Existence and Uniqueness of Non-linear, Possibly Degenerate Parabolic PDEs, with Applications to Flow in Porous Media

Master's Thesis in Applied and Computational Mathematics

Anders Westrheim

Department of Mathematics
University of Bergen
August 7, 2017

Contents

Acknowledgements

I have to first and foremost thank my dream-team of proof-readers: Erlend Storvik and Eirik Berge, and most of all, the team captains Erlend Raa Vågset and Krister Trandal. I will forever be grateful for the big effort you guys put in for little me.

To my mother and father. I love you very much. Even if I'm bad at expressing my feelings now and then. You have shown me your sincere love without ever wanting something in return. Again, I love you both, equally.

Ever since I was a little boy I have delayed my projects to the last possible starting date. I express my deepest gratitude to my advisor, Professor Florin A. Radu, for keeping his patience and showing his belief in me. There have been some dark times, but it is in the darkest hour in which show your true self.

I want to dedicate this thesis to all my friends at the department of mathematics. Pi Happy has been there for me since the beginning, and you have been like brothers and sisters to me.

Lastly, I want to thank Oda. The party is just getting started.

> The only girl I've ever loved
> Was born with roses in her eyes
>
> Jeff Mangum,
> *In the Aeroplane over the Sea*

Introduction

In the branch of mathematical analysis known as *functional analysis*, one mainly studies functions defined on vector spaces. For *partial differential equations* (PDEs), this analysis has proven to be a mighty resource of understanding and modelling the behavior of the equations. Throughout this thesis, the work will focus of theory of function spaces and existence and uniqueness theorems for variational formulations in normed vector spaces. We will recast PDEs as variational problems with operators acting on normed spaces, and further seek to prove the existence and uniqueness of a solution by assigning certain properties to the operator.

The outline of this thesis is as follows:

In Chapter 1, we summarize the **Basic Notions of Functional Analysis** relevant for the later work in the thesis. We define operators, discuss monotonicity, present the theory of Sobolev spaces, and illustrate the finite element method, giving short hints to the future relevancy of the described properties.

Linear Problems have been extensively studied in the past. In Chapter 2, we present three important theorems illustrating the conditions for existence and uniqueness of solutions for variational formulations of the type:

(i) Galerkin formulations in Hilbert spaces: *The Lax-Milgram Theorem*,

(ii) Petrov-Galerkin formulations in Hilbert spaces: *The Babuška-Lax-Milgram Theorem*,

(iii) Petrov-Galerkin formulations in Banach spaces: *The Banach-Nečas-Babuška Theorem*,

and give their proofs.

Chapter 3 is dedicated to the study of **Non-linear Problems**. We seek to extend the ideas of the previous chapter to variational formulations containing a non-linearity $b(\cdot)$ depending on the solution we seek. This has a major application in the analysis of non-linear PDEs, which in general may not possess analytical solutions. To attack these types of problems, we define a weak formulation of the main problem, and discretize the domain of where a solution is sought. Next, existence and uniqueness is established through fixed point theorems, which will be given with proof.

We will focus our study on two central problems: *The Richards equation* (a non-linear, possibly degenerate parabolic PDE) and a *transport equation* modelling reactive flow in porous media (two coupled PDEs). For the fully discrete (non-linear) formulation of Richards equation we show results for

(i) a Lipschitz continuous non-linearity. Here we consider three cases:

    (a) First, a linearization scheme is proposed. We prove existence and uniqueness by using the Lax-Milgram Theorem in combination with the Banach Fixed Point Theorem.
    (b) Second, we make the assumption that the non-linearity is strongly monotone. Here, existence is proven by the Brouwer Fixed Point Theorem
    (c) Third, we let the non-linearity be monotone and add a regularization term to the fully discrete formulation. Here, we prove existence as in the previous step, and lastly show convergence of the regularized scheme to the fully discrete scheme.

(ii) a Hölder continuous non-linearity. We give two results:

    (a) First, we prove existence for a monotone and bounded non-linearity.
    (b) Second, we state the result of existence for a strongly monotone and bounded non-linearity by the Brouwer Fixed Point Theorem.

In the applications of Brouwer Fixed Point Theorem, the uniqueness of the problem is proved by assuming there exists two solutions and obtaining a contradiction through inequalities by showing estimates that can not be true.

Lastly, in Chapter 4, a mathematical model of **Two-phase Flow** in porous media is studied. We discuss the case of a Lipschitz continuous saturation, and show for the first time a proof of existence and uniqueness of a solution for the fully discrete (non-linear) scheme, assuming the saturation to be Hölder continuous and strongly monotonically increasing. This is done by creating a regularization of the fully discrete scheme, further proving existence with the Brouwer Fixed Point Theorem, and finally showing convergence with the help of an a priori estimate.

Chapter 1

Basic Notions of Functional Analysis

In this chapter we will state and discuss the prerequisites for the future chapters. The purpose of this chapter is to provide preliminary knowledge of functional analysis, and to make the thesis fairly self-contained.

We start by defining operators in normed spaces (Section 1.1) and explain their most relevant properties, especially monotonicity (Section 1.2). This will be an important tool for future work in Chapters 3 and 4. In Section 1.3, we give a short but concise summary of the theory of the different function spaces considered. The results presented in this thesis will only require the reader to be fluent in the most fundamental facts regarding Sobolev Spaces.

In Section 1.4, the Eberlein-Šmuljan Theorem will be stated, and compact embeddings for Sobolev spaces will be discussed briefly. The most useful inequalities for proving estimates in chapters 3 and 4 can be found in Section 1.5. In Section 1.6, we give an example of a relevant problem for the theory in Chapter 2.

Finally, in Section 1.7, a short introduction to the Finite Element Method is given. This will be a motivation and a main application for the existence and uniqueness analysis explored in later chapters.

1.1 Operators between Normed Spaces

The definitions and results in this section are collected from [22].

A *normed vector space* is a *vector space* with a metric defined by a *norm*. We assume the definition of a vector space, norm, inner product and the other cornerstone definitions of functional analysis and set theory to be known by the reader. If X is a normed vector space, we will denote a norm on X by $\|\cdot\|_X$ or simply $\|\cdot\|$ if there is no room for confusion. If $x := (x_1, \dots, x_n) \in \mathbb{R}^n$, then we denote by

$$|x|_n := \sqrt{x_1^2 + \cdots + x_n^2},$$

the *Euclidean norm* in $\mathbb{R}^n$ (unless specified otherwise). A *Banach* space X is a *complete* normed vector space. That is, if $\{x_k\}_k$ is a sequence in X satisfying

$$\|x_n - x_m\|_X \to 0 \qquad \text{as} \qquad m, n \to \infty,$$

for $m, n \in \mathbb{N}$ (i.e. a *Cauchy sequence*), then $\{x_k\}$ converges to an element $x \in X$. A *Hilbert* space H is a Banach space with a norm induced by an *inner product*, denoted by $\langle \cdot, \cdot \rangle_H$, or simply $\langle \cdot, \cdot \rangle$ if we explicitly state so in the text.

Let X and Y be normed spaces. We define an *operator* T to be a mapping from a *domain* $D(T) \subset X$ into Y, and write

$$T : D(T) \to Y.$$

We will denote the action of an operator T on an element $x \in D(T)$ by Tx, or $T(x)$. The *kernel* (or *null space*) and the *range* (or *image*) of T are defined by

$$\begin{aligned} \mathrm{Ker}(T) &:= \{x \in D(T) \mid Tx = 0\}, \\ \mathrm{Im}(T) &:= \{y \in Y \mid Tx = y \text{ for } x \in D(T)\}. \end{aligned}$$

Moreover, if T is linear, $\mathrm{Ker}(T)$ and $\mathrm{Im}(T)$ form subspaces of X and Y, respectively. If T is mapped into $\mathbb{R}$, we call T a *functional.*

Definition 1.1. Let $T : D(T) \to Y$ be an operator between normed spaces X and Y, where $D(T) \subset X$. We say that

(i) T is *linear* if $T(\alpha x_1 + \beta x_2) = \alpha T x_1 + \beta T x_2$ for all $x_1, x_2 \in D(T)$ and for all $\alpha, \beta \in \mathbb{R}$.

(ii) T is *bounded* if there exists a constant $M > 0$ such that $\|Tx\|_Y \leq M\|x\|_X$. The smallest such M (if it exists) is called the *operator norm* of T, denoted $\|T\|_{\mathcal{L}(X,Y)}$. That is,

$$\|T\|_{\mathcal{L}(X,Y)} := \sup_{x \in D(T)} \frac{\|Tx\|_Y}{\|x\|_X}, \qquad x \neq 0.$$

(iii) T is *injective* if $Tx_1 = Tx_2$ implies that $x_1 = x_2$ $\forall x_1, x_2 \in D(T)$.

(iv) T is *surjective* if for all $y \in Y$ there exists $x \in D(T)$ such that $Tx = y$.

(v) Let $Y = \mathbb{R}$, then T is said to be *corecive* if $\|x\|_X \to \infty$ implies $Tx \to \infty$.

Definition 1.2 (Continuity of operators). Let T be an operator between normed spaces X and Y. T is said to be *continuous* if it is continuous at each $x \in X$, that is, if for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\|Tx_1 - Tx_2\|_Y < \epsilon \qquad \text{whenever} \qquad \|x_1 - x_2\|_X < \delta \qquad \forall x_1, x_2 \in X.$$

We say that T is *Hölder continuous* with exponent α if there exists $\alpha \in (0, 1]$ and $C > 0$ such that

$$\|Tx_1 - Tx_2\|_Y \leq C\|x_1 - x_2\|_X^\alpha, \qquad \forall x_1, x_2 \in X.$$

We say that T is *Lipschitz continuous* if it is Hölder continuous with exponent $\alpha = 1$, that is, There exists $L > 0$ such that

$$\|Tx_1 - Tx_2\|_Y \leq L\|x_1 - x_2\|_X, \qquad \forall x_1, x_2 \in X.$$

If $L < 1$, we call T a *contraction.*

It is then clear that the following chain of implications holds:

$$T \text{ is Lipschitz continuous} \implies T \text{ is Hölder continuous} \implies T \text{ is continuous.}$$

Let T be a linear operator. One can easily show that if T is bounded, then it is necessarily continous, and vice versa. We define $\mathcal{L}(X, Y)$ to be the space of all linear and continuous operators from X into Y.

Definition 1.3 (Dual space). The dual space of a normed space X is the collection of all linear and continuous functionals defined on X, and is denoted V^* $(= \mathcal{L}(X, \mathbb{R}))$.

Remark 1.0.1. For $T \in X^*$, the action (or *duality product*) between X and X^* will often be written as $\langle T, x\rangle_{X^*,X}$. To avoid confusion with the inner product of a Hilbert space, this notation will be specified beforehand.

Definition 1.4. Let X and Y be normed spaces. A *bilinear form* a on $X \times Y$ is a mapping

$$a : X \times Y \to \mathbb{R}$$

that is linear with respect to each arguments. That is, for $x, x_1, x_2 \in X$, $y, y_1, y_2 \in Y$ and $c, d \in \mathbb{R}$, we have

$$\begin{aligned} a(cx_1 + dx_2, y) &= ca(x_1, y) + da(x_2, y), \\ a(x, cy_1 + dy_2) &= ca(x, y_1) + da(x, y_2). \end{aligned}$$

If there exists a real number $M > 0$ such that for all $x \in X$ and $y \in Y$, we have

$$|a(x, y)| \leq M\|x\|_Y \|y\|_Y,$$

then a is said to be bounded. The smallest such M is called the norm of a, and is denoted $\|a\|_{\mathcal{L}(X\times Y,\mathbb{R})}$. Moreover, we define $\mathcal{L}(X \times Y, \mathbb{R})$ to be the space of all bounded linear operators defined on $X \times Y$.

Definition 1.5 (Adjoint operator). Let X and Y be normed spaces and let $T \in \mathcal{L}(X, Y)$. The *adjoint operator* $T^* : Y^* \to X^*$ of T is defined by

$$\langle T^*g, x\rangle_{X^*,X} = \langle g, Tx\rangle_{Y^*,Y} \qquad \text{for } g \in Y^*, x \in X.$$

Theorem 1.1. *Let X and Y be Banach spaces and let $T \in \mathcal{L}(X, Y)$ be an operator. Then there exists a unique linear and continuous adjoint operator T^* of T. Moreover, T^* satisfies*

$$\|T^*\|_{\mathcal{L}(Y^*,X^*)} = \|T\|_{\mathcal{L}(X,Y)}.$$

Definition 1.6 (Annihilator). Let M be a nonempty subset of a normed space X. The *annihilator* $M^\perp$ of M in X is the set of all $\phi \in X^*$ that are zero everywhere on M. That is,

$$M^\perp := \{\phi \in X^* \mid \phi(m) = 0 \; \forall m \in M\}.$$

In the case of a Hilbert space H and a closed subspace Y of H, we shall denote $Y^\perp$ the orthogonal complement of Y in H, defined as

$$Y^\perp := \{v \in H \mid \langle y, v\rangle_H = 0 \quad \forall y \in Y\}.$$

Proposition 1.1. *Let M be a subspace of a Banach space X with $M^\perp = \{0\}$. Then M is dense in X.*

Remark 1.1.1. This result also holds for orthogonal complements in Hilbert spaces. Also, if Y is a closed subspace of a Hilbert space H, then

$$H = Y \oplus Y^\perp.$$

That is, for $y \in Y, z \in Y^\perp$, we can represent an element $x \in H$ as $x = y + z$.

The next theorem is a very important result in functional analysis and will be used in the proof of the existence and uniqueness theorems in Chapter 2, and can be found in [22], p. 188 & 192):

Theorem 1.2 (The Riesz Representation Theorem). *Let W, V be Hilbert spaces, $a \in \mathcal{L}(W \times V, \mathbb{R})$, and $f \in V^*$. Then there exist an operator $A \in \mathcal{L}(W, V)$ and an element $z \in V$ such that*

$$a(u, v) = \langle Aw, v\rangle_V \quad \text{and} \quad f(v) = \langle z, v\rangle_V,$$

where A and z are uniquely determined by a and f, respectively, and have norms

$$\|A\|_{\mathcal{L}(W,V)} = \|a\|_{\mathcal{L}(W\times V,\mathbb{R})}, \quad \text{and} \quad \|z\|_V = \|f\|_{V^*}.$$

1.2 Monotone Operators

The definitions and results of this section are collected from [15, 39], and will be regularly referenced in chapters 3 and 4 when we discuss the behavior of operators that may be *nonlinear* (as opposed to Definition 1.1 (i)). Let X be a Banach space and $A : X \to X^*$ an operator. Consider the problem:

$$\textit{Given } f \in X^*, \textit{find } u \in X \textit{ such that } Au = f. \tag{P_1}$$

We shall now discuss the assumptions required to prove existence and uniqueness of Problem (P_1). As a simple example, we study the case where $X = \mathbb{R}$, given in [39] (page 471):

Let $f : \mathbb{R} \to \mathbb{R}$. Consider the problem:

$$\textit{Given } y \in \mathbb{R}, \textit{ find } x \in \mathbb{R} \textit{ such that } f(x) = y. \tag{P_2}$$

If f is continuous and $f \to \pm\infty$ as $x \to \pm\infty$, the *Intermediate Value Theorem* (as found in elementary texts on Calculus) states that f takes any value on the interval $(-\infty, \infty)$. This gives the existence of x. For the uniqueness of a solution we note that if we assume f to be *strictly increasing* (see Figure 1.2), i.e. if $f'(x) > 0 \ \forall x \in \mathbb{R}$, it will only pass through points in $\mathbb{R}$ once.

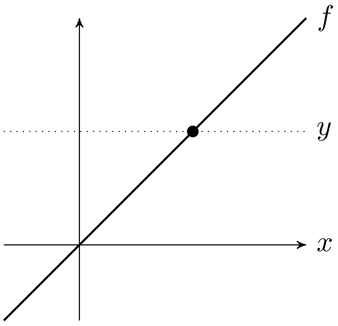


Figure 1.1: A strictly increasing function.

Therefore we conclude that if f is continuous, strictly increasing and if $f \to \pm\infty$ as $x \to \pm\infty$, then there *exists* a *unique* $x \in \mathbb{R}$ solving Problem (P_2). To extend this analysis to Problem (P_1), we first generalize the definition of an "increasing function":

Definition 1.7. Let X be a Banach space, $A : X \to X^*$ an operator, and $\langle \cdot, \cdot \rangle$ the duality product between X and X^*. Then

(i) A is called *monotone* if

$$\langle Au - Av, u - v \rangle \geq 0 \qquad \forall u, v \in X.$$

(ii) A is called *strictly monotone* if

$$\langle Au - Av, u - v \rangle > 0 \qquad \forall u, v \in X, \ u \neq v.$$

(iii) A is called *strongly monotone* if $\exists\, c > 0$ such that

$$\langle Au - Av, u - v \rangle \geq c\|u - v\|^2 \qquad \forall u, v \in X.$$

(iv) A is called coercive if $Au \in X^*$ is coercive. That is, if

$$\lim_{\|u\| \to \infty} \frac{\langle Au, u \rangle}{\|u\|} \to \infty.$$

In most of the results we prove in Chapter 3, we will assume X to be a Hilbert space, and the non-linearity to be strongly monotone as in Definition 1.7 (iii) and increasing. If we furthermore assume that $A(0) = 0$, we have the estimate

$$\langle Au, u \rangle \geq c\|u\|^2.$$

Let $Ax, Ay \in \mathrm{Im}(A)$ with $Ax = Ay$. Then $Ax - Ay \in X^*$ is the zero functional. We observe that if A is a strictly monotone operator as in (ii), then we must necessarily have $x = y$. This establishes the uniqueness.

The property $f \to \pm\infty$ as $x \to \pm\infty$ implies coercivity for Problem (P_2). We have $X = X^* = \mathbb{R}$, $\|\cdot\| = |\cdot|$ and thus

$$\frac{\langle f(x), x \rangle}{\|x\|} = \frac{f(x)x}{|x|} \to \infty \qquad \text{as } |x| \to \infty.$$

The next result we present is the *Minty-Browder* Theorem. This gives sufficient conditions for an operator to be surjective, which is equivalent to the existence of a solution of Problem (P_1). The proofs of the following theorems are in [15], sections 9.13 and 9.14.

Definition 1.8 (Hemicontinuity). Let X be a normed vector space. A mapping $A : X \to X^*$ is said to be *hemicontinuous* if, given any $u, v, w \in X$, there exist $t_0 = t_0(u, v, w) > 0$ such that the function

$$t \in (-t_0, t_0) \mapsto \langle A(u + tv), w \rangle \in \mathbb{R},$$

is continuous at $t = 0$.

Theorem 1.3 (The Minty-Browder Theorem). *Let X be a real separable Banach space and $A : X \to X^*$ a coercive and hemicontinuous monotone operator. Then there exists a solution of Problem (P_1) $\forall f \in X^*$. If A is strictly monotone, the solution is unique.*

Remark 1.3.1. Note that the existence implies surjectivity of A, while the uniqueness implies injectivity. Thus A is bijective.

Theorem 1.4. *Let X be a finite-dimensional normed vector space and let $A : X \to X^*$ be a hemicontinuous operator. Then A is continuous.*

The problems encountered in Chapter 3 will be attached to finite-dimensional spaces, so the choices of properties for the non-linearities will be motivated by the hypothesis of Theorem 1.4 in combination with Theorem 1.3.

1.3 Function Spaces

In this section the function spaces we will utilize later on is given. These definitions and results are from [17]. We will present some theory about the structural properties of *Sobolev spaces*, which will prove to be very useful for analysis of partial differential equations.

Definition 1.9. Let Ω be an open subset of $\mathbb{R}^n$ and $1 < p < \infty$. We define $L^p(\Omega)$ to be the space of all measurable functions $f : \Omega \to \mathbb{R}$ for which $\|f\|_{L^p(\Omega)} < \infty$, where

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f|^p dx \right)^{1/p}.$$

Remark 1.4.1. It can be shown that the space $L^2(\Omega)$ is a Hilbert space with inner product

$$\langle f, g \rangle := \int_\Omega fg \, dx.$$

Definition 1.10. We define the space of *test functions*

$$C_c^\infty(\Omega) := \big\{ f \in C^\infty(\Omega) \mid f \text{ has compact support} \big\}.$$

Furthermore, let $L^1_{\text{loc}}(\Omega)$ is the space of all integrable functions on every compact subset of Ω. Let $u, v \in L^1_{\text{loc}}(\Omega)$. The function v is the α^{th} *weak partial derivative* of u, written $D^\alpha u = v$, provided

$$\int_\Omega u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v \phi \, dx,$$

for all test functions $\phi \in C_c^\infty(\Omega)$. If so, we say that the α-th partial derivative of u exists in the *weak sense.*

Definition 1.11 (Mollifier). A sequence of *mollifiers* is any sequence $\{\rho_n\}_n$ of test functions on $\mathbb{R}^d$ satisfying

$$\text{supp}(\rho_n) \subset \overline{B_{1/n}(0)}, \qquad \int_{\mathbb{R}^d} \rho_n dx = 1, \quad \rho_n \geq 0 \text{ on } \mathbb{R}^d.$$

The sequence of *standard mollifiers* $\{\eta_j\}_j$ is defined as $\eta_j := j^d \eta(jx)$ for $x \in \mathbb{R}^d$, such that

$$\eta(x) := \begin{cases} Ce^{\left(\frac{1}{|x|^2-1}\right)}, & \text{if } |x| < 1, \\ 0, & \text{if } |x| \geq 1, \end{cases}$$

where $C \in \mathbb{R}$ is chosen such that $\{\eta_j\}_j$ is a sequence of mollifiers.

Remark 1.4.2. One can check that $\eta \in C_c^\infty(\mathbb{R}^d)$. Moreover, $\text{supp}(\eta_j) \subset \overline{B_{1/j}(0)}$.

Proposition 1.2. *Assume $f \in C(\Omega)$, and let $\{\rho_n\}_n$ be a sequence of mollifiers. Then $\rho_n * f \to f$ uniformly as $n \to \infty$ on every compact subset of $\mathbb{R}^d$. Moreover, let*

$$\Omega_\epsilon := \{x \in \Omega \mid \text{dist}(x, \mathbb{R}^n - \Omega) > \epsilon\}, \tag{1.1}$$

for all open $\Omega \subset \mathbb{R}^n$. Then $\rho * f \in C^\infty(\Omega_\epsilon)$ for all $\epsilon > 0$.

Definition 1.12 (Sobolev spaces). Let $k \in \mathbb{N}$ and $1 \leq p \leq +\infty$. The Sobolev space $W^{k,p}(\Omega)$ is the space of all functions $f : \Omega \to \mathbb{R}$ whose $1, \ldots, k$-th order partial derivatives belong to $L^p(\Omega)$ in the weak sense. That is,

$$W^{k,p}(\Omega) := \big\{ f \in L^p(\Omega) \mid \exists\, D^\alpha \widetilde{f} \in L^p(\Omega) \text{ for all multi-indices } |\alpha| \leq k \big\}.$$

For $u \in W^{k,p}(\Omega)$, a norm on $W^{k,p}(\Omega)$ is defined as

$$\|u\|_{W^{k,p}(\Omega)} := \Bigg(\sum_{|\alpha| \leq k} \int_\Omega \|D^\alpha u\|^p_{L^p(\Omega)} \Bigg)^{1/p},$$

for $1 \leq p < +\infty$, and

$$\|u\|_{W^{k,\infty}(\Omega)} := \sum_{|\alpha| \leq k} \operatorname*{ess\,sup}_{\Omega} |D^\alpha u|.$$

Remark 1.4.3. It can be shown that the space $H^k(\Omega) := W^{k,2}(\Omega)$ is a Hilbert space.

