}_D$ and $\tilde{u}_D$ is a continuous extension of u_D, with $\Delta\tilde{u}_D = 0$, to the entire domain Ω. Otherwise, we can include the Dirichlet boundary conditions to the solution space and solve as before.

- If we on the other hand have Neumann boundary conditions

$$\begin{cases} -\Delta u = f \ in \ \Omega \\ \nabla_{\mathbf{n}} u = g \quad on \ \Gamma_N \\ u = 0 \quad\quad on \ \Gamma_D, \end{cases}$$

where $\partial\Omega = \Gamma_D \cup \Gamma_N$, the variational formulation becomes find $u \in V$ such that

$$\int_\Omega \nabla u \nabla v dx = \int_\Omega f v dx + \int_{\Gamma_N} g v d\sigma$$

As the integral defines an inner product, $\langle \cdot, \cdot \rangle$, on the space of functions $C^1(\Omega)$ a common way to write problem (1.11) is; find $u \in V$ such that

$$\langle \nabla u, \nabla v \rangle = \langle f, v \rangle \tag{1.12}$$

for all $v \in V$. Another way of expressing problem (1.12) is by; find $u \in V$ such that

$$a(u, v) = L(v) \tag{1.13}$$

for all $v \in V$ where $a(u, v) = \langle \nabla u, \nabla v \rangle$ and $L(v) = \langle f, v \rangle$.

1.2.2 Sobolev spaces

Before we define the finite element approximation to the solution of (1.12) we need a better understanding of the solution space and test space as they will not end up being V from (1.10). There is a lot of literature on this matter which originates in the subject of functional analysis, or more specifically the analysis of Sobolev spaces. We will not go into full detail on Sobolev spaces, but refer instead to [19] for a full overview. Regardless, we state the essentials here.

Definition 8 (Hilbert space)**.** *A Hilbert space, X, is a complete normed vector space, where the norm, $\|\cdot\|$, has an associated inner product, $\langle\cdot,\cdot\rangle$, i.e. for any $x \in X$ we have $\|x\|^2 = \langle x, x\rangle$.*

Definition 9 (L^p spaces)**.** *The L^p-spaces (often called Lebesgue spaces) are defined as*

$$L^p(\Omega) := \{f \in C(\Omega) : \|f\|_p < \infty\}$$

where $\|f\|_p = \left(\int_\Omega |f|^p dx\right)^{\frac{1}{p}}$ and $p \in [1, \infty)$.

Theorem 1.2.1 (Riesz-Fisher theorem, [18] Chapter 8)**.** *The $L^p(\Omega)$ spaces are all Banach spaces. $L^2(\Omega)$ is also a Hilbert space.*

From now, every time a function norm is used with no subscript it is understood to be the L^2-norm.

Definition 10 (Sobolev spaces)**.** *We define the Sobolev spaces by*

$$W^{m,p}(\Omega) = \{f \in L^p(\Omega) : f \text{ has weak derivatives of order up to } m, \|f\|_{m,p} < \infty\}$$

where $\|f\|_{m,p} = \left(\sum_{|\alpha|\leq m} \int_\Omega |\partial_\alpha f|^p dx\right)^{\frac{1}{p}}$. The spaces where $p = 2$ are of special importance and are denoted by $H^m(\Omega) := W^{m,2}(\Omega)$.

Proposition 1.2.2 ([18] Chapter 8)**.** *The spaces $H^m(\Omega)$ are all Hilbert spaces through the inner product $\sum_{|\alpha|\leq m}\langle\partial_\alpha f, \partial_\alpha g\rangle$.*

If the boundary of the domain, $\partial\Omega$, is "good enough", e.g. Lipschitz continuous, we can define a trace operator

$$T : H^m(\Omega) \to H^{m-\frac{1}{2}}(\partial\Omega)$$

where $f \in H^m(\Omega)$ is sent to $Tf = f_{|\partial\Omega}$. This is of particular importance in the case where $m = 1$ so that we can define the space

$$H_0^1(\Omega) = \{f \in H^1(\Omega) \ : \ f|_{\partial\Omega} = 0\}.$$

Typically this, $H_0^1(\Omega)$, is the natural choice of test space V.

1.2.3 Existence and uniqueness of the solution to variational problems

Before searching for a solution to a PDE one should know that it exists and is unique. Here, two theorems for this purpose are presented, but first we need some important definitions.

Definition 11. *Let $a(\cdot,\cdot)$ be a bilinear form on some normed vector space V,*

- *we say that $a(\cdot,\cdot)$ is* **bounded** *(or continuous) w.r.t. $\|\cdot\|_V$ if there exists a constant $M \in \mathbb{R}$ such that $a(u,v) \leq M\|u\|_V\|v\|_V$ for all $u, v \in V$,*
- *we say that $a(\cdot,\cdot)$ is* **coercive** *w.r.t. $\|\cdot\|_V$ if there exists a constant $\alpha > 0$ such that $a(u,u) \geq \alpha\|u\|_V^2$ for all $u \in V$.*

Proposition 1.2.3 ([18] Chapter 1)**.** *Let Y be a closed subspace of the Hilbert space X. Then $X = Y \oplus Y^\perp$, where $Y^\perp$ is the orthogonal complement of Y, i.e.*

$$Y^\perp = \{x \in X \mid \langle x, y\rangle = 0 \quad \text{for all} \quad y \in Y\}$$

and $\oplus$ denotes the direct sum.

Proof. We notice at first that $Y^\perp$ is a subspace of X. Let $x, y \in Y^\perp$, then $\langle \alpha_1 x + \alpha_2 y, z\rangle = 0$ for all $z \in Y$, hence $\alpha_1 x + \alpha_2 y \in Y^\perp$. Also, the only element in $Y \cap Y^\perp$ is the null element, which implies $Y \oplus Y^\perp \subset X$. Now, to see that X is a subspace of $Y \oplus Y^\perp$ we let x be an arbitrary element in X. It follows from the fact that Y is closed that there exists an element $y \in Y$ such that y is the closest element in Y to x, i.e. $\|x - y\| \leq \|x - v\|$ for all $v \in Y$. Then we know from functional analysis that $x - y \perp Y$, hence $x - y \in Y^\perp$, which means that $x = y + (x - y) \in Y \oplus Y^\perp$. We have that $X = Y \oplus Y^\perp$. □

We now state the Riesz-Frechet representation theorem and prove existence and uniqueness for symmetric variational problems of the form (1.13).

Theorem 1.2.4 (Riesz-Frechet representation theorem, [18] Chapter 2)**.** *Let H be a Hilbert space and f be a bounded linear functional in the dual space, H'. Then there exists a unique element $g \in H$ such that $f(h) = \langle h, g\rangle$ for all $h \in H$. Moreover, $\|f\|_{H'} = \|g\|_H$.*

Remark 5. *Notice that the opposite is trivial; every element of H defines a continuous linear functional through the inner product. This means that we have an isomorphism between the spaces H and H'.*

Proof of Theorem 1.2.4. Let f be an element of H', and define the annihilator of f as $Y = \{x \in X : f(x) = 0\}$. If $Y = X$ then $f(x) = 0$ for all elements of X and $f(x) = \langle x, 0\rangle$. If on the other hand $Y \neq X$ we notice that Y is a closed subspace of X and define $Y^\perp$. By Proposition 1.2.3 we have that $X = Y \oplus Y^\perp$. Take an element $u \in Y^\perp \backslash \{0\}$, and define $g = u/\|u\|^2 \in Y^\perp$. Because for all $x \in X$ we have $f(x)u \in Y^\perp$, we know that $x - f(x)u \in Y$. Then we have

$$\langle x, g\rangle = \langle x - f(x)u, g\rangle + \langle f(x)u, g\rangle = f(x)\langle u, g\rangle = f(x).$$

Moreover,

$$\|f\|_{H'} = \sup_{0\neq x\in H} \frac{|f(x)|}{\|x\|_H} = \sup_{0\neq x\in H} \frac{|\langle x, g\rangle|}{\|x\|_H} \leq \frac{\|x\|_H \|g\|_H}{\|x\|_H} = \|g\|_H$$

and by inserting $x = g$

$$\|f\|_{H'} = \sup_{0\neq x\in H} \frac{|\langle x, g\rangle|}{\|x\|_H} \geq \frac{\|g\|_H^2}{\|g\|_H} = \|g\|_H$$

□

Now existence and uniqueness of the solution to the symmetric problem follows.

Lemma 1.2.5. *If $a(\cdot,\cdot)$ is symmetric, bounded and coercive on a Hilbert space V then the problem* (1.13) *has a unique solution, u, in V for any given bounded linear functional $L \in V'$. Moreover, $\|u\| \leq \frac{1}{\sqrt{\alpha}}\|L\|_{V'}$, where α is the coercivity constant of $a(\cdot,\cdot)$.*

Proof. Because $a(\cdot,\cdot)$ is coercive and bilinear it follows that $a(v,v) = 0$ if and only if $v = 0$. Together with the symmetry this implies that $a(\cdot,\cdot)$ actually defines an inner product on V. Define the norm $\|\cdot\|_a := \sqrt{a(\cdot,\cdot)}$. From coercivity and boundedness of $a(\cdot,\cdot)$ we get the equivalence of norms

$$\sqrt{\alpha}\|\cdot\| \leq \|\cdot\|_a \leq \sqrt{M}\|\cdot\|.$$

This implies that $(V, a(\cdot,\cdot))$ is in fact a Hilbert space itself. Now consider the bounded linear functional $L \in V'$. By the Riesz-Frechet representation theorem there exists a unique $u \in V$ such that $a(u,v) = L(v)$ for all $v \in V$. Moreover, $\|u\| \leq \frac{1}{\sqrt{\alpha}}\|u\|_a = \frac{1}{\sqrt{\alpha}}\|L\|_{V'}$. □

We state now the Lax-Milgram theorem which gives both existence and uniqueness of the solution to non-symmetric variational problems, (1.13).

Theorem 1.2.6 (Lax-Milgram, [17] Chapter 3)**.** *Let V be a Hilbert space, $a(\cdot,\cdot)$ be a bounded and coercive bilinear form and $L(\cdot)$ be a continuous linear functional in V'. Then there exists a unique $u \in V$ such that $a(u,v) = L(v)$ for all $v \in V$. Moreover $\|u\| \leq \frac{1}{\alpha}\|L\|_{V'}$, where α is the coercivity constant of $a(\cdot,\cdot)$.*

Proof. Because $a(u,v)$ is a bounded functional the Riesz-Frechet representation theorem defines an operator, $A : V \to V$, by $a(u,v) = \langle Au, v\rangle$. Linearity of A follows from linearity of $a(\cdot, v)$. We also get boundedness from the Riesz-Frechet theorem;

$$\|Au\|_V = \|a(u,\cdot)\|_{V'} = \sup_{0\neq v\in V} \frac{|a(u,v)|}{\|v\|_V} \leq M\|u\|_V.$$

Additionally, the Riesz-Frechet theorem defines a continuous linear function $B : V' \to V$ defined by $L(v) = \langle BL, v\rangle$. This reduces the problem of finding a solution to the equation $a(u, v) = L(v)$, to finding a solution to $Au = BL$. From the coercivity of $a(\cdot, \cdot)$ we get the inequality

$$\alpha\|v\|^2 \leq a(v, v) = \langle Av, v\rangle \leq \|Av\|\|v\|.$$

This implies that

$$\|Av\| \geq \alpha\|v\| \tag{1.14}$$

and it follows that A is injective. In other words, if we can find a solution, it will certainly be unique.

Now, if A is also surjective we have the existence of a solution to $a(u, v) = L(v)$. First we prove that the image of A, $im(A)$, is closed and that the orthogonal complement is the null set. From there Proposition 1.2.3 gives surjectivity. Let $\{Av_n\} \subset im(A)$ be a sequence that converges to $w \in cl(im(A))$, where $cl(im(A))$ is the closure of $im(A)$. From (1.14) we have that

$$\|Av_n - Av_m\| \geq \alpha\|v_n - v_m\|$$

which implies that $\{v_n\}$ is a Cauchy sequence in V. By the completeness of V, $\{v_n\}$ converges to some $v \in V$. Because A is continuous we know that Av_n converges to Av which means that $Av = w$, and that $w \in im(A)$. Hence, $im(A)$ is closed. Now let $y \in im(A)^\perp$. Coercivity gives the inequality

$$\|y\|^2 \leq \frac{1}{\alpha}a(y, y) = \frac{1}{\alpha}\langle Ay, y\rangle = 0,$$

which shows that $im(A)^\perp = \{0\}$. All in all, $im(A) = V$.

Finally we see that if u solves $Au = BL$ then

$$\|u\|^2 \leq \frac{1}{\alpha}a(u, u) = \frac{1}{\alpha}\langle Au, u\rangle = \frac{1}{\alpha}\langle BL, u\rangle \leq \frac{1}{\alpha}\|BL\|\|u\| = \frac{1}{\alpha}\|L\|_{V'}\|u\|.$$

This proves the Lax-Milgram theorem. □

Remark 6. *Notice that closed subspaces of Hilbert spaces are in their own right Hilbert spaces by restricting the inner product to functions of the subspaces. Therefore, since all finite dimensional subspaces of normed spaces are closed it is in fact enough to prove existence and uniqueness for the continuous problem to also get it in the discrete problem.*

We will finally give an example showing that the variational problem arising from Poisson's equation (1.9) has a unique solution in $H_0^1(\Omega)$. For this we will apply the Poincaré inequality.

Theorem 1.2.7 (Poincaré inequality)**.** *For $p \in [1, \infty)$ and a domain $\Omega \subset \mathbb{R}^n$ that is open and bounded there exists a constant $C(\Omega, p)$ such that for every function u in the Sobolev space with zero trace functions $W_0^{1,p}(\Omega) = H_0^1(\Omega)$ we have*

$$\|u\|_{L^p(\Omega)} \leq C(\Omega, p)\|\nabla u\|_{L^p(\Omega)}$$

Example 3 (Existence and uniqueness for Poisson's equation)**.** *We consider again the Poisson equation* (1.9) *in its variational form, $a(u, v) = L(v)$ where*

$$a(u, v) = \langle \nabla u, \nabla v \rangle \text{ and } L(v) = \langle f, v \rangle.$$

By the following arguments and Lemma 1.2.5, $a(u, v) = L(v)$ has a unique solution in the space $V = H_0^1(\Omega)$.

1. *Boundedness of $L(\cdot)$: Cauchy-Schwarz inequality implies that*

$$L(v) = \langle f, v \rangle \leq \|f\|_{H^{-1}(\Omega)}\|v\|_{H^1(\Omega)}.$$

   We see here that as long as f lives in the dual space of $H^1(\Omega)$ denoted by $H^{-1}(\Omega)$ we have boundedness of $L(\cdot)$.

2. *Boundedness of $a(\cdot, \cdot)$: Again due to the Cauchy-Schwarz inequality we have*

$$a(u, v) = \langle \nabla u, \nabla v \rangle \leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}$$

   where the last inequality follows directly from the definition of the $H^1(\Omega)$-norm.

3. *Coercivity of $a(\cdot, \cdot)$: Consider the Poincaré inequality with constant C_Ω, then*

$$\begin{aligned} \|u\|_{H^1(\Omega)}^2 &= \|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \\ &\leq C_\Omega\|\nabla u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \\ &=: \frac{1}{\alpha}\|\nabla u\|_{L^2(\Omega)}^2 = \frac{1}{\alpha}a(u, u) \end{aligned}$$

   proves coercivity of $a(\cdot, \cdot)$.

This proves existence and uniqueness of the variational formulation of Poisson's equation by Lemma 1.2.5.

Similar proofs can be done for other linear problems, however this is not the goal of this thesis, and what we have done here concludes the section of existence and uniqueness of variational formulations of PDEs.

1.2.4 The Galerkin method

Before we properly define the finite element method we introduce the more general Galerkin method for approximating solutions to variational problems. Consider a variational problem

$$a(u, v) = L(v). \tag{1.15}$$

Assuming that equation (1.15) has a solution in $V = H_0^1(\Omega)$ we can start to look for what is called discrete solutions (approximations), in finite dimensional subspaces of V. Let V_h be a finite dimensional subspace of V, where the h is a constant such that in the limit where h goes to 0, V_h goes to V. The discrete variational formulation then reads; find $u_h \in V_h$ such that

$$a(u_h, v_h) = L(v_h) \tag{1.16}$$

for all $v_h \in V_h$. Since V_h now is a finite dimensional space finding the solution explicitly is possible. Let $\{\varphi_i\}_{i=1}^N$ be a basis for V_h. Now we can write $u_h = \sum_{i=1}^N \eta_i \varphi_i$, substitute u_h in (1.16) and test with $v_h = \varphi_j$. This gives N equations

$$\sum_{i=1}^{N} \eta_i a(\varphi_i, \varphi_j) = L(\varphi_j) \tag{1.17}$$

for $j = 1, 2, ..., N$ where the sum can be taken on the outside of $a(\cdot, \cdot)$ because of its bilinearity. Now it is just a matter of calculation to find $a(\varphi_i, \varphi_j)$ and $L(\varphi_j)$, and we can therefore solve the system of N equations with the N unknowns $\{\eta_i\}_{i=1}^N$. Because equation (1.17) holds for all basis functions of V_h it holds for all functions, $v_h \in V_h$, due to the bilinearity of $a(\cdot, \cdot)$, and we have found our discrete solution, $u_h \in V_h$. In matrix form this would be the same as solving the linear equation

$$\mathbf{A}\boldsymbol{\eta} = \mathbf{b}, \tag{1.18}$$

where $\mathbf{A}_{i,j} = a(\varphi_j, \varphi_i)$, $\mathbf{b}_j = L(\varphi_j)$ and $\boldsymbol{\eta}_i = \eta_i$.

As seen in the previous section, Section 1.2.3, the important properties of $a(\cdot, \cdot)$ and $L(\cdot)$ are their linearity (bilinearity for $a(\cdot, \cdot)$), their boundedness and the coercivity of $a(\cdot, \cdot)$, with respect to V.

1.2.5 Finite elements

We will now be more precise on how to choose the finite dimensional subspace, V_h, in a manner that makes it possible to solve the set of equations (1.16) efficiently. There are many options, but only some of the most basic are presented here, for other we refer to [20]. Those are the spaces consisting of piecewise functions from

$\mathcal{P}_n$, where $\mathcal{P}_n$ is the space of polynomials of degree n. In Chapter 2 we are using $\mathcal{P}_1$ and in Chapter 3 we are using both $\mathcal{P}_1$ and $\mathcal{P}_2$. The main reason for choosing this type of functions is so that the system of equations becomes sparse and therefore easier to solve. It is also computationally less expensive to compute and store $\mathbf{A}$.

Suppose that we want to approximate the solution to some equation in the domain $\Omega \subset \mathbb{R}^n$ by piecewise $\mathcal{P}_1$ functions. First of all we need to make a mesh of our domain. This is equivalent to subdividing our domain into polytopes, called elements. Denote by $\mathcal{T}_h$ the set of elements and define the vertices to be the nodes of the domain. Now we can define our finite dimensional subspace of V consisting of piecewise linear functions as

$$V_h = \{v_h \in C(\Omega) : v_{h|T} \in \mathcal{P}_1 \textit{ for all } T \in \mathcal{T}_h\}.$$

Equivalently, one could have defined the space of piecewise quadratic functions or any other set of piecewise $\mathcal{P}_n$ functions. The next step is to find a basis for the vector space, V_h. A convenient basis when talking about piecewise linear functions are the hat functions. Pick a node, x_i, and let $\mathcal{T}_h^i$ be the subset of $\mathcal{T}_h$ that consists of elements with x_i as a vertex. We then define the hat function φ_i, corresponding to x_i, as the piecewise linear function that satisfies

$$\varphi_i(x_j) = \delta_{ij} \tag{1.19}$$

where $\{x_j\}$ is the set of nodes and δ_{ij} is the Kronecker delta, see Figure 1.1. To verify that the set of functions $\{\varphi_i\}$ actually defines a basis for the space V_h, one can easily see that they are linearly independent (only φ_i takes nonzero values at x_i), and then it is just a matter of comparing the dimension of V_h to the number of hat functions. Clearly, they are both equal to the number of nodes, because if we choose a value at each node we have uniquely defined a function that is piecewise linear on each element.

Although this is the basis that we use, it is not the most practical way to define the functions from the point of view of the implementation. The convention in the finite element method is to define the basis functions on a reference element (e.g. a triangle between the points (0,0), (0,1) and (1,0) if working on a triangular mesh) corresponding to its vertices. Then for each element, T, we define a linear transformation to the reference element and thereby inherit its basis functions. Each of these basis functions are defined as $\varphi_i|_T$. If we were considering piecewise $\mathcal{P}_2$ functions we would introduce in each element twice as many nodes (to make up for the dimension of the space) and then define the basis functions in the same manner (except with quadratic instead of linear). This logic holds for any basis of piecewise $\mathcal{P}_n$ functions. Looking back to our system of linear equations (1.18) it becomes apparent why this choice of basis (and space) is so convenient; only a part of the matrix entries, $a(\varphi_j, \varphi_i)$, will be nonzero making the system sparse.

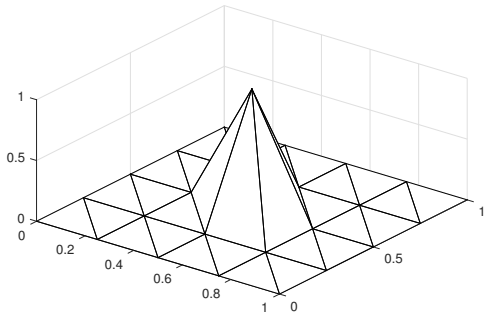


Figure 1.1: Basis function for V_h on triangulated unit square domain

1.2.6 Pseudocode

Here, an outline of the code for solving variational problems, used in the later chapters, is presented.

1. Define the domain and create a mesh. Creating meshes can be challenging if the domain is not very simple (for example the unit square), but there are many software packages that helps in doing this. Here one should structure the elements and nodes so that when running through the elements one knows exactly what the coordinates of the nodes are. For example if the elements are triangles, structure the element matrix as a "number of elements"$\times$3-matrix where each row contains the numbers assigned to each node. Then a matrix containing the information of the numbers of the nodes and their coordinates should be defined.

2. Create the basis functions (see (1.19)) on a reference element, e.g. a triangle through the points $(0,0)$, $(1,0)$ and $(0,1)$.

3. Choose a quadrature on the reference element, and make evaluations for the basis functions.

4. Go through all the elements one by one (e.g. through a for loop):

   - Extract the coordinates of the vertices.
   - Calculate the linear transformation from the physical element to the reference element.
   - Define the local stiffness matrix, $\mathbf{A_{loc}}$ and source vector $\mathbf{b}_{loc}$ by computing $a(\varphi_j, \varphi_i)$ and $L(\varphi_j)$ using the chosen quadrature rule. Take here full advantage of the reference element and linear transformation.

5. Assemble the global stiffness matrix $\mathbf{A}$ and source vector $\mathbf{b}$, using the numbering and matrices created in 1.

6. Go through the matrix and source vector and assign the Dirichlet boundary conditions. A way to do this is to go through all nodes in the boundary where we want to assign Dirichlet boundary conditions. For each of these nodes, x_j, set $\mathbf{A}_{j,k} = 0$ for all $k \neq j$, $\mathbf{A}_{j,j} = 1$ and $\mathbf{b}_j = u_D(x_j)$ where $u_D(x_j)$ is the Dirichlet boundary condition at x_j.

7. Finally, solve the linear system and get the approximated solution.

For the numerical tests in Section 2.3 and 3.5 a standalone MATLAB code for solving Richards' equation and the Biot equations has been implemented using this procedure. The time-dependence and non-linearities/coupling does however change the pseudocode slightly in the sense that another two loops are wrapped around it to iterate and move forward in time.

1.3 Flow in porous media

In this section we give a brief introduction to the basics of flow in porous media, specifically directed towards Richards' equation and Biot's equations that are analyzed in the next two chapters. The theory is from [21] and we refer to [22] for a detailed introduction to flow in porous media.

1.3.1 Porosity and saturation

A porous medium consists of a solid material, the matrix, and void spaces in between. In order to make sensible definitions and discussion regarding the physical properties of the medium we need to define a point in the space, not exactly as a point, but as a volume around the point. This volume is called the representative elementary volume, denoted by REV. It is important that the REV is large enough so that we never enter the situation where it only captures the properties of the void spaces or matrix, but small enough so that it still preserves the local properties of the medium.

An important concept of porous media, is the notion of porosity, i.e. the measure of how porous a medium is. The **porosity**, ϕ, is defined as the volume of void space in the REV divided by the total volume of the REV

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

Another concept is the **saturation** of some fluid the medium. It is defined as the volume of the specific fluid in the REV, divided by the volume of voids in the REV,

$$s_f := \frac{\text{vol(fluid } f \text{ in REV)}}{\text{vol(voids in REV)}}.$$

As a porous medium can contain several fluids (for example water, CO_2, oxygen, brine, etc.) it is important to keep track of the different ones and their specific saturations. In a fully saturated porous medium the sum of the saturations is equal to one,

$$\sum_f s_f = 1. \tag{1.20}$$

Observe that the volume of a fluid, θ_f, is described by the product of the porosity and the saturation of the fluid,

$$\theta_f = \phi s_f.$$

1.3.2 Energy and pressure

A fluid is usually described by its energy. The total energy is the sum of potential and kinetic energy. For flow in porous media we assume to have a slow flow rate and therefore the influence of kinetic energy is neglected. The effects of temperature and dissolved substances in the fluid are also disregarded at this point. Therefore, the energy of the fluid is simply described by its potential energy, which is affected by the pressure and the gravity. The pressure potential is described by pressure times volume, from the equations

$$p = \frac{F}{A} = \frac{F \cdot d}{A \cdot d} = \frac{\text{Work}}{V} = \frac{\text{Energy}}{V},$$

where F denotes force, A area, d distance, V volume and p pressure. This gives an equation for the total potential energy, in terms of hydraulic potential,

$$mgh = E_p + E_g = pV + mgz,$$

where h denotes the hydraulic head, m the mass, g gravity and z height above reference zero (called **datum**). If we divide by mg we get the more convenient expression

$$h = \frac{pV}{mg} + z = \frac{p}{\rho g} + z := \psi + z, \tag{1.21}$$

where ρ is the density and ψ is called the pressure head. Figure 1.2 shows an instructive picture regarding how these quantities can be measured using a Piezometer.