The work in this paper will mostly be focused on the Hilbert space $H^1(\Omega) := W^{1,2}(\Omega)$, which is the space of functions with a first-order weak derivative in $L^2(\Omega)$. This has norm which we from here on will denote by $\|\cdot\|_1 := \|\cdot\|_{H^1(\Omega)}$. For $u \in H^1(\Omega)$ we have

$$\|u\|_1 := \Bigg(\int_\Omega |u|^2 + |\nabla u|^2 \, dx \Bigg)^{1/2}.$$

Definition 1.13. We define the space $W^{k,p}_0(\Omega)$ as the closure of $C^\infty_c(\Omega)$ in $W^{k,p}(\Omega)$.

Remark 1.4.4. $W^{k,p}_0(\Omega)$ can and will be interpreted as the space of functions that have $D^\alpha u = 0$ on the boundary of Ω $\forall |\alpha| \leq k - 1$. So, $H^1_0(\Omega)$ will be defined as

$$H^1_0(\Omega) := \big\{ f \in H^1(\Omega) \mid f = 0 \text{ on } \partial\Omega \big\}.$$

This space is highly relevant for studying PDEs with homogeneous Dirichlet boundary conditions.

Definition 1.14. We denote by $H^{-1}(\Omega)$ the dual space of $H^1_0(\Omega)$.

Remark 1.4.5. $H^1_0 \subset L^2(\Omega) \subset H^{-1}(\Omega)$.

Next, Bochner spaces will create a Sobolev space structure for functions that also possess a time variable. As an illustration. let $T > 0$ be a real number. If $u = u(t, x) : [0, T] \times \Omega \to \mathbb{R}$ and $u(t, x) \in L^2(\Omega)$ for all $t \in [0, T]$, we look at u as a mapping from $[0, T]$ into $L^2(\Omega)$. Indeed, this generalizes the concept of L^p-spaces to functions with range in Banach spaces (not necessarily just the real numbers).

For a summary of measure theory, we refer to [17], Appendix E.

Definition 1.15 (Bochner spaces). Let X be a Banach space, and $T > 0$. The space $L^p(0, T; X)$ consists of all Bochner measurable functions $u : [0, T] \to X$ with

$$\|u\|_{L^p(0,T;X)} := \Bigg(\int_0^T \|u(t)\|^p_X \, dt \Bigg)^{1/p} < \infty,$$

for $1 \leq p < \infty$ and

$$\|u\|_{L^p(0,T;X)} := \operatorname*{ess\,sup}_{0 \leq t \leq T} \|u(t)\|_X < \infty.$$

In the context of Bochner spaces, a *weak derivative* of $L^1(0,T;X)$ means that $\exists\ v \in L^1(0,T;X)$ such that

$$\int_0^T \phi'(t)u(t)\,dt = -\int_0^T \phi(t)v(t)\,dt,$$

for all $\phi \in C_0^\infty(0,T)$.

Definition 1.16 (Weak derivatives in Bochner spaces). Let X be a Banach space. The Sobolev space $W^{1,p}(0,T;X)$ consists of all $u \in L^p(0,T;X)$ such that $\partial_t u$ exists in the weak sense and belongs to $L^p(0,T;X)$. Furthermore,

$$\|u\|_{W^{1,p}(0,T;X)} := \left(\int_0^T \|u(t)\|_X^p + \|\partial_t u(t)\|_X^p\,dt \right)^{1/p} < \infty,$$

for $1 \leq p < \infty$ and

$$\|u\|_{W^{1,p}(0,T;X)} := \operatorname*{ess\,sup}_{0\leq t\leq T} \left\{ \|u(t)\|_X + \|\partial_t u(t)\|_X \right\} < \infty.$$

1.4 Embeddings

The theory in this section is from [17].

Here we will provide a short discussion on convergence of sequences and some compactness arguments. This will be used in later chapters, where if we cannot prove existence and uniqueness with the methods we apply, we construct a similar problem. For this similar problem, we prove existence and uniqueness, and then apply the theory of this chapter to prove that there can be constructed a sequence of solutions for the similar problem which converge to a solution of the original problem.

Let X be a normed vector space with norm $\|\cdot\|_X$. We say that a sequence $\{x_n\}_n$ in X converges *weakly* to $x \in X$ if for every $\phi \in X^*$ we have $\phi(x_n) \to \phi(x)$ as $n \to \infty$. We denote this by

$$x_n \rightharpoonup x \in X \qquad \text{as} \qquad n \to \infty.$$

A sequence $\{x_n\}_n$ is *bounded* in X if there exists a constant $M > 0$ such that $\|x_n\|_X \leq M$ for all $n \in \mathbb{N}$.

Remark 1.4.6. It can be shown that if X is finite-dimensional, weak convergence is equivalent to strong convergence.

We begin with an essential result:

Theorem 1.5 (Eberlein-Šmuljan). *Let X be a reflexive normed space and $\{x_n\}_n$ a bounded sequence in X. Then there exists a subsequence $\{x_{n_k}\}_k \subset \{x_n\}_n$ and $x \in X$ such that*

$$x_{n_k} \rightharpoonup x \in X.$$

Definition 1.17 (Continuous embedding). Let X, Y be normed spaces. We say that X is *continuously embedded* in Y if $X \subset Y$ and $\exists\ c > 0$ such that

$$\|u\|_Y \leq c\|u\|_X \qquad u \in X.$$

We write

$$X \hookrightarrow Y.$$

Definition 1.18 (Compact embedding). Let X, Y be Banach spaces and $X \subset Y$. We say that X is *compactly embedded* in Y, written $X \subset\subset Y$, provided X is continuously embedded in Y and that each bounded sequence in X has a convergent subsequence in Y.

Theorem 1.6 (The Trace Theorem). *Let Ω be a bounded domain with Lipschitz continuous boundary $\partial\Omega$. Then there exists a $c > 0$ such that*

$$\|u\|_{L^p(\partial\Omega)} \leq c\|u\|_{W^{1,p}(\Omega)} \qquad \forall u \in C^1(\bar{\Omega}).$$

Remark 1.6.1. This tells us that there exists a linear and continuous mapping

$$\gamma : W^{1,p}(\Omega) \to L^p(\partial\Omega),$$

which we call the *trace operator.*

Remark 1.6.2. In the problems we will be studying, we usually consider a boundary condition for a PDE. When reducing the problem to a weak formulation, the trace operator allows us to go smoothly from the domain to its boundary even though the boundary may be of measure zero.

Next, the *Rellich-Kondrachov Theorem* is an interesting and useful tool for proving existence and uniqueness of continuous variational formulations in Sobolev spaces. As we will talk about in Section 1.7, the *Galerkin method* focuses on defining and finding solutions to discrete problems, and next, showing that there is a sequence converging to the solution of the original problem. To show convergence, one often show it in the $L^p(\Omega)$-norm for a relevant p and then use the Rellich-Kondrachov Theorem to say that the Sobolev space in which we seek a solution in is compactly embedded in $L^p(\Omega)$.

Theorem 1.7 (Rellich-Kondrachov). *Let $\Omega \in \mathbb{R}^d$ be open, bounded with $\partial\Omega \in C^1$. Suppose $1 \leq p < n$ and let $p^* := \frac{pn}{n-p}$. Then there holds*

$$W^{1,p}(\Omega) \subset\subset L^q(\Omega),$$

for each $1 \leq q < p^*$.

1.5 Inequalities

The theorems in this section are from [17], and will just be stated in short here and referenced frequently in chapters 3 and 4.

Theorem 1.8 (The Cauchy-Schwarz Inequality). *Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Then*

$$|\langle u, v\rangle| \leq \|u\|_H \|v\|_H \qquad \forall\, u, v \in H.$$

Theorem 1.9 (The Young Inequality). *Let $a, b \in \mathbb{R}$, $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then, for any $\epsilon > 0$,*

$$|ab| \leq \epsilon\frac{a^2}{p} + \epsilon^{-1}\frac{b^2}{q}.$$

Theorem 1.10 (The Poincaré Inequality). *Assume $\Omega \in \mathbb{R}^n$ is open and bounded. Let $p \in [1, \infty]$. Then there exists a constant $C = C(\Omega, p)$ such that for every $u \in W^{1,p}(\Omega)$,*

$$\|u\|_{L^p(\Omega)} \leq C\|Du\|_{L^p(\Omega)}.$$

1.6 A First Look at a Weak Formulation

To start off the study of existence and uniqueness of variational formulations motivated by partial differential equations, let us consider the Poisson equation as an example. Let $\Omega \subset \mathbb{R}^n$ be an open, bounded domain, and $f \in L^2(\Omega)$. We seek a solution to the problem:

$$\begin{cases} -\Delta u = f, & x \in \Omega, \\ u = 0, & x \in \partial\Omega. \end{cases} \tag{A}$$

Let V be some function space we have yet to define, and let $v \in V$. If we multiply both sides of the equation (A) by v and integrate over Ω, we get

$$-\int_\Omega (\Delta u) v \, dx = \int_\Omega f v \, dx.$$

Next, we integrate by parts (assuming this is well defined for v). If we furthermore suppose that v also satisfies the boundary condition $v = 0$ on $\partial\Omega$, we obtain the equation

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx, \tag{1.2}$$

where the boundary term is removed because of the properties of v. If we now define the space V to be

$$V := H_0^1(\Omega)$$

and say that we seek a solution u satisfying (1.2), the problem becomes equivalent to the variational formulation

$$\textit{Find } u \in H_0^1(\Omega) \quad \textit{such that} \quad a(u,v) = f(v) \quad \forall v \in H_0^1(\Omega), \tag{B}$$

where $a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx$ is a bilinear form (linear in each argument separately) and $f(v) := \int_\Omega f v \, dx$ is a linear functional.

It is interesting to note that the original Problem (A) has been reduced to a *weaker* statement in Problem (B): The functions in $C^2(\Omega)$ which takes zero as value on the boundary are included in $H_0^1(\Omega)$. From here on we will often call variational formulations derived in the same manner as (B) *weak formulations.* The space in which we seek a solution u is called the *solution space*, while the space of all v we call the *test space.* A problem for which the test and solution space are the same, as in (B)), is called a *Galerkin formulation.* If they are different, it is termed a *Petrov-Galerkin formulation.*

In the next chapter we will state the necessary and sufficent properties this type of problem must posess in order to prove the existence and uniqueness of such a function u. For now, we note the following properties for the bilinear form $a(\cdot,\cdot)$ in Problem (B):

(i) $a(\cdot,\cdot)$ is bounded: let $u, v \in H_0^1(\Omega)$. Then

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

which follows from the Cauchy-Schwarz inequality (Theorem 1.8).

(ii) $a(\cdot,\cdot)$ is coercive. We can show this by using the Poincaré inequality (Theorem 1.10). Let $v \in H_0^1(\Omega)$. Then $\exists\, m > 0$ such that

$$a(v,v) = \|\nabla v\|^2_{L^2(\Omega)} \geq m \|v\|^2_{H^1(\Omega)}.$$

1.7 The Finite Element Method

The theory of this section is extracted from [15, 16].

Consider the problem

$$\textit{Find } u \in V \textit{ such that } a(u,v) = f(v) \quad \forall v \in V, \tag{G}$$

where V is a Hilbert space, the bilinear form $a : V \times V \to \mathbb{R}$ is bounded and coercive, and $f \in V^*$. The standard way to find an approximate solution of a Galerkin formulation is the *Galerkin method.* In this method, we find a finite-dimensional subspace $V_h \subset V$ and consider the *discrete problem*:

$$\textit{Find } u_h \in V_h \textit{ such that } a(u_h,v_h) = f(v_h) \quad \forall v_h \in V_h. \tag{G_h}$$

as we will see in the next chapter, existence and uniqueness of Problem (G) can be proved with the Lax-Milgram Theorem (Theorem 2.4). For the discrete problem (G_h), we can also apply the same theorem, since a finite-dimensional subspace of a Hilbert space is in its own right a Hilbert space (see [22]).

Let us illustrate the convergence by means of assuming the Hilbert space H is separable (see [22]). If so, there exists a sequence of finite-dimensional subspaces $\{V_h\}_h \in H$ (of dimension h) such that $\bigcup_{h\in\mathbb{N}} V_h$ is dense in H (see [15], Theorem 2.2-7). Therefore, if we can find unique solutions u_h of Problem (G_h) for each V_h, then $\{u_h\}_h$ forms a sequence that may converge to a solution $u \in V$ for Problem (G). The Eberlein-Šmuljan Theorem (Theorem 1.5) will give existence of a subsequence of $\{u_h\}_h$ converging to some $u \in V$, and we further need to show that this is the solution we seek. Mind that this is just a sketch, and is only meant to be used as an ideal example. The expressions we consider in the next chapters will be of a different complexity than Problems (G) and (G_h).

The *Finite Element method* is related to the Galerkin method, where we specify the construction of the space V_h and focus on solving Problem (G_h). In the analysis presented later on in the thesis, $u := u(t,x)$ is a function of time and space. So our objective is to discretize a function space, which will require us to first partition the time and space, and next the functions defined on each element of the partition.

We discretize the time with step length $\tau > 0$. So for $t \in (0,T]$, where $T > 0$ is the final time, we characterize the time steps as $t_n := n\tau$, where $n \in \{1,\dots,N\}$ for $\tau = \frac{T}{N}$ (see Figure 1.7).

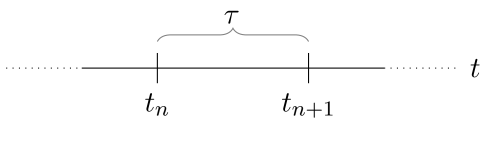


Figure 1.2: A time discretization.

We will usually discretize the possible time derivatives in the variational formulations derived from partial differential equations through the *Backward Euler method*, where we approximate

$$\partial_t u(t_n,x) \approx \frac{u(t_n,x) - u(t_{n-1},x)}{\tau}.$$

at time step t_n.

For the spatial discretization, we will assume that the domain $\Omega \in \mathbb{R}^d$ can be partitioned into *d-simplices.* We denote this as a triangulation $\mathcal{T}_h$ over the set $\overline{\Omega}$ (see Figure 1.7). That is, $\overline{\Omega}$ is subdivided into a finite number of subsets K (called *finite elements*), satisfying

(i) $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$

(ii) For each $K \in \mathcal{T}_h$, the set K is closed, and the interior of K is nonempty.

(iii) For each distinctive $K_1, K_2 \in \mathcal{T}_h$, the interiors of K_1 and K_2 do not intersect.

(iv) For each $K \in \mathcal{T}_h$, the boundary ∂K is Lipschitz continuous.

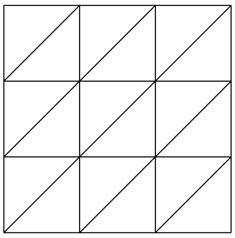


Figure 1.3: A triangularization of a square domain.

Next, one defines a function space over each K, called a *finite element space.* This space should be finite-dimensional. These are often referred to in later chapters as *discrete subspaces* of the test and solution spaces. The finite element spaces we will consider will either be spaces of constant functions or piecewise polynomial on each element K.

The solution is constructed from a set of *basis functions* for the finite element space, defined on each element K. We will not go into detail and show explicit calculations on how this is further analysed. For a more detailed explanation and concrete examples, we refer to [7, 9, 16].

In chapters 3 and 4, we will also encounter *mixed finite elements*, which will be based on the introduction of a new variable in the variational formulation motivated by a partial differential equation. There we will define two finite element spaces and seek solutions for our problems in both of them simultaneously.

Chapter 2

Linear Problems

In this chapter we will go through the necessary and sufficient conditions needed for proving existence and uniqueness for linear variational formulations in general. We will study problems of the type:

$$\textit{Find } u \in W \quad \textit{such that} \quad a(u,v) = \langle f, v\rangle_V \qquad \forall v \in V, \tag{2.1}$$

where $a(\cdot,\cdot) : W \times V \to \mathbb{R}$ is a bilinear form and $f \in V^*$ is a continuous linear functional. First, in Section 2.1, we present the Open Mapping Theorem and Closed Range Theorem, which are essential results for proving the existence and uniqueness theorems. Section 2.2 deals with the case of Problem (2.1) for which W and V are Hilbert spaces. We prove the theorems for $W = V$ (the Lax-Milgram Theorem) and $W \neq V$ (the Babuška-Lax-Milgram Theorem).

In the final Section 2.3, we prove a result for $W \neq V$, with W and V being Banach spaces (the Banach-Nečas-Babuška Theorem).

2.1 Central Results

To be able to construct the proofs of the existence and uniqueness theorems for linear variational formulations, we will revisit the cornerstones of functional analysis. First, we present the Baire Category Theorem, which will be used in the proof of the Open Mapping Theorem (Theorem 2.2). The results in this section are from [8].

Theorem 2.1 (the Baire Category Theorem). *Let X be a complete metric space and let $\{X_n\}_n$ be a sequence of closed subsets in X with empty interior. Then*

$$\mathrm{Int}(\bigcup_{n=1}^{\infty} X_n) = \emptyset.$$

Or, equivalently; let $\bigcup_{i=1}^{\infty} X_n = X$. Then there exists a non-empty closed subset $X_{n_0} \subset X$ for some $n_0 \in \mathbb{N}$.

Definition 2.1 (Open mapping). Let X, Y be metric spaces. Then $T : \mathcal{D}(T) \to Y$ with domain $\mathcal{D}(T) \subset X$ is called an *open mapping* if for every open set in $\mathcal{D}(T)$ the image is an open set in Y.

Theorem 2.2 (The Open Mapping Theorem). *Let E,F be Banach spaces, $T \in \mathcal{L}(E,F)$ surjective. Then T is an open mapping.*

Proof. It suffices to show that T maps the open unit ball to a neighbourhood of the origin of F. This proof will consist of two steps:

(i) Assume that $T \in \mathcal{L}(E,F)$ is surjective. Then there exists a $c > 0$ such that

$$\overline{T(B_E(0,1))} \supset B_F(0,2c).$$

(ii) Assume that T satisfies step 1. Then

$$T(B_E(0,1)) \supset B_F(0,c),$$

which is the desired result.

Proof of step 1. Define $U = B_E(0,1)$ and $X_n = \overline{nT(U)}$, $n \in \mathbb{N}$. Since T is surjective and linear,

$$F = T(E) = T\big(\bigcup_{n=1}^{\infty} nU\big) = \bigcup_{n=1}^{\infty} nT(U) = \bigcup_{n=1}^{\infty} X_n.$$

It is immaterial if we use the union of the closed sets because F is complete. Combining this with the fact that each X_n is closed, we make use of the Baire theorem: there exists $k \in \mathbb{N}$ such that $\mathrm{Int}(X_k) \neq \emptyset$. This implies that X_1 must contain an open ball. Let $v \in B_F(0,1)$. Pick $c > 0$ and $y_0 \in F$ satisfying

$$B_F(y_0, 4c) \subset X_1,$$

then $y_0, y_0 + cv \in B_F(y_0, 4c)$. By this and continuity, we get

$$4cv \in B_F(y_0, 4c) + B_F(y_0, 4c) \subset 2X_1,$$

and so $2cv \in X_1$. $v \in B_F(0,1)$ is equivalent to $2cv \in B_F(0,2c)$, which gives

$$B_F(0,2c) \subset X_1.$$

■

Proof of step 2. Choose $y \in B_F(0,c)$. Our goal is to find $x \in E$ such that

$$\|x\|_E < 1, \qquad Tx = y.$$

By the previous step, we can find $z \in B_E(0,\frac{1}{2})$ such that $\forall\, \epsilon > 0$, $\exists\, z \in E$ such that

$$\|z\|_E < \frac{1}{2}, \qquad \textit{and} \qquad \|y - Tz\|_F < \epsilon.$$

Let $\epsilon = \frac{c}{2}$. Then $\exists\, z_1 \in E$ such that

$$\|z_1\|_E < \frac{1}{2}, \qquad \textit{and} \qquad \|y - Tz_1\|_F < \frac{c}{2}.$$

We can keep this process going: since $Tz_1 - y \in B_F(0,\frac{c}{2})$, $\exists\, z_2 \in B_E(0,\frac{1}{4}) \subset E$ such that

$$\|z_1\|_E < \frac{1}{4}, \qquad \textit{and} \qquad \|(y - Tz_1) - Tz_2\|_F < \frac{c}{2}.$$

Hence we obtain a sequence $\{z_n\}_n$ satisfying

$$\|z_n\|_E < \frac{1}{2^n}, \qquad \textit{and} \qquad \|y - T(z_1 + z_2 + ... + z_n)\|_F < \frac{c}{2^n}.$$

It is easy to see that $x_n := z_1 + ... + z_n$ is a Cauchy sequence converging to some $x \in E$ with $\|x\|_E < 1$. Also, by continuity; $x_n \to x \implies Tx_n \to Tx = y$.

■

This concludes the proof.

□

The next result is an immediate consequence of Theorem 2.2:

Corollary 2.2.1 (Continuous inverse). *Let E,F be Banach spaces, $T \in \mathcal{L}(E,F)$ bijective. Then $T^{-1} : F \to E$ is continuous.*

Theorem 2.3 (Closed Range). *Let X, Y be Banach spaces, $A \in \mathcal{L}(X,Y)$. The following statements are equivalent*

(i) $\mathrm{Im}(A)$ *is closed,*

(ii) $\mathrm{Im}(A^*)$ *is closed,*

(iii) $\mathrm{Im}(A) = (\mathrm{Ker}(A^*))^{\perp}$,

(iv) $\mathrm{Im}(A^*) = (\mathrm{Ker}(A))^{\perp}$.

2.2 Existence and Uniqueness in Hilbert Spaces

Let V be a Hilbert space, $a(\cdot,\cdot) : W \times V \to \mathbb{R}$ a bilinear form, $f \in V^*$. Consider the *Petrov-Galerkin formulation*

$$\textit{Find } u \in W \quad \textit{such that} \quad a(u,v) = \langle f, v\rangle_V \quad \forall v \in V, \tag{2.2}$$

which can be deduced from a boundary-value problem like (A) in Chapter 1. Now, we may ask, what are the sufficient conditions for existence and uniqueness of a solution for this problem? We will observe that the facts explored in Section 1.6 with the Poisson equation will be exactly what is needed.

2.2.1 Lax-Milgram Theorem

In this subsection we assume $W = V$ in Problem (2.2). To write a proof of the Lax-Milgram Theorem, we first establish this simple (but powerful) result:

Lemma 2.1. *Let X, Y be Banach spaces and let $T \in \mathcal{L}(X,Y)$ be injective. Then T has closed range if and only if there exists $c > 0$ such that*

$$\|Tx\|_Y \geq c\|x\|_X \qquad \forall x \in X.$$

Proof. Assume the range of T to be closed in Y. Then by the continuous inverse Corollary (Corollary 2.2.1), $T : X \to \mathrm{Im}(T)$ admits a continuous inverse, that is, there exists $c > 0$ such that

$$\|T^{-1}y\|_X \leq c\|y\|_Y \qquad \forall\, y \in \mathrm{Im}(T).$$

This further implies

$$\|x\|_X \leq c\|Tx\|_Y \qquad \forall\, x \in X.$$

Conversely, let $\|Tx\|_Y \geq c\|x\|_X$ for all $x \in X$. Let $\{y_n\}_n \in \mathrm{Im}(T)$ be a sequence converging to $y \in Y$, $Tx_n = y_n$ for all $n \in \mathbb{N}$. Then

$$\|y_n - y_m\|_Y \geq c\|x_n - x_m\|_X \qquad \forall m, n \in \mathbb{N},$$

so $\{x_n\}_n$ is a Cauchy sequence in X. Thus $x_n \to x \in X$. By continuity of T, we have $Tx_n \to Tx \in \mathrm{Im}(T)$ and $Tx = y$. Hence $\mathrm{Im}(T)$ is closed, and the proof is complete.

□

The next theorem is the much celebrated *Lax-Milgram* Theorem. It is a remarkable result, and was proved by Peter Lax and Arthur Milgram in 1954 (see [25]).

Theorem 2.4 (The Lax-Milgram Theorem). *Let $W = V$ be a Hilbert space and $a \in \mathcal{L}(V \times V; \mathbb{R})$ a continuous bilinear form which is coercive, i.e.*

$$a(v, v) \geq m\|v\|_V^2 \qquad \forall\, v \in V,\ m \in \mathbb{R}.$$

Then for any $f \in V^*$, there exist a unique solution of (2.2)*. Furthermore, the following estimate holds:*

$$m\|u\|_V \leq \|f\|_{V^*}. \tag{2.3}$$

Proof. Pick $w \in V$. Define the operator A_w as $A_w(v) := a(w, v)$ for all $v \in V$. Then $A_w \in V^*$ because it is linear (a is bilinear) and bounded: for fixed $w \in V$, there exists $M > 0$ such that

$$|A_w(v)| = |a(w, v)| \leq M\|v\|_V.$$

Then $A_w(v) = f(v)$ for all $v \in V$. Now we can define $A : V \to V^*$ as $Av = A_v$ for all $v \in V$. The operator $A \in \mathcal{L}(V, V^*)$, because A is linear and by

$$\|Aw\|_{V^*} = \|A_w\|_{V^*} = \sup_{v \in V} \frac{|A_w(v)|}{\|v\|_V} \leq M\|w\|_V,$$

A is bounded.

The task now is to prove that A is bijective. First, for the error estimate in equation (2.3), we have

$$\|f\|_{V^*} = \|Aw\|_{V^*} = \sup_{v \in V} \frac{\langle Aw, v\rangle_{V^*,V}}{\|v\|_V} \geq \frac{a(w, w)}{\|w\|_V} \geq m\|w\|_V, \tag{2.4}$$

for fixed $w \in V$. From this follows injectivity: let $w_1, w_2 \in V$ be such that $Aw_1 = Aw_2$. Then by the result in equation (2.4) we have

$$0 = \|Aw_1 - Aw_2\|_{V^*} = \|A(w_1 - w_2)\|_{V^*} \geq m\|w_1 - w_2\|_V. \tag{2.5}$$

By Lemma 2.1, $\text{Im}(A)$ is closed.

The orthogonal complement of $\text{Im}(A)$ is given by

$$[\text{Im}(A)]^\perp := \{\psi \in V^* \mid \langle \psi, \phi\rangle_{V^*} = 0 \quad \forall \phi \in \text{Im}(A)\}.$$

Since V is a Hilbert space, there exist unique $v_\psi, v_\phi \in V$ related to ψ, ϕ such that $\langle \psi, \phi\rangle_{V^*} = \langle v_\psi, v_\phi\rangle_V = 0$. This implies that $v_\psi = 0$ and thus $\langle v_\psi, v\rangle_V = \psi(v) = 0 \quad \forall v \in V$, so ψ is the zero functional and therefore

$$[\text{Im}(A)]^\perp = \{0\}.$$

This implies that $\text{Im}(A)$ is dense in V^*.

□

2.2.2 Babuška-Lax-Milgram Theorem

The next result is due to Ivo Babuška (1971) [4], and provides a significant generalization to the Lax-Milgram Theorem (Theorem 2.4) to problems posed with a Petrov-Galerkin formulation in Hilbert spaces.

Theorem 2.5 (Babuška-Lax-Milgram). *Let W, V be Hilbert spaces and $a \in \mathcal{L}(W, V)$. Assume that* *(i) a is weakly coercive (inf-sup). That is, $\exists \alpha > 0$ such that*

$$\inf_{\substack{w \in W \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{|a(w,v)|}{\|w\|_W \|v\|_V} \geq \alpha. \tag{2.6}$$

(ii) Let $v \in V, v \neq 0$ be fixed. Then

$$\sup_{w \in W} |a(w,v)| > 0. \tag{2.7}$$

Then there exists a unique $u \in W$ that solves (2.2)*. Moreover, there exists $M > 0$ such that u satisfies*

$$\|u\|_W \leq M \|f\|_{V^*} \tag{2.8}$$

Proof. By the Riesz Representation Theorem (Theorem 1.2), there exists $A \in \mathcal{L}(W,V)$ such that

$$\langle Aw, v \rangle_V = a(w,v)$$

and a furthermore a representative $z \in V$ for $f \in V^*$. We obtain the equivalent problem of finding $u \in W$ such that

$$Au = z.$$

We need to prove that A is bijective. The inf-sup condition in (i) gives

$$\alpha \leq \inf_{\substack{w \in W \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{|a(w,v)|}{\|w\|_W \|v\|_V} \leq \inf_{\substack{w \in W \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{\|Aw\|_V \|v\|_V}{\|w\|_W \|v\|_V} = \inf_{\substack{w \in W \\ u \neq 0}} \frac{\|Aw\|_V}{\|w\|_W}, \tag{2.9}$$

which implies $\|Aw\|_V \geq \alpha \|w\|_W$ and injectivity follows as in equation (2.5). To show surjectivity, we prove that $\text{Im}(A) = V$. That is, $\text{Im}(A)$ is closed and dense in V. The fact that A has closed range follows directly from Lemma 2.1.

$\text{Im}(A)$ is a subspace of a Hilbert space V. Let $y \in V$ be chosen such that

$$\langle Aw, y \rangle_V = 0 \qquad \forall\, w \in W.$$

Then $a(w,y) = 0$ and $\sup_{w \in W} |a(w,y)| = 0$. This contradicts our hypothesis unless $y = 0$, and so $[\text{Im}(A)]^\perp = \{0\}$. Thus $\text{Im}(A)$ is dense in V.

The error estimate (2.8) is derived directly from (2.9): There exists $M > 0$ such that

$$\|u\|_W \leq \frac{1}{\alpha} \|Au\|_V = M \|f\|_{V^*}.$$

This concludes the proof.

□

2.3 Existence and Uniqueness in Banach Spaces

From the proofs presented in the previous existence and uniqueness theorems, it is clear that it is equivalent to consider the problem

$$\textit{Find } u \in W \quad \textit{such that} \quad Au = f, \tag{2.10}$$

where $A : W \to V^*$ and $f \in V^*$. Before we start to talk about existence and uniqueness for the case of V, W being Banach spaces, we want to find necessary conditions for A to be bijective. The Banach space version of our model problem reads:

$$\textit{Find } u \in W \quad \textit{such that} \quad a(u,v) = \langle f, v \rangle_{V^*,V} \qquad \forall v \in V. \tag{2.11}$$

Let $T \in \mathcal{L}(X, Y)$. By now it is clear that the bounding property in Lemma 2.1:

$$\|Tx\|_Y \geq c\|x\|_X \qquad \forall x \in X,$$

is equivalent to saying that T is injective and has closed range. If T has the bounding property, then $\operatorname{Ker}(T) = \{0\}$ and thus by the Closed Range Theorem (Theorem 2.3), T^* has closed range and $\operatorname{Im}(T^*) = [\operatorname{Ker}(T)]^\perp = \{0\}^\perp \implies \operatorname{Im}(T^*)$ dense in X^*, so T^* is surjective. This process can also be *reversed*: if the previous hold, then T^* injective $\implies$ T surjective. We obtain the result

$$T \text{ bounding and } T^* \text{ injective} \implies T \text{ bijective.}$$

We will apply this result to prove the next theorem. Condition (i) in Theorem 2.6 will imply that A is bounding, while condition (ii) will imply that A^* is injective.

We could of course have this reasoning in the previous theorems for Hilbert spaces, but this spices things up a bit, and it is fun and fruitful to construct different proofs. It is worth noting that we require V to be reflexive. We need this for the process to go smoothly when we define the adjoint operator $A^* : (V =)\ V^{**} \to W^*$ for Problem (2.10).

The next theorem can be found in [19], and is often referred to as the *Banach-Nečas-Babuška* Theorem, or in short the *BNB* Theorem. It was first stated in 1962 by Jindřich Nečas [27], and popularized by Ivo Babuška in 1972 [5]. The proof of this theorem will be a direct consequence of the Open Mapping Theorem (Theorem 2.2) and the Closed Range Theorem (Theorem 2.3), and this is from where Stefan Banach gets his name attatched to the results, which were proved in his 1932 groundbreaking work *Théorie des opérations linéaires* [6].

Theorem 2.6 (Banach-Nečas-Babuška). *Let W be a Banach space and V a reflexive Banach space. Let $a \in \mathcal{L}(W \times V, \mathbb{R})$ and $f \in V^*$. Then,* (2.11) *has a unique solution iff*

(i) $\exists\, \alpha > 0$ *such that*

$$\inf_{w \in W} \sup_{v \in V} \frac{a(w, v)}{\|w\|_W \|v\|_V} \geq \alpha.$$

(ii) Let $v \in V$. Then

$$a(w, v) = 0 \quad \forall\, w \in W \qquad \implies \qquad v = 0.$$

Moreover, the following a priori estimate holds

$$\|u\|_W \leq \frac{1}{\alpha}\|f\|_{V^*} \quad \forall\, f \in V^*.$$

Proof. In the same way as in the proof of the Lax-Milgram Theorem (Theorem 2.4), we construct the operators $A_w \in V^*$ and $A \in \mathcal{L}(W, V^*)$. This process should be seamless. By the Theorem on the existence of a unique adjoint operator (Theorem 1.1), there exists a unique $A^* \in \mathcal{L}(V^{**}, W^*) = \mathcal{L}(V, V^*)$ defined by

$$\langle Aw, v\rangle_{V^*,V} = \langle w, Av\rangle_{W,W^*} = a(w, v) \qquad \forall w \in W,\ \forall v \in V.$$

Assume $(i), (ii)$ holds. We begin the proof by showing that A is bijective. Statement (i) implies

$$\|Aw\|_{V^*} \geq \alpha\|w\|_W \qquad \forall w \in W,$$

and so A has the bounding property. By this and linearity of A, Lemma 2.1 implies that $\operatorname{Im}(A)$ is closed. (ii) implies that $\operatorname{Ker}(A^*) = \{0\}$. By the Closed Range Theorem (Theorem 2.3),

$\text{Im}(A) = \{0\}^{\perp} \implies \text{Im}(A)$ dense in V^* and thus A is bijective.

Conversely, assume A is bijective. We want to show that this proves that the statements (i) and (ii) hold. By Lemma 2.1, we have that

$$\left.\begin{array}{l}\text{A injective}\\ \text{Im}(A) = V^*\\ V^* \text{ Banach}\end{array}\right\} \implies \|Aw\|_{V^*} \geq \alpha \|w\|_W \qquad \forall w \in W.$$

Combining this with the dual norm

$$\|Aw\|_{V^*} = \sup_{v \in V} \frac{\langle Aw, v\rangle_{V^*,V}}{\|v\|_V} = \sup_{v \in V} \frac{a(w,v)}{\|v\|_V} \qquad \forall w \in W,$$

implies the inf-sup condition (i). A is surjective and thus by the Closed range Theorem (Theorem 2.3) we have $\text{Im}(A) = [\text{Ker}(A^*)]^{\perp} = V^* \implies \text{Ker}(A^*) = \{0\}$. A^* is therefore injective, so A^*v is the zero functional in W^* iff $v = 0$. Consequently,

$$\langle A^*v, w\rangle_{W^*,W} = a(w,v) = 0 \qquad \forall w \in W$$

implies that $v = 0$, which proves (ii).

This concludes the proof.

□

Remark 2.6.1. An interesting result which also follows from playing around with the Closed Range Theorem ideas is that if $T \in \mathcal{L}(X,Y)$ with X, Y Banach spaces, then

$$T \text{ bijective} \iff T^* \text{ bijective}.$$

Chapter 3

Non-linear Problems

Solving non-linear partial differential equations is extremely important as many of the real world mathematical models are based on non-linear equations. The focus of this chapter is to explain different methods for studying existence and uniqueness of such problems. Inspired by the ideas of the previous chapter, we focus on the following type of variational formulations:

Let V be a Hilbert space and $\langle \cdot, \cdot \rangle_V$ an inner product on V. Find $u \in V$ such that

$$a(u,v) = f(v) \; \forall v \in V, \quad \textit{where} \quad a(u,v) := \langle b(u), v \rangle_V. \tag{3.1}$$

As we will see later, many non-linear equations have this form (sometimes after applying a variable transformation). It is also possible to extend the analysis to operators $b : X \to X^*$, where X is a Banach space and $\langle \cdot, \cdot \rangle_X$ is the duality product between X^* and X.

The existence and uniqueness of problems like (3.1) relies on the properties of the function $b : V \to V$. In our analysis, we will assume b to be a continuous monotone operator (see definition 1.7), mapping $\mathbb{R}$ to $\mathbb{R}$ (thus giving a coefficient depending on the values of u). The monotonicity is motivated by the Minty-Browder Theorem (see Section 1.2, page 4). For the continuity, we look at the possibilities that follows if b is Lipschitz or Hölder continuous.

In section 3.1, we explore Fixed Point Theorems (Banach and Brouwer) and sketch how they can be used to prove existence of a solution of variational formulations. The main part of this chapter (section 3.2) is dedicated to an application of these ideas to a weak formulation of a non-linear, possibly degenerate partial differential equation: *the Richards equation*. The different types of properties for the coefficient function $b(\cdot)$ we will explore in this section are as follows:

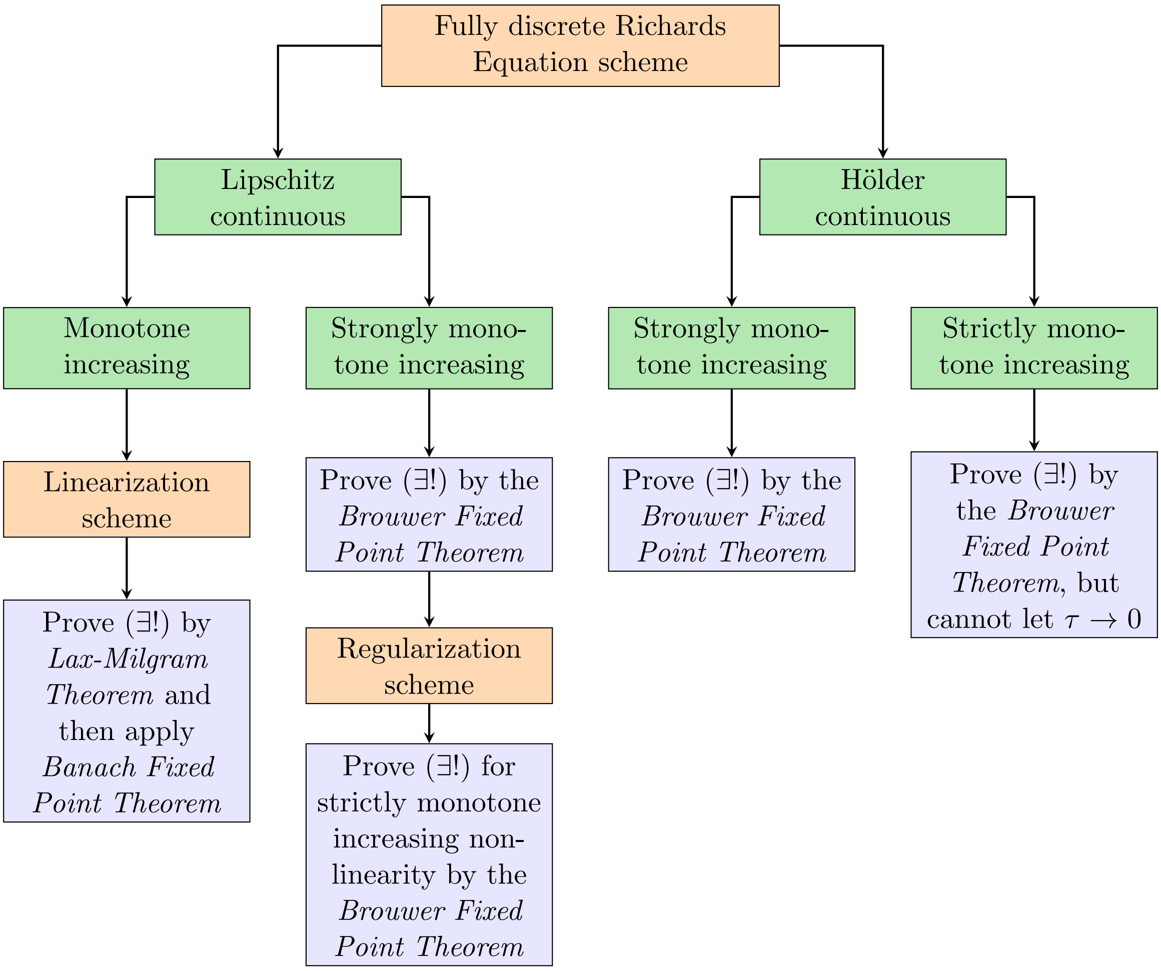
Case 1: $b(\cdot)$ linear with $b(u) = u$. We prove existence and uniqueness by using the Lax-Milgram Theorem (Theorem 2.4).

Case 2: $b(\cdot)$ Lipschitz continuous and monotone. First we consider a linearization of the variational formulation. Here we will prove ($\exists!$) by Lax-Milgram (Theorem 2.4) and show convergence to the original formulation by the Banach fixed point Theorem (Theorem 3.1). Secondly, we will assume strong monotonicity for $b(\cdot)$ and show how the Brouwer Fixed Point Theorem (Theorem 3.2) can be applied to prove existence directly of the non-linear formulation. Lastly, we will also look at how we can extend this to the weaker condition of $b(\cdot)$ being monotone increasing.

Case 3: $b(\cdot)$ Hölder continuous, monotone and bounded. We prove a similar result as in the previous step, using the Brouwer Fixed Point Theorem.

In Section 3.3, we extend the previous theory to the case of a coupling of two PDEs (the Richards equation coupled with a transport equation) discretized through mixed finite elements.

In Figure 3.1 below, a flowchart of the results and the assumptions from which they are derived from is given.



Fully discrete Richards Equation scheme
Lipschitz continuous
Hölder continuous
Monotone increasing
Strongly monotone increasing
Strongly monotone increasing
Strictly monotone increasing
Linearization scheme
Prove (∃!) by the *Brouwer Fixed Point Theorem*
Prove (∃!) by the *Brouwer Fixed Point Theorem*
Prove (∃!) by the *Brouwer Fixed Point Theorem*, but cannot let $\tau \to 0$
Prove (∃!) by *Lax-Milgram Theorem* and then apply *Banach Fixed Point Theorem*
Regularization scheme
Prove (∃!) for strictly monotone increasing nonlinearity by the *Brouwer Fixed Point Theorem*

Figure 3.1: Flowchart of results derived in this chapter. The orange boxes represent the schemes, the green assumptions, and the blue results.

3.1 Fixed Point Theorems

In this section we will give the proofs of two important fixed point theorems and discuss how their applications for variational formulations. The most celebrated result is the Banach Fixed Point Theorem:

Theorem 3.1 (The Banach Fixed Point Theorem). *Let X be a Banach space, U a closed subspace of X, and $f : U \to X$ Lipschitz continuous with Lipschitz constant $L < 1$. Assume $f(U) \subseteq U$. Then $(\exists!)$ $x^* \in U$ such that $f(x^*) = x^*$. Moreover, the sequence defined by $x_n = f(x_{n-1})$, $n \geq 1$, with x_0 chosen arbitrarily, converges to x^*. Further we have the a priori* *and a posteriori error estimates:*

$$\|x_k - x^*\|_X \leq \frac{L}{1-L}\|x_k - x_{k-1}\|_X,$$
$$\|x_k - x^*\|_X \leq \frac{L^k}{1-L}\|x_1 - x_0\|_X.$$

The proof follows the lines of [14].

Proof. Let $\|\cdot\| := \|\cdot\|_X$. Define a sequence $x_n := f(x_{n-1}),\ n \geq 1$ as in the hypothesis, with $x_0 \in U$ arbitrary. We will show that this is a Cauchy sequence. Let $k \geq 1$. Then, by the Lipschitz continuity of $f(\cdot)$,

$$\|x_{k+1} - x_k\| = \|f(x_k) - f(x_{k-1})\| \leq L\|x_k - x_{k-1}\| \leq \cdots \leq L^k\|x_1 - x_0\|. \tag{3.2}$$

It follows by (3.2), the Lipschitz continuity of $f(\cdot)$ and the triangle inequality that for any $m > n > 1 \in \mathbb{N}$,

$$\begin{aligned}
\|x_m - x_n\| &= \|(x_m - x_{m-1}) + (x_{m-1} - x_{m-2}) + \cdots + (x_{n+1} - x_n)\| \\
&\leq \|x_m - x_{m-1}\| + \|x_{m-1} - x_{m-2}\| + \cdots + \|x_{n+1} - x_n\| \\
&\leq (L^{m-1} + L^{m-2} + \cdots + L^n)\|x_1 - x_0\| \\
&= L^n(L^{m-n-1} + L^{m-n-2} + \cdots + L)\|x_1 - x_0\| \\
&\leq L^n \frac{1}{1-L}\|x_1 - x_0\|.
\end{aligned}$$

The last inequality is due to

$$L + \dots L^{m-n-1} \leq \sum_{k=0}^{\infty} L^k = \frac{1}{1-L},$$

for $L < 1$. Thus we obtain

$$\|x_m - x_n\| \to 0 \qquad \text{as} \qquad m, n \to \infty,$$

which implies that $\{x_k\}_k$ is a Cauchy sequence. Since $U \subseteq X$ is a closed subspace of a Banach space, U is itself a Banach space, and thus $x_k \in U\ \forall k \geq 1$ gives the existence of some $x^* \in U$ such that $x_n \to x^*$ in U. This proves the convergence of $\{x_k\}_k$ to a fixed point $x^* = f(x^*)$.

For the uniqueness, assume that there exists fixed points $x, y \in U$ with $x \neq y$. Then

$$\|x - y\| = \|f(x) - f(y)\| \leq L\|x - y\|$$

implies $L \geq 1$, which contradicts the hypothesis that $L < 1$.

For the error estimates, let $k \geq 1$. Then

$$\begin{aligned}
\|x_k - x^*\| &= \|(x_{k+1} - x_k) - (x_{k+1} - x^*)\| \\
&\leq \|x_{k+1} - x_k\| - \|x_{k+1} - x^*\| \\
&= \|f(x_k) - f(x_{k-1})\| + \|f(x_k) - f(x^*)\| \\
&\leq L\|x_k - x_{k-1}\| + L\|x_k - x^*\|,
\end{aligned}$$

which yields the a posteriori error estimate

$$\|x_k - x^*\| \leq \frac{L}{1-L}\|x_k - x_{k-1}\|. \tag{3.3}$$

The a priori error estimate is derived directly as a consequence of equations (3.2) and (3.3):

$$\|x_k - x^*\| \leq \frac{L^k}{1-L}\|x_1 - x_0\|.$$

this concludes the proof.

□

In the next section we will see an application of this theorem to a linearized version (an iterative method) of a non-linear variational formulation. The process of showing that the solution of the linear problem converges to the non-linear solution will follow two steps:

(i) Using the theory from Chapter 2 to prove existence and uniqueness for the linearized problem.

(ii) Assuming that such a solution exists, we consider two consecutive solutions u_n^i, u_n^{i+1} of the linearized problem. Define a function $\mathcal{F}(u_n^i) = u_n^{i+1}$, and show that it is Lipschitz continuous with $L_{\mathcal{F}} < 1$.

Then the Banach fixed point theorem implies existence of a unique u_n such that $u_n^i \to u_n$ as $i \to \infty$.

There will be times when we do not have Lipschitz continuity. As an example $\mathcal{F}$ might be Hölder continuous. The Brouwer Fixed Point Theorem gives sufficient conditions for the existence of fixed points of continuous functions:

Theorem 3.2 (The Brouwer Fixed Point Theorem). *Let K be a compact and convex subset of a finite-dimensional normed vector space, and $f : K \to K$ a continuous function. Then f has at least one fixed point.*

Proof. Let $B := B_1(0)$. Define $C^k(X,Y)$ as the space of functions $f : X \to Y$ with continuous k-th order partial derivatives. The proof of this theorem is due to [18] and is presented as in [15], p. 720. There are four steps:

(i) We show first that here is no mapping $v \in C^2(\overline{B}, \mathbb{R}^n)$ that satisfies

$$v(x) \in \partial B \quad \forall x \in \overline{B}, \qquad \text{and} \qquad v(x) = x \quad \forall x \in \partial B. \tag{3.4}$$

(ii) Second, we prove that there is no mapping $w \in C(\overline{B}, \mathbb{R}^n)$ that satisfies

$$w(x) \in \partial B \quad \forall x \in \overline{B}, \qquad \text{and} \qquad w(x) = x \quad \forall x \in \partial B. \tag{3.5}$$

(iii) Third, we show that any continuous mapping $g : \overline{B} \to \overline{B}$ has at least one fixed point in $\overline{B}$.