1.3.3 Darcy's law

Darcy's law of flow in porous media [23] gives a relation between the pressure and the flow of the fluid. This is an experimental law, where the flow was measured through a tube between two points. The observations were that the flow is proportional to the difference in hydraulic head in the two points and to the cross-sectional area, A, of the tube and inversely proportional to the length, l, of the tube. The equation then becomes

$$\boldsymbol{q}_1 = -k\frac{A\Delta h}{l}\mathbf{e},$$

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

$$\boldsymbol{q} = \frac{\boldsymbol{q}_1}{A},$$

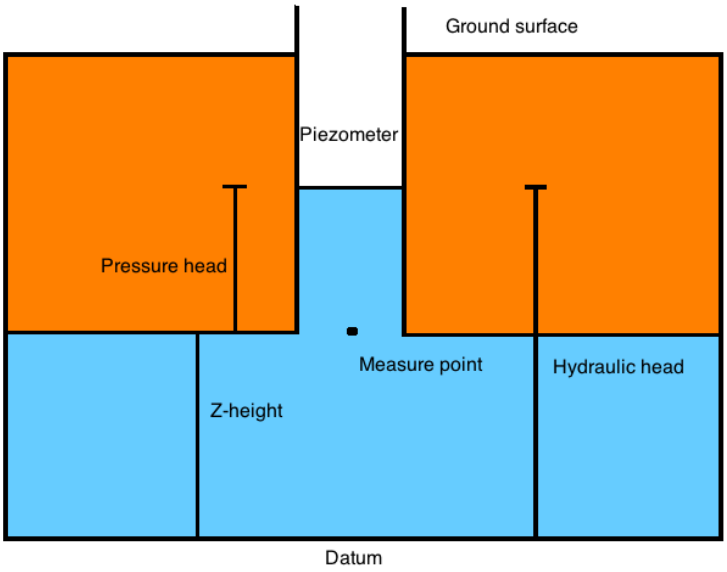


Figure 1.2: Taking measurements with a Piezometer

and taking the limit as l goes to zero we get the differential form of Darcy's law;

$$\boldsymbol{q} = -k\nabla h. \tag{1.22}$$

Substituting for h from equation (1.21) and assuming incompressibility (constant density, ρ) we get the pressure formulation

$$\boldsymbol{q} = -\frac{k}{\rho g}(\nabla p - \rho \mathbf{g}), \tag{1.23}$$

where $\mathbf{g} = -g\nabla z$.

1.3.4 Mass conservation

As the pressure of the fluid is unknown we still cannot describe the flow through a porous medium. Our system is closed by the mass conservation equation. The idea is; the change of mass through an arbitrary volume ω is balanced by the mass that flows through the boundary and the mass that is added to the system through sources or sinks not in the boundary;

$$\int_\omega \frac{\partial m}{\partial t} dx = -\int_{\partial\omega} \boldsymbol{F} \cdot \nu_n dx + \int_\omega f dx \quad \forall\omega,$$

where $\boldsymbol{F}$ is the flux though the boundary and f is the sources and sinks. The minus in front of the boundary flux comes from the notion of an outwards pointing normal

vector ν_n. Gauss' theorem in the boundary flux term gives the general local mass balance equation

$$\frac{\partial m}{\partial t} + \nabla \cdot \boldsymbol{F} = f \tag{1.24}$$

as ω was arbitrary. Interpreting this for flow in porous media we set the mass as the product of density and volume, $m = \rho\theta_f$, and the flux as the flow $\boldsymbol{F} = \boldsymbol{q}\rho$,

$$\frac{\partial \rho\theta_f}{\partial t} + \nabla \cdot (\boldsymbol{q}\rho) = f. \tag{1.25}$$

1.3.5 Two-phase flow

We now have a closed system of equations for single phase flow (flow with one fluid). Suppose then that we have two fluids, the wetting fluid w, and the non-wetting fluid n. We can assume that both equations follow the mass balance and Darcy's law in addition to the saturation equation (1.20). This gives the system of equations

$$\begin{cases} \boldsymbol{q}_\alpha = -\dfrac{k_\alpha}{\rho_\alpha g}(\nabla p_\alpha - \rho_\alpha \mathbf{g}), \\ \dfrac{\partial \rho_\alpha \theta_\alpha}{\partial t} + \nabla \cdot (\boldsymbol{q}_\alpha \rho_\alpha) = f_\alpha, \\ s_w + s_n = 1, \end{cases} \tag{1.26}$$

where $\alpha = \{w, n\}$. The hydraulic conductivity can be written

$$k_\alpha = \frac{\kappa_{r,\alpha}\hat{\kappa}}{\mu_\alpha}$$

where $\hat{\kappa}$ is the permeability, $\kappa_{r,\alpha}$ is the relative permeability (will be assumed to be a function of saturation), and μ_α is the viscosity. If we count the number of unknowns and equations in (1.26) we realize that we have $2d + 4$ unknowns (2d flow unknowns, 2 pressures and 2 saturations) and only $2d + 3$ equations where d is the dimension. The missing equation to close the system is called the capillary pressure equation, giving a relation between the two pressures

$$p_c(s_w) = p_n - p_w,$$

where $p_c(s_w)$ is a given function.

1.3.6 Introduction to Richards' equation

One of many special cases of two-phase flow in porous media is the Richards equation, which models flow when the two fluids are water and air. One assumes that air is not trapped by the water and therefore has constant pressure which

significantly simplifies the equations, in fact to the extent that one often calls it one and a half phase flow. Also the density of water is assumed constant, $\rho_w = 1$, giving the system of equations

$$\begin{cases} \boldsymbol{q}_w = -\dfrac{k_w}{g}(\nabla p_w - \mathbf{g}), \\ \dfrac{\partial \theta_w}{\partial t} + \nabla \cdot \boldsymbol{q}_w = f_w, \\ s_w + s_n = 1, \\ p_c(s_w) = -p_w. \end{cases} \tag{1.27}$$

Remembering that θ_w is the product of saturation and porosity, setting $p = p_w$, and substituting for the flow, we have the Richards equation

$$\partial_t(s_w(p)) - \nabla \cdot (\kappa(s_w(p))(\nabla p - \boldsymbol{g})) = f, \tag{1.28}$$

where we have divided by the porosity which we assume to be independent in time, and

$$\kappa(s_w(p)) := \frac{\kappa_{r,w}(s_w(p))\hat{\kappa}}{\mu_w g \phi} \qquad \text{and} \qquad f = f_w/\phi.$$

Equation (1.28) will be analyzed theoretically and numerically in Chapter 2. There are several parameterizations of $s_w(p)$ and $\kappa(s_w(p))$ to close the system. We mention two of the most famous, the Van Genuchten-Mualem [24] and Brooks-Corey [25]. We consider a case of the Van Genuchten-Mualem in Chapter 2.

1.3.7 Introduction to Biot's equations

In Chapter 3 we consider the quasi-static linear Biot model, the simplest model to describe flow in deformable porous media. It reads, find $(\boldsymbol{u}, p)$ such that

$$-\nabla \cdot (2\mu \boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \nabla \cdot \boldsymbol{u}\mathbf{I}) + \alpha \nabla p = \boldsymbol{f}, \tag{1.29}$$

$$\frac{\partial}{\partial t}\left(\frac{p}{M} + \alpha \nabla \cdot \boldsymbol{u}\right) - \nabla \cdot (\kappa(\nabla p - \rho \mathbf{g})) = S_f, \tag{1.30}$$

where $\boldsymbol{u}$ is the displacement, $\varepsilon(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top)$ is the (linear) strain tensor, μ, λ are the Lamé parameters, α is the Biot-Willis constant, p, ρ are fluids pressure and density, respectively, M is a compressibility constant, $\mathbf{g}$ the gravitational vector and κ is the permeability. The source terms $\boldsymbol{f}$ and S_f represent the density of applied body forces and a forced fluid extraction or injection process.

- Equation (1.29) models the mechanical behavior of the system through linear momentum balance under quasi-static conditions combined with an effective stress formulation,

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f}. \tag{1.31}$$

If we allow only small deformations we can apply the St. Venant Kirchhoff model for the effective stress, determining the poroelastic stress as

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\nabla\cdot\boldsymbol{u}\mathbf{I} - \alpha p\mathbf{I} \tag{1.32}$$

which together with (1.31) gives (1.29).

- Equation (1.30) describes the fluid flow through mass conservation and Darcy's law. Mass conservation is equivalent to volume conservation for incompressible fluids,

$$\partial_t V + \nabla\cdot\boldsymbol{q} = S_f, \tag{1.33}$$

where V is the volume of the fluid and $\boldsymbol{q}$ is the fluid flow. The volume of a fluid in a porous medium is given by the product of the porosity, ϕ, and the saturation, s_w, and since we consider fully saturated flow the volume is simply equal to the porosity. The porosity changes linearly with respect to the volumetric deformation, $\nabla\cdot\boldsymbol{u}$ and the pore pressure p,

$$V = \phi(\boldsymbol{u}, p) = \phi_0 + \alpha\nabla(\boldsymbol{u} - \boldsymbol{u}_0) + \frac{1}{M}(p - p_0), \tag{1.34}$$

where ϕ_0, $\boldsymbol{u}_0$ and p_0 are the initial porosity, displacement and pore pressure, respectively. Insert (1.34) and (1.23) into (1.33) to get (1.30).

Chapter 2

Richards' equation

In this chapter we consider Richards' equation (1.28), see Section 1.3 for the derivation of this equation. We neglect gravity in the following analysis. The problem then states, find $p \in C^2(\Omega)$ (see Definition 7) such that

$$\begin{cases} \partial_t s_w(p) - \nabla \cdot (\kappa(s_w(p))\nabla p) = f, \ x \in \Omega, \ t \in [0,T] \\ p(0,x) = p_0(x), \ x \in \Omega \\ p(t,x) = g(t,x), \ x \in \partial\Omega, \ t \in [0,T], \end{cases} \tag{2.1}$$

where Ω is a domain in $\mathbb{R}^a$ ($a = 2$ for the numerical experiments) and $g(t,x) = 0$ for the simplicity of the analysis, see Section 1.2. The equation is usually discretized by implicit Euler in time because of the low regularity of the solution. For the spatial discretizations, however, there are several options. In this thesis as well as in [5] conforming finite elements are applied. This is not locally mass conservative and we refer to [26, 27] where they apply a mixed finite element method with this property. There are several other examples of locally mass conservative discretizations.

Using an implicit temporal discretization we need a linearization scheme to deal with the two non-linearities, s_w and κ, see Section 1.3.6. Several options for iterative schemes are discussed and the L-scheme is particularly analyzed both in the sense of convergence and optimization of the stability parameter (optimization in the sense that we seek the lowest amount of iterations). Finally, we present a comparative numerical study of the different applied schemes.

2.1 Linearizations of the fully-discrete Richards' equation

We begin by defining a variational formulation of (2.1). Find $p \in H_0^1(\Omega)$ such that

$$\langle \partial_t s_w(p), q \rangle + \langle \kappa(s_w(p))\nabla p, \nabla q \rangle = \langle f, q \rangle \tag{2.2}$$

for all $q \in H_0^1(\Omega)$. Applying implicit Euler in time we define a uniform mesh of the interval $[0, T]$ with time step size τ. The time-discretized version of equation (2.2) then reads; given $p^{n-1} \in H_0^1(\Omega)$ find $p^n \in H_0^1(\Omega)$ such that

$$\langle s_w(p^n), q\rangle + \tau\langle\kappa(s_w(p^n))\nabla p^n, \nabla q\rangle = \langle \tau f^n + s_w(p^{n-1}), q\rangle \tag{2.3}$$

for all $q \in H_0^1(\Omega)$ where $f^n = f(\cdot, t^n)$. Let $\mathcal{T}_h$ be a regular decomposition of Ω, where h represents the mesh diameter. Consider the subspace of $H_0^1(\Omega)$

$$Q_h = \left\{q_h \in H_0^1(\Omega) \mid q_{h|_T} \in \mathcal{P}_m(T), T \in \mathcal{T}_h\right\}$$

where $\mathcal{P}_m(T)$ denotes the space of polynomials of degree m on the simplex (triangle in our case) T. In the numerical experiments we use $m = 1$, which corresponds to a linear finite element approximation. Now, look for solutions in Q_h instead of $H_0^1(\Omega)$; given $p_h^{n-1} \in Q_h$ find $p_h^n \in Q_h$ such that

$$\langle s_w(p_h^n), q_h\rangle + \tau\langle\kappa(s_w(p^n))\nabla p_h^n, \nabla q_h\rangle = \langle \tau f^n + s_w(p_h^{n-1}), q_h\rangle \tag{2.4}$$

for all $q_h \in Q_h$. The final step is to deal with the non-linearities $s_w(\cdot)$ and $\kappa(s_w(\cdot))$. We propose four different linearizations, three of which take similar form. Let i be the iteration index. Given $p_h^{n,i-1}, p_h^{n-1} \in Q_h$ find $p_h^{n,i} \in Q_h$ such that

$$\begin{aligned}&\langle s_w(p_h^{n,i-1}), q_h\rangle + \langle M(p_h^{n,i-1})(p_h^{n,i} - p_h^{n,i-1}), q_h\rangle \\ &+ \tau\langle\kappa(s_w(p^{n,i-1}))\nabla p_h^{n,i}, \nabla q_h\rangle = \langle \tau f^n + s_w(p_h^{n-1}), q_h\rangle\end{aligned} \tag{2.5}$$

for all $q_h \in Q_h$. We iterate until a user-defined stopping criterion is reached. Here, different realizations of M give rise to different schemes. We list them here:

- The Modified Picard method (MP): Define $M(p) = s_w'(p)$, see [4].
- The L-scheme (LS): Define $M(p) = L$ for some user-defined $L > 0$, see [5].
- The modified L-scheme (MS): Define $M(p) = max\{[s_w'(p)+m], 2m\}$ for some user-defined $m > 0$, see [7].

The last linearization scheme is

- the Newton-Raphson method (NR) in which we replace any non-linearities $b(p)$ by $b(p^i)+b'(p^{i-1})(p^i-p^{i-1})$, see Section 1.1.2. For this problem we have $b_1(p) = s_w(p)$ and $b_2(p) = \kappa(s_w(p))\nabla p$. For the derivative of the non-linearity $b_2(p) = \kappa(s_w(p))\nabla p$ we have to apply the theory of Frechet derivatives. The Frechet derivative with respect to p, D_p, of ∇p in the direction h is simply

∇h, i.e. $D_p(\nabla p)(h) = \nabla h$. Together with the product rule for differentiation we then get

$$D_p(\kappa(s_w(p))\nabla p)|_{p=p^{i-1}}(p^i - p^{i-1}) = \\ (\kappa \circ s_w)'(p^{i-1})\nabla p^{i-1}(p^i - p^{i-1}) + \kappa(s_w(p^{i-1}))\nabla(p^i - p^{i-1}).$$

The Newton-Raphson then reads; given $p_h^{n,i-1}, p_h^{n-1} \in Q_h$ find $p_h^{n,i} \in Q_h$ such that

$$\langle s_w(p_h^{n,i-1}), q_h\rangle + \langle s_w'(p_h^{n,i-1})(p_h^{n,i} - p_h^{n,i-1}), q_h\rangle + \tau\langle\kappa(s_w(p^{n,i-1}))\nabla p_h^{n,i}, \nabla q_h\rangle \\ + \tau\langle(\kappa \circ s_w)'(p^{n,i-1})\nabla p_h^{n,i-1}(p_h^{n,i} - p^{n,i-1}), \nabla q_h\rangle = \langle\tau f^n + s_w(p_h^{n-1}), q_h\rangle \tag{2.6}$$

for all $q_h \in Q_h$.

Remark 7. *Notice that (MS) is a mixture of (MP) and (LS). In particular, if $m = 0$ then (MS)=(MP) and if $m > max\{s_w'(p)\}$ then (MS)=(LS).*

Remark 8 (Kirchhoff Transformation)**.** *For homogeneous absolute permeability one can rewrite problem* (2.1) *through an invertible transformation called the Kirchhoff transformation. This would make the permeability term of the equation independent of p, see [6]. When that is the case, and we, like here, consider Richards' equation without gravity, the (MP) and (NR) coincide. Moreover, we suddenly get a huge numerical performance advantage when applying the (LS); we do not have to update the stiffness matrix every iteration.*

In all the linearization schemes we start at $p_h^{n,0} = p_h^{n-1}$, and stop when $\|p_h^{n,i} - p_h^{n,i-1}\| \leq \epsilon_a + \epsilon_r\|p_h^{n,i}\|$, where ϵ_a (absolute tolerance) and ϵ_r (relative tolerance) are user defined tolerances.

2.1.1 Convergence of the linearization methods applied to Richards' equation

It is well known that the Newton-Raphson is a locally convergent scheme of asymptotic quadratic order, see section 1.1.2. The Modified Picard is only of linear order and is also a locally convergent scheme. It does however preserve the symmetry of the problem (2.5) better than the Newton-Raphson method making the linear system faster to solve. Additionally, the permeability, $\kappa(\cdot)$, is often only Hölder continuous making the computations of its derivatives dangerous as they might become infinitely large. The modified L-scheme is also linearly convergent. Its convergence is, however, global due to its stabilizing term for Lipschitz continuous $\kappa \circ s_w$.

We will not make any proofs for these methods here, but rather focus theoretically on the L-scheme. However, they will all be tested numerically in comparison to the L-scheme and a similar method which we will call the locally optimized L-scheme based on the theory of the L-scheme.

2.2 The L-scheme and the optimization of its stabilization parameter

The goal of this section is to determine how to choose our L in the most optimal way for the L-scheme, i.e. the choice of L that gives the optimal theoretical rate of convergence.

2.2.1 Constant permeability

We start with a simpler form of Richards' equation (2.1) with constant permeability, see Remark 8. The non-discretized version is formulated in the following way; find $p \in C^2(\Omega)$ such that

$$\begin{cases} \partial_t s_w(p) - \nabla(\kappa \nabla p) = f, \ x \in \Omega, \ t \in [0,T] \\ p(0,x) = p_0(x), \ x \in \Omega \\ p(t,x) = 0, \ x \in \partial\Omega, \ t \in [0,T]. \end{cases} \tag{2.7}$$

The fully discretized version, recall (2.4), of this equation after applying implicit Euler in time and conforming finite elements in space reads; find $p_h \in Q_h$ such that

$$\langle s_w(p_h^n), q_h \rangle + \tau \langle \kappa \nabla p_h^n, \nabla q_h \rangle = \langle \tau f^n + s_w(p_h^{n-1}), q_h \rangle \tag{2.8}$$

for all $q_h \in Q_h$. Here, we consider only the L-scheme which reads; find $p_h \in Q_h$ such that

$$\langle s_w(p_h^{n,i-1}), q_h \rangle + L\langle (p_h^{n,i} - p_h^{n,i-1}), q_h \rangle + \tau \langle \kappa \nabla p_h^{n,i}, \nabla q_h \rangle = \langle \tau f^n + s_w(p_h^{n-1}), q_h \rangle \tag{2.9}$$

for all $q_h \in Q_h$ and $L > 0$.
For the following theorem we require the assumptions listed below.

Assumption 1. *The saturation $s_w(\cdot)$ is Lipschitz continuous and strictly monotone increasing with L_s and $s_{w,m} > 0$ being the Lipschitz constant and a lower bound for the derivative, respectively.*

Assumption 2. *The permeability κ is positive.*

Theorem 2.2.1. *Suppose that Assumption 1 and 2 hold true and define* $e_h^{n,i} := p_h^{n,i} - p_h^n$. *Then the scheme* (2.9) *converges for*

$$L = \frac{L_s}{2(1-\gamma)} \tag{2.10}$$

with rate of convergence

$$\text{rate}(\gamma) = \frac{L_s - 4(1-\gamma)s_{w,m}\gamma}{L_s + 4(1-\gamma)\frac{\tau\kappa}{C_\Omega}}, \tag{2.11}$$

through the inequality

$$\|e_h^{n,i}\|^2 \leq \text{rate}(\gamma)\|e_h^{n,i-1}\|^2 \tag{2.12}$$

where γ *is a constant that can be chosen arbitrarily in its domain in* $[0,1)$ *and* C_Ω *is the Poincaré constant depending on the domain* Ω.

Proof. Subtract (2.8) from (2.9) and set $q_h = e_h^{n,i}$

$$\begin{aligned}&\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle + \langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i} - e_h^{n,i-1}\rangle \\ &\quad + \tau\langle\kappa\nabla e_h^{n,i}, \nabla e_h^{n,i}\rangle + L\langle e_h^{n,i} - e_h^{n,i-1}, e_h^{n,i}\rangle = 0.\end{aligned}$$

Let $\gamma \in [0,1)$ and split the first term while applying the lower bound of κ

$$\begin{aligned}&\gamma\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle + (1-\gamma)\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle \\ &\quad + \langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i} - e_h^{n,i-1}\rangle + \tau\kappa\|\nabla e_h^{n,i}\|^2 \\ &\quad + L\langle e_h^{n,i} - e_h^{n,i-1}, e_h^{n,i}\rangle \leq 0.\end{aligned}$$

Using the monotonicity and Lipschitz continuity of s_w while applying the algebraic identity

$$L\langle e_h^{n,i} - e_h^{n,i-1}, e_h^{n,i}\rangle = \frac{L}{2}\|e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2 - \frac{L}{2}\|e_h^{n,i-1}\|^2, \tag{2.13}$$

we get the inequality

$$\begin{aligned}&s_{w,m}\gamma\|e_h^{n,i-1}\|^2 + \frac{1}{L_s}(1-\gamma)\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \tau\kappa\|\nabla e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i}\|^2 \\ &\quad + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2 \leq \frac{L}{2}\|e_h^{n,i-1}\|^2 - \langle s_w(p^{n,i-1}) - s_w(p^n), e_h^{n,i} - e_h^{n,i-1}\rangle.\end{aligned}$$

From Young's inequality we can further advance the inequality to

$$\begin{aligned}&s_{w,m}\gamma\|e_h^{n,i-1}\|^2 + \frac{(1-\gamma)}{L_s}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \tau\kappa\|\nabla e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i}\|^2 \\ &+ \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2 \leq \frac{L}{2}\|e_h^{n,i-1}\|^2 + \frac{1}{2L}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2\end{aligned}$$

which rearranges into

$$s_{w,m}\gamma\|e_h^{n,i-1}\|^2 + \frac{(1-\gamma)}{L_s}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \tau\kappa\|\nabla e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i}\|^2 \leq \frac{L}{2}\|e_h^{n,i-1}\|^2 + \frac{1}{2L}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2.$$

Let $L := \dfrac{L_s}{2(1-\gamma)}$ and apply the Poincaré inequality to finally get

$$\|e_h^{n,i}\|^2\left(\frac{L}{2} + \frac{\tau\kappa_m}{C_\Omega}\right) \leq \|e_h^{n,i-1}\|^2\left(\frac{L}{2} - s_{w,m}\gamma\right). \tag{2.14}$$

Dividing and inserting $L = \dfrac{L_s}{2(1-\gamma)}$ in (2.14) gives the rate

$$\text{rate}(\gamma) = \frac{\frac{L_s}{2(1-\gamma)} - 2s_{w,m}\gamma}{\frac{L_s}{2(1-\gamma)} + \frac{2\tau\kappa}{C_\Omega}}. \tag{2.15}$$

which is equivalent to

$$\text{rate}(\gamma) = \frac{L_s - 4(1-\gamma)s_{w,m}\gamma}{L_s + 4(1-\gamma)\frac{\tau\kappa}{C_\Omega}}. \tag{2.16}$$

This proves Theorem 2.2.1. □

2.2.2 Optimality of the stabilization parameter L for the L-scheme applied to Richards' equation with constant permeability

As we now have convergence we seek to optimize it. To do this we continue from the rate-function (2.16) and minimize it w.r.t. $\gamma \in [0, 1)$. We see that

$$\text{rate}(0) = \frac{L_s}{L_s + 4\frac{\tau\kappa}{C_\Omega}} \leq 1 = rate(1).$$

Also we have

$$\text{rate}\left(\frac{1}{2}\right) = \frac{L_s - s_{w,m}}{L_s + 2\frac{\tau\kappa}{C_\Omega}}$$

which can be both smaller or larger than rate(0) depending on our problem. Differentiating (2.16) with respect to γ and looking for roots in $[0, 1)$ yields the potential minimum

$$\gamma_{\text{opt}} = 1 + \frac{L_s s_{w,m} C_\Omega - \sqrt{L_s^2 s_{w,m}^2 C_\Omega^2 + 4L_s s_{w,m}^2 C_\Omega \kappa\tau + 4L_s s_{w,m}\kappa^2\tau^2}}{4s_{w,m}\kappa\tau}. \tag{2.17}$$

The other root will be greater than 1, and is therefore outside the domain of γ, $[0, 1)$. Keep in mind that we only have proved convergence for $\gamma \in [0, 1)$ so if $\gamma_{\text{opt}} \leq 0$ we put it equal to 0. Our theoretically obtained optimal choice of L becomes

$$L_{\text{opt}} = \frac{L_s}{2(1 - \gamma_{\text{opt}})},$$

where we should experience the rate of convergence

$$\text{rate}(\gamma_{\text{opt}}) = \frac{L_{\text{opt}} - 2s_{w,m}\gamma_{\text{opt}}}{L_{\text{opt}} + \frac{2\tau\kappa}{C_\Omega}}.$$

We emphasize here the situations

- when $\gamma_{\text{opt}} = 0$ we get $L_{\text{opt}} = \dfrac{L_s}{2}$
- when $\gamma_{\text{opt}} = \frac{1}{2}$ we get $L_{\text{opt}} = L_s$.

Remark 9. *A similar analysis could have been done with heterogeneous permeability $\kappa(x)$ as long as it is bounded from below by κ_m.*

2.2.3 The general case: Non-linear permeability

In this section we optimize the L-scheme for the general form of Richards' equation, where the permeability is non-linear. The L-scheme reads, see (2.5), given $p_h^{n,i-1}, p_h^{n-1} \in Q_h$ find $p_h^{n,i} \in Q_h$ such that

$$\begin{aligned} &\langle s_w(p_h^{n,i-1}), q_h\rangle + L\langle (p_h^{n,i} - p_h^{n,i-1}), q_h\rangle \\ &+ \tau\langle \kappa(s_w(p_h^{n,i-1}))\nabla p_h^{n,i}, \nabla q_h\rangle = \langle \tau f^n + s_w(p_h^{n-1}), q_h\rangle \end{aligned} \tag{2.18}$$

Before presenting the convergence result we require some assumptions.

Assumption 3. *The permeability, $\kappa(s_w(\cdot))$ is Lipschitz continuous, with Lipschitz constant L_κ, and bounded uniformly from below by $\kappa_m > 0$.*

Assumption 4. *The solution at each time step, p_h^n, satisfies the the bound $\|\nabla p_h^n\| \leq \eta$ uniformly in n.*

Assumption 5. *The time step is chosen so that the inequality*

$$\tau < \frac{2\kappa_m}{\eta^2 L_\kappa^2 L_s}$$

is satisfied.

Theorem 2.2.2. *Suppose that Assumption 1, 3, 4 and 5 hold true. Then for*

$$L \geq \frac{L_s \kappa_m}{2\kappa_m(1-\gamma) - \eta^2 L_\kappa^2 \tau L_s}. \tag{2.19}$$

the scheme (2.18) *converges with rate of convergence*

$$\text{rate}(L,\gamma) = \frac{L - 2\gamma s_{w,m}}{L + \frac{\tau \kappa_m}{C_\Omega}} \tag{2.20}$$

through the inequality

$$\|e_h^{n,i}\|^2 \leq \text{rate}(L,\gamma)\|e_h^{n,i-1}\|^2$$

where γ is an arbitrary constant satisfying

$$0 \leq \gamma < 1 - \tau\left(\frac{\eta^2 L_\kappa^2 L_s}{2\kappa_m}\right). \tag{2.21}$$

Proof. Notice first that the L defined in equation (2.19) is well-defined by Assumption 5 and equation (2.21). We then begin as in the linear case by subtracting the non-linearized equation (2.4) from the linearized one (2.18) and set $q_h = e_h^{n,i}$

$$\begin{aligned}&\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle + \langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i} - e_h^{n,i-1}\rangle \\ &+ \tau\langle \kappa(s_w(p_h^{n,i-1}))\nabla p_h^{n,i} - \kappa(s_w(p_h^n))\nabla p_h^n, \nabla e_h^{n,i}\rangle + L\langle e_h^{n,i} - e_h^{n,i-1}, e_h^{n,i}\rangle = 0.\end{aligned}$$

We again split the first term and now add a zero in the permeability part of the equation. Let some arbitrary γ satisfy (2.21)

$$\begin{aligned}&\gamma\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle + (1-\gamma)\langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i-1}\rangle \\ &+ \tau\langle \kappa(s_w(p_h^{n,i-1}))\nabla e_h^{n,i} + \{\kappa(s_w(p_h^{n,i-1})) - \kappa(s_w(p_h^n))\}\nabla p_h^n, \nabla e_h^{n,i}\rangle \\ &+ \langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i} - e_h^{n,i-1}\rangle + L\langle e_h^{n,i} - e_h^{n,i-1}, e_h^{n,i}\rangle = 0.\end{aligned}$$

Applying Lipschitz continuity of $s_w(\cdot)$ and the boundedness from below of s_w' (Assumption 1) and κ (Assumption 3) together with the algebraic identity (2.13) we get

$$\begin{aligned}&\frac{(1-\gamma)}{L_s}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \tau\kappa_m\|\nabla e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2 \\ &\quad\leq \frac{L}{2}\|e_h^{n,i-1}\|^2 - \tau\langle\{\kappa(s_w(p_h^{n,i-1})) - \kappa(s_w(p_h^n))\}\nabla p_h^n, \nabla e_h^{n,i}\rangle \\ &\qquad - \gamma s_{w,m}\|e_h^{n,i-1}\|^2 - \langle s_w(p_h^{n,i-1}) - s_w(p_h^n), e_h^{n,i} - e_h^{n,i-1}\rangle.\end{aligned}$$

Young's inequality together with the boundedness from above of ∇p_h^n (Assumption 4) and the Lipschitz continuity of κ (Assumption 3) now gives

$$\begin{aligned}\frac{(1-\gamma)}{L_s}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \tau\kappa_m\|\nabla e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i}\|^2 + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2 \\ \leq \frac{L}{2}\|e_h^{n,i-1}\|^2 + \frac{\eta^2 L_\kappa^2 \tau}{2\kappa_m}\|s_w(p_h^{n,i-1})) - (s_w(p_h^n)\|^2 + \frac{\tau\kappa_m}{2}\|\nabla e_h^{n,i}\|^2 \\ - \gamma s_{w,m}\|e_h^{n,i-1}\|^2 + \frac{1}{2L}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 + \frac{L}{2}\|e_h^{n,i} - e_h^{n,i-1}\|^2.\end{aligned}$$

Collecting terms and applying the Poincaré inequality yields

$$\begin{aligned}\|s_w(p_h^{n,i-1}) - s_w(p_h^n)\|^2 \left(\frac{(1-\gamma)}{L_s} - \frac{\eta^2 L_\kappa^2 \tau}{2\kappa_m} - \frac{1}{2L}\right) + \frac{\tau\kappa_m}{2C_\Omega}\|e_h^{n,i}\|^2 \\ + \frac{L}{2}\|e_h^{n,i}\|^2 \leq \|e_h^{n,i-1}\|^2 \left(\frac{L}{2} - \gamma s_{w,m}\right).\end{aligned}$$

Now, for L satisfying (2.19) we have that

$$\frac{(1-\gamma)}{L_s} - \frac{\eta^2 L_\kappa^2 \tau}{2\kappa_m} - \frac{1}{2L} \geq 0.$$

This implies that the scheme converges with a rate of convergence

$$\text{rate}(L,\gamma) = \frac{L - 2\gamma s_{w,m}}{L + \frac{\tau\kappa_m}{C_\Omega}}.$$

□

2.2.4 Optimality of the stabilization parameter L for the L-scheme applied to Richards' equation with nonlinear permeability

We now perform a similar optimality analysis as for the case with constant permeability. However, as one would expect, it becomes a bit more technical. We seek to minimize the rate of convergence (2.20).

By choosing L minimally in (2.19) and inserting it into (2.20) a new rate is obtained

$$\text{rate}(\gamma) = \frac{L_s\kappa_m - 4\gamma(1-\gamma)\kappa_m s_{w,m} + 2\gamma s_{w,m}\eta^2 L_\kappa^2 L_s \tau}{L_s\kappa_m + \frac{2\kappa_m^2(1-\gamma)\tau}{C_\Omega} - \frac{\tau^2\eta^2 L_\kappa^2 L_s\kappa_m}{C_\Omega}} \tag{2.22}$$

To simplify (2.22) we make the following notations

- $\alpha := L_s\kappa_m$
- $\beta := 4\kappa_m s_{w,m}$
- $\sigma := 2s_{w,m}\eta^2 L_\kappa^2 L_s\tau$
- $\delta := \dfrac{2\kappa_m^2\tau}{C_\Omega}$
- $\xi := \dfrac{\tau^2\eta^2 L_\kappa^2 L_s\kappa_m}{C_\Omega}$.

Then the rate function becomes

$$\text{rate}(\gamma) = \frac{\alpha - \beta\gamma(1-\gamma) + \sigma\gamma}{\alpha + \delta(1-\gamma) - \xi}. \tag{2.23}$$

Now, differentiating and solving for critical values gives the potential optimal choice of γ

$$\gamma^* = \frac{\alpha\beta + \beta\delta - \beta\xi + \sqrt{\beta(\alpha^2\beta + \alpha\beta(\delta - 2\xi) + \alpha\delta(\delta+\sigma) + (\delta-\xi)(-\beta\xi + \delta\sigma))}}{\beta\delta}. \tag{2.24}$$

The other critical value of rate(γ) is larger than 1 and is not included as it cannot satisfy (2.21). The proposed optimal choice of γ is then

$$\gamma_{\text{opt}} = \begin{cases} \gamma^*, & \text{if } 0 \le \gamma^* < 1 - \tau\left(\frac{\eta L_\kappa L_s}{2\kappa_m}\right) \\ 0, & \textit{otherwise.} \end{cases} \tag{2.25}$$

2.2.5 Summary

To summarize, if one would like to choose the optimal L according to the presented theory we propose a recipe:

1. Choose τ in accordance with Assumption 5.
2. Compute γ^* as in (2.24).
3. Set $\gamma = \gamma_{\text{opt}}$ from (2.25).
4. Set $L = \dfrac{L_s\kappa_m}{2\kappa_m(1-\gamma) - \eta^2 L_\kappa^2\tau L_s}$.

Similar steps hold also for the equation with constant permeability.

2.2.6 Locally optimized L-scheme

The theory presented above is done with respect to global and time and iteration independent coefficients. This is not always most efficient, and we present an alternative. However, we make one important note; when we consider (2.9) the iteration independent choice of L gives a huge numerical advantage in the sense

that we do not have to update the stiffness matrix every iteration. Therefore, one is simply able to do an LU decomposition once and have a very efficient linear solver. On the other hand, with the non-linear permeability (2.18) we would have to update the stiffness matrix every iteration anyways and therefore lose this advantage.

Local parameters

Although the convergence proofs are not done with respect to locally chosen parameters one could still try this numerically. By this we mean that in every element we compute from the previous iteration the maximal and minimal derivative of $s_w(p_h^{n,i-1})$, which corresponds to L_s and $s_{w,m}$, the maximal derivative and minima of $\kappa(p_h^{n,i-1})$, corresponding to L_κ and κ_m respectively, and the maxima of $\nabla p_h^{n,i-1}$ corresponding to η. Then one computes, for each iteration, the optimal γ to determine the optimal stabilization parameter L as discussed above. The only issue here is the value one should choose for C_Ω since it is a global parameter. If choosing it globally (as volume of domain) everywhere it would still be reasonable to assume that we get convergence. In terms of implementation this will be more like the Newton-Raphson scheme.

Time dependent

In the case of constant permeability it might be more efficient to compute from the previous time step, instead of the previous iteration the local parameters so that we only need to update the stiffness matrix once for each time step and not for every iteration. This is especially reasonable for small time step sizes.

2.3 Numerical experiments

In this section we test our theoretical results numerically by checking if the optimal stabilization parameter, L_{opt} from Section 2.2.2 and 2.2.5, is indeed optimal. Additionally, we compare the L-scheme with practical optimally chosen stabilization parameter to the L-scheme with the locally optimized stabilization parameter (Section 2.2.6), the modified L-scheme, the modified Picard and the Newton-Raphson. We consider four different setups for the Richards equation. In two of which we consider constant permeability (2.7) and in the other two we consider non-linear permeability (2.1). In all the setups a MATLAB code for solving 2D problems using P1 finite elements has been used, see Section 1.2.6.

- **Setup 1**: A polynomial saturation of the form

$$s_w(p) = s_{w,m}p + (L_s - s_{w,m})\text{pol}(p), \text{ for } p \in [0,1] \tag{2.26}$$

has been constructed for the ease of determining maximal and minimal derivatives. Here pol(p) is defined as a third degree polynomial in p with zero derivative at $p = 0$ and $p = 1$ and maximal and unitary derivative at $p = 0.5$ in the domain $p \in [0, 1]$, see Figure 2.1. This means that we experience minimal derivative, $s_{w,m} = 0.125$, at $p = 0$ and $p = 1$ and maximal derivative, $L_s = 1.33$, at $p = 0.5$ for $s_w(p)$. For $p \notin [0, 1]$ we consider $s_w(p)$ to be the continuous constant extension of $s_w(p)$ as it is defined in (2.26). The permeability κ is kept constant and the cases $\kappa = \{0.01, 0.1, 1, 10\}$ are tested while the source $f(x, y, t)$ is computed so that the continuous problem yields the solution

$$p = txy(x - 1)(y - 1).$$

For the mesh we fix a uniform triangular mesh of size $h = \{\frac{1}{2^2}, \frac{1}{2^3}, \frac{1}{2^4}, \frac{1}{2^5}\}$ on the unit square and apply zero Dirichlet boundary conditions on the entire boundary. The time step size is set to $\tau = \{0.001, 0.01, 0.1, 1\}$ while computing the single time step from $t_0 = 7.9$ to $T = t_0 + \tau$. The reason for this is that we want to experience the entire non-linearity, i.e. both its minimal and its maximal derivative. When $\tau = 0.1$ we have an exact solution, p, that varies from 0 to 0.5.

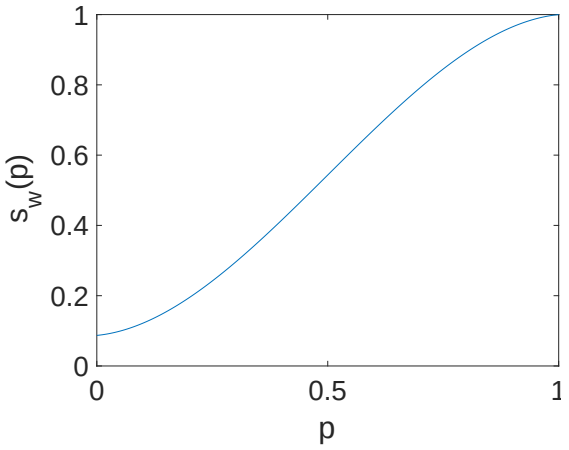


Figure 2.1: Saturation non-linearity (2.26) used in setup 1, 2 and 3 with $s_{w,m} = 0.125$ and $L_s = 1.33$.

- **Setup 2**: Exactly the same source, saturation and permeability as in setup 1 is considered, but now we impose homogeneous Neumann boundary conditions on the top, $\Gamma = [0, 1] \times \{1\}$.

- **Setup 3**: We apply the same source and saturation as in setup 1 and 2, but now with non-linear permeability

$$\kappa(s_w(p)) = 1 + p^2.$$

For this setup we consider homogeneous Dirichlet boundary conditions. We here express the permeability as an evaluation in the pressure. Although L_κ can be seen as the maximal derivative of κ with respect to s_w this is not an issue. The chain rule gives the formula for L_κ:

$$L_\kappa = max\left\{\frac{d\kappa}{ds_w}\right\} = max\left\{\frac{\frac{d\kappa}{dp}}{\frac{ds_w}{dp}}\right\} = max\left\{\frac{2p}{s_{w,m} + (L_s - s_{w,m})\text{pol}'(p)}\right\}.$$

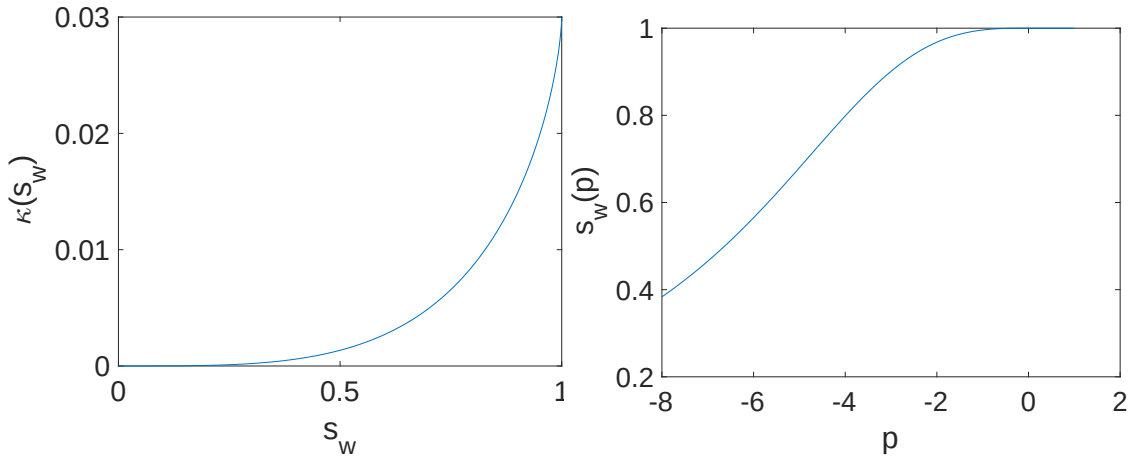
Although the source terms are not computed so that we get the continuous solution from setup 1 we remark that p will still be bounded giving implying the Lipschitz continuity of $\kappa(s_w(\cdot))$ that we need.

- **Setup 4**: The Van Genuchten-Mualem non-linearities are considered. They are given as

$$s_w(p) = \begin{cases} (1 + (-a_{\text{vG}}p)^{n_{\text{vG}}})^{-\frac{n_{\text{vG}}-1}{n_{\text{vG}}}}, & \text{for } p \leq 0 \\ 1, & \text{otherwise} \end{cases} \tag{2.27}$$

and

$$\kappa(s_w) = \frac{\kappa_{\text{abs}}}{\mu}\sqrt{s_w}\left(1 - \left(1 - s_w^{\frac{n_{\text{vG}}}{n_{\text{vG}}-1}}\right)^{\frac{n_{\text{vG}}-1}{n_{\text{vG}}}}\right)^2, \quad s_w \in [0,1]. \tag{2.28}$$



(a) Permeability (b) Saturation

Figure 2.2: Setup 4 - non-linearities

Symbol	Name	Value
a_{vG}	Inverse of air suction	0.1844
n_{vG}	Pore size distribution	3
κ_{abs}	Absolute permeability	$3 \cdot 10^{-2}$
μ	Fluid viscosity	1
f	Source	0
p_0	Initial pressure	-7.78
t_0	Start time	0
ε_a	Absolute tolerance	10^{-8}
ε_r	Relative tolerance	10^{-8}

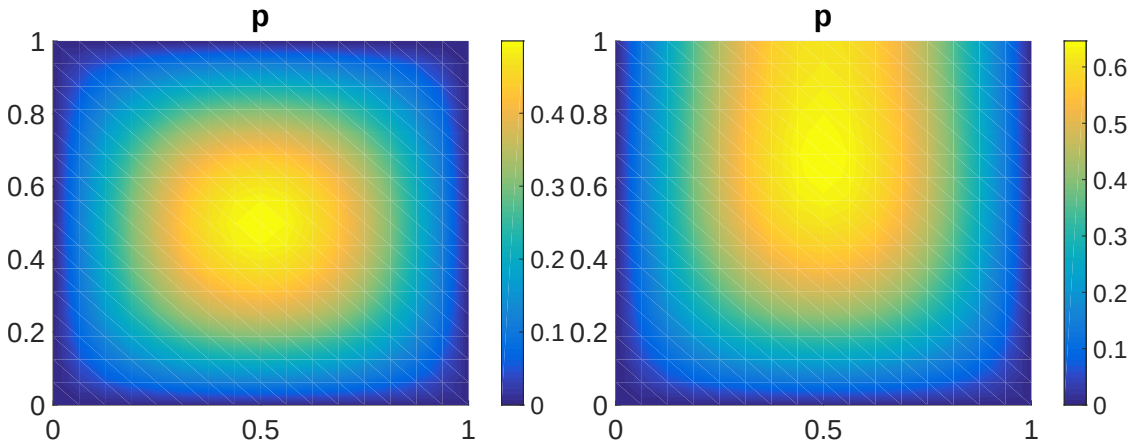
Table 2.1: Parameters for Setup 4

The constants a_{vG}, n_{vG}, κ_{abs} and μ can be found in Table 2.1, and the non-linearities are plotted in Figure 2.2a and 2.2b. Again we apply the mesh and boundary conditions of Setup 1. We consider zero source.

In all the setups we choose both the relative, ϵ_r, and absolute, ϵ_a, tolerance to be 10^{-8}. However, when we analyze the orders of convergence for different setups and schemes we set the tolerances equal to 10^{-12} to give a few more iterations of information. Solutions to all setups are plotted in Figure 2.3 and 2.4.

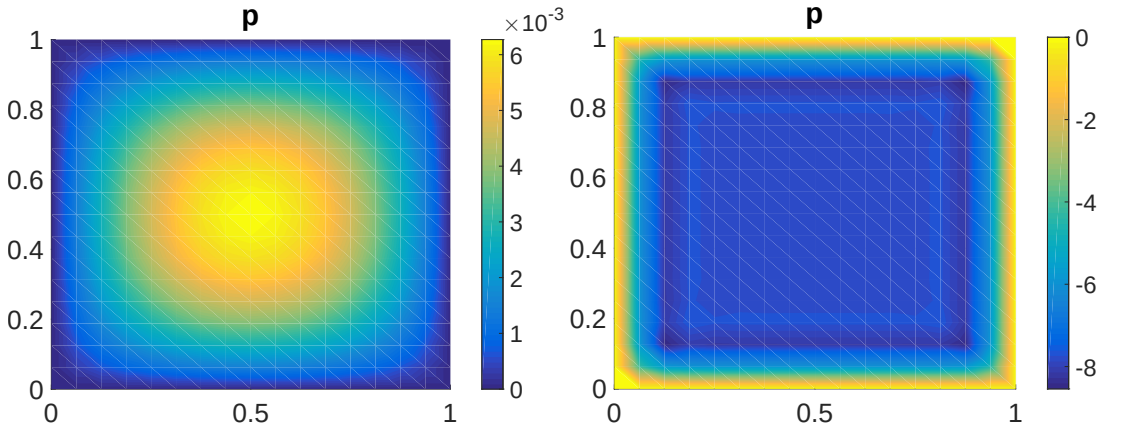
2.3.1 Solutions to test problems

Here some solutions to the different setups are displayed. All solutions are after one time step with a mesh-size $h = \frac{1}{16}$.



(a) Solution to setup 1. (b) Solution to setup 2

Figure 2.3: Solutions for setup 1 and 2 with $\kappa = 10$ at time $T = t_0 + \tau$ where $t_0 = 7.9$ and $\tau = 0.1$.



(a) Solution to setup 3. (b) Solution to setup 4

Figure 2.4: Solutions for setup 3 and 4 at time $T = t_0 + \tau$ where $t_0 = 0$ and $\tau = 0.1$.

2.3.2 Interpretation of the numerical schemes

We apply five different numerical schemes in this section:

- **L-scheme:** When referring to the L-scheme the formulations (2.9) and (2.18) are in mind. Choosing the stabilization parameter L is of vital importance and we will see that the number of iterations it takes to achieve convergence varies remarkably depending on the parameter. We always begin each setup

by testing for different values of L and compare them to the theoretically optimized values (marked by black stars) that we calculated in section 2.2.2 and 2.2.4. If the best value turns out not to be the theoretically calculated one we instead choose the practical best one when comparing to the other schemes.

- **Locally optimized L-scheme (Loc. opt. L-scheme):** The locally optimized L-scheme is the scheme discussed in section 2.2.6. We go into each element and compute locally the parameters needed to choose the optimal γ_{opt} in equations (2.17) and (2.24). Then inserting this optimal γ_{opt} to equations (2.10) and (2.19) gives a location dependent stabilization parameter. In the case of non-linear permeability we have dismissed this scheme as approximating the constants needed to calculate γ_{opt}, (2.24), turned out to be difficult and the resulting scheme was far to slow.

- **Newton-Raphson:** This is the usual Newton-Raphson scheme. The derivatives are computed exactly in all cases.

- **Modified Picard:** The modified Picard scheme is defined in (2.5). We remark that for setup 1 and 2 this scheme is omitted due to the fact that it completely coincides with the Newton-Raphson.

- **The modified L-scheme:** The modified L-scheme is a combination of the modified Picard scheme and the L-scheme. It defines a local and iteration dependent stabilization term, $M(p) = \max\{[b'(p)+m], 2m\}$ with user defined m. The scheme was introduced in [7] where the authors prove convergence for $m \geq \tau\Lambda\max\{|s_w''(p)|\}$. Here, Λ is a constant satisfying

$$\|p_h^n - p_h^{n-1}\|_{L^\infty(\Omega)} \leq \Lambda\tau.$$

  In our implementation of this scheme we define

$$\Lambda := \frac{\|p_h^{n,i-1} - p_h^{n-1}\|_{L^\infty(\Omega)}}{\tau}$$

  and choose $m = \tau\Lambda\max\{|s_w''(p)|\}$. However, to ensure convergence in the first iteration we choose $m = \max\{s_w'(p)\}$.

2.3.3 Setup 1: Polynomial solution

In Figure 2.5 we plot the number of iterations corresponding to different choices of stabilization parameters in the L-scheme for four different constant permeabilities. The observation we make is that the theoretically calculated optimal choice of L

(marked by black stars) is a bit away from the numerically observed optimal choice, though the difference in number of iterations is not huge. We also experience a contradictory phenomenon; as the permeability increases, the theoretically calculated optimal choice of L decreases while the practically observed optimal choice of L increases. It is still unknown what causes this contradiction. Nevertheless, we will be using the value $L = 0.8L_s$, which seems to be close to optimal for all the permeabilities for the further analysis of the schemes in this setup.