(iv) Finally, we extend the previous result to any compact and convex subset of $\mathbb{R}^n$.

Proof of step (i). Assume such a mapping exist. Define another mapping $\tilde{v} \in C^2(\overline{B}, \mathbb{R}^n)$ by $\tilde{v}(x) = x \; \forall x \in \overline{B}$. Now, since $v \equiv \tilde{v}$ on ∂B and for $F \in \mathbb{R}^{n \times n}$, the mapping $F \mapsto \det(F)$ is a null Lagrangian (Theorem 2, p. 463 in [17]), Theorem 2 (p. 461) in [17] implies

$$\int_B \det(\nabla v(x))dx = \int_B \det(\nabla \tilde{v}(x))dx = \int_B dx > 0,$$

where the last equality is due to the fact that $\nabla v(x)$ is the $n \times n$ identity matrix. We want to show that

$$\int_B \det(\nabla v(x))dx = 0,$$

thus yielding a contradiction. Let $\varphi : \overline{B} \to \mathbb{R}^n$ be defined by $\varphi(x) = |v(x)|^2$. φ is differentiable, and

$$\varphi'(x)h = 2\langle \nabla v(x)h, v(x)\rangle \qquad \forall h \in \mathbb{R}^n, \forall x \in B.$$

φ is also constant, because $v(x) \in \partial B \; \forall x \in \overline{B} \implies |v(x)| = 1$. Thus

$$0 = \varphi'(x)h = 2h^T(\nabla v(x))^T v(x) \qquad \forall h \in \mathbb{R}^n, \forall x \in B.$$

This further implies that

$$(\nabla v(x))^T v(x) = 0 \qquad \forall x \in B.$$

By (3.4), $v(x) \neq 0$ for all $x \in B$. Thus $v(x)$ is an eigenvector to $(\nabla v(x))^T$ with corresponding eigenvalue 0. An elementary linear algebra fact states that a matrix has non-zero determinant if and only if 0 is not an eigenvalue. Thus

$$\det((\nabla v(x))^T) = \det(\nabla v(x)) = 0 \qquad \forall x \in B.$$

This yields a contradiction.

■

Proof of step (ii). Define $w \in C(\overline{B}, \mathbb{R}^n)$ such that

$$w(x) \in \partial B \quad \forall x \in \overline{B}, \qquad \text{and} \qquad w(x) = x \quad \forall x \in \partial B. \tag{3.6}$$

We will show that this implies the existence of a mapping in $C^2(\overline{B}, \mathbb{R}^n)$ with the same properties, thus contradicting step *(i)*.

Extend the domain of w to $\mathbb{R}^n$ by $w(x) := x$ for $|x| > 1$. The mapping now satisfies

$$w \in C(\mathbb{R}^n, \mathbb{R}^n), \qquad w(x) = x \text{ for } |x| \geq 1, \qquad |w(x)| = 1 \text{ for } |x| < 1. \tag{3.7}$$

For each $i \in \{1, \dots, n\}$, let $w_{i,\epsilon}$ be the convolution of the i-th component of w and a sequence $\{\eta_\epsilon\}_\epsilon$ of standard mollifiers (as in Definition 1.11). Then

$$w_{i,\epsilon}(x) := \int_{B_\epsilon(0)} \eta_\epsilon(x-y)w_i(y)dy \qquad \forall x \in \mathbb{R}^n$$

resides in $C^\infty(\mathbb{R}^n)$ for all $\epsilon > 0$, and one can show that there exists $0 < \epsilon_0 \leq 1$ such that $w_\epsilon := (w_{1,\epsilon}, \dots, w_{n,\epsilon}) \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ satisfies

$$|w_\epsilon(x)| > 0 \qquad \forall \epsilon \leq \epsilon_0 \text{ and } \forall \; |x| \leq 2$$

since $|w(x)| \geq 1$ for all $|x| \leq 2$. Moreover, we have

$$w_{i,\epsilon}(x_i) = \int_{B_\epsilon(0)} \eta_\epsilon(y)w_i(x_i - y)dy = x_i \qquad \forall \; |x| \geq 2,$$

so $w_\epsilon(x) = x$ for all $|x| \geq 2$ and $\epsilon \leq \epsilon_0$. Based on this, define

$$v(x) = \frac{w_{\epsilon_0}(2x)}{|w_{\epsilon_0}(2x)|} \qquad \forall |x| \leq 1.$$

This mapping is in $C^\infty(\mathbb{R}^n; \mathbb{R}^n)$ and satisfies

$$|v(x)| = 1 \quad \forall x \in \overline{B} \qquad \text{and} \qquad v(x) = x \quad \forall x \in \partial B,$$

which is a contradiction to (3.4).

■

Proof of step (iii). Let $g : \overline{B} \to \overline{B}$ be a continuous map. Assume per reducto ad absurdum that g has no fixed point. Pick $x \in \overline{B}$. There exists a uniquely defined point $w(x)$ and a real number $\alpha(x) \geq 1$ such that

$$w(x) \in \partial B \qquad \text{and} \qquad w(x) = g(x) + \alpha(x)(x - g(x)).$$

If $x \in \partial B$, we specify $\alpha(x) = 1$, so $w(x) = x$. We want to show that this $w(x)$ is a continuous mapping. If so, $w \in C(\overline{B}; \mathbb{R}^n)$ and satisfies

$$w(x) \in \partial B \quad \forall x \in \overline{B} \qquad \text{and} \qquad w(x) = x \quad \forall x \in \partial B,$$

thus giving a contradiction to step *(ii)*. $\alpha : \overline{B} \to \mathbb{R}$ is continuous, since $\forall\ x \in \overline{B}$, $\alpha(x)$ is the unique root ≥ 1 of the polynomial

$$\lambda \in \mathbb{R} \quad \mapsto \quad \lambda^2 |x - g(x)|^2 + 2\lambda(x - g(x)) \cdot g(x) + |g(x)|^2 - 1.$$

This is due to:

$$\begin{aligned} &\alpha^2|x-g|^2 + 2\alpha(x-g)\cdot g + |g|^2 - 1 \\ &\quad = |w-g|^2 + 2(w-g)\cdot g + |g|^2 - 1 \\ &\quad = |w|^2 - 2w\cdot g + |g|^2 + 2w \cdot g - 2|g|^2 + |g|^2 - 1 \\ &\quad = 0. \end{aligned}$$

The coefficients of this polynomial are continuous functions of $x \in \overline{B}$. Consequently w is continuous, because it it composed of continuous functions. This contradicts (3.5), so g must have at least one fixed point.

■

Proof of step (iv). The result in the previous step will also hold if B is replaced by any ball $B_r(0)$ of radius $r > 0$ centered at the origin. Let K be a compact and convex subset of $\mathbb{R}^n$. Then $\exists\ r > 0$ such that $K \subset \overline{B_r(0)}$. Let $P : \mathbb{R}^n \to K$ be a projection operator. That is, $P(x) \in K$ is defined as

$$|x - P(x)| = \inf_{z \in K} |x - z| = \text{dist}(x, K).$$

By Theorem 4.3-1 (p. 183) in [15], P is continuous (it is here that we make use of the convexity of K). Let $f : K \to K$ be a continuous function. Then the mapping

$$g := f \circ P : \overline{B_r(0)} \to \overline{B_r(0)}$$

is continuous and by step *(iii)*, g has at least one fixed point $x_0 \in K$. Thus

$$x_0 = g(x_0) = f(P(x_0)) = f(x_0),$$

which implies that $f(x_0) = x_0$. Therefore f has at least one fixed point.

■

This concludes the proof of the Brouwer Fixed Point Theorem.

□

The next result is of great importance in proving existence for variational formulations of PDEs in Hilbert spaces.

Corollary 3.2.1. *Let H be a finite-dimensional Hilbert space. Let $f : H \to H$ be continuous with the following property*

$$\forall\, v \in H,\ \|v\| = M, \qquad \langle f(v), v\rangle \geq 0. \tag{3.8}$$

Then $\exists\, v_0 \in H$ with $\|v_0\| \leq M$ such that

$$f(v_0) = 0.$$

Proof. Assume $f(v) \neq 0\ \forall v \in H$ satisfying $\|v\| \leq M$. Define a mapping $F : \overline{B_M(0)} \to \overline{B_M(0)}$ (where $\overline{B_M(0)} \subset H$) by

$$F(v) = \frac{-Mf(v)}{\|f(v)\|}.$$

The function F is continuous because f is. By the Brouwer Fixed Point Theorem (Theorem 3.2) there exists $v_0 \in \overline{B_M(0)}$ such that

$$F(v_0) = \frac{-Mf(v_0)}{\|f(v_0)\|} = v_0.$$

Thus

$$\langle f(v_0), v_0\rangle = \langle f(v_0), \frac{-Mf(v_0)}{\|f(v_0)\|}\rangle = -M\|f(v_0)\|^2 < 0,$$

which contradicts (3.8). Hence $f(v) = 0$.

□

As stated previously, this corollary will be essential in proving existence whenever we are working with variational formulations on Hilbert spaces. To illustrate this, consider the problem:

$$\textit{Find } u \in V_h \textit{ such that } \langle Au, v_h\rangle = \langle f, v_h\rangle \qquad \forall v_h \in V_h, \tag{3.9}$$

where V_h is a finite-dimensional subspace of $L^2(\Omega)$, $\langle\cdot,\cdot\rangle$ is the $L^2(\Omega)$ inner product, $A : V_h \to V_h$ and f is either an element of V_h or $\langle f, v_h\rangle$ is f acting on v_h for $f : V_h \to \mathbb{R}$. We form an orthonormal basis $\{\varphi_1, \dots, \varphi_k\}$ on $V_h \subset L^2(\Omega)$. Assuming $\dim(V_h) = k$, we can construct some $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k$ such that $\|\bar{u}\|_{L^2(\Omega)} = |\alpha|_k$ for any $\bar{u} = \sum_{i=1}^k \alpha_i\varphi_i \in V_h$. We define $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}^k$ as

$$\mathcal{F}_i(\alpha) := \langle A\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - f, \varphi_i\rangle = \langle A\bar{u} - f, \varphi_i\rangle. \tag{3.10}$$

for $i = 1, \dots, k$. Next, we want to get a lower bound with respect to $|\alpha|_k$ on

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} := \sum_{i=1}^k \langle A\bar{u} - f, \varphi_i\rangle\alpha_i = \langle A\bar{u} - f, \bar{u}\rangle,$$

where $\langle\cdot,\cdot\rangle_{\mathbb{R}^k}$ is the Euclidean inner product. To satisfy the hypothesis of Corollary 3.2.1, we pick M based on the previous estimate. If this holds true for the Euclidean inner product and norm on $\mathbb{R}^k$, there exists a bounded $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_k) \in \mathbb{R}^k$ such that $u := \sum_{i=1}^k \hat{\alpha}_i\varphi_i$ satisfies $\mathcal{F}(\hat{\alpha}) = 0$ and thus is a solution to the variational formulation (3.9). We will prove a result based on this in the next section.

3.2 The Richards Equation

Let $\Omega \in \mathbb{R}^d$ be an open and bounded domain with Lipschitz continuous boundary Γ, and $t \in (0,T]$, where $T > 0$ is the final time. In this section we will consider a non-linear variational formulation that can be motivated from the *Richards equation*

$$\partial_t \Theta(\psi) - \nabla \cdot (K(\Theta(\psi))\nabla(\psi + z)) = f \qquad \text{in } (0,T] \times \Omega, \tag{3.11}$$

which is a non-linear, possibly degenerate parabolic PDE. This equation is used to model water flow in saturated/unsaturated porous media, where $\psi = \psi(t,x)$ is the pressure head, Θ the water content, f a source term, K the hydraulic conductivity, and z the height in the gravitational direction. To achieve more regular unknowns and reduce the non-linearities to a single one, it is useful to apply the *Kirchhoff transform*:

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\to \mathbb{R} \\ \psi &\mapsto \int_0^{\psi} K(\Theta(s))ds. \end{aligned}$$

If we now define $u := \mathcal{K}(\psi)$ and let

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

the equation (3.11) becomes

$$\partial_t b(u) - \nabla \cdot (\nabla u + k(b(u))e_z) = f \quad \text{in} \quad (0,T] \times \Omega.$$

We refer to [1, 3, 26, 30, 33] for results concerning the mathematical analysis of this problem. Here, we will study the Richards equation without the gravity term. After imposing initial conditions for time and homogeneous Dirichlet boundary conditions for space, we want to find u such that

$$\begin{cases} \partial_t b(u) - \Delta u = f & \text{in } (0,T] \times \Omega, \\ u = u_0 & \text{on } 0 \times \Omega, \\ u = 0 & \text{on } (0,T] \times \Gamma. \end{cases} \tag{3.12}$$

In this section, we will study five variational formulations that can be related to this problem. These are denoted as follows:

(P) : Continuous non-linear weak formulation.

(P_τ^n) : Semidiscrete non-linear weak formulation (discrete in time).

$(P_\tau^{n,h})$: Fully discrete non-linear weak formulation (discrete in time and space).

$(P_{\tau,i}^{n,h})$: Fully discrete linearized weak formulation.

$(P_{\tau,\epsilon}^{n,h})$: Fully discrete regularized weak formulation.

We let $\langle \cdot, \cdot \rangle$ be the inner product in $L^2(\Omega)$ or the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, $\| \cdot \|_1 := \| \cdot \|_{H^1(\Omega)}$, $\| \cdot \|_{-1} := \| \cdot \|_{H^{-1}(\Omega)}$ and $\| \cdot \| := \| \cdot \|_{L^2(\Omega)}$. Next, we define the weak formulation of the Problem (3.11) using the notation of Bochner spaces (see Definition 1.15):

Problem (P)**:** Let $f \in L^2(0,T;H^{-1}(\Omega))$ and $u_0 \in L^2(\Omega)$ be given. A function u is called a *weak solution* of (3.12) iff $b(u) \in H^1(0,T;H^{-1}(\Omega))$, $u \in L^2(0,T;H_0^1(\Omega))$ and $u(0) = u_0$ satisfies

$$\int_0^T \langle \partial_t b(u), \phi \rangle + \langle \nabla u, \nabla \phi \rangle \, dt = \int_0^T \langle f, \phi \rangle \, dt$$

for all $\phi \in H^1(0,T;H_0^1(\Omega))$.

The requirement $b(u) \in H^1(0,T;H^{-1}(\Omega))$ comes from the fact that we need $\partial_t b(u) \in L^2(0,T;H^{-1}(\Omega))$ (see definition 1.16). Next, we discretize in time by using the Backward Euler method: Partition the interval $[0,T]$ into $N+1$ time steps t_n with uniform step length $\tau := t_n - t_{n-1}$ $\forall n \in \{1,\dots,N\}$. Define $u_n := u(t_n)$. We approximate

$$\partial_t b(u) \approx \frac{b(u_n) - b(u_{n-1})}{\tau}.$$

If $f \in C(0,T;H^{-1}(\Omega))$, we let $f_n := f(t_n)$. If not, we use the time average over $(t_{n-1}, t_n]$,

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

This alternative definition is not needed for u_n, because Theorem 3 (p. 303) in [17] implies $u \in C([0,T], L^2(\Omega))$ (after possibly being redefined on a set of measure zero). At time step t_n, we define the time discrete Problem (P_τ^n):

Problem (P_τ^n)**:** Let $n \in \{1,\dots N\}$, $u_{n-1} \in H_0^1(\Omega)$ be given. Find $u_n \in H_0^1(\Omega)$ such that

$$\langle b(u_n) - b(u_{n-1}), v \rangle + \tau \langle \nabla u_n, \nabla v \rangle = \tau \langle f_n, v \rangle \qquad \forall v \in H_0^1(\Omega),$$

where τ is the step length, so $t_n = n\tau$ and u_n is the solution at time step t_n.

Now we will prove a result for existence and uniqueness of the linear case $b(u) = u$ of (P_τ^n) by using Lax-Milgram (theorem 2.4):

Proposition 3.1. *Let $b(u) := u$. Then there exists a unique solution of problem (P_τ^n) for $\tau > 0$.*

Proof. Define

$$l(v) := \tau \langle f, v \rangle + \langle u_{n-1}, v \rangle,$$
$$a(u,v) := \langle u, v \rangle + \tau \langle \nabla u, \nabla v \rangle.$$

It is trivial that $l(\cdot)$ is linear and $a(\cdot,\cdot)$ is bilinear.

(i) $a(\cdot,\cdot)$ *bounded:*

$$|a(u,v)| \leq \|u\|\|v\| + \tau \|\nabla u\|\|\nabla v\| \leq C\|u\|_1\|v\|_1$$

where $C := \max\{1, \tau\}$.

(ii) $a(\cdot,\cdot)$ *coercive:*

$$a(u,u) = \|u\|^2 + \tau\|\nabla u\|^2 \geq m\|u\|_1^2$$

where $m := \min\{1, \tau\}$.

(iii) $l(\cdot)$ *bounded*:

$$|l(v)| \leq \tau\|f\|_{-1}\|v\|_1 + \|u_{n-1}\|\|v\| \leq C\|v\|_1,$$

where $C := \max\{\tau\|f\|_{-1}, \|u_{n-1}\|\}$.

Thus by the Lax-Milgram Theorem (Theorem 2.4) there exists a unique $u_n \in H_0^1(\Omega)$ solving (P_τ^n).

□

3.2.1 A Linearization Scheme

If $b(\cdot)$ is not linear, existence and uniqueness of problem (P_τ^n) can no longer be studied with the theory we discussed earlier, because what we have now resembles

$$a(u,v) := \langle b(u), v\rangle + \tau\langle \nabla u, \nabla v\rangle.$$

To be able to look at this in the eyes of Chapter 1, we have to perform a linearization. The *L-scheme* is a linearization scheme proposing to create an iterative method $u_n^{i+1} = \mathcal{F}(u_n^i)$ where we add one extra term

$$L\langle u_n^{i+1} - u_n^i, v\rangle$$

and then instead evaluate $b(\cdot)$ at the solution in the previous iteration, u_n^i. The first iteration at each time step will be given as the value at the previous time step; that is, the initial guess is $u_{n+1}^0 := u_n$. Define the difference between two consecutive solutions of the iterative method as $e_n^{i+1} := u_n^{i+1} - u_n^i$. For the convergence, we want to get an inequality resembling

$$\|e_n^{i+1}\| \leq C\|e_n^i\| \tag{3.13}$$

for $C < 1$. This will imply

$$\|\mathcal{F}(u_n^i) - \mathcal{F}(u_n^{i-1})\| \leq C\|u_n^i - u_n^{i-1}\|, \tag{3.14}$$

for which we can apply the Banach fixed point Theorem (Theorem 3.1).

From here on we also substitute $H_0^1(\Omega)$ with a finite-dimensional subspace $V_h \subset H_0^1(\Omega)$. This is not necessary for proving existence and uniqueness, but is in harmony with the finite element method, which requires the finite-dimensionality. We define the fully discrete (non-linear) problem $(P_\tau^{n,h})$:

Problem $(P_\tau^{n,h})$: Let $n \in \{1,\dots N\}$, $V_h \subset H_0^1(\Omega)$ be finite-dimensional, $u_{n-1} \in V_h$ be given. Find $u_n \in V_h$ such that

$$\langle b(u_n) - b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n, \nabla v_h\rangle = \tau\langle f_n, v_h\rangle \qquad \forall v_h \in V_h \tag{3.15}$$

where τ is the step length, so $t_n = n\tau$ and u_n is the solution at time step t_n.

Assuming $u_n^i \to u_n$ as $i \to \infty$, we define the linearized fully discrete problem $(P_{\tau,i}^{n,h})$:

Problem $(P_{\tau,i}^{n,h})$: Let $n \in \{1,\dots,N\}$ and $u_n^i, u_{n-1} \in V_h$ be given, with $u_n^0 := u_{n-1}$. Let $L > 0$ be a constant. Find $u_n^{i+1} \in V_h$ such that

$$\langle b(u_n^i) - b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n^{i+1}, \nabla v_h\rangle + L\langle u_n^{i+1} - u_n^i, v_h\rangle = \tau\langle f_n, v_h\rangle \tag{3.16}$$

for all $v_h \in V_h$.

Now we wish to prove two results:

(i) The existence of a unique solution of $(P_{\tau,i}^{n,h})$ for each i.

(ii) That $(P_{\tau,i}^{n,h})$ converges to $(P_\tau^{n,h})$ as $i \to \infty$. That is, the sequence of solutions $\{u_n^i\}_i$ of $(P_{\tau,i}^{n,h})$ converges to a unique $u_n \in V_h$ solving $(P_\tau^{n,h})$ as $i \to \infty$.

Proposition 3.2. *Let $n \in \{1,\dots,N\}$ and $i \in \mathbb{N}$ be fixed. Suppose $b(\cdot)$ is Lipschitz continuous with Lipschitz constant $L_b > 0$. Then there exists a unique solution of problem $(P_{\tau,i}^{n,h})$.*

Proof. We will apply the Lax-Milgram Theorem (Theorem 2.4) to show this: Let

$$\begin{aligned}\alpha(u,v) &:= \tau\langle \nabla u, \nabla v\rangle + L\langle u, v\rangle,\\ \beta(v) &:= \langle b(u_{n-1}) - b(u_n^{i-1}), v\rangle + L\langle u_n^{i-1}, v\rangle + \tau\langle f_n, v\rangle.\end{aligned}$$

$\alpha(\cdot,\cdot)$ is linear (trivial) and bounded:

$$|\alpha(u,v)| \leq \tau\|\nabla u\|\|\nabla v\| + L\|u\|\|v\| \leq M\|u\|_1\|v\|_1,$$

with $m := \tau + L$, and coercive

$$\alpha(v,v) = \tau\|\nabla v\|^2 + L\|v\|^2 \geq m\|v\|_1^2,$$

where $m := \min\{\tau, L\}$. At last, we show that β is bounded (it is clear that it is linear). There holds:

$$\begin{aligned}|\beta(v)| &\leq \big(\|b(u_{n-1}) - b(u_n^{i-1})\| + L\|u_n^{i-1}\| + \tau\|f\|\big)\|v\|\\ &\leq \big(\|u_{n-1} - u_n^{i-1}\| + L\|u_n^{i-1}\| + \tau\|f\|\big)\|v\|\\ &\leq m\|v\|_1.\end{aligned}$$

Thus by the Lax-Milgram Theorem, there exists a unique solution of Problem $(P_{\tau,i}^{n,h})$.

□

Proposition 3.3. *Let $n \in \{1,\dots,N\}$. Assume $u_n^i \in V_h$ solves $(P_{\tau,i}^{n,h})$ and $u_n \in V_h$ solves $(P_\tau^{n,h})$. Suppose $b(\cdot)$ is monotone increasing and Lipschitz continuous with Lipschitz constant $L_b > 0$. Then $u_n^i \to u_n$ as $i \to \infty$ whenever*

$$L \geq \frac{L_b}{2}.$$

Proof. The proof is based on the Banach Fixed Point Theorem (Theorem 3.1). Let $e_n^{i+1} := u_n^{i+1} - u_n^i$ be the difference between two iterations. We want to show that there exists $c < 1$ such that

$$\|e_n^{i+1}\| \leq c\|e_n^i\| \qquad \forall\, i \in \mathbb{N}.$$

We look at two consecutive solutions $u_n^{i+1}, u_n^i \in V$ of problem $(P_{\tau,i}^{n,h})$. Subtract the two equations (3.16) from each other to obtain

$$\begin{aligned}\langle b(u_n^i)-&b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n^{i+1}, \nabla v_h\rangle + L\langle u_n^{i+1} - u_n^i, v_h\rangle\\ &= \langle b(u_n^{i-1}) - b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n^i, \nabla v_h\rangle + L\langle u_n^i - u_n^{i-1}, v_h\rangle,\end{aligned}$$

for all $v_h \in V_h$. This is equivalent to

$$L\langle e_n^{i+1} - e_n^i, v_h\rangle + \tau\langle \nabla e_h^{i+1}, \nabla v_h\rangle + \langle b(u_n^i) - b(u_n^{i-1}), v_h\rangle = 0 \qquad \forall v_h \in V_h. \tag{3.17}$$

Choose now $v_h = e_n^{i+1}$ as test function in (3.17) above. Then, after expanding the last term,

$$\begin{aligned}&L\langle e_n^{i+1} - e_n^i, e_n^{i+1}\rangle + \tau\|\nabla \cdot e_n^{i+1}\|^2\\ &\quad + \langle b(u_n^i) - b(u_n^{i-1}), e_n^{i+1} - e_n^i\rangle = -\langle b(u_n^i) - b(u_n^{i-1}), e_n^i\rangle.\end{aligned}$$

Since we assumed $b(\cdot)$ to be Lipschitz and monotonically increasing, we have

$$\begin{aligned}\frac{1}{L_b}\|b(u_n^i) - b(u_n^{i-1})\|^2 &= \frac{1}{L_b}\int_\Omega |b(u_n^i(x)) - b(u_n^{i-1}(x))|^2\,dx \\ &\leq \int_\Omega |b(u_n^i(x)) - b(u_n^{i-1}(x))||u_n^i(x) - u_n^{i-1}(x)|\,dx \\ &\leq \langle b(u_n^i) - b(u_n^{i-1}), u_n^i - u_n^{i-1}\rangle.\end{aligned}$$

There holds also the obvious algebraic identity

$$2\langle e_n^{i+1} - e_n^i, e_n^i\rangle = \|e_n^{i+1}\|^2 + \|e_n^{i+1} - e_n^i\|^2 - \|e_n^i\|^2.$$

Next we apply the Cauchy-Schwarz inequality (Theorem 1.8), the Young inequality (Theorem 1.9), and the Poincaré inequality (Theorem 1.10) to get

$$\begin{aligned}\frac{L}{2}\|e_n^{i+1}\|^2 + \frac{L}{2}\|e_n^{i+1} - e_n^i\|^2 + \tau C_\Omega\|e_n^{i+1}\|^2 + \frac{1}{L_b}\|b(u_n^i) - b(u_n^{i-1})\|^2 \\ \leq \frac{L}{2}\|e_n^i\|^2 + \frac{1}{2L}\|b(u_n^i) - b(u_n^{i-1})\|^2 + \frac{L}{2}\|e_n^{i+1} - e_n^i\|^2.\end{aligned}$$

And thus for $L \geq \frac{L_b}{2}$, we have

$$\|e_n^{i+1}\|^2 \leq \frac{L}{L + 2\tau C_\Omega}\|e_n^i\|^2.$$

Which is equivalent to

$$\|\mathcal{F}(u_n^i) - \mathcal{F}(u_n^{i-1})\| \leq L_F\|u_n^i - u_n^{i-1}\|, \tag{3.18}$$

with $L_F < 1$. Hence $u_n^{i+1} = \mathcal{F}(u_n^i)$ is a Lipschitz continuous contraction for any $L \geq \frac{L_b}{2}$.

□

3.2.2 A First Application of the Brouwer Fixed Point Theorem

In the previous section, we performed a linearization scheme and studied existence and uniqueness for the new version of the problem, and proved that it did in fact converge to the solution of problem $(P_\tau^{n,h})$. Now we will look at an application of the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1); we wish to tackle the nonlinearity directly. Assuming the previous hypotheses hold true, there are two cases we need to split the problem into for proving existence:

(i) b is strongly monotone increasing, i.e. there exists $b_0 > 0$ such that

$$\langle b(u) - b(v), u - v\rangle \geq b_0\|u - v\|^2,$$

(ii) b is monotone increasing, but not necessarily strongly:

$$\langle b(u) - b(v), u - v\rangle \geq 0,$$

for all $u, v \in V_h$. In our study, we assume $b : \mathbb{R} \to \mathbb{R}$, so for $b(0) = 0$ these cases are imply (i) $b'(u) \geq b_0 > 0$ for all $u \in V_h$ and (ii) $b'(u) \geq 0$ for all $u \in V_h$.

To apply Corollary 3.2.1, note that we need also assume $V_h \subset H_0^1(\Omega)$ to be a finite-dimensional subspace. We recall the definition of Problem $(P_\tau^{n,h})$:

Problem $(P_\tau^{n,h})$: Let $n \in \{1, \dots N\}$, $V_h \subset H_0^1(\Omega)$ be finite-dimensional, $u_{n-1} \in V_h$ be given. Find $u_n \in V_h$ such that

$$\langle b(u_n) - b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n, \nabla v_h\rangle = \tau\langle f_n, v_h\rangle \qquad \forall v_h \in V_h \tag{3.19}$$

where τ is the time step length, so $t_n = n\tau$ and u_n is the solution at time step t_n.

Proposition 3.4. *Suppose $b(\cdot)$ is Lipschitz continuous with Lipschitz constant $L_b > 0$, strongly monotone increasing with $b' \geq b_0 > 0$, and $b(0) = 0$. Then there exists a unique solution of problem $(P_\tau^{n,h})$ for all $\tau > 0$.*

Proof. Existence: We will apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1). Let $\dim(V_h) = k$. Let $\{\varphi_1, \dots, \varphi_k\}$ be an orthogonal basis for V_h as a subspace of $H_0^1(\Omega)$ and orthonormal as a subspace of $L^2(\Omega)$. That is,

$$\begin{aligned}&\langle \varphi_i, \varphi_j\rangle_{L^2(\Omega)} = \delta_{ij},\\ &\langle \varphi_i, \varphi_j\rangle_{H^1(\Omega)} = 0 \qquad \text{for } i \neq j.\end{aligned}$$

Then for some element $\bar{u} \in V_h$ there exists $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k$ such that

$$\bar{u} = \sum_{i=1}^k \alpha_i \varphi_i, \qquad \text{and} \qquad \|\bar{u}\| = |\alpha|_k,$$

where $|\cdot|_k$ is the Euclidean norm defined as $|x|_k := \sqrt{\sum_{i=1}^k x_i^2}$ for $x = (x_1, \dots, x_k) \in \mathbb{R}^k$. Let $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}^k$ be defined by $\mathcal{F}(\alpha) = \hat{\alpha}$, where

$$\hat{\alpha}_i := \langle b(\bar{u}) - b(u_{n-1}), \varphi_i\rangle + \tau\langle \nabla\bar{u}, \nabla\varphi_i\rangle - \tau\langle f_n, \varphi_i\rangle, \qquad i = 1, \dots, k.$$

$\mathcal{F}$ is continuous:

$$\begin{aligned}|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2 &= \sum_{i=1}^k |\mathcal{F}_i(\alpha) - \mathcal{F}_i(\beta)|^2\\ &= \sum_{i=1}^k \Big|\langle b\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - b\big(\sum_{j=1}^k \beta_j\varphi_j\big), \varphi_i\rangle + \tau(\alpha_i - \beta_i)\|\nabla\varphi_i\|^2\Big|^2.\end{aligned}$$

Apply the inequality $\sum |a_n + b_n|^2 \leq 2\sum |a_n|^2 + 2\sum |b_n|^2$ for real sequences $\{a_n\}_n, \{b_n\}_n$; the Cauchy-Schwarz inequality (Theorem 1.8) and use the Lipschitz continuity of $b(\cdot)$ to obtain

$$\begin{aligned}&|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2\\ &\quad\leq 2L_b \sum_{i=1}^k \|\textstyle\sum_{j=1}^k(\alpha_j - \beta_j)\varphi_j\|^2 + 2\tau^2 \displaystyle\sum_{i=1}^k |\alpha_i - \beta_i|^2\|\nabla\varphi_i\|^4.\end{aligned}$$

We have

$$\begin{aligned}\big\|\textstyle\sum_{j=1}^k(\alpha_j - \beta_j)\varphi_j\big\|^2 &= \displaystyle\int_\Omega \big|\textstyle\sum_{j=1}^k(\alpha_j - \beta_j)\varphi_j\big|^2 dx\\ &\leq \int_\Omega \big(\textstyle\sum_{j=1}^k |\alpha_j - \beta_j|^2\big)\big(\sum_{j=1}^k |\varphi_j|^2\big) dx\\ &= |\alpha - \beta|_k^2 \int_\Omega \big(\sum_{j=1}^k |\varphi_j|^2\big) dx \leq M|\alpha - \beta|_k^2,\end{aligned}$$

where $M < \infty$. The second to last inequality is the Cauchy-Schwarz inequality (Theorem 1.8). The integral is bounded because $\dim(V_h) < +\infty$. Thus we get

$$|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2 \quad \leq \quad 2ML_b|\alpha - \beta|_k^2 + 2\tau^2 \max_{j \in \{1,\dots,k\}} \|\nabla\varphi_j\|^4 |\alpha - \beta|_k^2,$$

which implies that $\mathcal{F}$ is Lipschitz continuous and furthermore that $\mathcal{F}$ is continuous.

Next, we want to show that there exists a ball of radius M in $\mathbb{R}^k$ for which

$$\langle \mathcal{F}(\alpha), \alpha \rangle_k \geq 0 \qquad \forall \alpha \in \mathbb{R}^k \text{ such that } |\alpha|_k = M.$$

where $\langle \cdot, \cdot \rangle_k$ is the Euclidean inner product defined by $\langle x, y \rangle := \sum_{i=1}^k x_n y_n$ for $x, y \in \mathbb{R}^k$. We have

$$\begin{aligned}\langle \mathcal{F}(\alpha), \alpha \rangle_k &= \langle b(\bar{u}) - b(u_{n-1}), \bar{u} \rangle + \tau \|\nabla \bar{u}\|^2 - \tau \langle f_n, \bar{u} \rangle \\ &\geq \langle b(\bar{u}), \bar{u} \rangle - \|b(u_{n-1})\| \|\bar{u}\| + \tau C_\Omega \|\bar{u}\|^2 - \tau \|f_n\| \|\bar{u}\|,\end{aligned}$$

by Cauchy-Schwarz inequality (Theorem 1.8) and the Poincaré inequalitity (Theorem 1.10). Next, using the strong monotonicity of $b(\cdot)$ and the fact that $b(0) = 0$ we have $\langle b(u), u \rangle \geq b_0 \|u\|^2 \ \forall u \in V_h$. We also apply Young inequality (Theorem 1.9) to get

$$\langle \mathcal{F}(\alpha), \alpha \rangle_k \geq \frac{1}{2}(b_0 + \tau C_\Omega)|\alpha|_k^2 - m_1,$$

where $m_1 := \dfrac{1}{2b_0}\|b(u_{n-1})\|^2 + \dfrac{\tau}{2C_\Omega}\|f_n\|^2$. Indeed, for all $\alpha \in \mathbb{R}^k$ satisfying

$$|\alpha|_k = \sqrt{\frac{2m_1}{b_0 + \tau C_\Omega}},$$

we have

$$\langle \mathcal{F}(\alpha), \alpha \rangle_k \geq 0.$$

Thus there exists a bounded $\alpha^0 \in \mathbb{R}^k$ such that $\mathcal{F}(\alpha^0) = 0$. Thus $\hat{\alpha}_i^0 = 0$ for all $i \in \{1, \dots, k\}$, and therefore for all $v_h \in V_h$. Consequently, we obtain that

$$u := \sum_{i=1}^k \alpha_i^0 \varphi_i$$

is a solution to Problem $(P_\tau^{n,h})$, proving the desired existence.

Uniqueness: Assume $u_{n,1}, u_{n,2} \in V_h$ solve $(P_\tau^{n,h})$. Then if we subtract eq. (3.15) with $u_{n,1}$, $u_{n,2}$, and pick $v_h = u_{n,1} - u_{n,2}$, we obtain

$$\langle b(u_{n,1}) - b(u_{n,2}), u_{n,1} - u_{n,2} \rangle + \tau \|\nabla(u_{n,1} - u_{n,2})\|^2 = 0$$

which, by the monotonicity of $b(\cdot)$, only holds if $u_{n,1} = u_{n,2}$.

□

Remark 3.2.1. Note the fact that $b_0 > 0$ was essential in order to prove the result. We had to apply the Young inequality to absorb the negative terms from the inner products with $b(u_{n-1})$ and f_n.

In the next proposition we will see that we actually can study the case when $b_0 \geq 0$, by adding a regularization term to problem $(P_\tau^{n,h})$ of a factor of some $\epsilon > 0$, giving a regularized problem $(P_{\tau,\epsilon}^{n,h})$. The strategy further on is to prove existence and uniqueness of a solution of $(P_{\tau,\epsilon}^{n,h})$, and then check that this converges to a solution of problem $(P_\tau^{n,h})$. We define the regularized problem $(P_{\tau,\epsilon}^{n,h})$:

Problem $(P_{\tau,\epsilon}^{n,h})$: Let $n \in \{1, \dots N\}$, $V_h \subset H_0^1(\Omega)$ be finite-dimensional, $u_{n-1}^\epsilon \in V_h$ be given, and $\epsilon > 0$. Find $u_n^\epsilon \in V_h$ such that

$$\epsilon\langle u_n^\epsilon, v_h\rangle + \langle b(u_n^\epsilon) - b(u_{n-1}^\epsilon), v_h\rangle + \tau\langle \nabla u_n^\epsilon, \nabla v_h\rangle = \tau\langle f_n, v_h\rangle \qquad \forall v_h \in V_h$$

where τ is the time step length, so $t_n = n\tau$ and u_n is the solution at time step t_n.

Remark 3.2.2. The existence of a solution of problem $(P_{\tau,\epsilon}^{n,h})$ is ensured by applying Proposition 3.4 for $\epsilon > 0$ and then using Proposition 3.5 to show convergence as $\epsilon \to 0$. The details are presented in the proof of Proposition 3.6.

First we derive an a priori estimate for the solution of problem $(P_{\tau,\epsilon}^{n,h})$.

Proposition 3.5. *Assume u_n^ϵ solves problem $(P_{\tau,\epsilon}^{n,h})$. Suppose $b(\cdot)$ is monotone increasing, Lipschitz continuous with Lipschitz constant $L_b > 0$, and $b(0) = 0$. Then there exists a constant $C > 0$ such that*

$$\epsilon\|u_n^\epsilon\|^2 + \frac{\tau}{2}\|\nabla u_n^\epsilon\|^2 \leq C.$$

Proof. Let $v_h = u_n^\epsilon$. Then

$$\begin{aligned}\epsilon\|u_n^\epsilon\|^2 + &\langle b(u_n^\epsilon) - b(u_{n-1}^\epsilon), u_n^\epsilon - u_{n-1}^\epsilon\rangle + \tau\|\nabla u_n^\epsilon\|^2 \\ &= \tau\langle f_n, u_n^\epsilon\rangle - \langle b(u_n^\epsilon) - b(u_{n-1}^\epsilon), u_{n-1}^\epsilon\rangle.\end{aligned}$$

Using the Cauchy-Schwarz inequality (Theorem 1.8), the Poincaré inequality (Theorem 1.10), monotonicity and Lipschitz continuity of $b(\cdot)$, we obtain

$$\begin{aligned}\epsilon\|u_n\|^2 + &\frac{1}{L_b}\|b(u_n^\epsilon) - b(u_{n-1}^\epsilon)\|^2 + \tau\|\nabla u_n^\epsilon\|^2 \\ &\leq \tau C_\Omega\|f_n\|\|\nabla u_n^\epsilon\| + \|b(u_n^\epsilon) - b(u_{n-1}^\epsilon)\|\|u_{n-1}^\epsilon\|.\end{aligned}$$

In conclusion, applying the Young inequality (Theorem 1.9) yields

$$\epsilon\|u_n^\epsilon\|^2 + \frac{\tau}{2}\|\nabla u_n^\epsilon\|^2 \leq \frac{\tau C_\Omega^2}{2}\|f_n\|^2 + \frac{L_b}{4}\|u_{n-1}^\epsilon\|^2.$$

The right-hand side of this inequality is bounded by $C > 0$ from our previous assumptions.

□

Proposition 3.6. *Suppose $b(\cdot)$ is monotone increasing, Lipschitz continuous with Lipschitz constant $L_b > 0$, and $b(0) = 0$. Then there exists a unique solution u_n^ϵ of problem $(P_{\tau,\epsilon}^{n,h})$ for fixed $\tau > 0$. Moreover, the sequence $\{u_n^\epsilon\}_\epsilon$ converges to a unique solution $u_n \in V_h$ of problem $(P_\tau^{n,h})$ as $\epsilon \to 0$.*

Proof. Uniqueness: Assume $u_{n,1}^\epsilon, u_{n,2}^\epsilon \in V_h$ both solve $(P_{\tau,\epsilon}^{n,h})$. By the same manner as in the proof of Proposition 3.4, we get that $u_{n,1}^\epsilon = u_{n,2}^\epsilon$.

Existence: Let $\widetilde{b}(u) := b(u) + \epsilon u$. Then it follows that $\widetilde{b}$ is strongly monotone:

$$\langle \widetilde{b}(u) - \widetilde{b}(v), u - v\rangle = \langle b(u) - b(v), u - v\rangle + \epsilon\|u - v\|^2 \geq \epsilon\|u - v\|^2,$$

and we have $\widetilde{b}(0) = 0$. Furthermore, it is trivial that $\widetilde{b}(\cdot)$ is Lipschitz continuous. In these terms, we seek u_n^ϵ such that

$$\langle \widetilde{b}(u_n^\epsilon) - b(u_{n-1}^\epsilon), v_h\rangle + \tau\langle \nabla u_n^\epsilon, \nabla v_h\rangle = \tau\langle f_n, v_h\rangle \qquad \forall v_h \in V_h.$$

The existence of such a u_n^ϵ can be proved by the same method as in Proposition 3.4.

Convergence: From Proposition 3.5, we get that the sequence $\{\nabla u_n^\epsilon\}_\epsilon$ is bounded independently of ϵ. By the Eberlein-Šmuljan Theorem (Theorem 1.5) there exists a subsequence that converges weakly to some ∇u_n. Since V_h is finite-dimensional, we obtain strong convergence. By the Poincaré inequality, we get

$$\|u_n^\epsilon - u_n\| \leq C_\Omega \|\nabla(u_n^\epsilon - u_n)\|,$$

which goes to zero as $\epsilon \to 0$. We also have

$$\epsilon\langle u_n^\epsilon, v_h\rangle \leq \sqrt{\epsilon}(\sqrt{\epsilon}\|u_n^\epsilon\|)\|v_h\| \leq \sqrt{\epsilon}\, C\|v_h\| \to 0 \qquad \text{as} \quad \epsilon \to 0,$$

by Proposition 3.5. At last, we have

$$\langle b(u_n^\epsilon) - b(u_n), v_h\rangle \leq L_b\|u_n^\epsilon - u_n\|\|v_h\| \to 0 \qquad \text{as} \quad \epsilon \to 0.$$

Thus the solution of problem $(P_{\tau,\epsilon}^{n,h})$ converges to the solution of $(P_\tau^{n,h})$ as $\epsilon \to 0$.

□

3.2.3 The Case of a Hölder Continuous Non-linearity $b(\cdot)$

In this subsection we investigate the third and last of the cases from the introduction (page 21), where we equip $b(\cdot)$ with the weaker condition of being Hölder continuous. That is, there exists $C_b > 0$ and $\gamma \in (0, 1)$ such that

$$\|b(v_1) - b(v_2)\| \leq C_b\|v_1 - v_2\|^\gamma \qquad \forall v_1, v_2 \in V_h$$

We recall problem $(P_\tau^{n,h})$:

Problem $(P_\tau^{n,h})$: Let $n \in \{1, \ldots N\}$, $V_h \subset H_0^1(\Omega)$ be finite-dimensional, $u_{n-1} \in V_h$ be given. Find $u_n \in V_h$ such that

$$\langle b(u_n) - b(u_{n-1}), v_h\rangle + \tau\langle \nabla u_n, \nabla v_h\rangle = \tau\langle f_n, v_h\rangle \qquad \forall v_h \in V_h$$

where τ is the time step length, so $t_n = n\tau$ and u_n is the solution at time step t_n.

Proposition 3.7. *Assuming that the following properties hold for $b(\cdot)$:*

$$\begin{cases} b(\cdot) \text{ is Hölder continuous,} \\ b(\cdot) \text{ is monotonically increasing,} \\ b(0) = 0, \\ b(\cdot) \text{ is bounded.} \end{cases}$$

Then Problem $(P_\tau^{n,h})$ has a unique solution for fixed $\tau > 0$.

Proof. Uniqueness: Assume $u_{n,1}, u_{n,2} \in V_h$ solves $(P_\tau^{n,h})$. It follows that

$$\langle b(u_{n,1}) - b(u_{n,2}), v_h\rangle + \tau\langle \nabla(u_{n,1} - u_{n,2}), v_h\rangle = 0 \qquad \forall v_h \in V_h.$$

Let $v_h := u_{n,1} - u_{n,2}$. Then

$$\langle b(u_{n,1}) - b(u_{n,2}), u_{n,1} - u_{n,2}\rangle + \tau\|\nabla(u_{n,1} - u_{n,2})\|^2 = 0$$

only holds for $u_{n,1} = u_{n,2}$ because the first term is greater than or equal to zero by the monotonicity of $b(\cdot)$.

Existence: Let $\dim(V_h) = k$. Let $\{\varphi_1, \dots, \varphi_k\}$ be an orthogonal basis for V_h as a subspace of $H_0^1(\Omega)$ and orthonormal as a subspace of $L^2(\Omega)$. Then we can represent an element $\bar{u} \in V_h$ as

$$\bar{u} = \sum_{j=1}^{k} \alpha_j \varphi_j, \qquad \text{with} \qquad \|\bar{u}\| = |\alpha|_k$$

where $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k$. Define $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}^k$ as $\mathcal{F}(\alpha) = \hat{\alpha}$, where

$$\begin{aligned}\hat{\alpha}_i : &= \langle b\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - b(u_{n-1}), \varphi_i\rangle + \tau\langle \nabla(\textstyle\sum_{j=1}^k \alpha_j\varphi_j), \nabla\varphi_i\rangle - \tau\langle f_n, \varphi_i\rangle \\ &= \langle b\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - b(u_{n-1}), \varphi_i\rangle + \tau\alpha_i\|\nabla\varphi_i\|^2 - \tau\langle f_n, \varphi_i\rangle,\end{aligned}$$

for $i \in \{1, \dots, k\}$. We will now prove that $\mathcal{F}$ is continuous. Pick $\alpha, \beta \in \mathbb{R}^k$. Then

$$\begin{aligned}|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2 &= \sum_{i=1}^k |\mathcal{F}_i(\alpha) - \mathcal{F}_i(\beta)|^2 \\ &= \sum_{i=1}^k \Big|\langle b\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - b\big(\textstyle\sum_{j=1}^k \beta_j\varphi_j\big), \varphi_i\rangle + \tau(\alpha_i - \beta_i)\|\nabla\varphi_i\|^2\Big|^2.\end{aligned}$$

Apply the inequality $\sum_n |a_n + b_n|^2 \leq 2\sum_n |a_n|^2 + 2\sum_n |b_n|^2$ for real sequences $\{a_n\}_n, \{b_n\}_n$ and the Cauchy-Schwarz inequality (Theorem 1.8) to obtain

$$\begin{aligned}|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2 \leq 2\sum_{i=1}^k &\|b\big(\textstyle\sum_{j=1}^k \alpha_j\varphi_j\big) - b\big(\textstyle\sum_{j=1}^k \beta_j\varphi_j\big)\|^2 \\ &+ 2\tau^2\sum_{i=1}^k |\alpha_i - \beta_i|^2\|\nabla\varphi_i\|^4.\end{aligned}$$

It follows, since $b(\cdot)$ is Hölder continuous, that

$$\begin{aligned}&\Big\|b\big(\sum_{j=1}^k \alpha_j\varphi_j\big) - b\big(\sum_{j=1}^k \beta_j\varphi_j\big)\Big\|^2 \\ &\qquad \leq C_b\|\sum_{j=1}^k(\alpha_j - \beta_j)\varphi_i\|^{2\gamma} \\ &\qquad = C_b\Big(\int_\Omega |\sum_{j=1}^k(\alpha_j - \beta_j)\varphi_i|^2 dx\Big)^\gamma\end{aligned}$$

$$\leq C_b\Big(\int_\Omega \sum_{j=1}^{k} |\alpha_j - \beta_j|^2 \sum_{j=1}^{k} |\varphi_i|^2 dx\Big)^\gamma$$

$$= C_b|\alpha - \beta|_k^{2\gamma}\Big(\int_\Omega \sum_{j=1}^{k} |\varphi_i|^2 dx\Big)^\gamma \leq M|\alpha - \beta|_k^{2\gamma},$$

where $M < \infty$. The second to last inequality is the Cauchy-Schwarz inequality for the Euclidean inner product and norm in $\mathbb{R}^k$. Thus we arrive at

$$|\mathcal{F}(\alpha) - \mathcal{F}(\beta)|_k^2 \quad \leq \quad 2M|\alpha - \beta|_k^{2\gamma} + 2\tau^2 \max_{j\in\{1,\ldots,k\}} \|\nabla\varphi_j\|^4 |\alpha - \beta|_k^2. \tag{3.20}$$

The continuity of $\mathcal{F}$ now follows.