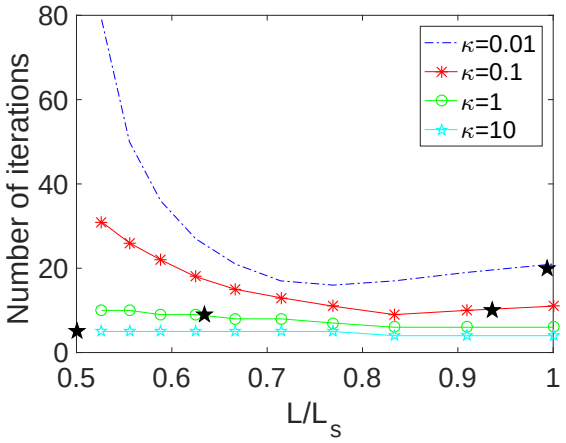
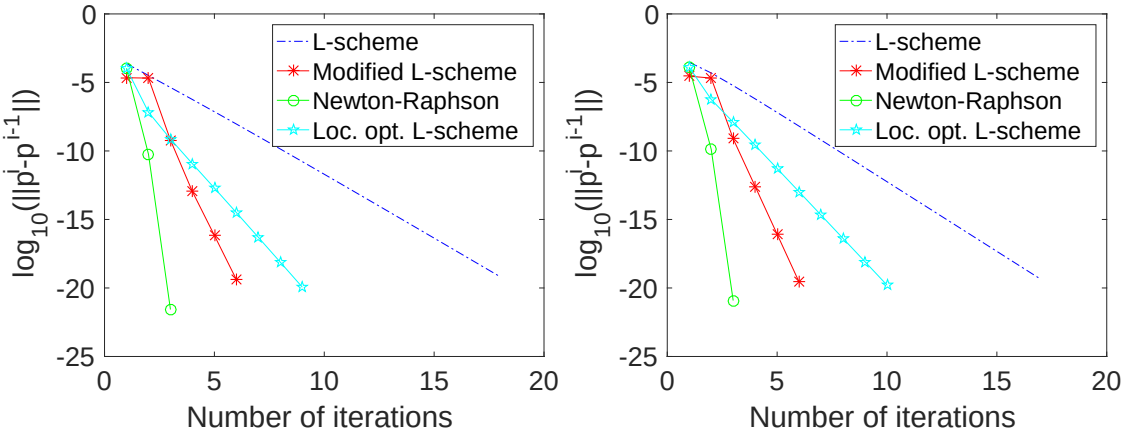


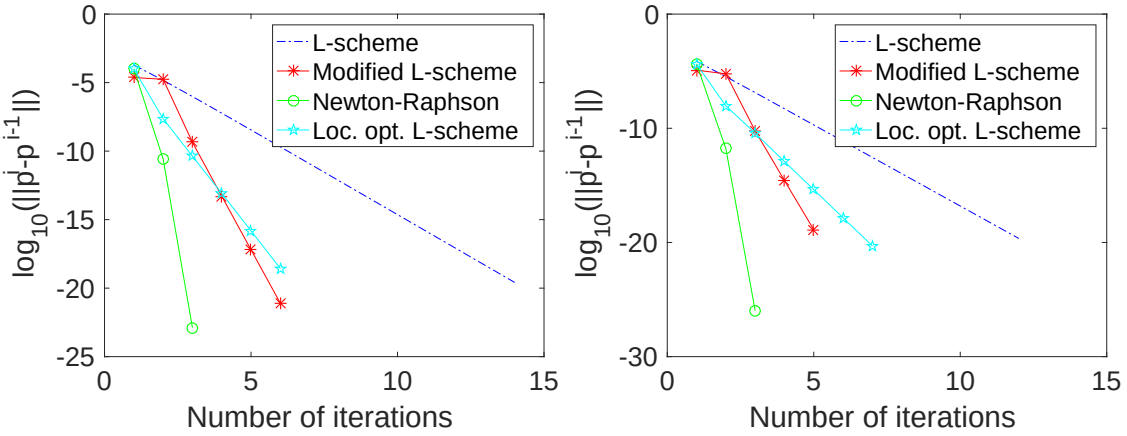
Figure 2.5: Setup 1 - Different stabilization parameters for the L-scheme.

In Figures 2.6 – 2.9 we compare the number of iterations to the decline in errors in logarithmic scale of the L-scheme (with $L = 0.8L_s$), the modified L-scheme, the Newton-Raphson and the locally optimized L-scheme for different permeabilities, κ, and mesh sizes. In all figures, (a) displays the comparison when we have a relatively fine mesh, $h = \frac{1}{16}$, and (b) the coarser $h = \frac{1}{4}$. We observe that as expected the Newton-Raphson is the fastest in all the situations. However, it is not much faster and when we consider the cost of assembling the matrix at every iteration, the L-scheme is likely to be faster for fine meshes.



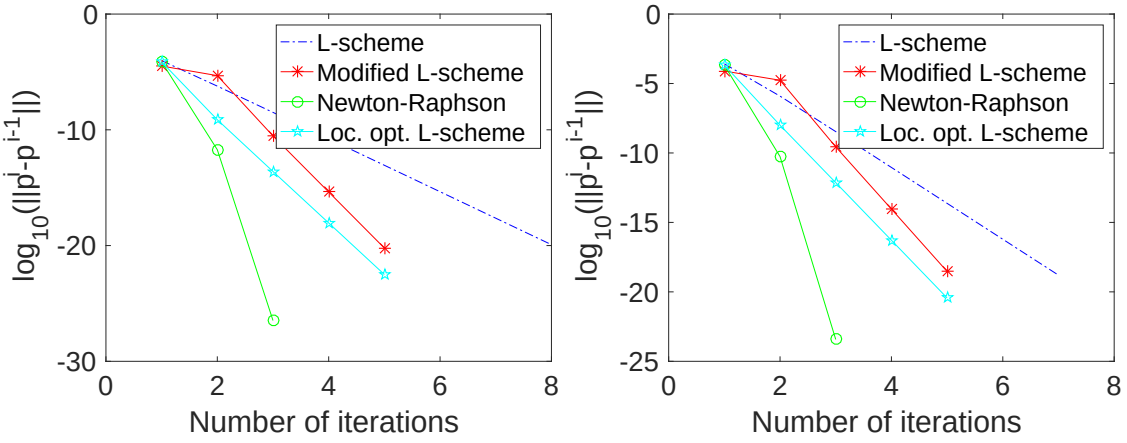
(a) $\kappa = 0.01$ and $h = 1/16$ (b) $\kappa = 0.01$ and $h = 1/4$

Figure 2.6: Setup 1 – Comparison between different numerical schemes for $\kappa = 0.01$.



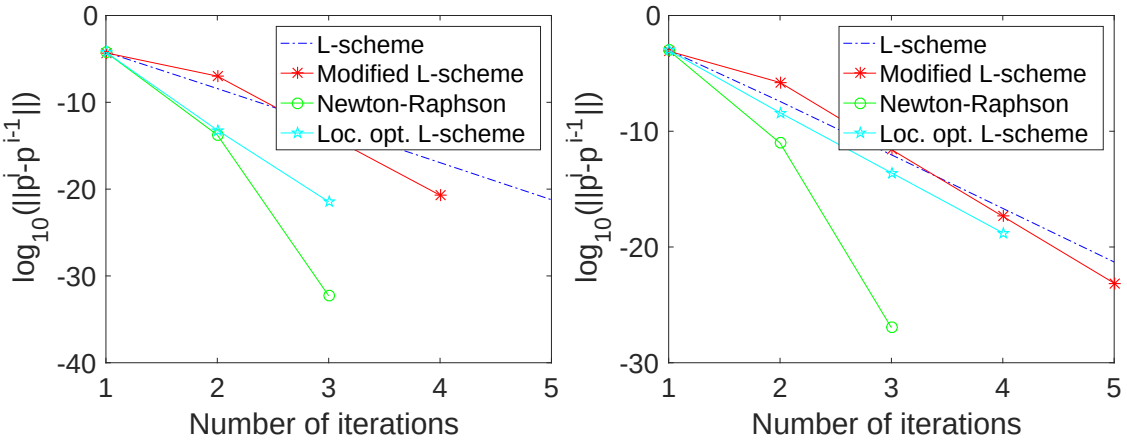
(a) $\kappa = 0.1$ and $h = 1/16$ (b) $\kappa = 0.1$ and $h = 1/4$

Figure 2.7: Setup 1 – Comparison between different numerical schemes for $\kappa = 0.1$.



(a) $\kappa = 1$ and $h = 1/16$ (b) $\kappa = 1$ and $h = 1/4$

Figure 2.8: Setup 1 – Comparison between different numerical schemes for $\kappa = 1$.

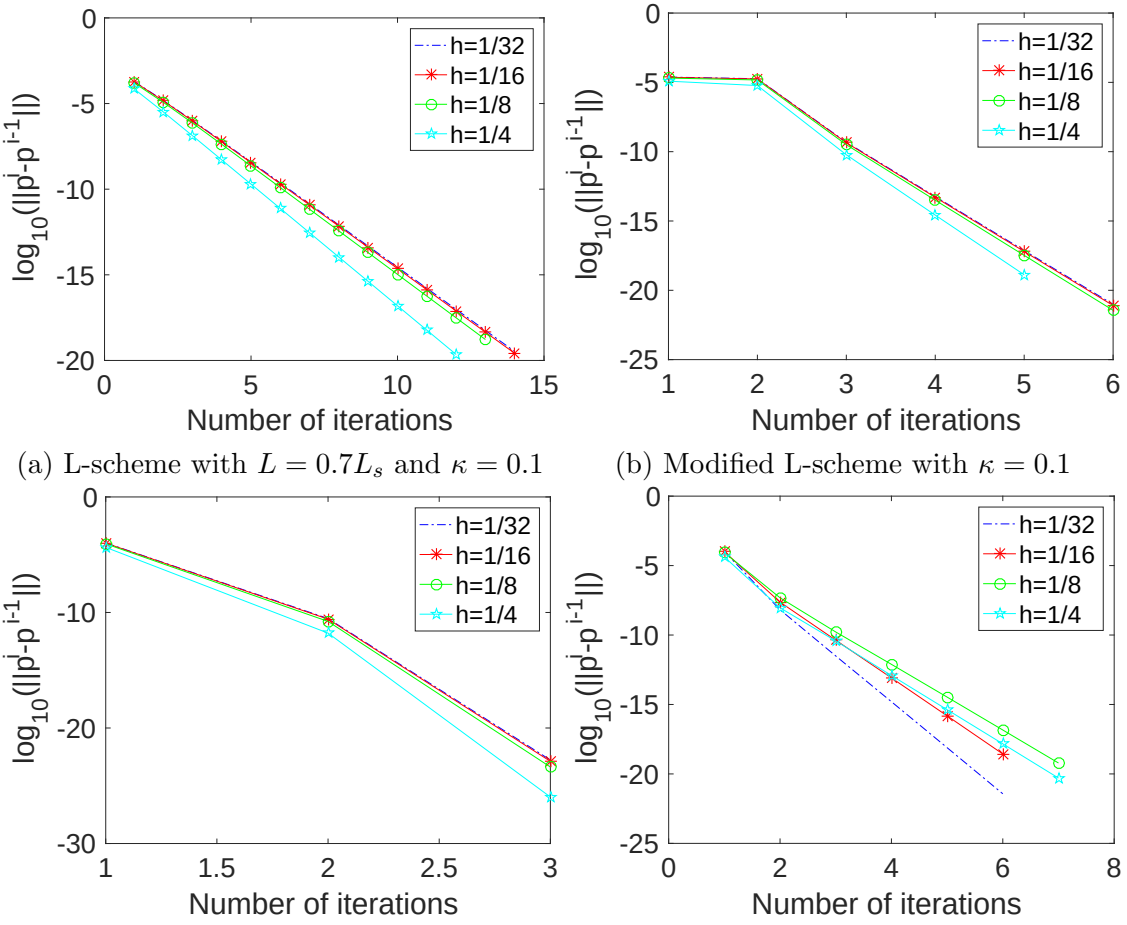


(a) $\kappa = 10$ and $h = 1/16$ (b) $\kappa = 10$ and $h = 1/4$

Figure 2.9: Setup 1 – Comparison between different numerical schemes for $\kappa = 10$.

Dependence on mesh size

Although we have not included the dependency of the mesh in our analysis, this is an interesting property of a linearization scheme. We analyze the different schemes separately for different mesh sizes, $h = \{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}\}$. In the plots of Figure 2.10 we see the results for $\kappa = 0.1$. We observe for the L-scheme, the modified L-scheme and the Newton-Raphson that the schemes are more or less independent of the mesh size. However, for the Locally optimized L-scheme we see some dependence with the trend that coarser meshes gives a slower scheme.



(a) L-scheme with $L = 0.7L_s$ and $\kappa = 0.1$

(b) Modified L-scheme with $\kappa = 0.1$

(c) Newton-Raphson with $\kappa = 0.1$

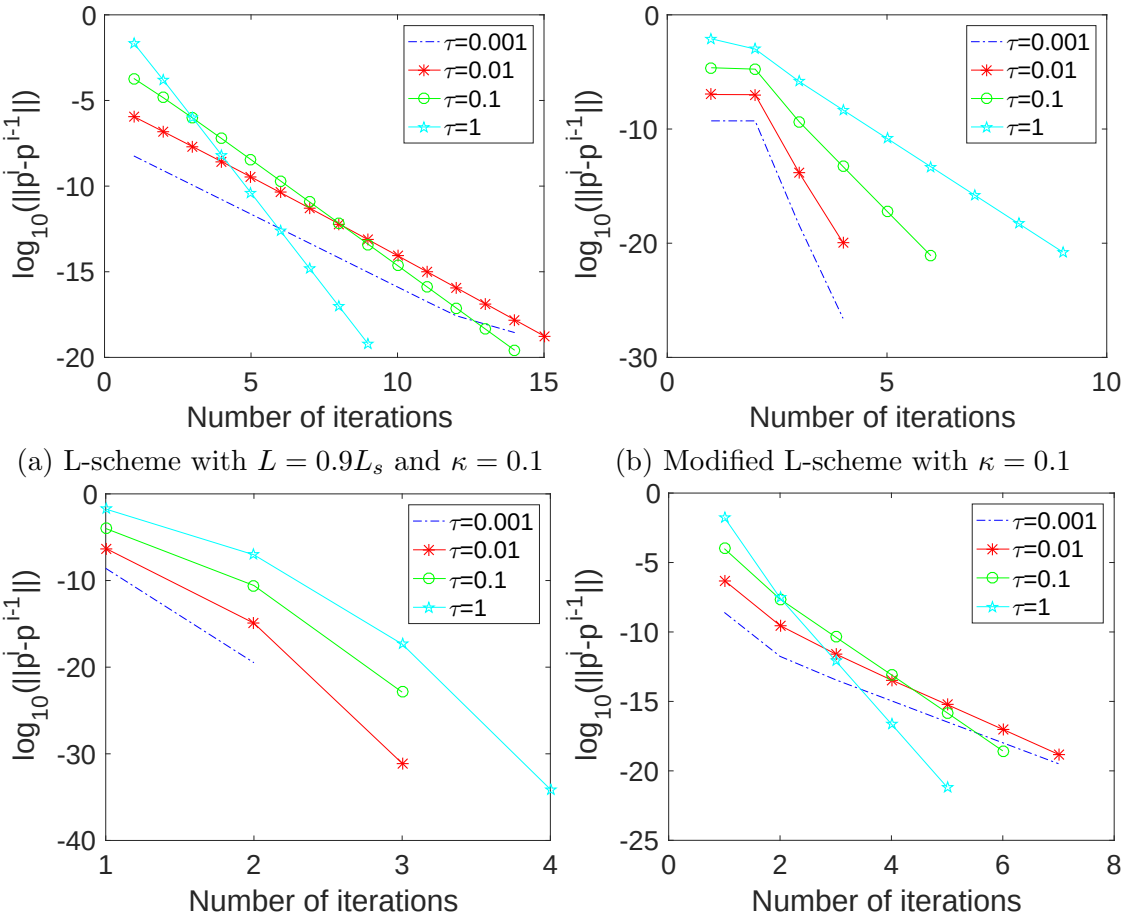
(d) Locally optimized L with $\kappa = 0.1$

Figure 2.10: Setup 1 – Dependence on mesh size.

Dependence on time step size

Another dependency we analyze within each scheme is the time step size. We plot in Figure 2.11 the decline in error for the different time step sizes, $\tau = \{0.001, 0.01, 0.1, 1\}$ in each scheme separately. We observe for the L-scheme that, exactly as theory predicts, larger time step sizes ultimately lead to faster convergence and it is clear that the rate is decreasing for increasing time step sizes, see Figure 2.11a. For the modified L-scheme, however, we see in Figure 2.11b that this is, interestingly, completely opposite. This is consistent with the theory in [7]. Figure 2.11c shows that for the Newton-Raphson method smaller time steps give faster convergence which is exactly what we would expect due to the fact that the natural initial guess (the solution at the previous time step) in this case is closer to the solution. Lastly, for the locally optimized L-scheme, see

Figure 2.11d, it seems like the rate is decreasing for increasing τ exactly as for the L-scheme.



(a) L-scheme with $L = 0.9L_s$ and $\kappa = 0.1$

(b) Modified L-scheme with $\kappa = 0.1$

(c) Newton-Raphson with $\kappa = 0.1$

(d) Locally optimized L with $\kappa = 0.1$

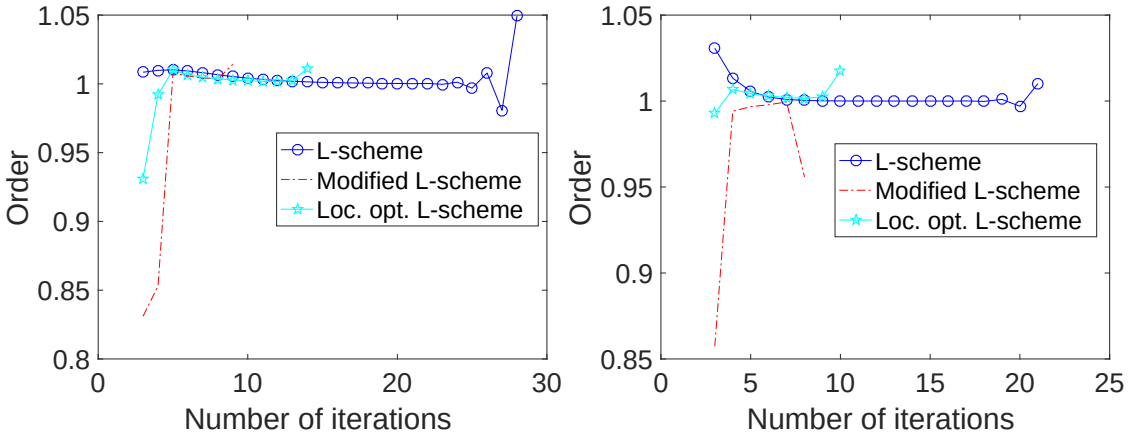
Figure 2.11: Setup 1 – Dependence on time step size.

Order of convergence

Finally for setup 1, we compare the rate of convergence for the different schemes with respect to different constant permeabilities. Here we let the time step size and the mesh size be $\tau = 0.1$ and $h = \frac{1}{16}$, respectively. In the horizontal axis we have as before the iteration number, while in the vertical axis we have the order of convergence. The formula for this number is described in Section 1.1 and is given by

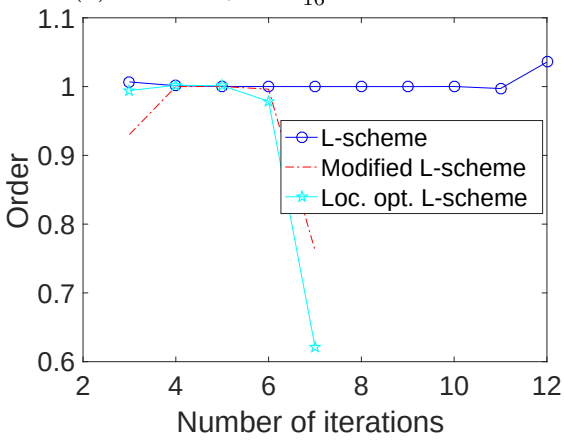
$$\text{Order} = \frac{log_{10}\left(\frac{\|p^{n,i}-p^n\|}{\|p^{n,i-1}-p^n\|}\right)}{log_{10}\left(\frac{\|p^{n,i-1}-p^n\|}{\|p^{n,i-2}-p^n\|}\right)}$$

where p^n is precomputed for every scheme. We see in Figure 2.12 that all the schemes for all the different permeabilities are varying around being linear except the Newton-Raphson which converges to fast to take into account. For the largest value of permeability $\kappa = 10$ also the modified L-scheme and the locally optimized L-scheme is converging too fast to consider.

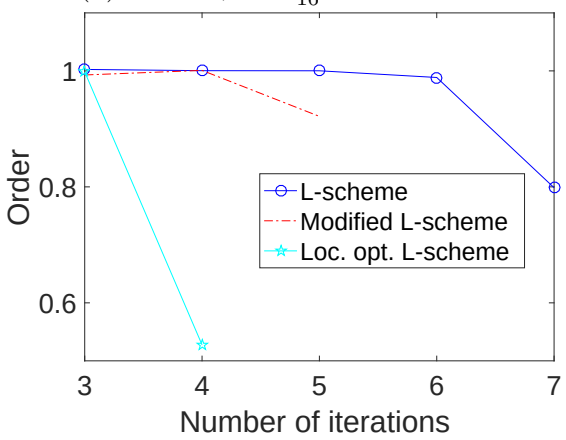


(a) $\kappa = 0.01$, $h = \frac{1}{16}$ and $\tau = 0.1$.

(b) $\kappa = 0.1$, $h = \frac{1}{16}$ and $\tau = 0.1$.



(c) $\kappa = 1$, $h = \frac{1}{16}$ and $\tau = 0.1$.



(d) $\kappa = 10$, $h = \frac{1}{16}$ and $\tau = 0.1$.

Figure 2.12: Setup 1 – Orders of convergence for the different schemes.

2.3.4 Setup 2: Natural boundary conditions on top

We perform similar numerical experiments to those in Section 2.3.3, but now for setup 2. In Figure 2.13 we again see the number of iterations for different values of the stabilization parameter, L. The theoretically calculated optimal choices of stabilization parameter L are marked as stars. We see that (like for setup 1) the theoretical and practical optimal value of L move in opposite directions. We will

for the rest of the numerical examples choose the value $L = 0.7L_s$ as our choice of stabilization parameter in the L-scheme as this seems to be the optimal parameter when choosing one for all the permeabilities.

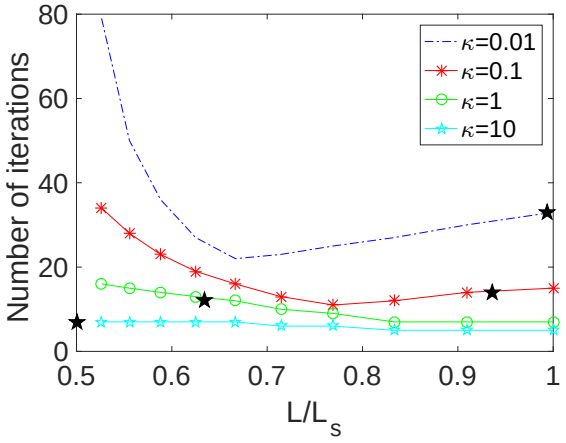
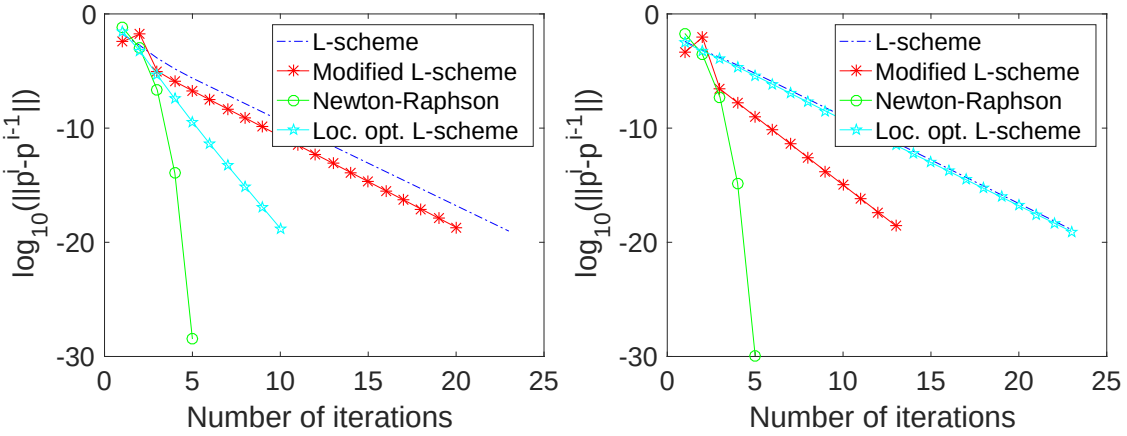


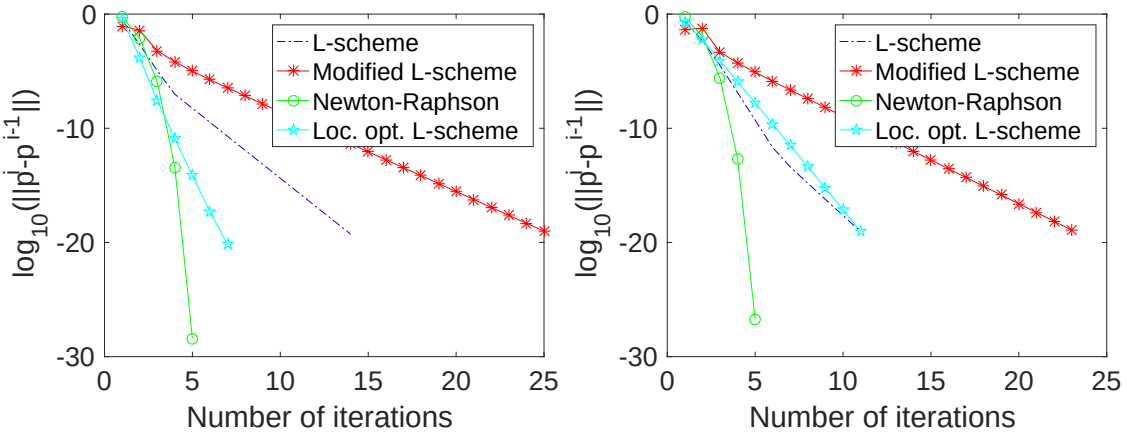
Figure 2.13: Setup 2 – Different stabilization parameters for the L-scheme.