Next, we want to apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) to prove the existence of a solution for problem $(P_\tau^{n,h})$. We want to show that there exists a constant $M < +\infty$ such that

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} \geq 0 \qquad \forall \alpha \in \mathbb{R}^k \text{ such that } |\alpha|_k \leq M.$$

We have

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} = \sum_{i=1}^{k} \Big(\langle b\big(\textstyle\sum_{j=1}^{k} \alpha_j\varphi_j\big) - b(u_{n-1}), \alpha_i\varphi_i\rangle$$

$$+ \tau\langle \nabla(\textstyle\sum_{j=1}^{k} \alpha_j\varphi_j), \nabla\alpha_i\varphi_i\rangle - \tau\langle f_n, \alpha_i\varphi_i\rangle\Big).$$

Let $\bar{u} \in V_h$ be defined as before,

$$\bar{u} := \sum_{j=1}^{k} \alpha_j\varphi_j.$$

Then

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} = \langle b(\bar{u}) - b(u_{n-1}), \bar{u}\rangle + \tau\|\nabla\bar{u}\|^2 - \tau\langle f_n, \bar{u}\rangle.$$

By the Poincaré inequality (Theorem 1.10) and the Cauchy-Schwarz inequality (Theorem 1.8), we have

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} \geq \langle b(\bar{u}), \bar{u}\rangle - \|b(u_{n-1})\|\|\bar{u}\| + \frac{\tau}{C_\Omega^2}\|\bar{u}\|^2 - \tau\|f_n\|\|\bar{u}\|, \tag{3.21}$$

and then applying the Young inequality (Theorem 1.9) gives

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} \geq \langle b(\bar{u}), \bar{u}\rangle + \frac{\tau}{2C_\Omega^2}\|\bar{u}\|^2 - \mathcal{R}(\tau),$$

where

$$\mathcal{R}(\tau) := \frac{C_\Omega^2}{\tau}\|b(u_{n-1})\|^2 + \tau C_\Omega^2\|f_n\|^2.$$

By the assumption of monotonicity of $b(\cdot)$ and $b(0) = 0$, we have $\langle b(\bar{u}), \bar{u}\rangle \geq 0$. Since $\|\bar{u}\| = |\alpha|_k$, we arrive at

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} \geq 0 \qquad \forall \alpha \in \mathbb{R}^k \text{ with } |\alpha|_k^2 = \frac{2C_\Omega^2\mathcal{R}(\tau)}{\tau}.$$

By the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1), $\exists\ \widetilde{\alpha} \in \mathbb{R}^k$ such that

$$\mathcal{F}(\widetilde{\alpha}) = 0 \qquad \text{and} \qquad |\widetilde{\alpha}|_k \leq C_\Omega\sqrt{\frac{2\mathcal{R}(\tau)}{\tau}},$$

which proves the existence of a solution of problem $(P_\tau^{n,h})$.

□

Remark 3.2.3. This does not hold for $\tau \to 0$. Moreover, without the assumption of $b(\cdot)$ being bounded, we would need to recompute $\mathcal{R}(\tau)$ at each time step, thus getting a dependence on n. In the Lipschitz case we got boundedness directly.

Remark 3.2.4. In the case where would we would assume strong monotonicity of $b(\cdot)$, yielding $\langle b(\bar{u}), \bar{u}\rangle \geq b_0\|\bar{u}\|^2$, we could use $b_0\|\bar{u}\|^2$ in eq. (3.21) to absorb the negative $\|\bar{u}\|^2$ terms and get a a bound on a solution $\widetilde{\alpha} \in \mathbb{R}^k$ not blowing up as $\tau \to 0$. Here we would obtain (after some calculations),

$$\langle \mathcal{F}(\alpha), \alpha\rangle_{\mathbb{R}^k} \geq \frac{1}{2}\Big(b_0 + \frac{\tau}{C_\Omega}\Big)|\alpha|_k^2 - \mathcal{R}(\tau).$$

which implies the existence of a solution $u = \sum_{i=1}^k \alpha_i\hat{\varphi}_i$ satisfying

$$|\hat{\alpha}|_k^2 \leq C_\Omega \frac{\|b(u_{n-1})\|^2 + \tau C_\Omega^2\|f_n\|^2}{(b_0+\tau)b_0}.$$

We give the theorem below without proof.

Proposition 3.8. *Assuming that the following properties hold for $b(\cdot)$:*

$$\begin{cases} b(\cdot) \text{ is Hölder continuous,} \\ b(\cdot) \text{ is strongly monotone increasing,} \\ b(0) = 0, \\ b(\cdot) \text{ is bounded.} \end{cases}$$

Then there exists a unique solution of problem $(P_\tau^{n,h})$ for all $\tau > 0$.

3.3 The Transport Equation

For our next example, we wish to look at how we can apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) to prove existence and uniqueness for a coupled problem.

Let $\Omega \subset \mathbb{R}^d$ be an open, bounded domain with Lipschitz continuous boundary Γ. The reaction-convection-diffusion equation reads as

$$\begin{cases} \partial_t u + Lu = f & \text{on } (0,T] \times \Omega, \\ u = 0 & \text{on } (0,T] \times \Gamma, \\ u = g & \text{on } 0 \times \Omega. \end{cases} \tag{3.22}$$

We define the operator L as

$$Lu := -\nabla \cdot (A\nabla u) + \vec{b}\nabla u + cu,$$

where $A = A(t,x) \in \mathbb{R}^{d\times d}$, $\vec{b} = \vec{b}(t,x) \in \mathbb{R}^d$ and $c = c(t,x) \in \mathbb{R}$. For our application, we present an equation describing reactive transport in saturated/unsaturated porous media. The result in Theorem 3.3 is recited from [29]. For a relatively recent review on numerical methods for flow and reactive transport in saturated/unsaturated porous media we refer to [37]. In [23, 24], compactness arguments are used for proving the existence and uniqueness of solutions.

We want to find $c = c(t, x)$ on $(0, T] \times \Omega$ satisfying

$$\begin{cases} \partial_t(\Theta(\psi)c) + \nabla \cdot \mathbf{q} = \Theta(\psi)r(c), \\ \qquad\qquad\qquad\quad\ \mathbf{q} = -\nabla c + \mathbf{Q}c, \end{cases}$$

along with the initial and boundary conditions

$$c = c_I \text{ in } 0 \times \Omega, \qquad \text{and} \qquad c = 0 \text{ on } (0, T] \times \Gamma,$$

where $\nabla \cdot \mathbf{q}$ is the diffusion and convection, and $r(\cdot)$ is the reactive term. In porous media terminology, $\Theta = \Theta(\psi)$ describes the water content (as a fraction of the total volume), ψ is the pressure head, $\mathbf{Q}$ is the water flux, and c the concentration. We obtain Θ and $\mathbf{Q}$ by solving Richards equation (which we discussed in the previous section), where $\mathbf{Q} = -K(\Theta(\psi))\nabla(\psi + z)$.

As before, we let $\langle \cdot, \cdot \rangle$ be the $L^2(\Omega)$ inner product or the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. We discretize in time with the Backward Euler method, and define the spatial discretizations: Let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into d-simplices, assuming $\Omega = \bigcup_{T \in \mathcal{T}_h} T$. We will use here the Raviart-Thomas spaces [9]:

$$\begin{aligned} W_h &:= \big\{ p \in L^2(\Omega) \mid p \text{ is constant on each } T \in \mathcal{T}_h \big\}, \\ V_h &:= \big\{ \mathbf{q} \in H(\mathrm{div}; \Omega) \mid \mathbf{q}_{|T} = \mathbf{a} + b\mathbf{x} \quad \forall\, T \in \mathcal{T}_h, \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R} \big\}. \end{aligned}$$

For a detailed description of a mixed formulation for multi-component transport in porous media, see [34, 35]. For other spatial discretizations we refer to [36], where also a discussion on numerical diffusion for the different formulations can be found.

Furthermore, we define the projections

$$\begin{aligned} P_h : L^2(\Omega) \to W_h, \qquad &\text{with} \qquad \langle P_h w - w, w_h \rangle = 0, \\ \Pi_h : H(\mathrm{div}; \Omega) \to V_h, \qquad &\text{with} \qquad \langle \nabla \cdot (\Pi_h \mathbf{v} - \mathbf{v}), w_h \rangle = 0, \end{aligned}$$

for all $w \in L^2(\Omega), \mathbf{v} \in H(\mathrm{div}; \Omega)$ and $w_h \in W_h$. We make the following assumptions:

(A1) $0 < \Theta_R \leq \Theta(x) \leq \Theta_s \leq 1\ \forall x \in \Omega$.

(A2) $r : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous and $r(c) = 0$ for $c \leq 0$.

(A3) $\mathbf{Q}_h^n$, a discrete approximation of $\mathbf{Q}$ (see [29], page 4), belongs to $L^\infty(\Omega)\ \forall n$. Thus $\exists\, M < +\infty$ such that $\|\mathbf{Q}_h^n\| < M$

Let us define the fully discrete Problem (PC_h^n):

Problem (PC_h^n)**:** Let $n \geq 1$ be fixed and $\Theta(\psi_h^n), \Theta(\psi_h^{n-1}), \mathbf{Q}_h^n, c_h^{n-1}$ be given. Find $(c_h^n, \mathbf{q}_h^n) \in W_h \times V_h$ such that

$$\begin{aligned} \langle \Theta(\psi_h^n)c_h^n - \Theta(\psi_h^{n-1})c_h^{n-1}, w_h \rangle + \tau \langle \nabla \cdot \mathbf{q}_h^n, w_h \rangle &= \tau \langle \Theta(\psi_h^n) r(c_h^n), w_h \rangle \\ \langle \mathbf{q}_h^n, \mathbf{v}_h \rangle - \langle c_h^n, \nabla \cdot \mathbf{v}_h \rangle - \langle c_h^n \mathbf{Q}_h^n, \mathbf{v}_h \rangle &= 0 \end{aligned}$$

for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$. The initial guess is $c_h^0 = P_h c_I$.

Theorem 3.3. *Assuming (A1)-(A3) hold true, there exists a unique solution of Problem (PC_h^n) for τ sufficiently small.*

Proof. The proof follows in the lines of [29].
Uniqueness: Assume that there exist two sets of solutions $(c^n_{h,1}, \mathbf{q}^n_{h,1}) \in W_h \times V_h$ and $(c^n_{h,2}, \mathbf{q}^n_{h,2}) \in W_h \times V_h$. Let $c^n_h := c^n_{h,1} - c^n_{h,2}$ and $\mathbf{q}^n_h := \mathbf{q}^n_{h,1} - \mathbf{q}^n_{h,2}$. Then there holds $\forall w_h \in W_h, \mathbf{v}_h \in V_h$:

$$\langle \Theta(\psi^n_h) c^n_h, w_h \rangle + \tau \langle \nabla \cdot \mathbf{q}^n_h, w_h \rangle = \tau \langle \Theta(\psi^n_h)[r(c^n_{h,1}) - r(c^n_{h,2})], w_h \rangle, \tag{3.23}$$

$$\langle \mathbf{q}^n_h, \mathbf{v}_h \rangle - \langle c^n_h, \nabla \cdot \mathbf{v}_h \rangle - \langle c^n_h \mathbf{Q}^n_h, \mathbf{v}_h \rangle = 0. \tag{3.24}$$

Now pick $w_h = c^n_h$ and $\mathbf{v}_h = \tau \mathbf{q}^n_h$ in (3.23) and (3.24), respectively. Add the resulting equalities to obtain

$$\langle \Theta(\psi^n_h) c^n_h, c^n_h \rangle + \tau \|\mathbf{q}^n_h\|^2 = \tau \langle \Theta(\psi^n_h)[r(c^n_{h,1}) - r(c^n_{h,2})], c^n_h \rangle + \tau \langle c_h \mathbf{Q}^n_h, \mathbf{q}^n_h \rangle.$$

by the Cauchy-Schwarz Inequality (Theorem 1.8) and assumptions (A1)-(A3), we get

$$\Theta_R \|c^n_h\|^2 + \tau \|\mathbf{q}^n_h\|^2 \leq \tau \Theta_S L_r \|c^n_h\|^2 + \tau M \|c^n_h\| \|q^n_h\|. \tag{3.25}$$

Next, using the Young inequality (Theorem 1.9) on the rightmost term of eq. (3.25), implies

$$\frac{\Theta_R}{2} \|c^n_h\|^2 + \tau \|\mathbf{q}^n_h\|^2 \leq \tau \Theta_S L_r \|c^n_h\|^2 + \tau^2 \frac{M^2}{2\Theta_R} \|\mathbf{q}^n_h\|^2.$$

Thus $c^n_{h,1} = c^n_{h,2}$ for τ sufficiently small. This further implies that $\mathbf{q}^n_{h,1} = \mathbf{q}^n_{h,2}$ for τ sufficiently small. This concludes the proof of uniqueness.

Existence: We will now go use the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) to prove the existence of a solution of Problem (PC^n_h).

Let $\{w_1, \dots, w_{k_1}\}$ and $\{v_1, \dots, v_{k_2}\}$ be orthonormal bases for W_h and V_h such that

$$\langle w_i, w_j \rangle_{L^2(\Omega)} = \langle v_i, v_j \rangle_{L^2(\Omega)} = \delta_{ij},$$
$$\langle w_i, w_j \rangle_{H^1(\Omega)} = \langle v_i, v_j \rangle_{H^1(\Omega)} = 0 \qquad \text{if } i \neq j.$$

Then we can represent elements $\bar{w} \in W_h$ and $\bar{\mathbf{v}} \in V_h$ as

$$\bar{w} := \sum_{j=1}^{k_1} \alpha_j w_j, \qquad \|\bar{w}\| = |\alpha|_{k_1},$$
$$\bar{\mathbf{v}} := \sum_{j=1}^{k_2} \beta_j \mathbf{v}_j, \qquad \|\bar{\mathbf{v}}\| = |\beta|_{k_2},$$

for $\alpha = (\alpha_1, \dots, \alpha_{k_1}) \in \mathbb{R}^{k_1}$ and $\beta = (\beta_1, \dots, \beta_{k_2}) \in \mathbb{R}^{k_2}$. $|\cdot|_k$ is the Euclidean norm in $\mathbb{R}^k$, defined as $|x|_k := \sqrt{\sum_{j=1}^k x_j^2}$. Let $\xi, \hat{\xi} \in \mathbb{R}^{k_1+k_2}$. We define $\mathcal{F} : \mathbb{R}^{k_1+k_2} \to \mathbb{R}^{k_1+k_2}$ as $\mathcal{F}(\xi) = \hat{\xi}$, with $\xi := (\alpha, \beta), \hat{\xi} := (\hat{\alpha}, \hat{\beta})$ for $\alpha, \hat{\alpha} \in \mathbb{R}^{k_1}$ and $\beta, \hat{\beta} \in \mathbb{R}^{k_2}$, where

$$\hat{\alpha}_i := \langle \Theta(\psi^n_h)\bar{w} - \Theta(\psi^{n-1}_h) c^{n-1}_h, w_i \rangle + \tau \langle \nabla \cdot \bar{\mathbf{v}}, w_i \rangle - \tau \langle \Theta(\psi^n_h) r(\bar{w}), w_i \rangle,$$

for all $i \in \{1, \dots, k_1\}$, and

$$\hat{\beta}_i := \langle \bar{\mathbf{v}}, \mathbf{v}_i \rangle - \langle \bar{w}, \nabla \cdot \mathbf{v}_i \rangle - \langle \bar{w} \mathbf{Q}^n_h, \mathbf{v}_i \rangle,$$

for all $i \in \{1, \dots, k_2\}$. Let $\kappa \in \{k_1, k_2\}$. We define an inner product on $\mathbb{R}^{k_1+k_2}$ as

$$\langle \xi_1, \xi_2 \rangle_{\mathbb{R}^{k_1+k_2}} := \langle \alpha_1, \alpha_2 \rangle_{k_1} + \tau \langle \beta_1, \beta_2 \rangle_{k_2},$$

$\forall\, \xi_1, \xi_2 \in \mathbb{R}^{k_1+k_2}$ with $\xi_\kappa = (\alpha_\kappa, \beta_\kappa)$ for $\alpha_\kappa \in \mathbb{R}^{k_1}$ and $\beta_\kappa \in \mathbb{R}^{k_2}$. $\langle \cdot, \cdot \rangle_\kappa$ is the Euclidean inner product in $\mathbb{R}^\kappa$, defined by $\langle x, y \rangle_\kappa := \sum_{i=1}^{\kappa} x_i y_i$, $\forall x, y \in \mathbb{R}^\kappa$. These inner products induce a norm on $\mathbb{R}^{k_1+k_2}$,

$$\|\xi\|^2_{\mathbb{R}^{k_1+k_2}} := |\alpha|^2_{k_1} + \tau |\beta|^2_{k_2}.$$

We will now show that $\mathcal{F}$ is continuous. Let $\xi^n := (\alpha^n, \beta^n)$ and $\hat{\xi}^n := (\hat{\alpha}^n, \hat{\beta}^n)$ in $\mathbb{R}^{k_1+k_2}$ satisfy $\mathcal{F}(\xi^n) = \hat{\xi}^n$ for $n = 1, 2$. Then

$$\begin{aligned}
&\big\|\mathcal{F}(\xi^1) - \mathcal{F}(\xi^2)\big\|^2_{\mathbb{R}^{k_1+k_2}} \\
&\quad = \big\|\hat{\xi}^1 - \hat{\xi}^2\big\|^2_{\mathbb{R}^{k_1+k_2}} = \big|\hat{\alpha}^1 - \hat{\alpha}^2\big|^2_{k_1} + \tau \big|\hat{\beta}^1 - \hat{\beta}^2\big|^2_{k_2} \\
&\quad = \sum_{i=1}^{k_1} |\hat{\alpha}^1_i - \hat{\alpha}^2_i|^2 + \tau \sum_{i=1}^{k_2} \big|\hat{\beta}^1_i - \hat{\beta}^2_i\big|^2 \\
&\quad = \sum_{i=1}^{k_1} \Big| \langle \Theta(\psi_h^n) \textstyle\sum_{j=1}^{k_1} (\alpha^1_j - \alpha^2_j) w_j, w_i \rangle + \tau \langle \nabla \cdot (\textstyle\sum_{j=1}^{k_2} (\beta^1_j - \beta^2_j) \mathbf{v}_j), w_i \rangle \\
&\qquad\qquad - \langle \Theta(\psi_h^n) \big[r\big(\textstyle\sum_{j=1}^{k_1} \alpha^1_j w_j \big) - r\big(\textstyle\sum_{j=1}^{k_1} \alpha^2_j w_j \big) \big], w_i \rangle \Big|^2 \\
&\qquad + \tau \sum_{i=1}^{k_2} \Big| \langle \textstyle\sum_{j=1}^{k_2} (\beta^1_j - \beta^2_j) \mathbf{v}_j, \mathbf{v}_i \rangle - \langle \textstyle\sum_{j=1}^{k_1} (\alpha^1_j - \alpha^2_j) w_j, \nabla \cdot \mathbf{v}_i \rangle \\
&\qquad\qquad - \langle \big[\textstyle\sum_{j=1}^{k_1} (\alpha^1_j - \alpha^2_j) w_j \big] \mathbf{Q}^n_h, \mathbf{v}_i \rangle \Big|^2.
\end{aligned}$$

Further we use the inequality

$$\sum_{n=1}^{k} |a_n + b_n + c_n|^2 \leq 2 \sum_{n=1}^{k} |a_n|^2 + 4 \sum_{n=1}^{k} |b_n|^2 + 4 \sum_{n=1}^{k} |c_n|^2,$$

for real sequences $\{a_n\}_{n=1}^k, \{b_n\}_{n=1}^k, \{c_n\}_{n=1}^k$, along with Cauchy-Schwarz inequality (Theorem 1.8) and the assumptions (A1)-(A3) to get

$$\begin{aligned}
&\big\|\mathcal{F}(\xi^1) - \mathcal{F}(\xi^2)\big\|^2_{\mathbb{R}^{k_1+k_2}} \\
&\quad \leq 2\Theta_s^2 \sum_{i=1}^{k_1} \big|\alpha^1_i - \alpha^2_i\big|^2 + 4k_1\tau^2 \sum_{j=1}^{k_2} \big|\beta^1_j - \beta^2_j\big|^2 \|\nabla \cdot \mathbf{v}_j\|^2 + 4\Theta_s^2 L_r^2 \sum_{i=1}^{k_1} \big|\alpha^1_i - \alpha^2_i\big|^2 \\
&\qquad + 2\tau \sum_{j=1}^{k_2} \big|\beta^1_j - \beta^2_j\big|^2 + 4\tau \sum_{j=1}^{k_1} \big|\alpha^1_i - \alpha^2_i\big|^2 \sum_{i=1}^{k_2} \|\nabla \cdot \mathbf{v}_i\|^2 + 4M^2 k_2 \tau \sum_{j=1}^{k_1} \big|\alpha^1_j - \alpha^2_j\big|^2 \\
&\quad \leq \Big(2\Theta_s^2 + 4\Theta^2 L_r^2 + 4k_2\tau \max_{i=1,\dots,k_1} \|\nabla \cdot \mathbf{v}_i\|^2 + 4M^2 k_2 \tau \Big) \big|\alpha^1 - \alpha^2\big|^2_{k_1} \\
&\qquad + \Big(4k_1\tau \max_{j=1,\dots,k_2} \|\nabla \cdot \mathbf{v}_j\|^2 + 2 \Big) \tau \big|\beta^1 - \beta^2\big|^2_{k_2}.
\end{aligned}$$

Thus there exists $C < +\infty$ such that

$$\|\mathcal{F}(\xi^1) - \mathcal{F}(\xi^2)\|^2_{\mathbb{R}^{k_1+k_2}} \leq C\|\xi^1 - \xi^2\|^2_{\mathbb{R}^{k_1+k_2}},$$

for all $\xi^1, \xi^2 \in \mathbb{R}^{k_1+k_2}$, which implies that $\mathcal{F}$ is continuous.

Next we want to prove the existence of a solution for Problem (PC^n_h). We will apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) for the previously defined function $\mathcal{F}$, which we already have shown to be continuous. We want to show that there exists $M < +\infty$ such that

$$\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} \geq 0 \qquad \forall\, \xi \in \mathbb{R}^{k_1+k_2} \text{ satisfying } \|\xi\|_{\mathbb{R}^{k_1+k_2}} = M,$$

and this will imply the existence.

Let $\xi := (\alpha, \beta) \in \mathbb{R}^{k_1+k_2}$ and $(\bar{w}, \bar{\mathbf{v}}) \in W_h \times V_h$ be defined as

$$\bar{w} := \sum_{j=1}^{k_1} \alpha_j w_j, \qquad \bar{\mathbf{v}} := \sum_{j=1}^{k_2} \beta_j \mathbf{v}_j.$$

for $\alpha := (\alpha_1, \ldots, \alpha_{k_1}) \in \mathbb{R}^{k_1}$ and $\beta := (\beta_1, \ldots, \beta_{k_2}) \in \mathbb{R}^{k_2}$. Then

$$\begin{aligned}\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} &= \langle \hat{\alpha}, \alpha \rangle_{k_1} + \tau \langle \hat{\beta}, \beta \rangle_{k_2} \\ &= \langle \Theta(\psi^n_h)\bar{w} - \Theta(\psi^n_h)c^{n-1}_h, \bar{w} \rangle + \tau \langle \nabla \cdot \bar{\mathbf{v}}, \bar{w} \rangle - \tau \langle \tau(\psi^n_h) r(\bar{w}), \bar{w} \rangle \\ &\quad + \tau \|\bar{\mathbf{v}}\|^2 - \tau \langle \bar{w}, \nabla \cdot \bar{\mathbf{v}} \rangle - \tau \langle \bar{w}\mathbf{Q}^n_h, \bar{\mathbf{v}} \rangle.\end{aligned}$$

By the Cauchy-Schwarz inequality (Theorem 1.8) and assumptions (A1)-(A3), we obtain

$$\begin{aligned}\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} \geq{}& \Theta_R \|\bar{w}\|^2 - \|\Theta(\psi^{n-1}_h c^{n-1}_h)\| \|\bar{w}\| - \tau L_r \|\bar{w}\|^2 \\ &+ \tau \|\bar{\mathbf{v}}\|^2 - \tau M \|\bar{w}\| \|\bar{\mathbf{v}}\|.\end{aligned}$$

Furthermore, the Young inequality (Theorem 1.9) and the fact that the basis vectors are unitary implies that

$$\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} \geq \left(\frac{\Theta_R}{2} - \tau \left[L_r \Theta_s + \frac{M^2}{2} \right] \right) |\alpha|^2_{k_1} + \frac{1}{2}\tau |\beta|^2_{k_2} - \frac{1}{2}C,$$

where $C = \dfrac{\|\Theta(\psi^{n-1}_h)c^{n-1}_h\|^2}{\Theta_R}$. For τ small enough, there exists $m < +\infty$ such that $\Theta_R - \tau(2L_r\Theta_S + M^2) \geq m > 0$, which implies that

$$\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} \geq \frac{1}{2}(\min\{m, 1\} \|\xi\|^2_{\mathbb{R}^{k_1+k_2}} - C).$$

Thus $\langle \mathcal{F}(\xi), \xi \rangle_{\mathbb{R}^{k_1+k_2}} \geq 0$ for all $\xi \in \mathbb{R}^{k_1+k_2}$ satisfying

$$\|\xi\|_{\mathbb{R}^{k_1+k_2}} = \sqrt{\frac{C}{\min\{m, 1\}}}.$$

By the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1), we obtain the existence of a $\widetilde{\xi} \in \mathbb{R}^{k_1+k_2}$ such that

$$\mathcal{F}(\widetilde{\xi}) = 0 \qquad \text{and} \qquad \|\widetilde{\xi}\|_{\mathbb{R}^{k_1+k_2}} \leq \sqrt{\frac{C}{\min\{m, 1\}}},$$

which concludes the proof of existence.

□

Chapter 4

Two-phase Flow

In this final chapter we will look at existence and uniqueness for a mathematical model of two-phase flow in porous media. There are many societal relevant applications modelled by multi-phase flow in porous media. This includes enhanced oil recovery, groundwater extraction and contamination, and geological storage of CO_2 [28]. In order to form robust discretizations and develop good linear solvers of the physical problems, the questions of existence and uniqueness are indeed a very important part of the research performed.

The work in this chapter will complement the results in [32]. In the paper, for the fully discrete (non-linear) formulation, a Lipschitz continuous saturation is considered. The question of proving existence of a solution if the saturation is assumed Hölder continuous is left open. It is suggested to apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) as in [29], which we earlier discussed and recited in Theorem 3.3 (page 40). Motivated by the results we proved for Hölder continuous non-linearities of the Richards equation in Proposition 3.7 (page 36) and Proposition 3.8 (page 39), and furthermore the regularization technique analysed in Subsection 3.2.2 (page 32), we will in this chapter prove a result for the existence and uniqueness of a fully discrete formulation in Theorem 4.2.

The application of these techniques to two-phase in porous media enables for the first time, to our knowledge, the proof of existence for the case of a non-Lipschitz saturation.

Let $\Omega \subset \mathbb{R}^d, (d > 1)$ be a bounded domain with Lipschitz continous boundary Γ. Let $T > 0$ be an upper bound for the time. We will assume immiscible and incompressible fluids, and a non-deformable solid matrix. Let $\alpha = w, n$ be the wetting and non-wetting phase, s_α the saturation, p_α the pressure, $\mathbf{q}_\alpha$ the flux and ρ_α the density of phase α. The model combines the mass balance law in equation (4.1) with the Darcy law in equation (4.2):

$$\left\{ \begin{array}{rcll} \dfrac{\partial(\phi\rho_\alpha s_\alpha)}{\partial t} + \nabla \cdot (\rho_\alpha \mathbf{q}_\alpha) &=& 0, & (4.1) \\ \mathbf{q}_\alpha &=& -k\dfrac{k_{r,\alpha}}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha \mathbf{g}), & (4.2) \\ s_w + s_n &=& 1, & (4.3) \\ p_n - p_w &=& p^{cap}(s_w), & (4.4) \end{array} \right.$$

for $\alpha = w, n$, where $\mathbf{g}$ denotes the constant gravitational vector. We have assumed an algebraic evidence expressing the pores to be filled with the two fluids in (4.3), and a relationship between the capillary pressure and the pressure for each phase (4.4) (we assume p^{cap} to be known). The permeability k is a scalar. The porosity ϕ and the viscosities μ_α are given constants, and the relative permeabilities $k_{r,\alpha}(\cdot)$ are given functions of s_w.

4.1 Discretization

In this section, the goal is to derive a fully discrete (non-linear) scheme to be used for simulation of two-phase flow in porous media. First, we wish to define two new unknowns as in [3, 10, 11, 12, 13]: a global pressure defined as

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

and a complementary pressure defined by a Kirchhoff transformation

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

where $\lambda_\alpha := \dfrac{k_{r,\alpha}}{\mu_\alpha}$ is the mobility of phase α and $f_w := \dfrac{\lambda_w}{\lambda_w + \lambda_n}$ is the fractional flow function. In the new unknowns, we obtain the system:

$$\begin{cases} \partial_t s(\Theta) + \nabla \cdot \mathbf{q} = 0, & (4.5) \\ \mathbf{q} = -\nabla\Theta + f_w(s)\mathbf{u} + \mathbf{f}_1(s), & (4.6) \\ \nabla \cdot \mathbf{u} = f_2(s), & (4.7) \\ a(s)\mathbf{u} = -\nabla p - \mathbf{f}_3(s). & (4.8) \end{cases}$$

where $s := s_w$, $\mathbf{q}$ is the (wetting) flux, and $\mathbf{u}$ is the total flux. The equations are defined on $\Omega \times (0, T]$.

We are now in possession of a coupling of two non-linear partial differential equations. The equations (4.5) and (4.6) form a degenerate parabolic equation (which degenerates as the derivative of $s(\cdot)$ possibly vanishes or blows up), while equations (4.7) and (4.8) form an elliptic equation. For the computational details and exact expressions for the coefficient functions $s(\cdot)$, $a(\cdot)$, $\mathbf{f}_1(\cdot)$, $f_2(\cdot)$, $\mathbf{f}_3(\cdot)$ and $f_w(\cdot)$ in equations (4.5)-(4.8), we refer to [11, 12, 13]. We adapt the initial and homogeneous Dirichlet boundary conditions

$$\Theta(0, \cdot) = \Theta_I \text{ in } \Omega \qquad \text{and} \qquad \Theta = 0, p = 0 \text{ on } (0, T] \times \Gamma \tag{4.9}$$

Let $\langle \cdot, \cdot \rangle$ be the $L^2(\Omega)$ inner product (as in Remark 1.4.1) or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. $\| \cdot \|$ is the $L^2(\Omega)$ norm induced by $\langle \cdot, \cdot \rangle$. We define time steps $t_n = n\tau$ for $n \in \{1, \dots, N\} \subset \mathbb{N}$ with step length τ.

First, a continuous mixed variational formulation is obtained by integration in time and space. The existence and uniqueness of the continuous variational formulation for a mixed finite element formulation of two-phase flow was proved in [31] by an equivalence with the conformal formulation used in [11]. For the analysis of two-phase flow with dynamic capillarity, including a linearization algorithm we refer to [20], [21].

Second, the Backward Euler method is applied in time to get a semi-discrete mixed variational formulation. Third, we give the discrete subspaces W_h, V_h of $L^2(\Omega)$ and $H(\text{div}; \Omega)$, respectively. Let $\mathcal{T}_h$ be a regular decomposition of $\Omega \in \mathbb{R}^d$ into closed d-simplices T with mesh size h (see [16], Chapter 2), assuming $\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} T$. We define the Raviart-Thomas spaces [9]

$$W_h := \{ p \in L^2(\Omega) \mid p \text{ is constant on each } T \in \mathcal{T}_h \}, \tag{4.10}$$

$$V_h := \{ \mathbf{q} \in H(\text{div}; \Omega) \mid \mathbf{q}(\mathbf{x}) := \mathbf{a_T} + b_T \mathbf{x} \text{ on each } T \in \mathcal{T}_h, \mathbf{a_T} \in \mathbb{R}^d, b_T \in \mathbb{R} \}. \tag{4.11}$$

We define the fully discrete (non-linear) variational formulation (P_h^n):

Problem (P_h^n): Let $n \in \mathbb{N},\ n \geq 1,\ s_h^n := s(\Theta_h^n)$, and assume Θ_h^{n-1} is known. Find $\Theta_h^n, p_h^n \in W_h$ and $\mathbf{q}_h^n, \mathbf{u}_h^n \in V_h$ such that

$$\begin{aligned}
\langle s_h^n - s_h^{n-1}, w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^n, w_h\rangle &= 0, && (4.12)\\
\langle \mathbf{q}_h^n, \mathbf{v}_h\rangle - \langle \Theta_h^n, \nabla \cdot \mathbf{v}_h\rangle - \langle f_w(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle &= \langle \mathbf{f}_1(s_h^n), \mathbf{v}_h\rangle, && (4.13)\\
\langle \nabla \cdot \mathbf{u}_h^n, w_h\rangle &= \langle f_2(s_h^n), w_h\rangle, && (4.14)\\
\langle a(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle - \langle p_h^n, \nabla \cdot \mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^n), \mathbf{v}_h\rangle &= 0 && (4.15)
\end{aligned}$$

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

We make the following assumptions as stated in [32] (page 7):

(A1) The function $s : \mathbb{R} \to \mathbb{R},\ s(0) = 0$ is strongly monotonically increasing: there exists $s_0 > 0$ such that

$$\langle s(\Theta_1) - s(\Theta_2), \Theta_1 - \Theta_2\rangle \geq s_0|\Theta_1 - \Theta_2|^2,$$

and Hölder continuous with exponent $\alpha \in (0, 1]$, that is, $\exists\ L_s > 0$ such that

$$|s(\Theta_1) - s(\Theta_2)| \leq L_s|\Theta_1 - \Theta_2|^\alpha, \qquad \forall\ \Theta_1, \Theta_2 \in \mathbb{R}.$$

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

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

and there exists $a_\star, a^\star > 0$ such that

$$0 < a_\star \leq a(y) \leq a^\star < \infty, \qquad \forall\ y \in \mathbb{R}.$$

(A3) The coefficient functions $\mathbf{f}_1, \mathbf{f}_3 : \mathbb{R} \to \mathbb{R}^d$ and $f_2, f_w : \mathbb{R} \to \mathbb{R}$ are bounded, satisfy $F(0) = 0$ and the growth condition

$$|F(s(\Theta_1)) - F(s(\Theta_2))|^2 \leq C\langle s(\Theta_1) - s(\Theta_2), \Theta_1 - \Theta_2\rangle, \qquad \forall\ \Theta_1, \Theta_2 \in \mathbb{R},$$

for $C > 0$ and where F is any of the functions above. We specify the constants as C_1, C_2, C_3 and C_w.

Furthermore, to prove existence with the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1), it is necessary for us to define a similar problem $(P_h^{n,\epsilon})$ and then show afterwards that a possibly unique solution of $(P_h^{n,\epsilon})$ converges to (P_h^n) as $\epsilon \to 0$. A regularization term is added to equation (4.14), and the reason for this will become apparent when we seek to satisfy the hypothesis of Corollary 3.2.1 with a norm defined as in equation (4.54). We define the regularized problem $(P_h^{n,\epsilon})$:

Problem $(P_h^{n,\epsilon})$: Let $n \in \mathbb{N},\ n \geq 1,\ s_h^{n,\epsilon} := s(\Theta_h^{n,\epsilon}),\ \epsilon > 0$, and assume $\Theta_h^{n-1,\epsilon}$ is known. Find $\Theta_h^{n,\epsilon}, p_h^{n,\epsilon} \in W_h$ and $\mathbf{q}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon} \in V_h$ such that

$$\begin{aligned}
\langle s_h^{n,\epsilon} - s_h^{n-1}, w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,\epsilon}, w_h\rangle &= 0, && (4.16)\\
\langle \mathbf{q}_h^{n,\epsilon}, \mathbf{v}_h\rangle - \langle \Theta_h^{n,\epsilon}, \nabla \cdot \mathbf{v}_h\rangle - \langle f_w(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{v}_h\rangle &= \langle \mathbf{f}_1(s_h^{n,\epsilon}), \mathbf{v}_h\rangle, && (4.17)\\
\epsilon\langle p_h^{n,\epsilon}, w_h\rangle + \langle \nabla \cdot \mathbf{u}_h^{n,\epsilon}, w_h\rangle &= \langle f_2(s_h^{n,\epsilon}), w_h\rangle, && (4.18)\\
\langle a(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{v}_h\rangle - \langle p_h^{n,\epsilon}, \nabla \cdot \mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^{n,\epsilon}), \mathbf{v}_h\rangle &= 0 && (4.19)
\end{aligned}$$

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

4.2 Existence and Uniqueness

In this section an a priori estimate of the regularized problem $(P_h^{n,\epsilon})$ will be shown in Proposition 4.1. Further, we prove existence and uniqueness of a solution of $(P_h^{n,\epsilon})$. It will become necessary to assume $f_2 \equiv 0$ to be able to apply Corollary 3.2.1. Combining these two results, along with the Eberlein-Šmuljan Theorem (Theorem 1.5), will give convergence of the sequence $\{p_h^{n,\epsilon}\}_\epsilon$ to a solution p_h^n of Problem (P_h^n) as $\epsilon \to 0$. Thus proving the existence and uniqueness of Problem (P_h^n).

For the next proposition, we need the following lemma, which was proven in [38]:

Lemma 4.1. *Given a $w_h \in W_h$, there exists a $\mathbf{v}_h \in V_h$ satisfying*

$$\nabla \cdot \mathbf{v}_h = w_h \qquad \text{and} \qquad \|\mathbf{v}_h\| \leq C_{\Omega,d}\|w_h\| \tag{4.20}$$

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

Proposition 4.1. *Let $\Theta_h^{n,\epsilon}, p_h^{n,\epsilon} \in W_h$ and $\mathbf{q}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon} \in V_h$ be the solution components of Problem $(P_h^{n,\epsilon})$, assuming $f_2 \equiv 0$. Then the following a priori estimate holds:*

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 + \|p_h^{n,\epsilon}\|^2 + \|\mathbf{u}_h^{n,\epsilon}\|^2 \leq C \tag{4.21}$$

for all $\epsilon > 0$ and some positive constant $C < +\infty$.

Proof. Throughout this proof, we let $0 < C < +\infty$ be a generic positive constant. Pick $w_h := p_h^{n,\epsilon}$ in (4.18) and $\mathbf{v}_h := \mathbf{u}_h^{n,\epsilon}$ in (4.19). Then

$$\epsilon\|p_h^{n,\epsilon}\|^2 + \langle \nabla \cdot \mathbf{u}_h^{n,\epsilon}, p_h^{n,\epsilon}\rangle = 0, \tag{4.22}$$

$$\langle a(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}\rangle - \langle p_h^{n,\epsilon}, \nabla \cdot \mathbf{u}_h^{n,\epsilon}\rangle + \langle f_3(s_h^{n,\epsilon}), \mathbf{u}_h^{n,\epsilon}\rangle = 0. \tag{4.23}$$

Adding equations (4.22) and (4.23) gives

$$\langle a(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}\rangle + \epsilon\|p_h^{n,\epsilon}\|^2 + \langle f_3(s_h^{n,\epsilon}), \mathbf{u}_h^{n,\epsilon}\rangle = 0. \tag{4.24}$$

By the Cauchy-Schwarz inequality (Theorem 1.8), assumption (A2), and lastly the Young inequality (Theorem 1.9), we get the bound

$$\|u_h^{n,\epsilon}\|^2 \leq C. \tag{4.25}$$

By Lemma 4.1, there exists $\mathbf{v}_h \in V_h$ such that $\nabla \cdot \mathbf{v}_h = p_h^{n,\epsilon}$ and

$$\|\mathbf{v}_h\| \leq C_{\Omega,d}\|p_h^{n,\epsilon}\|. \tag{4.26}$$

Pick $\mathbf{v}_h \in V_h$ satisfying this for equation (4.19). Then we get

$$\langle a(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{v}_h\rangle - \|p_h^{n,\epsilon}\|^2 + \langle \mathbf{f}_3(s_h^{n,\epsilon}), \mathbf{v}_h\rangle = 0. \tag{4.27}$$

Next, the Cauchy-Schwarz inequality (Theorem 1.8), the relation in (4.26) and the assumptions (A2) and (A3) imply that

$$\|p_h^{n,\epsilon}\|^2 \leq a^* C_{\Omega,d}\|\mathbf{u}_h^{n,\epsilon}\|\|p_h^{n,\epsilon}\| + C_3 C_{\Omega,d}\|p_h^{n,\epsilon}\|.$$

By the Young inequality (Theorem 1.9), we obtain the estimate

$$\|p_h^{n,\epsilon}\| \leq 2(a^*)^2 C_{\Omega,d}^2\|\mathbf{u}_h^{n,\epsilon}\|^2 + C_3^2 C_{\Omega,d}^2. \tag{4.28}$$

We can combine (4.28) with (4.25) to get

$$\|p_h^{n,\epsilon}\|^2 + \|\mathbf{u}_h^{n,\epsilon}\|^2 \leq C. \tag{4.29}$$

Now test equation (4.16) with $w_h := \Theta_h^{n,\epsilon}$ and equation (4.17) with $\mathbf{v}_h := \tau \mathbf{q}_h^{n,\epsilon}$. Then

$$\langle s_h^{n,\epsilon} - s_h^{n-1,\epsilon}, \Theta_h^{n,\epsilon}\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle = 0, \tag{4.30}$$

$$\tau\|\mathbf{q}_h^{n,\epsilon}\|^2 - \tau\langle \Theta_h^{n,\epsilon}, \nabla \cdot \mathbf{q}_h^{n,\epsilon}\rangle - \tau\langle f_w(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}\rangle = \tau\langle \mathbf{f}_1(s_h^{n,\epsilon}), \mathbf{q}_h^{n,\epsilon}\rangle. \tag{4.31}$$

Adding these two equations yields

$$\langle s_h^{n,\epsilon} - s_h^{n-1,\epsilon}, \Theta_h^{n,\epsilon}\rangle + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 = \tau\langle f_w(s_h^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}\rangle + \tau\langle \mathbf{f}_1(s_h^{n,\epsilon}), \mathbf{q}_h^{n,\epsilon}\rangle. \tag{4.32}$$

By the Cauchy-Schwarz inequality (Theorem 1.8), the estimate (4.29) and the assumptions (A1) and (A3), we obtain the existence of a positive constant C such that

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 \leq \|s_h^{n-1,\epsilon}\|\|\Theta_h^{n,\epsilon}\| + \tau C\|\mathbf{q}_h^{n,\epsilon}\|. \tag{4.33}$$

Lastly, the Young inequality (Theorem 1.9) implies

$$\frac{s_0}{2}\|\Theta_h^{n,\epsilon}\|^2 + \frac{\tau}{2}\|\mathbf{q}_h^{n,\epsilon}\|^2 \leq \frac{1}{2s_0}\|s_h^{n-1,\epsilon}\|^2 + \tau C^2. \tag{4.34}$$

Hence there exists $C > 0$ such that

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 \leq C. \tag{4.35}$$

This, added to the inequality (4.29), proves the statement in Proposition 4.1.

□

Remark 4.0.1. Note that the condition $f_2 \equiv 0$ was not necessary to prove Proposition 4.1. It is only used to show the bound for $u_h^{n,\epsilon}$. The reason for including the assumption is because we want to apply the proposition directly in Theorem 4.1.