Next we compare the different schemes with respect to their decline in error for different mesh sizes and permeabilities, see Figure 2.14–2.17. We see that for the larger permeability values, $\kappa \geq 1$, the modified L-scheme is pretty slow compared to the other schemes. Probably this is due to the stabilizing parameter, m, being too large. This has not been investigated further as the goal here is to optimize the stabilization parameter of the L-scheme and not of the modified L-scheme. We observe that as the permeability is increasing the L-scheme becomes more competitive to the Newton-Raphson method which is close to independent of permeability. The Locally optimized L-scheme is a bit faster than the L-scheme for all the different permeabilities, but where both the L-scheme and the Newton-Raphson is independent of the mesh size, it seems as though the locally optimized L-scheme requires more iterations for coarser meshes. One thing we can say through a comparison of the performances in setup 1 and 2 is that the performance of the linearization schemes are very much dependent on the domain and boundary conditions of the problems, even though this is not a factor yet in our theory for optimizing the L-scheme.



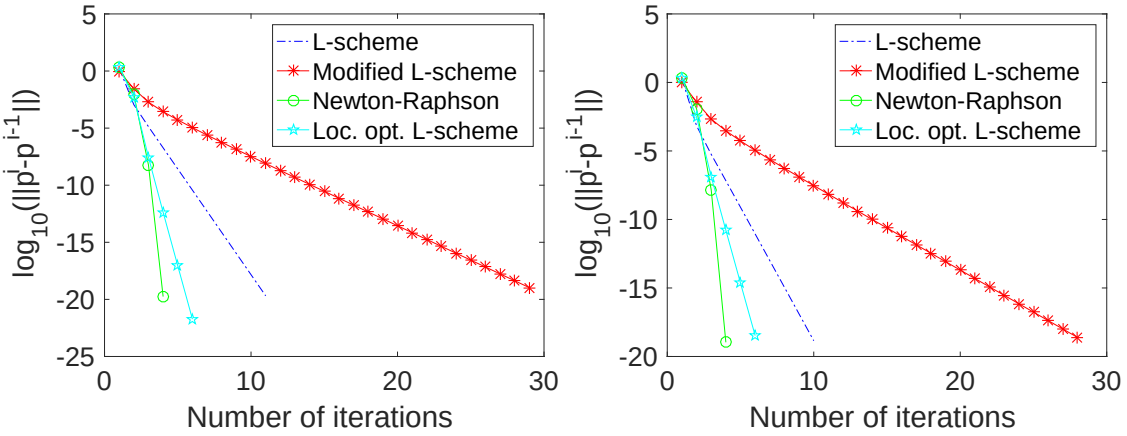
(a) $\kappa = 0.01$ and $h = 1/16$ (b) $\kappa = 0.01$ and $h = 1/4$

Figure 2.14: Setup 2 – Comparison between different numerical schemes for $\kappa = 0.01$.



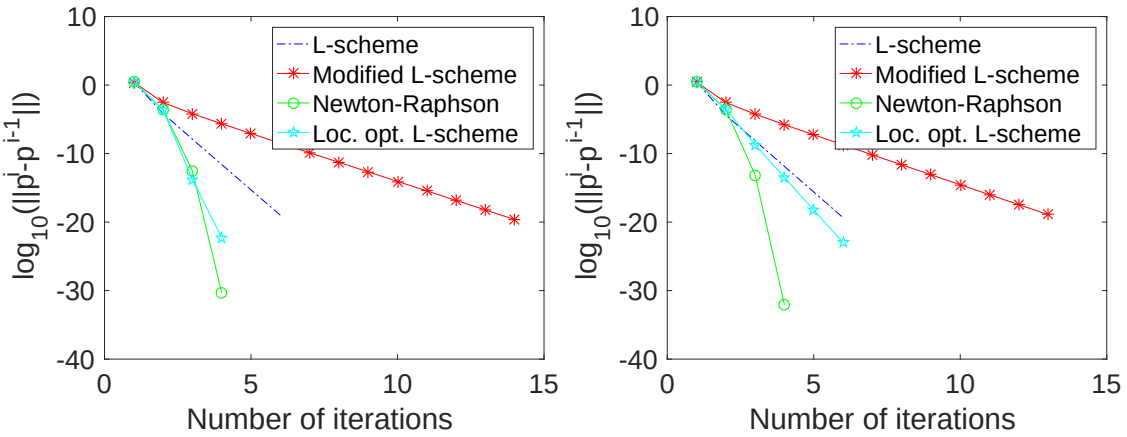
(a) $\kappa = 0.1$ and $h = 1/16$ (b) $\kappa = 0.1$ and $h = 1/4$

Figure 2.15: Setup 2 – Comparison between different numerical schemes for $\kappa = 0.1$.



(a) $\kappa = 1$ and $h = 1/16$ (b) $\kappa = 1$ and $h = 1/4$

Figure 2.16: Setup 2 – Comparison between different numerical schemes for $\kappa = 1$.

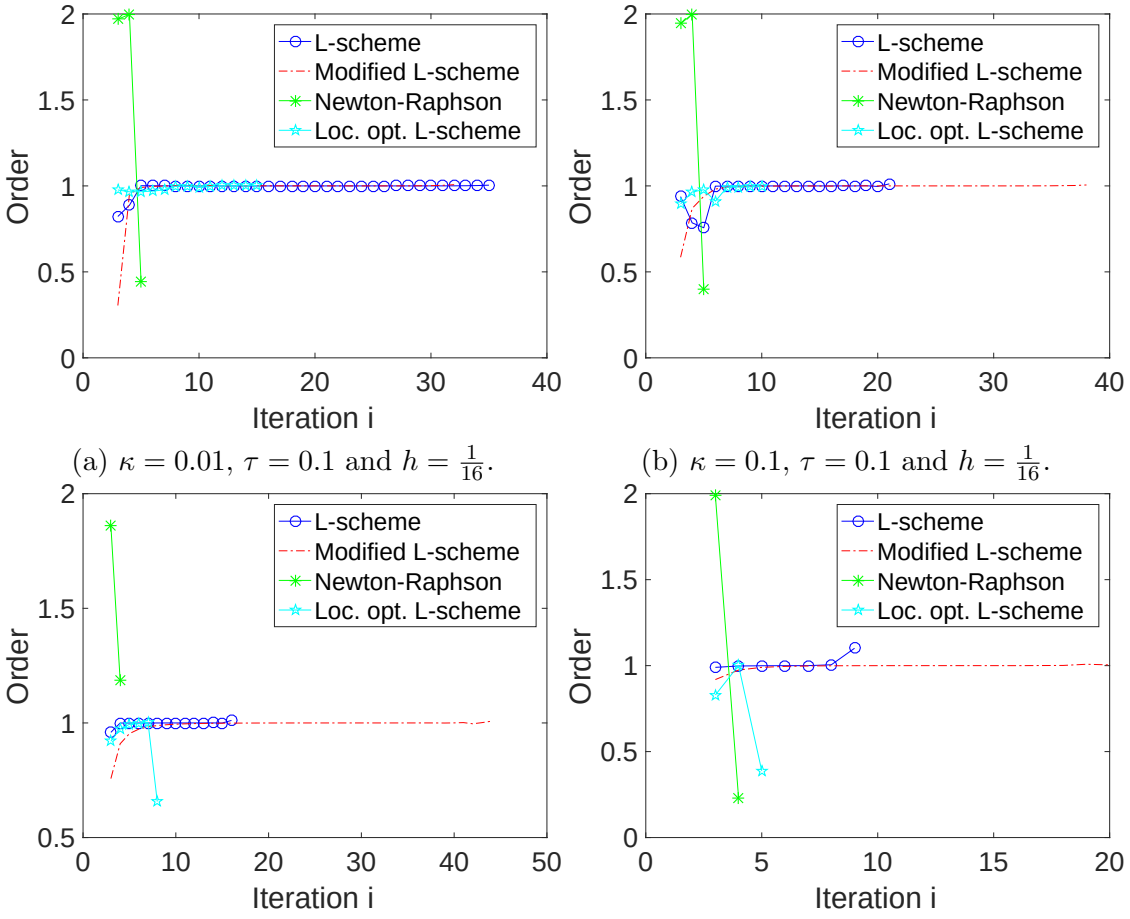


(a) $\kappa = 10$ and $h = 1/16$ (b) $\kappa = 10$ and $h = 1/4$

Figure 2.17: Setup 2 – Comparison between different numerical schemes for $\kappa = 10$.

Order of convergence

Also for this setup we compare the orders of the numerical schemes applying the same tools as in setup 1. This can be seen in Figure 2.18. We (still) clearly experience linear convergence for all schemes except the Newton-Raphson which is of quadratic order.



(a) $\kappa = 0.01$, $\tau = 0.1$ and $h = \frac{1}{16}$.

(b) $\kappa = 0.1$, $\tau = 0.1$ and $h = \frac{1}{16}$.

(c) $\kappa = 1$, $\tau = 0.1$ and $h = \frac{1}{16}$.

(d) $\kappa = 10$, $\tau = 0.1$ and $h = \frac{1}{16}$.

Figure 2.18: Setup 2 – Orders of convergence for the different schemes

2.3.5 Setup 3: Non-linear permeability

Now we enter the part of the numerical experiments with non-linear permeability. For this setup we have $\kappa(s_w(p)) = p^2 + 1$. In Figure 2.19 we see the number of iterations required for convergence for different values of the stabilization parameter, L. We clearly see in Figure 2.19 how sensitive this setup is to the choice of stabilization parameter. As $L = L_s$ seems to be the best choice we use that one when comparing the convergence of the L-scheme to the other schemes.

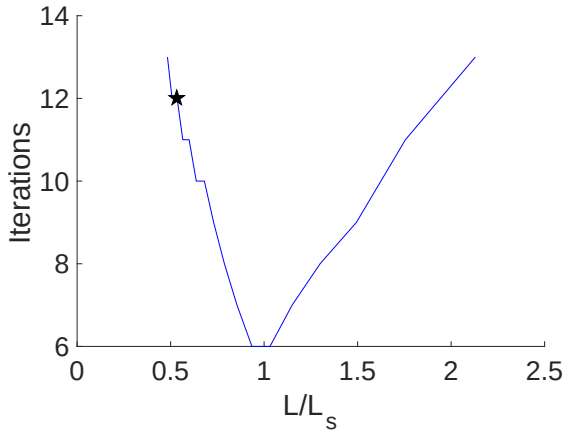
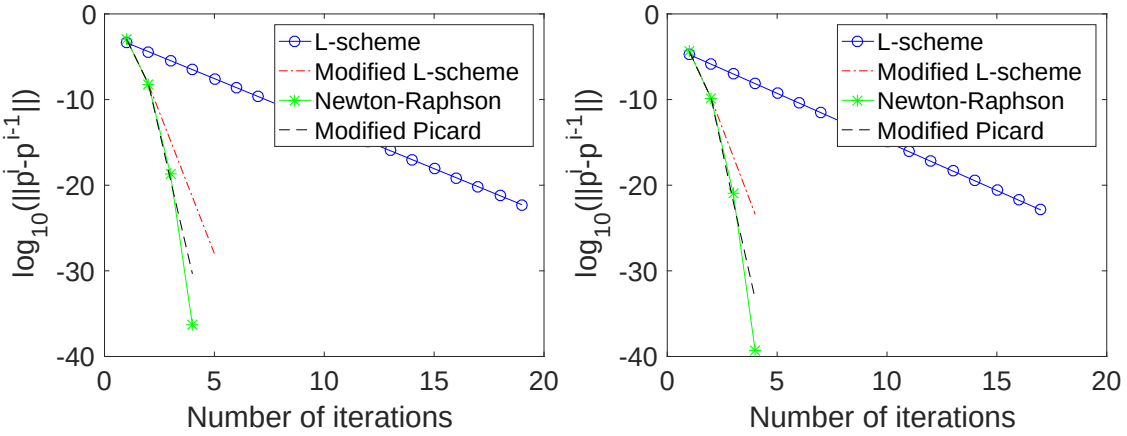


Figure 2.19: Setup 3 – Different stabilization parameters for the L-scheme.

Figure 2.20 shows a comparison of the numerical schemes for different mesh sizes. We see that the Newton-Raphson method, the modified Picard method and the modified L-scheme perform almost equally good while the L-scheme is falling behind. The trend and the numbers of iterations are the same for both the coarse and the finer mesh.



(a) $h = 1/16$ (b) $h = 1/4$

Figure 2.20: Setup 3 – Comparison between different numerical schemes, $\tau = 0.1$ and $L = L_s$ for the L-scheme.

Order of convergence

We compare the order of the schemes in Figure 2.21. Again we observe linear convergence for all the schemes except the Newton-Raphson which is closer to quadratic order.

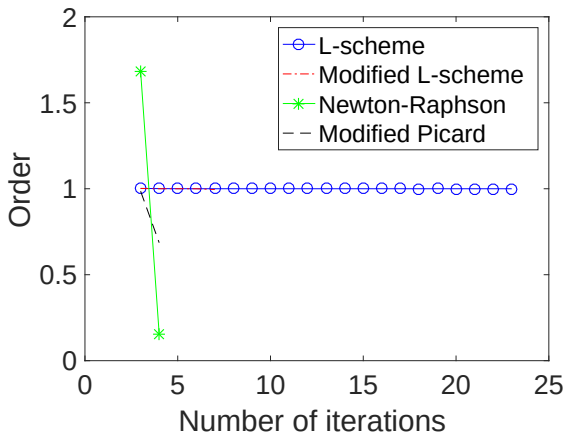


Figure 2.21: Setup 3 – Orders of convergence for the different schemes, $\tau = 0.1$, $h = \frac{1}{16}$ and $L = L_s$ for the L-scheme.

2.3.6 Setup 4: Van Genuchten-Mualem

In this section we do similar experiments as before, now with the van Genuchten-Mualem non-linearities. We remark that the theory for the L-scheme does not really hold for this particular non-linear permeability, in the sense that it is both equal to zero (for $p = 0$) and its derivative becomes infinitely large. We therefore do not plot the star in Figure 2.22. We choose $L = 0.8L_s$ when comparing to the other schemes, as this seems like the optimal practical choice. We do not, as we have done in the other setups, experiment with the schemes on different mesh sizes because the nature of the problem with homogeneous Dirichlet boundary conditions and constant negative initial condition gives a mesh dependent solution where finer meshes lead to solutions with larger derivatives than those of coarser meshes.

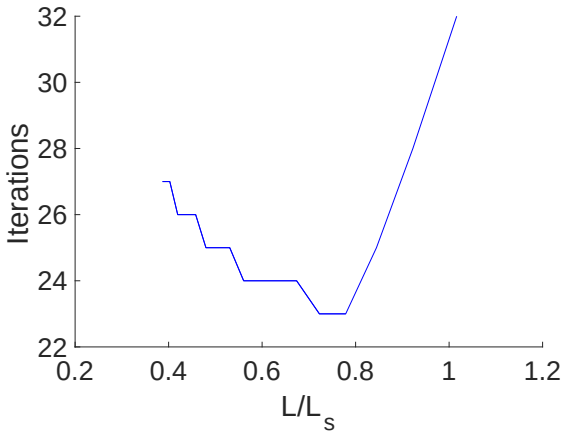


Figure 2.22: Setup 4 – Different stabilization parameters for the L-scheme.

In Figure 2.23, we see a comparison of the performance of the schemes. Notice that in this section the stabilization constant in the modified L-scheme has been scaled by 0.1 because the original one turned out to be far too large. We remark also that the time step size τ has been set to 0.01 because 0.1 is too large for the Newton-Raphson method to converge. This is an example where the local convergence of the Newton-Raphson becomes a problem. As the other schemes still converge for larger time step sizes they are clearly more robust. On the other hand we see that when the Newton-Raphson converges it converges faster than the other schemes. The L-scheme performs as good as the modified Picard for this setup while the Modified L-scheme is a bit slower.

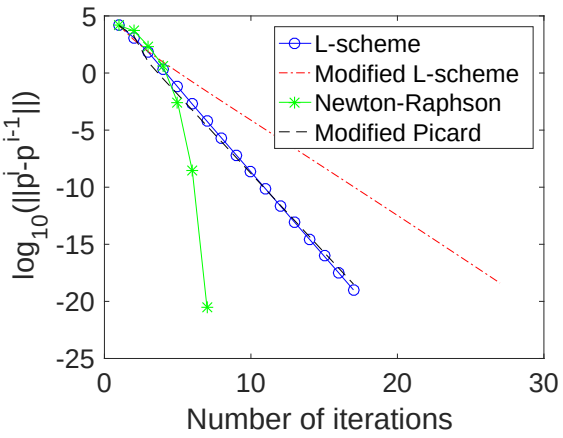
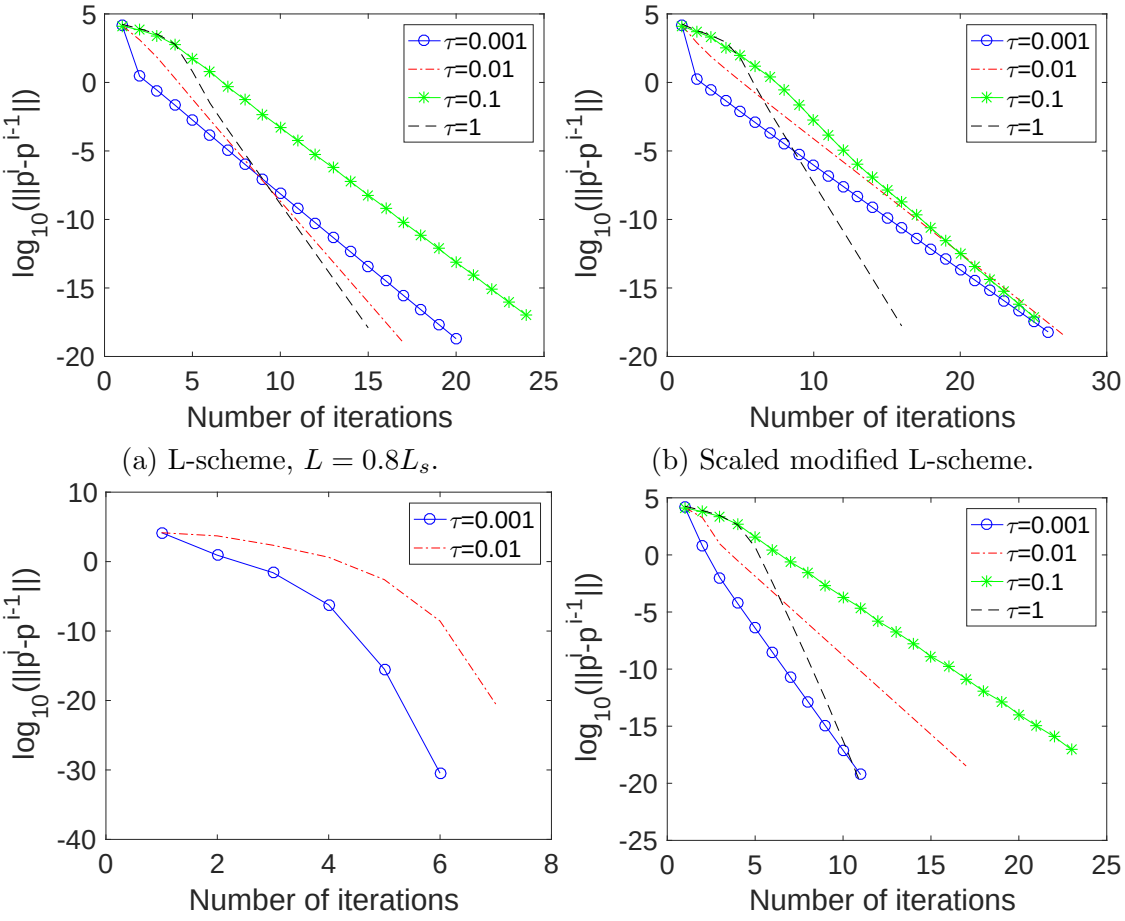


Figure 2.23: Setup 4 – Comparison between different numerical schemes, scaled modified L-scheme, $h = 1/16$, $\tau = 0.01$ and $L = 0.8L_s$ for the L-scheme.

Dependence on time step size

In Figure 2.24 we compare the different schemes for varying time step sizes, while the mesh size is fixed to $h = \frac{1}{16}$. For the L-scheme we see that larger time step sizes result in faster convergence, except for $\tau = 0.1$. For the modified L-scheme it seems as in the limit we will have faster convergence for larger time steps, this is similar to the L-scheme but contradictory to theory [7]. Most likely the stabilizing parameter is too large. For the Newton-Raphson smaller time steps give faster convergence exactly as theory predicts, however for $\tau \geq 0.1$ it diverges. The Modified Picard acts in accordance with the theory with larger time steps resulting in slower convergence except for $\tau = 1$ where it suddenly becomes much faster.



(a) L-scheme, $L = 0.8L_s$.

(b) Scaled modified L-scheme.

(c) Newton-Raphson

(d) Modified Picard

Figure 2.24: Setup 4 – Dependence on time step size.

Order of convergence

At last in Figure 2.25 we plot the observed orders of convergence for the different schemes. We see here that we observe linear convergence for the L-scheme, modified L-scheme and the modified Picard. The Newton-Raphson expresses close to quadratic behavior.

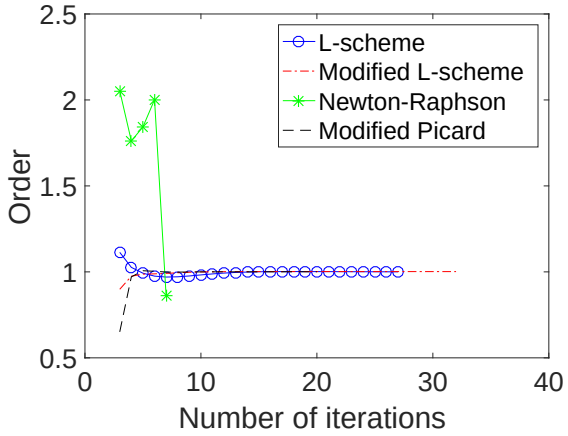


Figure 2.25: Setup 4 – Orders of convergence for the different schemes, $\tau = 0.01$, $h = \frac{1}{16}$ and $L = 0.8L_s$ for the L-scheme.

2.4 Conclusions

In this chapter we have considered the Richards equation for flow in porous media and studied theoretically and numerically the convergence properties of the L-scheme. Convergence proofs have been provided for the L-scheme applied to Richards' equation with both constant and non-linear permeability. Following the convergence proofs are optimality analyses with proposed optimal choices of the stabilization parameter L. However, this choice of stabilization parameter seems not to be the best practical choice, as showed in the numerical examples.

We also compared the L-scheme to a modified L-scheme, a newly proposed locally optimized L-scheme, the Newton-Raphson method and the modified Picard method. In all the setups, if the stabilization parameter L is chosen in a good (near optimal) way the L-scheme is certainly competitive with the other schemes. The locally optimized L-scheme even performs better in some cases and seems to be a solid choice for a numerical scheme. We do not have a rigorous convergence proof of this scheme, but it does give some indication to that the optimality analysis done for the L-scheme is going in the right direction. It might be that the setups at hand are of a character where the optimality analysis of the L-scheme is not suitable, or the theory miss some ingredients that define the setups. For example, the theory does not yet take into account the boundary conditions of the problems. We summarize in the following bullet-points:

- We proved convergence for the L-scheme applied to Richards' equation both with constant and non-linear permeability.

- A theoretical analysis of the optimal stabilization parameter L has been provided.
- The numerical experiments showed that the theory for optimal stabilization parameters is not completely sound.
- A comparative study of different linearization schemes was performed and we conclude that the Newton-Raphson clearly is the fastest scheme, but is not as robust as the other schemes, in particular the L-scheme.

Chapter 3

Biot's equations

We now consider the quasi-static linear Biot model, see Section 1.3.7, which models flow in deformable porous media. In this section we will perform an analysis (of similar nature to the analysis of the optimal stabilization parameter for the L-scheme in Chapter 2) of the fixed-stress splitting scheme applied to this system of equations. The Biot equations read, find $(\boldsymbol{u}, p)$ such that

$$-\nabla \cdot (2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\nabla \cdot \boldsymbol{u}\mathbf{I}) + \alpha\nabla p = \boldsymbol{f}, \tag{3.1}$$

$$\frac{\partial}{\partial t}\Big(\frac{p}{M} + \alpha\nabla \cdot \boldsymbol{u}\Big) - \nabla \cdot (\kappa(\nabla p - \mathbf{g}\rho)) = S_f, \tag{3.2}$$

where $\boldsymbol{u}$ is the displacement, $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\top)$ is the (linear) strain tensor, μ, λ are the Lamé parameters, α is the Biot-Willis constant, p, ρ are the fluids pressure and density, respectively, M is a compressibility constant, $\mathbf{g}$ the gravitational vector and κ is the permeability. The source terms $\boldsymbol{f}$ and S_f represent the density of applied body forces and a forced fluid extraction or injection process.

There are plenty of works concerning the discretization of Biot's equations (3.1)–(3.2). The most common temporal discretization is based on implicit Euler, see e.g. [16, 13]. Many combinations of spatial discretizations have been proposed and analyzed, e.g. cell-centered finite volumes [28], continuous Galerkin for the mechanics and mixed finite elements for the flow [29, 30, 13, 31], mixed finite elements for flow and mechanics [30, 32], non-conforming finite elements [33], the MINI element [34], continuous or discontinuous Galerkin [35, 36] or multiscale methods [37, 38, 39]. Continuous and discontinuous higher-order Galerkin space time elements were proposed in [14]. Adaptive computations were considered e.g. in [40]. For a discussion on the stability of the different spatial discretizations we refer to the recent papers [41, 42]. We discretize in time by the implicit Euler, and apply conforming finite elements for the spatial discretization.

After the discretization we are essentially left with two options for solving the system: monolithically or by using an iterative splitting algorithm. The former has

the advantage of being unconditionally stable, while a splitting scheme is much easier to implement, typically building on already available, separate numerical codes for porous media flow and for mechanics. However, a naive splitting of Biot's equations will lead to an unstable scheme [9]. To overcome this, one adds a stabilization term in either the mechanics equation (the so-called *undrained split scheme* [10]) or in the flow equation (the *fixed-stress splitting scheme* [11]). The splitting methods have very good convergence properties, making them a valuable alternative to monolithic solvers for simulation of the linear Biot model, see e.g. [11, 9, 12, 13]. Here, we discuss the fixed-stress splitting scheme, but we remark that a similar analysis can be performed for the undrained split scheme.

The initial derivation of the fixed-stress splitting scheme had a physical motivation [11, 9]: one 'fixes the (volumetric) stress', i.e. imposes $K_{dr}\nabla \cdot \mathbf{u}^i - \alpha p^i = K_{dr}\nabla \cdot \mathbf{u}^{i-1} - \alpha p^{i-1}$ and uses this to replace $\alpha \nabla \cdot \mathbf{u}^i$ in the flow equation. Here, K_{dr} is the physical, drained bulk modulus, defined as $K_{dr} = \frac{2\mu}{d} + \lambda$. The resulting stabilization parameter L, called from now on the *physical* parameter, is $L_{phys} = \dfrac{\alpha^2}{K_{dr}}$ (it depends on the mechanics and the coupling coefficient). Consequently, L_{phys} was the recommended value for the stabilization parameter, and the general opinion was that the method is not converging (it is not stable) for $L < L_{phys}$. In 2013, a rigorous mathematical analysis of the fixed-stress splitting scheme was for the first time performed in [12], where the authors show that the scheme is a contraction for any stabilization parameter $L \geq \dfrac{L_{phys}}{2}$. This analysis was confirmed in [13] for heterogeneous media, and by using a simpler technique. Noticeable, the same result was obtained also for both continuous or discontinuous Galerkin, higher order space-time elements in [14, 43], implying that the values of the tuning parameter are not depending on the order of the used elements. A legitimate question arises immediately: is now L_{phys} or $\dfrac{L_{phys}}{2}$ the optimal stabilization parameter, in the sense that the number of iterations is smallest? The question is relevant, because the number of iterations can differ considerably depending on the choice of the stabilization parameter [14, 13, 15, 16].

In a recent study [15], the authors considered different numerical settings and looked at the convergence of the fixed-stress splitting scheme. They determined numerically the optimal stabilization parameter for each considered case. This study, together with the previous results presented in [16] and [13] is suggesting that the optimal parameter is actually a value in the interval $\left[\dfrac{L_{phys}}{2}, L_{phys}\right]$, depending on the data. Especially, the optimal parameter depends on the boundary conditions and also on the flow parameters, not only on the mechanics and coupling coefficient.

Here, we show that the optimal stabilization parameter for the fixed-stress

scheme is neither $\frac{L_{phys}}{2}$ nor L_{phys}, but depends also on the flow parameters. The values $\frac{L_{phys}}{2}$, L_{phys} are obtained as limit situations. We prove first that the fixed-stress scheme converges linearly and then derive a theoretical optimal parameter, by minimizing the rate of convergence. The proof techniques in [13] are improved to reach the new results. For this we require the discretization to be inf-sup stable. Essentially, this allows for the control of errors in the pressure by those in the stress. A consequence of our theoretical result is that the fixed-stress splitting scheme also converges in the limit case of low-compressible fluids and low-permeable porous media. Finally, we perform numerical computations to test the optimized parameter. As can be seen in Section 3.5, the numerical results are sustaining the theory. In particular, we remark the connection between inf-sup stability and the performance of the fixed-stress splitting scheme: a not inf-sup stable discretization leads to non-monotonic behavior of the splitting scheme with respect to the problems parameters (e.g. the permeability).

To summarize; after applying implicit Euler in time to (3.1)–(3.2) and discretizing in space (using conforming finite elements), one has to solve a fully coupled, discrete system at each time step. For this, we apply the iterative fixed-stress splitting scheme [11]. If we denote by i the iteration index, one looks to find a pair $(\boldsymbol{u}^i, p^i)$ to converge to the solution $(\boldsymbol{u}, p)$, when $i \to +\infty$. Algorithmically, one solves first the flow equation (3.2) using the displacement from the previous iteration, then solves the mechanics equation (3.1) with the updated pressure and iterates until convergence is achieved. To ensure convergence [9, 12, 13], one adds a term $L(p^i - p^{i-1})$ to the flow equation (3.2). The free to be chosen parameter $L \geq 0$ is called the stabilization or tuning parameter. The choice of L is the deciding element for the success of the algorithm, because the number of iterations (and therefore the speed of the algorithm) strongly depends on its value, see [14, 13, 15, 16, 44]. Moreover, a too small or too big L will lead to no convergence.

The main results of this chapter are:

- an improved, theoretical convergence result for the fixed-stress splitting scheme under the assumption of an inf-sup stable discretization,
- the derivation of an optimized tuning parameter depending on both mechanics and fluid flow parameters, and
- the numerical evidence that not inf-sup stable discretizations lead to non-monotonic behavior of the fixed-stress splitting scheme w.r.t. to data (e.g. the permeability).

We mention that the fixed-stress splitting scheme also can be used for non-linear extensions of Biot's equations, see [45, 46] for non-linear water compressibility and

[47, 48, 49, 50] for unsaturated flow and mechanics. In these cases, one combines a linearization technique, e.g. the L-scheme [5, 6] with the splitting algorithm. The convergence of the resulting scheme can be proved rigorously [45, 47]. Furthermore, the fixed-stress splitting scheme has been applied in connection with fracture propagation [51] and phase field models [52]. There are several valuable variants of the fixed-stress splitting scheme: the multirate fixed-stress splitting scheme [53], the multiscale fixed-stress splitting scheme [44] and the parallel-in-time fixed-stress splitting scheme [54]. For a future analysis we mention the work in [55, 56] where the Biot equations are considered together with thermodynamical effects.

3.1 Fixed-stress splitting scheme applied to the fully discrete Biot Equations

We aim to solve the Biot equations (3.1)–(3.2) on $\Omega \times (0,T)$ together with, for simplicity, homogeneous Dirichlet boundary conditions and some given initial condition. We discretize in time using the implicit Euler method considering a uniform grid, with time step size $\tau := \frac{T}{N}, N \in \mathbb{N}$ and $t_n := n\tau, n \in \mathbb{N}$. Here T denotes the final time. The index n will throughout this section refer to the time step.

For the spatial discretization denote by $\mathcal{T}_h$ a regular triangulation of Ω and let $\mathcal{P}_1(\Omega)$ and $\mathcal{P}_2(\Omega)$ be the spaces of linear and quadratic polynomials on Ω. We then introduce two spaces $\boldsymbol{V}_h$ and Q_h where

$$Q_h := \{q_h \in H_0^1(\Omega) \mid q_{h|_K} \in \mathcal{P}_1(K) \; \forall \, K \in \mathcal{T}_h\} \tag{3.3}$$

and

$$\boldsymbol{V}_h := \{\boldsymbol{v}_h \in H_0^1(\Omega)^d \mid \left(v_{h|_K}\right)_i \in \mathcal{P}_2(K) \; \forall \, K \in \mathcal{T}_h \text{ for } i = 1, ..., d\}. \tag{3.4}$$

We proceed by solving the problem in a two field formulation. Given $\left(\boldsymbol{u}_h^{n-1}, p_h^{n-1}\right) \in \boldsymbol{V}_h \times Q_h$ find $(\boldsymbol{u}_h^n, p_h^n) \in \boldsymbol{V}_h \times Q_h$ such that for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and $q_h \in Q_h$

$$\begin{aligned} 2\mu\langle \boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\rangle + \lambda\langle \nabla\cdot\boldsymbol{u}_h^n, \nabla\cdot\boldsymbol{v}_h\rangle - \alpha\langle p_h^n, \nabla\cdot\boldsymbol{v}_h\rangle &= \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle, \qquad (3.5)\\ \frac{1}{M}\langle p_h^n - p_h^{n-1}, q_h\rangle + \alpha\langle \nabla\cdot\left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right), q_h\rangle & \\ +\tau\langle \kappa\nabla p_h^n, \nabla q_h\rangle - \tau\langle \kappa\rho\boldsymbol{g}, \nabla q_h\rangle &= \tau\langle S_f, q_h\rangle \qquad (3.6) \end{aligned}$$

where the functions $(\boldsymbol{u}_h^0, p_h^0)$ are obtained through the initial condition. Remark that this it is not a locally mass conservative discretization, as opposed to the mixed formulation used in for example [13].

We now introduce the fixed-stress splitting scheme [11, 9, 16, 13]. Denote by i the iteration index and let $L \geq 0$ be the stabilization parameter. Given

$\left(\boldsymbol{u}_h^{n-1}, p_h^{n-1}\right)$ and $\left(\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}\right)$ in $\boldsymbol{V}_h \times Q_h$ find $\left(\boldsymbol{u}_h^{n,i}, p_h^{n,i}\right) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned} 2\mu\langle\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right), \boldsymbol{\varepsilon}\left(\boldsymbol{v}_h\right)\rangle + \lambda\langle\nabla\cdot\boldsymbol{u}_h^{n,i}, \nabla\cdot\boldsymbol{v}_h\rangle - \alpha\langle p_h^{n,i}, \nabla\cdot\boldsymbol{v}_h\rangle &= \langle\boldsymbol{f}^n, \boldsymbol{v}_h\rangle, \quad (3.7)\\ \frac{1}{M}\langle p_h^{n,i} - p_h^{n-1}, q_h\rangle + \alpha\langle\nabla\cdot\left(\boldsymbol{u}_h^{n,i-1} - \boldsymbol{u}_h^{n-1}\right), q_h\rangle & \\ + L\langle p_h^{n,i} - p_h^{n,i-1}, q_h\rangle + \tau\langle\kappa\nabla p_h^{n,i}, \nabla q\rangle - \tau\langle\kappa\mathbf{g}\rho, \nabla q_h\rangle &= \tau\langle S_f, q_h\rangle \quad (3.8) \end{aligned}$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h, q_h \in Q_h$. We start the iterations with the solution at the last time step and the initial solution for the first time step. The system (3.7)–(3.8) is now solved decoupled starting with equation (3.8) and then we iterate between the two equations.

3.2 Algebraic approach

In this subsection we make a connection between the L-scheme and the fixed-stress splitting scheme. From the discrete problem (3.5)–(3.6) one could right away write the problem in matrix form and reduce it to a pure pressure formulation instead of applying the fixed-stress splitting scheme. As we look for solutions in the spaces $\boldsymbol{V}_h$ and Q_h, let $\{\boldsymbol{\varphi}_{h,i}^V\}_{i=1}^{d\cdot M}$ be a basis for the space $\boldsymbol{V}_h$ and $\{\varphi_{h,i}^Q\}_i^M$ be a basis for the space Q_h where M is the number of nodes and d is the spatial dimension. Now, let

$$\boldsymbol{u}_h^n = \sum_{i=1}^{d\cdot M} \eta_{h,i}^n \boldsymbol{\varphi}_{h,i}^V \quad \textit{and} \quad p_h^n = \sum_{i=1}^{M} \xi_{h,i}^n \varphi_{h,i}^Q \tag{3.9}$$

be the solutions to (3.5)–(3.6). Since equation (3.5) holds for any $\boldsymbol{v}_h \in \boldsymbol{V}_h$ it holds in particular for $\boldsymbol{v}_h = \boldsymbol{\varphi}_{h,j}^V$ for any of the basis functions of $\boldsymbol{V}_h$. The same goes for $q_h = \varphi_{h,j}^Q$ in equation (3.6). Making the substitutions in (3.9) into the equations (3.5)–(3.6) gives the equations

$$\begin{aligned} 2\mu\langle\boldsymbol{\varepsilon}\left(\textstyle\sum_{i=1}^{d\cdot M}\eta_{h,i}^n\boldsymbol{\varphi}_{h,i}^V\right), \boldsymbol{\varepsilon}\left(\boldsymbol{\varphi}_{h,j}^V\right)\rangle + \lambda\langle\nabla\cdot\textstyle\sum_{i=1}^{d\cdot M}\eta_{h,i}^n\boldsymbol{\varphi}_{h,i}^V, \nabla\cdot\boldsymbol{\varphi}_{h,j}^V\rangle & \\ -\alpha\langle\textstyle\sum_{i=1}^{M}\xi_{h,i}^n\varphi_{h,i}^Q, \nabla\cdot\boldsymbol{\varphi}_{h,j}^V\rangle = \langle\mathbf{f}^n, \boldsymbol{\varphi}_{h,j}^V\rangle, & \quad (3.10) \end{aligned}$$

$$\begin{aligned} \tfrac{1}{M}\langle\textstyle\sum_{i=1}^{M}\xi_{h,i}^n\varphi_{h,i}^Q - \sum_{i=1}^{M}\xi_{h,i}^{n-1}\varphi_{h,i}^Q, \varphi_{h,j}^Q\rangle & \\ +\alpha\langle\nabla\cdot\left(\textstyle\sum_{i=1}^{d\cdot M}\eta_{h,i}^n\boldsymbol{\varphi}_{h,i}^V - \sum_{i=1}^{d\cdot M}\eta_{h,i}^{n-1}\boldsymbol{\varphi}_{h,i}^V\right), \varphi_{h,j}^Q\rangle & \quad (3.11)\\ +\tau\langle\kappa\nabla\textstyle\sum_{i=1}^{M}\xi_{h,i}^n\varphi_{h,i}^Q, \nabla\varphi_{h,j}^Q\rangle - \tau\langle\kappa\mathbf{g}\rho, \nabla\varphi_{h,j}^Q\rangle = \tau\langle S_f, \varphi_{h,j}^Q\rangle. & \end{aligned}$$

Exploiting the linearity of $\boldsymbol{\varepsilon}(\cdot)$, the inner products and the derivatives we can write (3.10)–(3.11) in matrix form

$$\boldsymbol{A}\boldsymbol{\eta}_h^n - \alpha \boldsymbol{D}\boldsymbol{\xi}_h^n = \boldsymbol{l}_m^n \tag{3.12}$$

$$\frac{1}{M}\boldsymbol{M}_p\left(\boldsymbol{\xi}_h^n - \boldsymbol{\xi}_h^{n-1}\right) + \alpha \boldsymbol{D}^T\left(\boldsymbol{\eta}_h^n - \boldsymbol{\eta}_h^{n-1}\right) + \tau \boldsymbol{K}\xi_h^n = \boldsymbol{l}_f^n \tag{3.13}$$

where $[\boldsymbol{A}]_{i,j} = 2\mu\langle\boldsymbol{\varepsilon}\left(\boldsymbol{\varphi}_{h,i}^V\right), \boldsymbol{\varepsilon}\left(\boldsymbol{\varphi}_{h,j}^V\right)\rangle + \lambda\langle\nabla\cdot\boldsymbol{\varphi}_{h,i}^V, \nabla\cdot\boldsymbol{\varphi}_{h,j}^V\rangle$, $[\boldsymbol{D}]_{i,j} = \langle\nabla\cdot\boldsymbol{\varphi}_{h,i}^V, \varphi_{h,j}^Q\rangle$, $[\boldsymbol{l}_m^n]_j = \langle\mathbf{f}^n, \boldsymbol{\varphi}_{h,j}^V\rangle$, $[\boldsymbol{M}_p]_{i,j} = \langle\varphi_{h,i}^Q, \varphi_{h,j}^Q\rangle$ and $[\boldsymbol{l}_f^n]_j = \tau\langle S_f, \varphi_{h,j}^Q\rangle + \tau\langle\kappa\mathbf{g}\rho, \nabla\varphi_{h,j}^Q\rangle$. If we subtract (3.12) in time step $n-1$ from (3.12) in time step n and solve for $\boldsymbol{\eta}_h^n - \boldsymbol{\eta}_h^{n-1}$, we can substitute this into equation (3.13) and get the single equation

$$\frac{1}{M}\boldsymbol{M}_p\left(\boldsymbol{\xi}_h^n - \boldsymbol{\xi}_h^{n-1}\right) + \alpha^2\boldsymbol{D}^T\boldsymbol{A}^{-1}\boldsymbol{D}\left(\boldsymbol{\xi}_h^n - \boldsymbol{\xi}_h^{n-1}\right) + \tau\boldsymbol{K}\xi_h^n = \boldsymbol{l}_f^n + \boldsymbol{A}^{-1}\left(\boldsymbol{l}_f^n - \boldsymbol{l}_f^{n-1}\right). \tag{3.14}$$

To simplify this expression we write

$$\boldsymbol{B}\left(\boldsymbol{\xi}_h^n - \boldsymbol{\xi}_h^{n-1}\right) + \tau\boldsymbol{K}\xi_h^n = \boldsymbol{l}_t^n, \tag{3.15}$$

where $\boldsymbol{B} = \frac{1}{M}\boldsymbol{M}_p + \alpha^2\boldsymbol{D}^T\boldsymbol{A}^{-1}\boldsymbol{D}$ and $\boldsymbol{l}_t^n = \boldsymbol{l}_f^n + \boldsymbol{A}^{-1}\left(\boldsymbol{l}_f^n - \boldsymbol{l}_f^{n-1}\right)$.

3.2.1 L-scheme

We propose an L-scheme to solve (3.15). Let L be a positive real number and consider the iterative scheme

$$L\boldsymbol{M}_p\left(\boldsymbol{\xi}_h^{n,i} - \boldsymbol{\xi}_h^{n,i-1}\right) + \boldsymbol{B}\left(\boldsymbol{\xi}_h^{n,i-1} - \boldsymbol{\xi}_h^{n-1}\right) + \tau\boldsymbol{K}\boldsymbol{\xi}_h^{n,i} = \boldsymbol{l}_t^n \tag{3.16}$$

where $\boldsymbol{\xi}_h^{n,0} = \boldsymbol{\xi}_h^{n-1}$.

Proposition 3.2.1. *The L-scheme* (3.16) *is equivalent to a corresponding transformation to a matrix equation starting from equations* (3.7)–(3.8)*.*

3.2.2 Optimization as a fixed-point iteration

The equation (3.16) can be rewritten in the form

$$\boldsymbol{\xi}_h^{n,i} = \boldsymbol{\xi}_h^{n,i-1} + (L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}\left(\boldsymbol{l}_t^n + \boldsymbol{B}\boldsymbol{\xi}_h^{n-1} - (\boldsymbol{B} + \tau\boldsymbol{K})\,\boldsymbol{\xi}_h^{n,i-1}\right) \tag{3.17}$$

which is similar to the modified Richardson iteration,

$$\boldsymbol{x}^k = \boldsymbol{x}^{k-1} + \hat{\boldsymbol{A}}^{-1}\left(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{k-1}\right).$$

We can analyze the rate of convergence by looking at this as a fixed-point iteration and finding the contraction constant. In this spirit we define

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} + (L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}\left(\boldsymbol{l}_t^n + \boldsymbol{B}\boldsymbol{\xi}_h^{n-1} - (\boldsymbol{B} + \tau\boldsymbol{K})\,\boldsymbol{x}\right).$$

Simple computations give

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| = \|\left(\boldsymbol{I} - (L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}(\boldsymbol{B} + \tau\boldsymbol{K})\right)(\boldsymbol{x} - \boldsymbol{y})\|.$$

Following from the definition of the induced matrix norm we have

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| \leq \|\boldsymbol{I} - (L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}(\boldsymbol{B} + \tau\boldsymbol{K})\|\|\boldsymbol{x} - \boldsymbol{y}\|.$$

An application of the inequality $\|\boldsymbol{A}\| \leq \rho(A)$ gives

$$\begin{aligned}\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| &\leq \rho\left(\boldsymbol{I} - (L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}(\boldsymbol{B} + \tau\boldsymbol{K})\right)\|\boldsymbol{x} - \boldsymbol{y}\| \\ &= 1 - \lambda_{min}\left((L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}(\boldsymbol{B} + \tau\boldsymbol{K})\right)\|\boldsymbol{x} - \boldsymbol{y}\|.\end{aligned}$$

Unfortunately the minimization of

$$1 - \lambda_{min}\left((L\boldsymbol{M}_p + \tau\boldsymbol{K})^{-1}(\boldsymbol{B} + \tau\boldsymbol{K})\right)$$

with respect to L is a more difficult approach, and not as applicable, compared to what comes in the following sections, so we will not do anything about this last computation. It is however interesting to realize the fixed-stress splitting scheme as an L-scheme.

3.3 Convergence analysis

In this section we analyze the convergence of the scheme (3.7)–(3.8). We are in particular interested in finding an *optimal* stabilization parameter L, in the sense that the scheme requires the least amount of iterations. Before we proceed with the main result we need some preliminaries.

First, we require the inequality

$$2\mu\|\varepsilon(\boldsymbol{u})\|^2 + \lambda\|\nabla\cdot\boldsymbol{u}\|^2 \geq K_{dr}\|\nabla\cdot\boldsymbol{u}\|^2 \tag{3.18}$$

to hold for some $K_{dr} > 0$ for all $\boldsymbol{u} \in \boldsymbol{V}_h$. It does in particular hold for $K_{dr} = \frac{2\mu}{d} + \lambda$ which is often known as the drained bulk modulus, where d is the spatial dimension. The following argument proves this statement through the root square mean - arithmetic mean inequality,

$$\sum_{i=1}^{d}\frac{x_i^2}{d} \geq \left(\sum_{i=1}^{d}\frac{x_i}{d}\right)^2, \quad \forall x_i \in \mathbb{R},\ i = 1, 2, ..., d.$$

Simple computations give the inequality

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|^2 = 2\mu\int_\Omega\left(\sum_{i=1}^d\left(\frac{\partial u_i}{\partial x_i}\right)^2+\sum_{i\neq j,i,j=1}^d\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)^2\right)d\mathbf{x}$$
$$\geq\frac{2\mu}{d}\int_\Omega\left(\sum_{i=1}^d\frac{\partial u_i}{\partial x_i}\right)^2 d\mathbf{x}=\frac{2\mu}{d}\|\nabla\cdot\boldsymbol{u}\|^2$$

which proves that (3.18) holds when $K_{dr}=\frac{2\mu}{d}+\lambda$. This constant K_{dr} turns out to be important in the optimization of the stabilization parameter L. In practice, for effectively lower-dimensional situations, e.g. one-dimensional compression, d can be chosen smaller than the spatial dimension, as (3.18) is assumed to hold only for a relevant subset of displacements $\boldsymbol{u}$, cf. proof of Theorem 3.3.3. Consistent with the literature, despite the discrepancy between K_{dr} and the physically well-defined drained bulk modulus, we continue calling K_{dr} the drained bulk modulus independent of its value. For a detailed discussion on the values of K_{dr} see [15].