Theorem 4.1. *Assuming (A1)-(A3) hold true and $f_2 \equiv 0$, there exists a unique solution of problem $(P_h^{n,\epsilon})$ for τ sufficiently small.*

Proof. Uniqueness: Assume there exists two sets of solutions $\Theta_{h,i}^{n,\epsilon}, p_{h,i}^{n,\epsilon} \in W_h$ and $\mathbf{q}_{h,i}^{n,\epsilon}, \mathbf{u}_{h,i}^{n,\epsilon} \in V_h$ for $i = 1, 2$. Define

$$\Theta_h^{n,\epsilon} := \Theta_{h,1}^{n,\epsilon} - \Theta_{h,2}^{n,\epsilon}, \qquad \mathbf{q}_h^{n,\epsilon} := \mathbf{q}_{h,1}^{n,\epsilon} - \mathbf{q}_{h,2}^{n,\epsilon} \tag{4.36}$$

$$p_h^{n,\epsilon} := p_{h,1}^{n,\epsilon} - p_{h,2}^{n,\epsilon}, \qquad \mathbf{u}_h^{n,\epsilon} := \mathbf{u}_{h,1}^{n,\epsilon} - \mathbf{u}_{h,2}^{n,\epsilon} \tag{4.37}$$

Moreover, we define $s_{h,i}^{n,\epsilon} := s(\Theta_{h,i}^{n,\epsilon})$ for $i = 1, 2$ and $s_h^{n,\epsilon} := s_{h,1}^{n,\epsilon} - s_{h,2}^{n,\epsilon}$. Pick $w_h := \Theta_h^{n,\epsilon}$ and $\mathbf{v}_h := \tau\mathbf{q}_h^{n,\epsilon}$ in (4.16) and (4.17). Subtract the equations with the two solutions to obtain

$$\langle s_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle = 0, \tag{4.38}$$

$$\begin{aligned} \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 - \tau\langle \Theta_h^{n,\epsilon}, \nabla \cdot \mathbf{q}_h^{n,\epsilon}\rangle - \tau\langle f_w(s_{h,1}^{n,\epsilon})\mathbf{u}_{h,1}^{n,\epsilon} - f_w(s_{h,2}^{n,\epsilon})\mathbf{u}_{h,2}^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}\rangle \\ = \langle \mathbf{f}_1(s_{h,1}^{n,\epsilon}) - \mathbf{f}_1(s_{h,2}^{n,\epsilon}), \mathbf{q}_h^{n,\epsilon}\rangle. \end{aligned} \tag{4.39}$$

Add the above equations (4.38), (4.39), and expand the $f_w(\cdot)$-term to get

$$\langle s_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 = \tau\langle \big[f_w(s_{h,1}^{n,\epsilon}) - f_w(s_{h,2}^{n,\epsilon})\big]\mathbf{u}_{h,1}^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}\rangle$$
$$+ \tau\langle f_w(s_{h,2}^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}\rangle + \tau\langle \mathbf{f}_1(s_{h,1}^{n,\epsilon}) - \mathbf{f}_1(s_{h,2}^{n,\epsilon}), \mathbf{q}_h^{n,\epsilon}\rangle.$$

By assumptions (A1), (A3) and the Cauchy-Schwarz inequality (Theorem 1.8),

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \tau\|\mathbf{q}_h^{n,\epsilon}\|^2 \le \tau M_{\mathbf{u}}\|f_w(s_{h,1}^{n,\epsilon}) - f_w(s_{h,2}^{n,\epsilon})\|\|\mathbf{q}_h^{n,\epsilon}\|$$
$$+ \tau M_w\|\mathbf{u}_h^{n,\epsilon}\|\|\mathbf{q}_h^{n,\epsilon}\| + \tau\|\mathbf{f}_1(s_{h,1}^{n,\epsilon}) - \mathbf{f}_1(s_{h,2}^{n,\epsilon})\|\|\mathbf{q}_h^{n,\epsilon}\|.$$

Furthermore, the Young inequality (Theorem 1.9) and assumption (A3) implies that there exists a positive constant $C < +\infty$ such that

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \frac{\tau}{4}\|\mathbf{q}_h^{n,\epsilon}\| \le \tau M_w^2\|\mathbf{u}_h^{n,\epsilon}\|^2 + \tau C\langle s_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle \tag{4.40}$$

Next, pick $w_h := p_h^{n,\epsilon}$ and $\mathbf{v}_h := \mathbf{u}_h^{n,\epsilon}$. Subtracting the equations (4.18) and (4.19) from themselves with the two solutions yields

$$\epsilon\|p_h^{n,\epsilon}\|^2 + \langle\nabla\cdot\mathbf{u}_h^{n,\epsilon}, p_h^{n,\epsilon}\rangle = 0, \tag{4.41}$$

$$\langle a(s_{h,1}^{n,\epsilon})\mathbf{u}_h^{n,\epsilon} - a(s_{h,2}^{n,\epsilon})\mathbf{u}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}\rangle - \langle p_h^{n,\epsilon}, \nabla\cdot\mathbf{u}_h^{n,\epsilon}\rangle + \langle \mathbf{f}_3(s_{h,1}^{n,\epsilon}) - \mathbf{f}_3(s_{h,2}^{n,\epsilon}), \mathbf{u}_h^{n,\epsilon}\rangle = 0. \tag{4.42}$$

After adding (4.41) and (4.42), and expanding the $a(\cdot)$-term, we obtain

$$\epsilon\|p_h^{n,\epsilon}\|^2 + \langle a(s_{h,2})\mathbf{u}_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}\rangle = -\langle\big[a(s_{h,1}^{n,\epsilon}) - a(s_{h,2}^{n,\epsilon})\big]\mathbf{u}_{h,1}^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}\rangle - \langle \mathbf{f}_3(s_{h,1}^{n,\epsilon}) - \mathbf{f}_3(s_{h,2}^{n,\epsilon}), \mathbf{u}_h^{n,\epsilon}\rangle.$$

It follows from the Cauchy-Schwarz inequality (Theorem 1.8), the Young inequality (Theorem 1.9), Proposition 4.1, and the assumptions (A2), (A3), that $\exists C > 0$ such that

$$\epsilon\|p_h^{n,\epsilon}\|^2 + \frac{a_*}{4}\|\mathbf{u}_h^{n,\epsilon}\|^2 \le C\langle s_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle. \tag{4.43}$$

Combining this with the inequality (4.40), we have

$$s_0\|\Theta_h^{n,\epsilon}\|^2 + \frac{\tau}{4}\|\mathbf{q}_h^{n,\epsilon}\|^2 + \frac{4M_w^2\epsilon}{a_*}\tau\|p_h^{n,\epsilon}\|^2 \le \tau C\langle s_h^{n,\epsilon}, \Theta_h^{n,\epsilon}\rangle, \tag{4.44}$$

for some $C < +\infty$. For τ sufficiently small, we get $\Theta_h^{n,\epsilon} = 0$. This further implies that $\mathbf{q}_h^{n,\epsilon} = \mathbf{0}$ and $p_h^{n,\epsilon} = 0$. $\mathbf{u}_h^{n,\epsilon} = \mathbf{0}$ follows from (4.43). This concludes the proof of uniqueness.

Existence: We will apply the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1) to prove the existence of a solution of problem $(P_h^{n,\epsilon})$. Let $\{w_1, \dots, w_{k_1}\}$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_{k_2}\}$ be orthonormal bases for W_h, V_h, respectively. Then we can represent elements of $\bar\Theta, \bar p \in W_h$ and $\bar{\mathbf{q}}, \bar{\mathbf{u}} \in V_h$ as

$$\bar\Theta := \sum_{j=1}^{k_1}\alpha_j^1 w_j, \quad \text{with } \|\bar\Theta\| = |\alpha^1|_{k_1}, \tag{4.45}$$

$$\bar p := \sum_{j=1}^{k_1}\alpha_j^2 w_j, \quad \text{with } \|\bar p\| = |\alpha^2|_{k_1}, \tag{4.46}$$

$$\bar{\mathbf{q}} := \sum_{j=1}^{k_2}\beta_j^1 \mathbf{v}_j, \quad \text{with } \|\bar{\mathbf{q}}\| = |\beta^1|_{k_2}, \tag{4.47}$$

$$\bar{\mathbf{u}} := \sum_{j=1}^{k_2}\beta_j^2 \mathbf{v}_j, \quad \text{with } \|\bar{\mathbf{u}}\| = |\beta^2|_{k_1}, \tag{4.48}$$

for $\alpha^n := (\alpha^n_1, \dots, \alpha^n_{k_1}) \in \mathbb{R}^{k_1}$ and $\beta^n := (\beta^n_1, \dots, \beta^n_{k_2}) \in \mathbb{R}^{k_2}$, with $n = 1, 2$. To prove existence with the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1), we consider the finite-dimensional Hilbert space $H := \mathbb{R}^{k_1+k_2} \times \mathbb{R}^{k_1+k_2}$ with inner product defined as the sum of two different inner products on $\mathbb{R}^{k_1+k_2}$. From here on, we will denote elements of H by $\eta := (\xi^1, \xi^2)$, for $\xi^n := (\alpha^n, \beta^n) \in \mathbb{R}^{k_1+k_2}$ with $\alpha^n \in \mathbb{R}^{k_1}, \beta^n \in \mathbb{R}^{k_2}$, $n = 1, 2$. If we consider more than one element of H, we will separate them by subscripts:

$$\eta_m := (\xi^1_m, \xi^2_m) := \Big((\alpha^1_m, \beta^1_m), (\alpha^2_m, \beta^2_m)\Big), \qquad m \in \mathbb{N}, \tag{4.49}$$

for $\xi^1_m, \xi^2_m \in \mathbb{R}^{k_1+k_2}$ and $\alpha^1_m, \alpha^2_m \in \mathbb{R}^{k_1}$, $\beta^1_m, \beta^2_m \in \mathbb{R}^{k_2}$. For elements of $\mathbb{R}^{k_1+k_2}$, we specify $\xi_m := (\alpha_m, \beta_m)$ for $\alpha_m \in \mathbb{R}^{k_1}$ and $\beta_m \in \mathbb{R}^{k_2}$ with $m \in \mathbb{N}$.

We define two inner products on $\mathbb{R}^{k_1+k_2}$ by

$$\langle \xi_1, \xi_2 \rangle_1 := \langle \alpha_1, \alpha_2 \rangle_{k_1} + \tau \langle \beta_1, \beta_2 \rangle_{k_2}, \tag{4.50}$$

$$\langle \xi_1, \xi_2 \rangle_2 := \langle \alpha_1, \alpha_2 \rangle_{k_1} + \langle \beta_1, \beta_2 \rangle_{k_2}. \tag{4.51}$$

For $\kappa = k_1, k_2$, $\langle \cdot, \cdot \rangle_\kappa$ is the Euclidean inner product in $\mathbb{R}^\kappa$, defined by

$$\langle x, y \rangle_\kappa := \sum_{j=1}^{\kappa} x_j y_j, \qquad \forall x, y \in \mathbb{R}^\kappa. \tag{4.52}$$

The motivation behind the different choices of inner products will become apparent when we develop an estimate satisfying the hypothesis of Corollary 3.2.1. See Remark 4.1.1 for further details. Let $\eta_1, \eta_2 \in H$. We define an inner product on H as

$$\langle \eta_1, \eta_2 \rangle_H := \langle \xi^1_1, \xi^1_2 \rangle_1 + \langle \xi^2_1, \xi^2_2 \rangle_2. \tag{4.53}$$

For $\eta \in H$, this inner product induces a norm

$$\left\|\eta\right\|^2_H \;:=\; \left\|\xi^1\right\|^2_1 + \left\|\xi^2\right\|^2_2 \;:=\; \left|\alpha^1\right|^2_{k_1} + \tau\left|\beta^1\right|^2_{k_2} + \left|\alpha^2\right|^2_{k_1} + \left|\beta^2\right|^2_{k_2}. \tag{4.54}$$

Next, define $\mathcal{F} \to H$ as $\mathcal{F}(\eta) := \hat{\eta}$, where the notation of $\hat{\eta} \in H$ will follow $\eta \in H$ as defined previously. Recall the sample elements of H in equations (4.45)-(4.48). $\hat{\eta} := ((\hat{\alpha}^1, \hat{\beta}^1), (\hat{\alpha}^2, \hat{\beta}^2))$ is given componentwise as

$$\hat{\alpha}^1_i := \langle s(\bar{\Theta}) - s^{n-1}_h, w_i \rangle + \tau \langle \nabla \cdot \bar{\mathbf{q}}, w_i \rangle, \tag{4.55}$$

$$\hat{\alpha}^2_i := \epsilon \langle \bar{p}, w_i \rangle + \langle \nabla \cdot \bar{\mathbf{u}}, w_i \rangle, \tag{4.56}$$

for $i \in \{1, \dots, k_1\}$, and

$$\hat{\beta}^1_i := \langle \bar{\mathbf{q}}, \mathbf{v}_i \rangle - \langle \bar{\Theta}, \nabla \cdot \mathbf{v}_i \rangle - \langle f_w(s(\bar{\Theta}))\bar{\mathbf{u}}, \mathbf{v}_i \rangle - \langle \mathbf{f}_1(s(\bar{\Theta})), \mathbf{v}_i \rangle, \tag{4.57}$$

$$\hat{\beta}^2_i := \langle a(s(\bar{\Theta}))\bar{\mathbf{u}}, \mathbf{v}_i \rangle - \langle \bar{p}, \nabla \cdot \mathbf{v}_i \rangle + \langle \mathbf{f}_3(s(\bar{\Theta})), \mathbf{v}_i \rangle, \tag{4.58}$$

for $i \in \{1, \dots, k_2\}$. It can be shown that $\mathcal{F}$ is continuous. Next, we deduce the estimate for the inner product:

$$\begin{aligned}
\langle \mathcal{F}(\eta), \eta \rangle_H &= \langle \hat{\eta}, \eta \rangle_H \\
&= \langle \hat{\alpha}^1, \alpha^1 \rangle_{k_1} + \tau \langle \hat{\beta}^1, \beta^1 \rangle_{k_2} + \langle \hat{\alpha}^2, \alpha^2 \rangle_{k_1} + \langle \hat{\beta}^2, \beta^2 \rangle_{k_2} \\
&= \langle s(\bar{\Theta}) - s^{n-1}_h, \bar{\Theta} \rangle + \tau \langle \nabla \cdot \bar{\mathbf{q}}, \bar{\Theta} \rangle + \tau \|\bar{\mathbf{q}}\|^2 - \tau \langle \bar{\Theta}, \nabla \cdot \bar{\mathbf{q}} \rangle - \langle f_w(s(\bar{\Theta}))\bar{\mathbf{u}}, \bar{\mathbf{q}} \rangle \\
&\quad + \langle \mathbf{f}_1(s(\bar{\Theta})), \bar{\mathbf{q}} \rangle + \epsilon \|\bar{p}\|^2 + \langle \nabla \cdot \bar{\mathbf{u}}, \bar{p} \rangle + \langle a(s(\bar{\Theta}))\bar{\mathbf{u}}, \bar{\mathbf{u}} \rangle \\
&\qquad - \langle \bar{p}, \nabla \cdot \bar{\mathbf{u}} \rangle + \langle \mathbf{f}_3(s(\bar{\Theta})), \bar{\mathbf{u}} \rangle.
\end{aligned}$$

Using the Cauchy-Schwarz inequality (Theorem 1.8), the assumptions (A2)-(A3), and Proposition 4.1,

$$\begin{aligned}\langle \mathcal{F}(\eta), \eta\rangle_H \geq s_0\|\bar{\Theta}\|^2 - \|s_h^{n-1}\|\|\bar{\Theta}\| + \tau\|\bar{\mathbf{q}}\|^2 - C_w\|\bar{u}\|\|\bar{\mathbf{q}}\| - \tau C_1\|\bar{\mathbf{q}}\| \\ + \epsilon\|\bar{p}\|^2 + a_*\|\bar{\mathbf{u}}\|^2 - C_3\|\bar{\mathbf{u}}\|.\end{aligned}$$

At last, after applying the Young inequality (Theorem 1.9) and using the norm relations in (4.45)-(4.48), we obtain the estimate

$$\begin{aligned}\langle \mathcal{F}(\eta), \eta\rangle_H \geq \frac{s_0}{2}\left|\alpha^1\right|_{k_1}^2 + \frac{1}{4}\tau\left|\beta^1\right|_{k_2}^2 + \epsilon\left|\alpha^2\right|_{k_1}^2 + \frac{1}{2}(a_* - \tau C_w^2)\left|\beta^2\right|_{k_2}^2 \\ - \Big(\frac{1}{2s_0}\|s_h^{n-1}\|^2 + \tau C_1^2 + \frac{1}{2a_*}C_3^2\Big).\end{aligned}$$

For τ sufficiently small, we have $a_* - \tau C_w^2 > 0$ Let $m := \min\{\frac{s_0}{2}, \frac{1}{4}, \epsilon, \frac{1}{2}(a_* - \tau C_w^2)\}$. Then

$$\langle \mathcal{F}(\eta), \eta\rangle_H \geq m\|\eta\|_H^2 - C, \qquad \forall \eta \in H, \tag{4.59}$$

where $C := \dfrac{1}{2s_0}\|s_h^{n-1}\|^2 + \tau C_1^2 + \dfrac{C_3^2}{2a_*}$. Therefore, for all $\eta \in H$ satisfying

$$\|\eta\|_H^2 = \frac{C}{m} \tag{4.60}$$

we have

$$\langle \mathcal{F}(\eta), \eta\rangle_H \geq 0. \tag{4.61}$$

By the Corollary of the Brouwer Fixed Point Theorem (Corollary 3.2.1), $\exists\, \widetilde{\eta} \in H$ such that

$$\mathcal{F}(\widetilde{\eta}) = 0 \qquad \text{and} \qquad \|\widetilde{\eta}\|_H \leq \sqrt{\frac{C}{m}} \tag{4.62}$$

which proves the existence of a solution of problem $(P_h^{n,\epsilon})$.

□

Remark 4.1.1. The choice of inner products in the proof of existence were directed at the removal of the terms $\tau\langle\nabla\cdot\bar{\mathbf{q}}, w_i\rangle$ in (4.55), $\langle\nabla\cdot\bar{\mathbf{u}}, w_i\rangle$ in (4.56), $\langle\bar{\Theta}, \nabla\cdot\mathbf{v}_i\rangle$ in (4.57), and $\langle\bar{p}, \nabla\cdot\mathbf{v}_i\rangle$ in (4.58). In the Galerkin formulations we considered in Chapter 2, we could simply apply the Poincaré inequality directly, but in the mixed case considered here, such a statement (a relation between W_h and V_h) does not necessarily exist.

Remark 4.1.2. It can be shown that the result of uniqueness in Theorem 4.1 is still valid without the assumption $f_2 \equiv 0$.

Here is the final result:

Theorem 4.2. *Assuming (A1)-(A3) hold true and $f_2 \equiv 0$, there exists a unique solution of problem (P_h^n) for τ sufficiently small.*

Proof. By Theorem 4.1, there exists a unique set of solutions $(\Theta_h^{n,\epsilon}, \mathbf{q}_h^{n,\epsilon}, p_h^{n,\epsilon}, \mathbf{u}_h^{n,\epsilon}) \in W_h \times V_h \times W_h \times V_h$. Moreover by Proposition 4.1, these are bounded sequences in ϵ. Therefore, by the Eberlein-Šmuljan Theorem (Theorem 1.5) there exists $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n) \in W_h \times V_h \times W_h \times V_h$ and subsequences of $\{\Theta_h^{n,\epsilon}\}_\epsilon$, $\{\mathbf{q}_h^{n,\epsilon}\}_\epsilon$, $\{p_h^{n,\epsilon}\}_\epsilon$ and $\{\mathbf{u}_h^{n,\epsilon}\}_\epsilon$ converging weakly to $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n)$. V_h and W_h are finite-dimensional, and therefore weak convergence is equivalent with strong convergence.

□

4.3 Summary

In this chapter we have proved a result for existence and uniqueness of a fully discrete formulation (P_h^n) derived from a mathematical model of two-phase flow in porous media, and based on the assumptions (A1)-(A3).

The objective of this chapter was to explore which properties the saturation $s(\cdot)$ could possess in order for us to show a result. We assumed $s(\cdot)$ to be Hölder continuous and strongly monotonically increasing, which was essential to derive a result.

In [32], it is discussed that a proof of existence and uniqueness of a solution of (P_h^n) with $s(\cdot)$ Lipschitz continuous and monotonically increasing is based on the Banach fixed point theorem (Theorem 3.1) along with the Lax-Milgram Theorem (Theorem 2.4). The construction of such a proof is similar to the result we showed in Proposition 3.2 followed by Proposition 3.3. There, we first proposed a linearization scheme, then proved existence and uniqueness of a solution of the scheme, and finally used the Banach fixed point Theorem to show for which values of $L > 0$ in the L-scheme we obtained convergence to a solution of the main problem.

For future studies, it would be interesting to see if it is possible to derive a result for a Hölder continuous and monotonically increasing saturation. The method using the Corollary of the Brouwer Fixed Point Theorem could also be applied to other physical problems, possibly giving satisfactory results for existence and uniqueness of other variational formulations.

Bibliography

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311-341.

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

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

[4] I. Babuška, *Error-Bounds for Finite Element Method*, Numerische Mathematik, Volume: 16 (1971), pp. 322-333.

[5] I. Babuka, A. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, (1972), pp. 1-359.

[6] S. Banach, *Théorie des opérations linéaires*, Monografie Matematyczne 1, Warszawa (1932).

[7] S. C. Brenner, L. R. Scott, *The Mathematical Theory of Finite Element Methods*, third edition, Springer-Verlag New York (2008).

[8] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer-Verlag New York (2011).

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Verlag, New York (1991).

[10] G. Chavent and J. Jaffre, *Mathematical models and finite elements for reservoir simulation*, Elsevier (1991).

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

[12] Z. Chen and R. Ewing, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology*, SIAM J. Numer. Anal. (1997), pp. 2228-2253.

[13] Z. Chen and R. Ewing, *Degenerate two-phase incompressible flow III. Sharp error estimates*, Numer. Mat. 90 (2001), pp. 215-240.

[14] W. Cheney, *Analysis for Applied Mathematics*, Springer-Verlag (2001).

[15] P. G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, Vol. 130. SIAM (2013).

[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1978).

[17] L. C. Evans, *Partial differential equations*, American Mathematical Society (2010).

[18] Y. Kannai, *An Elementary Proof of the No-Retraction Theorem*, The American Mathematical Monthly, vol. 88, no. 4 (1981), pp. 264268.

[19] A. Ern and J. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag New York (2004).

[20] S. Karpinski, I. S. Pop, *Analysis of an interior penalty discontinuous Galerkin scheme for two phase flow in porous media with dynamic capillarity effects*, Numerische Mathematik 136 (2017), pp. 249-286.

[21] S Karpinski, I. S. Pop, F. A. Radu, *Analysis of a linearization scheme for an interior penalty discontinuous Galerkin method for two phase flow in porous media with dynamic capillarity effects* Internat. J. Numer. Methods Engrg. (2017), DOI: 10.1002/nme.5526.

[22] E. Kreyszig, *Introductory functional analysis with applications*, Vol. 1. Wiley, New York (1989).

[23] K. Kumar, I.S. Pop and F.A. Radu, *Convergence analysis of mixed numerical schemes for reactive flow in a porous medium*, SIAM J. Num. Anal. 51 (2013), pp. 2283-2308.

[24] K. Kumar, I.S. Pop and F.A. Radu, *Convergence analysis for a conformal discretization of a model for precipitation and dissolution in porous media*, Numerische Mathematik 127 (2014), pp. 715-749.

[25] P. D. Lax and A. N. Milgram, *Parabolic Equations*, Ann. Math. Studies, 33 (1954), pp. 167190.

[26] F. List and F. A. Radu, *A study on iterative methods for Richards' equation*, Computational Geosciences 20 (2016), pp. 341-353.

[27] J. Nečas, *Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle*, Annali della Scuola Normale Superiore di Pisa - Classe di Scienze, Volume: 16, Issue: 4 (1962), pp. 305-326.

[28] J. M. Nordbotten and M. A. Celia, *Geological Storage of CO2. Modeling Approaches for Large-Scale Simulation*, John Wiley & Sons, (2012).

[29] F. A. Radu, I. S. Pop and S. Attinger, *Analysis of an Euler implicit - mixed finite element scheme for reactive solute transport in porous media*, Numerical Methods for Partial Differential Equations, Volume 26, Issue 2, (2010), pp. 320-344, DOI:10.1002/num.20436.

[30] F. A. Radu, I. S. Pop and P. Knabner, *Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation*, SIAM J. Numer. Anal. 42 (2004), pp. 1452-1478.

[31] F. A. Radu, K. Kumar, J. M. Nordbotten and I. S. Pop, *A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media*, arXiv:1512.08387, (2015).

[32] F. A. Radu, K. Kumar, J. M. Nordbotten and I. S. Pop *A robust, mass conservative scheme for two-phase flow in porous media including Hoelder continuous nonlinearities*, IMA Journal of Numerical Analysis (2017).

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

[34] F. A. Radu, *Mixed finite element discretization of Richards' equation: error analysis and application to realistic infiltration problems*, PhD Thesis, University of Erlangen, Germany (2004).

[35] F. A. Radu, S. Attinger, M. Bause and A. Prechtel, *A mixed hybrid finite element discretization scheme for reactive transport in porous media.* In Numerical Mathematics and Advanced Applications, K. Kunisch, G. Of, O. Steinbach (editors), Springer (2008), pp. 513-520.

[36] F. A. Radu, N. Suciu, J. Hoffmann, A. Vogel, O. Kolditz, C-H. Park and S. Attinger, *Accuracy of numerical simulations of contaminant transport in heterogeneous aquifers: a comparative study*, Advances in Water Resources, Volume 34, Issue 1 (2011), pp. 47-61.

[37] F. A. Radu, *Convergent mass conservative schemes for flow and reactive solute transport in variably saturated porous media*, Habilitation Thesis, University of Erlangen, Germany (2013).

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

[39] E. Zeidler, *Nonlinear Functional Analysis and its Applications: II/B: Nonlinear Monotone Operators*, Springer-Verlag New York (1990).