We further make the following assumptions for Theorem 3.3.3:

Assumption 6. *All the constants $\mu,\lambda,M,K_{dr},\kappa$ are strictly positive.*

Assumption 7. *The discretization, $\boldsymbol{V}_h\times Q_h$, is inf-sup stable.*

From Assumption 7 follows Lemma 3.3.2 by applying Corollary 4.1.1 in [20], which states:

Corollary 3.3.1. *Let V and Q be Hilbert spaces, and let B be a linear continuous operator from V to Q'. Then, the following statements are equivalent:*

- *B^t is bounding: $\exists\gamma>0$ such that $\|B^tq\|_{V'}\geq\gamma\|q\|_Q\ \forall q\in Q$*
- *$\exists L_B\in\mathcal{L}(Q',V)$ such that $B(L_B(g))=g\ \forall g\in Q'$ with $\|L_b\|=\dfrac{1}{\gamma}$*

Lemma 3.3.2. *Let Assumption 7 hold true. There exists $\beta>0$ such that for any $p_h\in Q_h$ there exists $\boldsymbol{u}_h\in\boldsymbol{V}_h$ satisfying $\langle\nabla\cdot\boldsymbol{u}_h,q_h\rangle=\langle p_h,q_h\rangle$ for all $q_h\in Q_h$ and*

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2+\lambda\|\nabla\cdot\boldsymbol{u}_h\|^2\leq\beta\|p_h\|^2. \tag{3.19}$$

Proof. Consider Corollary 3.3.1 and define a continuous linear function from $\boldsymbol{V}_h$ to Q_h' by $B(\boldsymbol{u}_h)(q_h)=\langle\nabla\cdot\boldsymbol{u}_h,q_h\rangle$. The first statement in Corollary 3.3.1 is the characterization of an inf-sup stable discretization, with stability constant γ. Considering the second statement we have the existence of a linear function $L_B\in$

$\mathcal{L}(Q_h', \boldsymbol{v}_h)$ such that $B(L_B(\langle p_h, \cdot\rangle)) = \langle p_h, \cdot\rangle$ for all $p_h \in Q_h$ with $\|L_B\| = 1/\gamma$. Hence L_B is giving for each $p_h \in Q_h$ the corresponding $\boldsymbol{u}_h \in \boldsymbol{V}_h$ such that

$$\langle \nabla \cdot \boldsymbol{u}_h, q_h\rangle = B(L_B(\langle p_h, q_h\rangle)) = \langle p_h, q_h\rangle$$

for all $q_h \in Q_h$. Now the following chain of inequalities holds true,

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2 + \lambda\|\nabla \cdot \boldsymbol{u}_h\|^2 \leq C\|\boldsymbol{u}_h\|^2_{H^1(\Omega)} \leq C\|L_B\|^2\|p_h\|^2$$

where the first one follows from Young's inequality with C depending only on the Lamé parameters, and the second inequality results from the operator norm,

$$\|L_B\| = \sup_{p_h \in Q_h, p_h \neq 0} \frac{\|L_B(\langle p_h, \cdot\rangle)\|_{H^1(\Omega)}}{\|\langle p_h, \cdot\rangle\|_{Q_h'}} = \sup_{p_h \in Q_h, p_h \neq 0} \frac{\|\boldsymbol{u}_h\|_{H^1(\Omega)}}{\|p_h\|_{Q_h}}$$

with $\|\cdot\| = \|\cdot\|_{Q_h}$. Then we have our desired inequality,

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2 + \lambda\|\nabla \cdot \boldsymbol{u}_h\|^2 \leq \frac{C}{\gamma^2}\|p_h\|^2 = \beta\|p_h\|^2.$$

□

Remark 10. *The constant C is dependent on both μ and λ while γ is dependent on the domain Ω and on the choice of the finite dimensional spaces $\mathbf{V}_h$ and Q_h. For more information see for example [57].*

Theorem 3.3.3. *Let Assumption 6 and Assumption 7 hold true and let $\delta \in (0, 2]$. Define the iteration errors as $\boldsymbol{e}_{\boldsymbol{u}}^{n,i} := \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^n$ and $e_p^{n,i} := p_h^{n,i} - p_h^n$ where $\boldsymbol{u}_h^{n,i}, p_h^{n,i}$ are solutions to (3.7)–(3.8) and $\boldsymbol{u}_h^n, p_h^n$ are solutions to (3.5)–(3.6). The fixed-stress splitting scheme (3.7)–(3.8) converges linearly for any $L \geq \frac{\alpha^2}{\delta K_{dr}}$, with a convergence rate given by*

$$\text{rate}(L, \delta) = \frac{L}{L + \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\frac{\alpha^2}{\beta}}, \tag{3.20}$$

through the error inequalities

$$\|e_p^{n,i}\|^2 \leq \text{rate}(L, \delta)\|e_p^{n,i-1}\|^2 \tag{3.21}$$

$$2\mu\|\varepsilon(\boldsymbol{e}_{\boldsymbol{u}}^{n,i})\|^2 + \lambda\|\nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}\|^2 \leq \frac{\alpha^2}{K_{dr}}\|e_p^{n,i}\|^2 \tag{3.22}$$

where C_Ω is the Poincaré constant and β is the constant from (3.19).

Proof. Subtract (3.7)–(3.8) from (3.5)–(3.6), respectively, to obtain the error equations for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and $q_h \in Q_h$

$$\begin{cases} (i) \;\; 2\mu\langle\varepsilon(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}),\varepsilon(\boldsymbol{v}_h)\rangle + \lambda\langle\nabla\cdot\boldsymbol{e}_{\boldsymbol{u}}^{n,i},\nabla\cdot\boldsymbol{v}_h\rangle - \alpha\langle e_p^{n,i},\nabla\cdot\boldsymbol{v}_h\rangle = 0 \\ (ii) \;\; \frac{1}{M}\langle e_p^{n,i},q_h\rangle + \alpha\langle\nabla\cdot\boldsymbol{e}_{\boldsymbol{u}}^{n,i-1},q_h\rangle + L\langle e_p^{n,i}-e_p^{n,i-1},q_h\rangle + \tau\langle\kappa\nabla e_p^{n,i},\nabla q_h\rangle = 0. \end{cases} \tag{3.23}$$

To prove (3.22) test (3.23)(i) with $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i}$, and apply the Cauchy Schwarz inequality and Young's inequality to the pressure term to obtain

$$2\mu\|\varepsilon(\boldsymbol{e}_{\boldsymbol{u}}^{n,i})\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\|^2 \leq \frac{\alpha^2}{2K_{dr}}\|e_p^{n,i}\|^2 + \frac{K_{dr}}{2}\|\nabla\cdot\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\|^2. \tag{3.24}$$

We now get (3.22) by applying (3.18).

In order to prove (3.21) test (3.23) with $q_h = e_p^{n,i}$ and $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i}$, add the resulting equations and use the algebraic identity

$$\langle e_p^{n,i}-e_p^{n,i-1},e_p^{n,i}\rangle = \frac{1}{2}\left(\|e_p^{n,i}-e_p^{n,i-1}\|^2 + \|e_p^{n,i}\|^2 - \|e_p^{n,i-1}\|^2\right)$$

to get

$$\begin{aligned} &2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2 - \alpha\langle e_p^{n,i},\nabla\cdot\boldsymbol{e}_u^{n,i}\rangle \\ &+\frac{1}{M}\|e_p^{n,i}\|^2 + \alpha\langle\nabla\cdot\boldsymbol{e}_u^{n,i-1},e_p^{n,i}\rangle + \tau\langle\kappa\nabla e_p^{n,i},e_p^{n,i}\rangle \\ &+\frac{L}{2}\|e_p^{n,i}-e_p^{n,i-1}\|^2 + \frac{L}{2}\|e_p^{n,i}\|^2 = \frac{L}{2}\|e_p^{n,i-1}\|^2. \end{aligned}$$

A rearrangement and an application of equation (3.23)(i) with $\boldsymbol{v}_h = \boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}$ yields

$$\begin{aligned} &2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2 + \frac{1}{M}\|e_p^{n,i}\|^2 + \tau\langle\kappa\nabla e_p^{n,i},e_p^{n,i}\rangle + \frac{L}{2}\|e_p^{n,i}\|^2 \\ &+\frac{L}{2}\|e_p^{n,i}-e_p^{n,i-1}\|^2 = \frac{L}{2}\|e_p^{n,i-1}\|^2 + 2\mu\langle\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right),\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}-\boldsymbol{e}_u^{n,i-1}\right)\rangle \\ &+\lambda\langle\nabla\cdot\boldsymbol{e}_u^{n,i},\nabla\cdot\left(\boldsymbol{e}_u^{n,i}-\boldsymbol{e}_u^{n,i-1}\right)\rangle. \end{aligned} \tag{3.25}$$

From Young's inequality we have that for any $\delta > 0$

$$\begin{aligned} &2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2 + \frac{1}{M}\|e_p^{n,i}\|^2 + \tau\langle\kappa\nabla e_p^{n,i},e_p^{n,i}\rangle + \frac{L}{2}\|e_p^{n,i}\|^2 \\ &+\frac{L}{2}\|e_p^{n,i}-e_p^{n,i-1}\|^2 = \frac{L}{2}\|e_p^{n,i-1}\|^2 + \frac{\delta}{2}\left(2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2\right) \\ &+\frac{1}{2\delta}\left(2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}-\boldsymbol{e}_u^{n,i-1}\right)\|^2 + \lambda\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i}-\boldsymbol{e}_u^{n,i-1}\right)\|^2\right). \end{aligned} \tag{3.26}$$

To take care of the last term in (3.26) consider equation (3.23)(i). Subtract iteration $i-1$ from iteration i and let $\boldsymbol{v}_h = \boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}$

$$\begin{aligned} 2\mu\langle\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i-1}\right), \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i-1}\right)\rangle \\ + \lambda\langle\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right), \nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\rangle \\ = \alpha\langle e_p^{n,i} - e_p^{n,i-1}, \nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\rangle \end{aligned}$$

or simply

$$\begin{aligned} 2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i-1}\right)\|^2 + \lambda\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\|^2 \\ = \alpha\langle e_p^{n,i} - e_p^{n,i-1}, \nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\rangle. \end{aligned}$$

The Cauchy-Schwarz inequality implies

$$\begin{aligned} 2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i-1}\right)\|^2 + \lambda\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\|^2 \\ \leq \alpha\|e_p^{n,i} - e_p^{n,i-1}\|\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\|. \end{aligned} \tag{3.27}$$

Finally equation (3.18) gives

$$K_{dr}\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\| \leq \alpha\|e_p^{n,i} - e_p^{n,i-1}\|. \tag{3.28}$$

In equation (3.27) apply inequality (3.28) to obtain

$$2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i-1}\right)\|^2 + \lambda\|\nabla\cdot\left(\boldsymbol{e}_u^{n,i} - \boldsymbol{e}_u^{n,i-1}\right)\|^2 \leq \frac{\alpha^2}{K_{dr}}\|e_p^{n,i} - e_p^{n,i-1}\|^2. \tag{3.29}$$

An application of (3.29) to (3.26) yields

$$\begin{aligned} \left(1-\frac{\delta}{2}\right)\left(2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2\right) + \frac{1}{M}\|e_p^{n,i}\|^2 \\ + \tau\langle\kappa\nabla e_p^{n,i}, e_p^{n,i}\rangle + \frac{L}{2}\|e_p^{n,i}\|^2 + \frac{L}{2}\|e_p^{n,i} - e_p^{n,i-1}\|^2 \\ \leq \frac{L}{2}\|e_p^{n,i-1}\|^2 + \frac{\alpha^2}{2K_{dr}\delta}\|e_p^{n,i} - e_p^{n,i-1}\|^2. \end{aligned}$$

Under the assumption that $L \geq \dfrac{\alpha^2}{K_{dr}\delta}$ we get

$$\begin{aligned} \left(1-\frac{\delta}{2}\right)\left(2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla\cdot\boldsymbol{e}_u^{n,i}\|^2\right) \\ + \left(\frac{1}{M} + \frac{L}{2} + \frac{\tau\kappa}{C_\Omega}\right)\|e_p^{n,i}\|^2 \leq \frac{L}{2}\|e_p^{n,i-1}\|^2. \end{aligned} \tag{3.30}$$

by applying Poincaré's inequality. In previous works, e.g. [13], the conclusion at this point was that $L = \dfrac{\alpha^2}{2K_{dr}}$ is the optimal parameter. However, this does not consider the influence of the first term in (3.30).

By Lemma 3.3.2 we get that there exists $\hat{\boldsymbol{v}}_h \in V_h$ such that $\langle e_p^{n,i}, q_h\rangle = \langle \nabla \cdot \hat{\boldsymbol{v}}_h, q_h\rangle$ for all $q_h \in Q_h$ and

$$2\mu\|\varepsilon\left(\hat{\boldsymbol{v}}_h\right)\|^2 + \lambda\|\nabla \cdot \hat{\boldsymbol{v}}_h\|^2 \leq \beta\|e_p^{n,i}\|^2. \tag{3.31}$$

Considering equation (3.23)(i) with $\boldsymbol{v}_h = \hat{\boldsymbol{v}}_h$ we get

$$\alpha\|e_p^{n,i}\|^2 = 2\mu\langle \boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right), \boldsymbol{\varepsilon}\left(\hat{\boldsymbol{v}}_h\right)\rangle + \lambda\langle \nabla \cdot \boldsymbol{e}_u^{n,i}, \nabla \cdot \hat{\boldsymbol{v}}_h\rangle \tag{3.32}$$

which by the Cauchy-Schwarz inequality and neglecting some terms becomes

$$\alpha\|e_p^{n,i}\|^2 \leq \left(2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla \cdot \boldsymbol{e}_u^{n,i}\|^2\right)^{\frac{1}{2}} \left(2\mu\|\boldsymbol{\varepsilon}\left(\hat{\boldsymbol{v}}_h\right)\|^2 + \lambda\|\nabla \cdot \hat{\boldsymbol{v}}_h\|^2\right)^{\frac{1}{2}}. \tag{3.33}$$

Inequality (3.31) and rearrangements give us

$$\frac{\alpha^2}{\beta}\|e_p^{n,i}\|^2 \leq 2\mu\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_u^{n,i}\right)\|^2 + \lambda\|\nabla \cdot \boldsymbol{e}_u^{n,i}\|^2 \tag{3.34}$$

which we can apply in inequality (3.30) to obtain

$$\left(\frac{1}{M} + \frac{L}{2} + \frac{\tau\kappa}{C_\Omega} + \left(1 - \frac{\delta}{2}\right)\left(\frac{\alpha^2}{\beta}\right)\right)\|e_p^{n,i}\|^2 \leq \frac{L}{2}\|e_p^{n,i-1}\|^2. \tag{3.35}$$

This gives the rate of change

$$rate\left(L, \delta\right) = \frac{L}{L + \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2 - \delta)\left(\frac{\alpha^2}{\beta}\right)}. \tag{3.36}$$

for all $\delta \in (0, 2]$. □

Remark 11. *One can easily extend the result for a heterogeneous media, i.e.* $\kappa = \kappa\left(\mathbf{x}\right)$ *as long as* κ *is bounded from below by a constant* $\kappa_m > 0$*. Also any of the other parameters can be chosen spatially dependent as long as they are bounded by appropriate constants larger than zero.*

3.4 Optimality

Given the rate obtained through Theorem 3.3.3 our next goal will be to minimize it. It follows directly that L should be as small as possible, while still satisfying the

inequality $L \geq \frac{\alpha^2}{\delta K_{dr}}$. Hence, in all instances where L appears in equation (3.20) substitute L with $\frac{\alpha^2}{\delta K_{dr}}$. Then the rate function becomes

$$rate\,(\delta) = \frac{\frac{\alpha^2}{K_{dr}}}{\frac{\alpha^2}{K_{dr}} + \delta\left(\frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\left(\frac{\alpha^2}{\beta}\right)\right)}. \tag{3.37}$$

Minimizing (3.37) with respect to δ corresponds to maximizing

$$\delta\left(\frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\left(\frac{\alpha^2}{\beta}\right)\right).$$

Let $A = \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + 2\frac{\alpha^2}{\beta}$ and $B = \frac{\alpha^2}{\beta}$ and we can see that the maximum of $\delta\,(A - \delta B)$ lies at $\delta = \frac{A}{2B}$. Note that $\frac{A}{2B} > 1$. As δ belongs to $(0, 2]$ the optimal choice of δ is

$$\delta_{opt} = min\left\{\frac{A}{2B}, 2\right\} \in (1, 2]\,. \tag{3.38}$$

This implies the optimal choice for L is

$$L = \frac{\alpha^2}{K_{dr}\ \min\left\{\frac{A}{2B}, 2\right\}} \in \left[\frac{\alpha^2}{2K_{dr}}, \frac{\alpha^2}{K_{dr}}\right] = \left[\frac{L_{phys}}{2}, L_{phys}\right]. \tag{3.39}$$

Remark 12 (Extremal cases)**.** *We consider the two extremal cases, $\delta_{opt} = 1$ and $\delta_{opt} = 2$.*

- *When M is large and $\tau\kappa$ is small we are in the case $\delta_{opt} \sim 1$, which gives $L = \frac{\alpha^2}{K_{dr}} = L_{phys}$.*
- *When M is small or $\tau\kappa$ is large we are in the case $\delta_{opt} = 2$, which gives $L = \frac{\alpha^2}{2K_{dr}} = \frac{L_{phys}}{2}$.*

Remark 13 (Consequence for low-compressible and low-permeable porous media)**.** *Previous convergence results in the literature for the fixed-stress splitting scheme have not predicted or guaranteed any convergence in the limit case $M \to \infty$ and $\kappa \to 0$. However, by Theorem 3.3.3, for inf-sup stable discretizations, convergence of the fixed-stress splitting scheme is guaranteed, even in the limit case.*

3.5 Numerical examples

In this section we verify numerically the theoretical results of Theorem 3.3.3. In particular we show that for constant material properties, the practical optimal value of δ increases for increasing permeability, κ, as the theory predicts. We also emphasize that this does not hold for not inf-sup stable discretizations, e.g. P1-P1.

Three test cases are considered:

1. An experiment in the unit square domain with source terms giving parabolas as analytical solution to the continuous problem, (3.1)–(3.2), for homogeneous Dirichlet boundary conditions.

2. An L-shaped domain with source terms from test case 1.

3. Mandel's problem.

We are using a MATLAB code for solving the problem in a two field formulation both in a P2-P1 stable discretization and in a P1-P1 not inf-sup stable discretization, see Section 1.2.6.

In all the plots we consider several permeabilities, κ. For each of them we solve (3.7)–(3.8) with a range of stabilization parameters $L = \frac{\alpha^2}{\delta K_{dr}}$. This is visualized through plots showing total numbers of iterations in the y-axis and δ in the x-axis. The domain of δ is varying slightly over the different test cases, but always contains the interval $(1, 2]$ which the theory predicts to contain the optimal value through subsection 3.4. The stars in each plot denote the theoretically calculated optimal value of δ from (3.38).

As stopping criterion we apply the relative errors in infinity-norm, $\frac{\|\boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n,i-1}\|_\infty}{\|\boldsymbol{u}_h^{n,i}\|_\infty} < \epsilon_{\boldsymbol{u},r}$ and $\frac{\|p_h^{n,i} - p_h^{n,i-1}\|_\infty}{\|p_h^{n,i}\|_\infty} < \epsilon_{p,r}$ where $\epsilon_{\boldsymbol{u},r}$ and $\epsilon_{p,r}$ are defined separately for the different test cases.

Remark 14 (Choice of K_{dr})**.** *If one knows the drained bulk modulus, K_{dr}, choosing the optimal stabilization parameter should be possible. However, as already mentioned in Section 3.3, this is problem dependent; finding the correct one might not be trivial. For our computations, we choose K_{dr} so that the theoretical optimal stabilization parameter is actually the practical optimal one for the smallest considered permeability. We experience that it also fits quite nicely for the remaining permeabilities for that particular setup. For all problems we set $\beta = K_{dr}$. However, we stress that this actually is not a realistic choice of β, which in reality is larger than K_{dr}.*

3.5.1 Unit square domain

In this test case we consider two setups on a unit square domain. For the first setup we apply homogeneous Dirichlet boundary conditions and zero initial data for both displacement and pressure. We employ source terms corresponding to the analytical solution of the continuous problem

$$u_1(x, y, t) = u_2(x, y, t) = \tfrac{1}{p_{\text{ref}}} p(x, y, t) = txy(1 - x)(1 - y),$$

$$(x, y) \in (0, 1)^2,\ t \in (0, 0.1),$$

regardless of permeability, Lamé parameters and the Biot-Willis constant, see Table 3.1. The pressure, p, is scaled by $p_{\text{ref}} = 10^{11}$ in order to balance the orders of magnitude of the mechanical and fluid stresses for the chosen physical parameters. In the second setup we keep the initial data and source terms from the first setup while assigning homogeneous Dirichlet boundary conditions for the displacement everywhere but at the top, $\Gamma_N = (0,1) \times \{1\}$, where homogeneous natural boundary conditions are applied. For the pressure homogeneous boundary conditions are applied on the entire boundary. For both setups, we compute one single time step from 0 to 0.1, and discretize the domain using a regular triangular mesh with mesh size $h = 1/8$. Numerical tests have showed that multiple time steps and different mesh diameters yield similar performance results. The tolerances $\epsilon_{\boldsymbol{u},r}$ and $\epsilon_{p,r}$ are set to 10^{-12}. Solutions for both setups are plotted for $\kappa = 10^{-10}$ in Figure 3.1. To summarize, we have

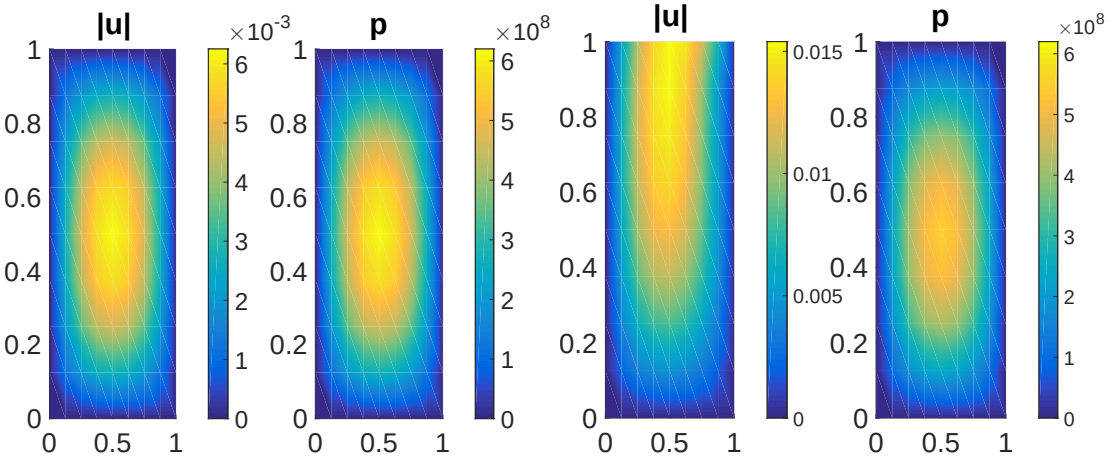
- Setup 1: Homogeneous Dirichlet data on the entire boundary for displacement and pressure.
- Setup 2: Homogeneous Dirichlet data for the pressure. Homogeneous Neumann data on top in the mechanics equation, homogeneous Dirichlet data everywhere else for the displacement.

The drained bulk modulus is set to $K_{dr} = 1.6\mu + \lambda$ for setup 1 and $K_{dr} = 1.1\mu + \lambda$ for setup 2.

Symbol	Name	Value
λ	Lamé parameter 1	$27.778 \cdot 10^9$
μ	Lamé parameter 2	$41.667 \cdot 10^9$
κ	Permeability	$10^{-15}, 10^{-14}, ..., 10^{-10}$
M	Compressibility coefficient	10^{11}
α	Biot-Willis constant	1
u_0, p_0	Initial data	0
h	Mesh diameter	$\frac{1}{8}$
τ	Time step size	0.1
t_0	Initial time	0
T	Final time	0.1
$\epsilon_{\boldsymbol{u},r}$ and $\epsilon_{p,r}$	Tolerances	10^{-12}

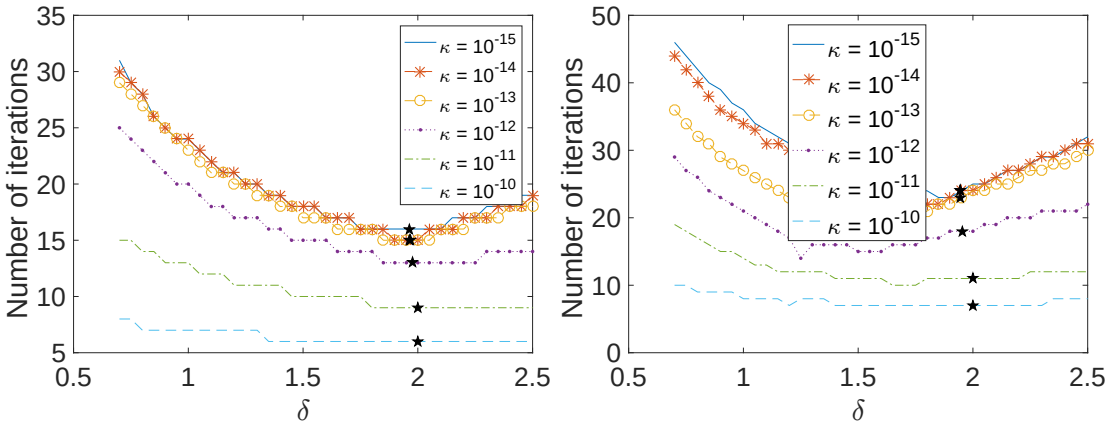
Table 3.1: Coefficients for test case 1 and 2

We experience for the inf-sup stable discretizations, Figure 3.2a and 3.3a, that as κ increases so does the optimal δ which is in accordance with Theorem 3.3.3.



(a) Setup 1 (b) Setup 2

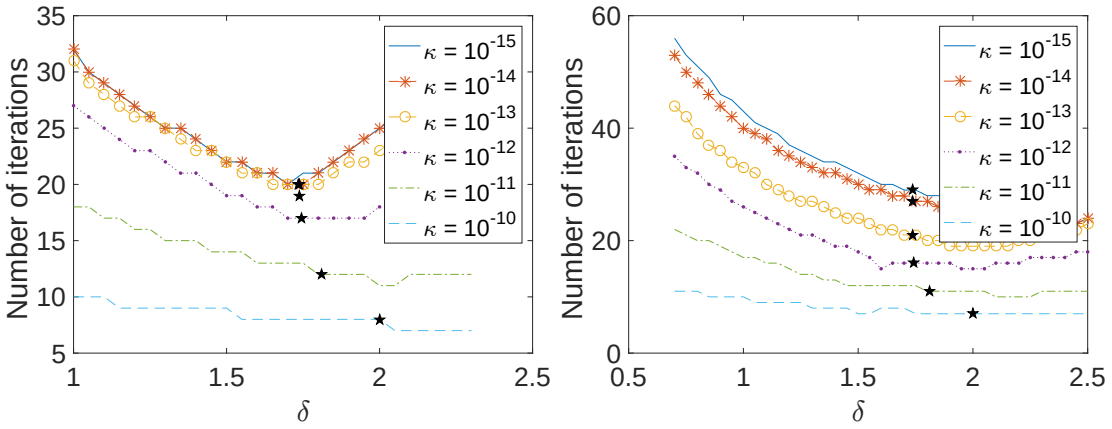
Figure 3.1: Displacement (Left) and Pressure (Right) for test case 1 with $\kappa = 10^{-10}$ at time step $t = 0.1$



(a) P2-P1 discretization (b) P1-P1 discretization

Figure 3.2: Test case 1 (Unit square domain) setup 1: Total iteration count for one time step applying stabilization parameter $L = \frac{\alpha^2}{\delta K_{dr}}$ with $K_{dr} = 1.6\mu + \lambda$. The star represents the theoretically calculated optimal δ.

However, when we have a not inf-sup stable discretization, Figure 3.2b and 3.3b, the behavior does not follow the same trend. In particular, we observe that for the first three permeability values, $\kappa = 10^{-15}$, $\kappa = 10^{-14}$ and $\kappa = 10^{-13}$, the optimal stabilization parameter is moving in the opposite direction to the theoretically calculated one. This is due to the instability that occurs when the diffusion term becomes of insignificant magnitude in the P1-P1 discretization.



(a) P2-P1 discretization (b) P1-P1 discretization

Figure 3.3: Test case 1 (Unit square domain) setup 2: Total iteration count for one time step applying stabilization parameter $L = \frac{\alpha^2}{\delta K_{dr}}$ with $K_{dr} = 1.1\mu + \lambda$. The star represents the theoretically calculated optimal δ.

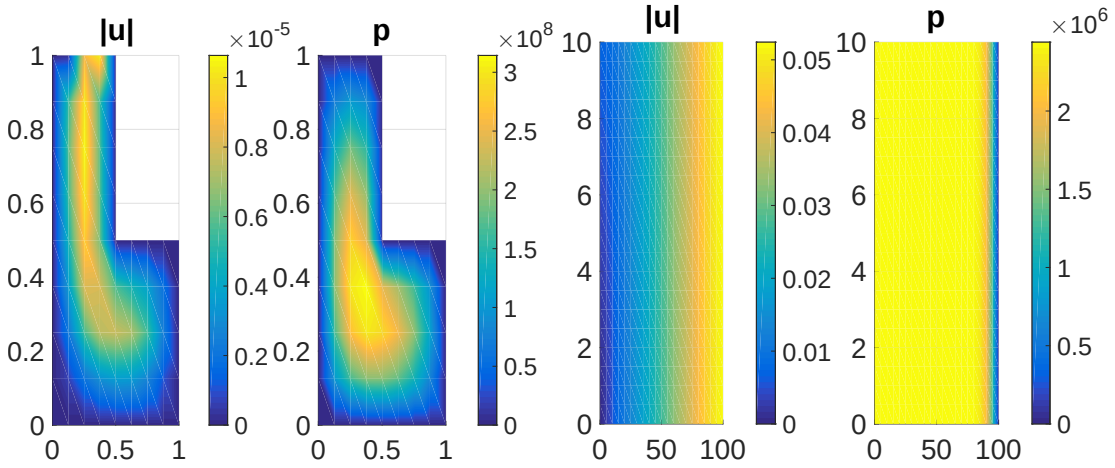
3.5.2 L-shaped domain

For this test case we consider an L-shaped domain with edges, $\Gamma_1 = \{0\} \times [0, 1]$, $\Gamma_2 = [0, 1] \times \{0\}$, $\Gamma_3 = \{1\} \times [0, 0.5]$, $\Gamma_4 = [0.5, 1] \times \{0.5\}$, $\Gamma_5 = \{0.5\} \times [0.5, 1]$ and $\Gamma_6 = (0, 0.5) \times \{1\}$. We are considering the same source terms and apply the same parameters, spatial and temporal discretization, initial data and stopping criterion as in test case 1, see Table 3.1. Similar to setup 2 above, for the pressure homogeneous Dirichlet boundary conditions are applied on the entire boundary, and for the displacement, homogeneous Dirichlet boundary conditions are considered everywhere except at the top, Γ_6. On the top, we apply zero Neumann boundary conditions in the mechanics equation, (3.1). The solution for $\kappa = 10^{-10}$ is displayed in Figure 3.4a. For the computations, we set $K_{dr} = 1.4\mu + \lambda$.

Again, for the stable discretization, Figure 3.5a, we observe that as the permeability increases so does the optimal choice of δ. For the not inf-sup stable discretization, Figure 3.5b, however, we experience that the optimal choice lies outside the theoretical interval of $(1, 2]$. And again the trend is inconsistent with the theory in the sense that the optimal δ is not increasing with increasing κ.

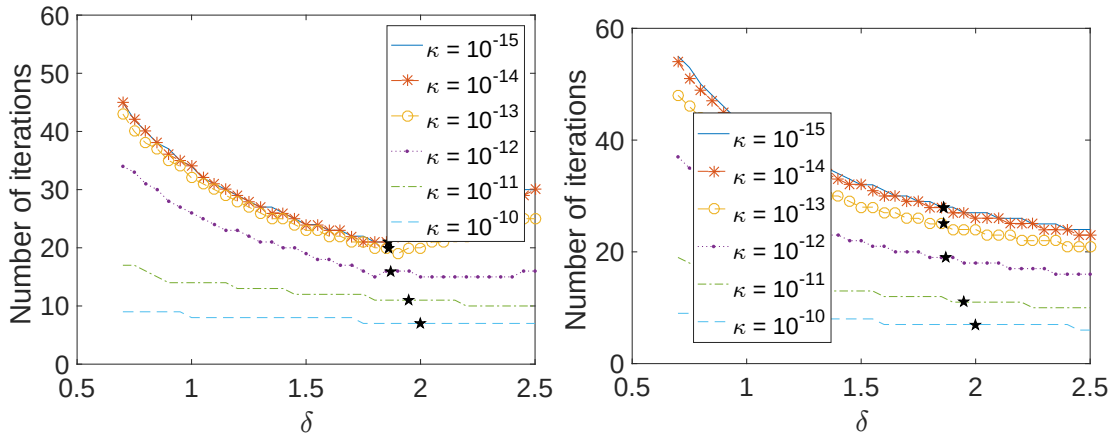
3.5.3 Mandel's Problem

In this section we consider Mandel's problem, a relevant 2D problem with known analytical solution that is derived in [58, 8]. The problem is often used as a benchmark problem for discretizations. The analytical expressions for pressure



(a) Test case 2, L-shaped domain (b) Test case 3, Mandel's problem

Figure 3.4: Displacement (Left) and Pressure (Right) for test case 2 and 3



(a) P2-P1 discretization (b) P1-P1 discretization

Figure 3.5: Test case 2 (L-shaped domain): Total iteration count for one time step applying stabilization parameter $L = \frac{\alpha^2}{\delta K_{dr}}$ with $K_{dr} = 1.4\mu + \lambda$. The star represents the theoretically calculated optimal δ.

and displacement are given by

$$p = \frac{2FB(1+\nu_u)}{3a} \sum_{n=1}^{\infty} \frac{sin(\alpha_n)}{\alpha_n - sin(\alpha_n)cos(\alpha_n)} \left(cos\left(\frac{\alpha_n x}{a}\right) - cos\left(\alpha_n\right)\right) e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \quad (3.40)$$

$$u_x = \left[\frac{F\nu}{2\mu a} - \frac{F\nu_u}{\mu a} \sum_{n=1}^{\infty} \frac{sin(\alpha_n)cos(\alpha_n)}{\alpha_n - sin(\alpha_n)cos(\alpha_n)} e^{-\frac{\alpha_n^2 c_f t}{a^2}}\right] x + \frac{F}{\mu} \sum_{n=1}^{\infty} \frac{cos(\alpha_n)}{\alpha_n - sin(\alpha_n)cos(\alpha_n)} sin\left(\frac{\alpha_n x}{a}\right) e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \quad (3.41)$$

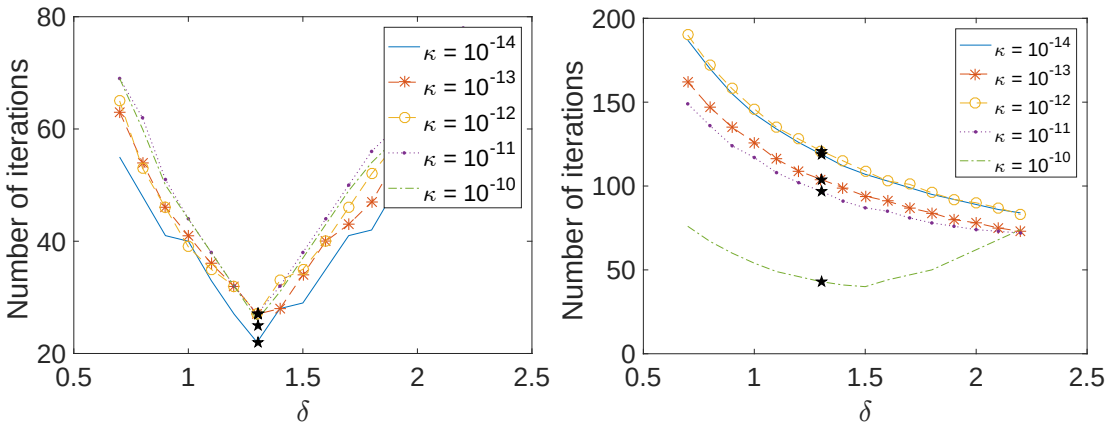
$$u_y = \left[\frac{-F(1-\nu)}{2\mu a} + \frac{F(1-\nu_u)}{\mu a} \sum_{n=1}^{\infty} \frac{sin(\alpha_n)cos(\alpha_n)}{\alpha_n - sin(\alpha_n)cos(\alpha_n)} e^{-\frac{\alpha_n^2 c_f t}{a^2}}\right] y, \quad (3.42)$$

where α_n, $n \in \mathbb{N}$, correspond to the positive solutions of the equation

$$tan\,(\alpha_n) = \frac{1-\nu}{\nu_u - \nu}\alpha_n,$$

and ν_u, F, B, c_f and a are input parameters, partially depending on the physical problem parameters. Here, we employ the values listed in Table 3.2, also used in [29]. For a thorough explanation of the problem and the coefficients in (3.40)–(3.42) we refer to [8, 29].

We consider the domain, $\Omega = (0, 100) \times (0, 10)$, discretized by a regular triangular mesh with mesh sizes dx = 5 and dy = 0.5. An equidistant partition of the time interval is applied with time step size $\tau = 10$ from $t_0 = 0$ to $T = 50$. Initial conditions are inherited from the analytic solutions, (3.40)–(3.42). As boundary conditions, we apply exact Dirichlet boundary conditions for the normal displacement on the top, left and bottom boundary. For the pressure, we apply homogeneous boundary conditions on the right boundary. On the remaining boundaries homogeneous, natural boundary conditions are applied. The tolerances, $\epsilon_{\boldsymbol{u},r}$ and $\epsilon_{p,r}$, are set to 10^{-6}. Our approximated solution for $\kappa = 10^{-10}$ is displayed in Figure 3.4b.



(a) P2-P1 discretization (b) P1-P1 discretization

Figure 3.6: Test case 3 (Mandel's problem): Total iteration count for five time steps applying stabilization parameter $L = \frac{\alpha^2}{\delta K_{dr}}$ with $K_{dr} = 1.35\mu + \lambda$. The star represents the theoretically calculated optimal δ.

Exactly as the theory predicts we observe that there is a fixed minimum for all the different permeabilities for the stable discretization, see Figure 3.6a. For the unstable discretization, Figure 3.6b, however, we experience the same non-monotonic behavior as before. There is also a clear difference in performance for the two discretizations. The inf-sup stable one performs much better, in terms of number of iterations. This is consistent with remark 13.

Symbol	Name	Value
λ	Lamé parameter 1	$1.650 \cdot 10^9$
μ	Lamé parameter 2	$2.475 \cdot 10^9$
ν	Poisson's ratio	0.2
B	Skempton coefficient	0.833
ν_u	Undrained Poisson's ratio	0.44
F	Applied force	$6 \cdot 10^8$
α	Biot-Willis constant	1
M	Compressibility coefficient	$1.650 \cdot 10^{10}$
c_f	Fluid diffusivity constant	0.47
κ	Permeability	$10^{-14}, 10^{-13}, ..., 10^{-10}$
a	Width of domain	100
b	Height of domain	10
dx	Horizontal mesh diameter	5
dy	Vertical mesh diameter	0.5
τ	Time step size	10
t_0	Initial time	0
T	Final time	50
$\epsilon_{\boldsymbol{u},r}$ and $\epsilon_{p,r}$	Tolerances	10^{-12}

Table 3.2: Coefficients for test case 3 (Mandel's Problem)

3.6 Conclusions

In this chapter we have considered the quasi-static, linear Biot model for poromechanics and studied theoretically and numerically the convergence of the fixed-stress splitting scheme. We have determined a formula for computing the optimal stabilization/tuning parameter, $L \in [L_{phys}/2, L_{phys}]$, depending also on the fluid flow properties and not only on the mechanics and the coupling term. Moreover, we identified limit cases when the physical parameter L_{phys} is the optimal one and cases when $L_{phys}/2$ should be taken. Both the numerical examples and the provided theory for inf-sup stable discretizations show that for increasing permeabilities the optimal stabilization parameter is decreasing (δ is increasing).

Furthermore, we have showed that the performance of the fixed-stress splitting scheme can be altered by a not inf-sup stable discretization in the sense that the scheme performs worse (more iterations), and the trend that increasing permeability implies decreasing stabilization parameter does not hold. Illustrative numerical examples have been performed including a well-known benchmark problem, Mandel's problem.

Chapter 4

Summary

The thesis started with a chapter providing the theoretical foundations for the next chapters. Section 1.1 considered iterative linearization schemes and splitting schemes, introducing theory concerning their convergence. The following section, Section 1.2, discussed discretization of PDEs with respect to the finite element method. The variational problems were defined and some theory on Sobolev spaces and the Lax-Milgram theorem were stated to deal with their solutions' existence and uniqueness. In Section 1.3 Darcy's law and mass balance equation were introduced for flow in porous media. General two-phase flow was discussed and a special case, Richards' equation, was derived. The section ended with an introduction to the Biot equations.

In Chapter 2 the work concerned the Richards equation. It was discretized in space by conforming finite elements and in time by the implicit Euler method. As the equation contains non-linearities different linearization schemes were discussed. In particular, the convergence of the L-scheme was proved using a similar technique to that in [5] for both constant and non-linear permeability. A theoretical optimal choice of the stabilization parameter L was discussed by minimizing the rate of convergence, and formulas were proposed for both constant and non-linear permeability. This theoretical choice of the optimal stabilization parameter L was tested in several numerical experiments. In these experiments the practical and theoretical optimal stabilization parameter do not coincide. Moreover, the theoretical and practical optimal stabilization parameter move in opposite directions for varying constant permeabilities showing that the theory is not sound yet. A numerical comparative study of different linearization schemes was performed in which the Newton-Raphson method showed the fastest convergence of all the schemes. However, when considering constant permeability we have a numerical advantage when applying the L-scheme in that we do not have to update the stiffness matrix every iteration, and as the L-scheme did not require many more iterations than the Newton-Raphson it is certainly competitive. For non-linear

permeabilities we lose this advantage, but the robustness of the L-scheme still makes it a strong alternative to the Newton-Raphson. In particular, we experienced that the Newton-Raphson method diverged for too large time step sizes when applying a case of the Van Genuchten-Mualem non-linearities. In contrast, linearly convergent schemes as the L-scheme did converge.

The final chapter, Chapter 3, considers the quasi static linear Biot equations which model flow in deformable porous media. After discretizing them in space by conforming finite elements and in time by the implicit Euler method we applied the fixed-stress splitting scheme [11]. Convergence was proved using the same techniques as in [13], but now, by going further, we obtained a theoretical optimal stabilization parameter by minimizing the rate of convergence. This was done under the assumption of an inf-sup stable discretization. We tested the theoretical results numerically applying both inf-sup stable (P2-P1) and not inf-sup stable (P1-P1) discretizations for several test cases, including the well-known benchmark problem, Mandel's problem. For the inf-sup stable discretization the theory coincides with the numerical experiments in the sense that for increasing permeability we experience decreasing optimal stabilization parameter both theoretically and practically. Additionally, when choosing the "mathematical" drained bulk modulus (see Remark 14) so that the theoretical and practical optimal stabilization parameter coincide for one permeability, the choices coincide also for the other permeabilities in all the test cases. However, when applying a not inf-sup stable discretization neither the trend nor the coinciding stabilization parameters hold true. Moreover, the fixed-stress splitting scheme converges in fewer iterations for the P2-P1 stable discretization than for the P1-P1 not inf-sup stable discretization for all the test cases. To our knowledge, the connection of robust performance of the fixed-stress splitting scheme and inf-sup stability of the underlying discretization has not been reported in the literature, yet.

Bibliography

[1] L.A. Richards. CAPILLARY CONDUCTION OF LIQUIDS THROUGH POROUS MEDIUMS. *Physics*, 1(5):318–333, 1931.

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

[3] F. Lehmann and P.H. Ackerer. Comparison of Iterative Methods for Improved Solutions of the Fluid Flow Equation in Partially Saturated Porous Media. *Transport in Porous Media*, 31(3):275–292, 1998.

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

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

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

[7] K. Mitra and I.S. Pop. A modified L-scheme to solve nonlinear diffusion problems. *Computers & Mathematics with Applications*, 2018. https://doi.org/10.1016/j.camwa.2018.09.042.

[8] O. Coussy. *Poromechanics*. John Wiley & Sons, 2004.

[9] J. Kim, H.A. Tchelepi, and R. Juanes. Stability and convergence of sequential methods for coupled flow and geomechanics: Fixed-stress and fixed-strain splits. *Computer Methods in Applied Mechanics and Engineering*, 200(13):1591 – 1606, 2011.

[10] J. Kim, H.A. Tchelepi, and R. Juanes. Stability and convergence of sequential methods for coupled flow and geomechanics: Drained and undrained splits.

Computer Methods in Applied Mechanics and Engineering, 200(23):2094 – 2116, 2011.

[11] A. Settari and F.M. Mourits. A Coupled Reservoir and Geomechanical Simulation System. *Society of Petroleum Engineers*, 3:219 – 226, 1998.

[12] A. Mikelić and M.F. Wheeler. Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*, 17(3):455–461, 2013.

[13] J.W. Both, M. Borregales, J.M. Nordbotten, K. Kumar, and F.A. Radu. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Applied Mathematics Letters*, 68:101 – 108, 2017.

[14] M. Bause, F.A. Radu, and U. Köcher. Space-time finite element approximation of the Biot poroelasticity system with iterative coupling. *Computer Methods in Applied Mechanics and Engineering*, 320:745 – 768, 2017.

[15] J.W. Both and U. Köcher. Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter. arXiv:1801.08352, 2018.

[16] A. Mikelić, B. Wang, and M.F. Wheeler. Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*, 18(3):325–341, 2014.

[17] L. Angermann and P. Knabner. *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*. Springer, 2003.

[18] W. Cheney. *Analysis for applied mathematics*. Springer, 2001.

[19] R.A Adams and J.J.F. Fournier. *Sobolev Spaces 2nd Edition*, volume 140. Academic Press, 2003.

[20] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44. Springer, 2013.

[21] F.A. Radu. MAT254 - Flow in Porous Media. Course at the University of Bergen, https://people.uib.no/fra001/radu/files/curriculum_MAT254.pdf, 2017.

[22] J.M. Nordbotten and M.A. Celia. *Geological storage of CO2: modeling approaches for large-scale simulation*. John Wiley & Sons, 2011.

[23] H. Darcy. *Les fontaines publiques de la ville de Dijon: exposition et application...* Victor Dalmont, 1856.

[24] M.T. Van Genuchten. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. *Soil science society of America journal*, 44(5):892–898, 1980.

[25] R. Brooks and T. Corey. Hydraulic properties of porous media. *Hydrology Papers, Colorado State University*, 24:37, 1964.

[26] M. Bause and P. Knabner. Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods. *Advances in Water Resources*, 27(6):565 – 581, 2004.

[27] J.W. Both, K. Kumar, J.M. Nordbotten, I.S. Pop, and F.A. Radu. Linear iterative schemes for doubly degenerate parabolic equations. arXiv:1801.00846v1, 2018.

[28] J.M. Nordbotten. Stable Cell-Centered Finite Volume Discretization for Biot Equations. *SIAM Journal on Numerical Analysis*, 54(2):942–968, 2016.

[29] P. J. Phillips and M.F. Wheeler. A coupling of mixed and continuous Galerkin finite element methods for poroelasticity i: the continuous in time case. *Computational Geosciences*, 11(2):131, 2007.

[30] S. Yi and M.L. Bean. Iteratively coupled solution strategies for a four-field mixed finite element method for poroelasticity. *International Journal for Numerical and Analytical Methods in Geomechanics*, 41(2):159–179, 2016.

[31] L. Berger, R. Bordas, D. Kay, and S. Tavener. A stabilized finite element method for finite-strain three-field poroelasticity. *Computational Mechanics*, 60(1):51–68, 2017.

[32] A. Elyes, F.A. Radu, and J.M. Nordbotten. A posteriori error estimates and adaptivity for fully mixed finite element discretizations for Biot's consolidation model. hal-01687026, 2018.

[33] X. Hu, C. Rodrigo, F.J. Gaspar, and L.T. Zikatanov. A nonconforming finite element method for the Biot's consolidation model in poroelasticity. *Journal of Computational and Applied Mathematics*, 310:143 – 154, 2017.

[34] C. Rodrigo, F.J. Gaspar, X. Hu, and L.T. Zikatanov. Stability and monotonicity for some discretizations of the Biot's consolidation model. *Computer Methods in Applied Mechanics and Engineering*, 298:183 – 204, 2016.

[35] N. Chaabane and B. Riviére. A splitting-based finite element method for the Biot poroelasticity system. *Computers & Mathematics with Applications*, 75(7):2328 – 2337, 2018.

[36] N. Chaabane and B. Rivière. A Sequential Discontinuous Galerkin Method for the Coupling of Flow and Geomechanics. *Journal of Scientific Computing*, 74(1):375–395, 2018.

[37] S. Dana and M.F. Wheeler. Convergence analysis of fixed stress split iterative scheme for anisotropic poroelasticity with tensor Biot parameter. *Computational Geosciences*, 22(5):1219–1230, 2018.

[38] N. Castelletto, S. Klevtsov, H. Hajibeygi, and H.A. Tchelepi. Multiscale two-stage solver for Biot's poroelasticity equations in subsurface media. *Computational Geosciences*, 2018. https://doi.org/10.1007/s10596-018-9791-z.

[39] N. Castelletto, H. Hajibeygi, and H.A. Tchelepi. Multiscale finite-element method for linear elastic geomechanics. *Journal of Computational Physics*, 331:337 – 356, 2017.

[40] A. Ern and S. Meunier. A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(2):353–375, 2009.

[41] J.B. Haga, H. Osnes, and H.P. Langtangen. On the causes of pressure oscillations in low-permeable and low-compressible porous media. *International Journal for Numerical and Analytical Methods in Geomechanics*, 36(12):1507–1522, 2012.

[42] C. Rodrigo, X. Hu, P. Ohm, J. H. Adler, F.J. Gaspar, and L.T. Zikatanov. New stabilized discretizations for poroelasticity and the Stokes' equations. arXiv:1706.05169, 2017.

[43] M. Bause. Iterative Coupling of Mixed and Discontinuous Galerkin Methods for Poroelasticity. arXiv:1802.03230, 2018.

[44] S. Dana, B. Ganis, and M.F. Wheeler. A multiscale fixed stress split iterative scheme for coupled flow and poromechanics in deep subsurface reservoirs. *Journal of Computational Physics*, 352:1 – 22, 2018.

[45] M. Borregales, F.A. Radu, K. Kumar, and J.M. Nordbotten. Robust iterative schemes for non-linear poromechanics. *Computational Geosciences*, 22(4):1021–1038, 2018.

[46] F.A. Radu, M. Borregales, F. Gaspar, K. Kumar, and C. Rodrigo. L-scheme and Newton based solvers for a nonlinear Biot model. In *ECCOMAS*, 2018.

[47] 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 & Mathematics with Applications*, 2018. https://doi.org/10.1016/j.camwa.2018.07.033.

[48] J.W. Both, K. Kumar, J.M. Nordbotten, and F.A. Radu. Iterative Methods for Coupled Flow and Geomechanics in Unsaturated Porous Media. In *Poromechanics VI*, pages 411–418, 2017.

[49] J.J. Lee, E. Piersanti, K.A. Mardal, and M.E. Rognes. A mixed finite element method for nearly incompressible multiple-network poroelasticity. arXiv:1804.07568v1, 2018.

[50] Q. Hong, J. Kraus, M. Lymbery, and F. Philo. Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelastic models. arXiv: 1806.00353v2, 2018.

[51] B. Giovanardi, L. Formaggia, A. Scotti, and P. Zunino. Unfitted FEM for Modelling the Interaction of Multiple Fractures in a Poroelastic Medium. In *Geometrically Unfitted Finite Element Methods and Applications*, pages 331–352, Cham, 2017. Springer International Publishing.

[52] S. Lee, M.F. Wheeler, and T. Wick. Iterative coupling of flow, geomechanics and adaptive phase-field fracture including level-set crack width approaches. *Journal of Computational and Applied Mathematics*, 314:40 – 60, 2017.

[53] T. Almani, K. Kumar, A. Dogru, G. Singh, and M.F. Wheeler. Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics. *Computer Methods in Applied Mechanics and Engineering*, 311:180 – 207, 2016.

[54] M. Borregales, K. Kumar, F.A. Radu, C. Rodrigo, and F.J. Gaspar. A partially parallel-in-time fixed-stress splitting method for biot's consolidation model. *Computers & Mathematics with Applications*, 2018. https://doi.org/10.1016/j.camwa.2018.09.005.

[55] M.K. Brun, I. Berre, J.M. Nordbotten, and F.A. Radu. Upscaling of the Coupling of Hydromechanical and Thermal Processes in a Quasi-static Poroelastic Medium. *Transport in Porous Media*, 124(1):137–158, 2018.

[56] M.K. Brun, E. Ahmed, J.M. Nordbotten, and F.A. Radu. Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport. *Journal of Mathematical Analysis and Applications*, 2018. https://doi.org/10.1016/j.jmaa.2018.10.074.

[57] S. Zsuppán. On the domain dependence of the inf-sup and related constants via conformal mapping. *Journal of Mathematical Analysis and Applications*, 382(2):856 – 863, 2011.

[58] Y. Abousleiman, A.H.D. Cheng, L. Cui, E. Detournay, and J.C. Roegiers. Mandel's problem revisited. *Géotechnique*, 46(2):187–195, 1996.