MPFA methods for Richards' equation

Truls Moholt

Master thesis in Applied and Computational Mathematics,
Institute of Mathematics,
University of Bergen,
Autumn 2021

Contents

Abstract

In this thesis, we review spatial discretization methods for parabolic problems, with applications to Richards' equation. In particular, we discretize Richards' equation after Kirchhoff transform with the MPFA-L-method in space, backward Euler in time and L-scheme for linearization. Then, we apply the techniques in [5] to prove a convergence rate estimate. We also compare the spatial discretization techniques numerically on different grids. Moreover, we do numerical experiments involving the fully discretized Richards' equation, verifying our theoretical findings. All the numerical experiments are done using our code, implemented with Python and Numpy, see https://github.com/trulsmoholt/masterthesis.

Acknowledgements

I would like to thank my supervisors Adrian Florin Radu and Erlend Storvik for their guidance. I have learned a lot from you during the last two years, both during our weekly meetings and in the courses you have taught at UIB. I am very grateful for your detailed feedback and advice. Without Erlend's help with debugging, I would have ended up using someone else's code.

I would also want to thank fellow master students and the members of the Porous media group for helpful discussions and motivation.

Introduction

Understanding porous media and how fluid flows through it has many useful applications, for example predicting the spread of some contaminant in an aquifer. Other examples include CO_2-storage, geothermal energy extraction and brain modelling. One way of understanding the processes involved in porous media flow is to describe it using partial differential equations, which may be time dependent, non-linear, degenerate and almost always impossible to solve analytically. We therefore, want numerical algorithms that solve these PDEs approximately, and at the same time, respect important properties of the equations/problems we consider, such as mass conservation.

In this thesis, we focus on numerical techniques for solving Richards' equation, introduced in [20], in two spatial dimensions, i.e., we want to find the pressure head ψ such that

$$\frac{\partial \theta(\psi)}{\partial t} - \nabla \cdot (\boldsymbol{\kappa}(\theta(\psi))(\nabla \psi + \boldsymbol{e}_z)) = f, \tag{0.1}$$

with boundary and initial conditions. The above equation models groundwater flow in partially saturated reservoirs. The non-linear functions $\theta(\cdot)$ and $\kappa(\cdot)$ are determined experimentally and corresponds to saturation and hydraulic conductivity respectively, see [24] for a commonly used example. The vector $\boldsymbol{e}_z$ corresponds to the gravitational force which we will neglect in this thesis, and f represents any sources or sinks.

Equation (0.1) is time dependent, parabolic, involves two non-linearities and possibly degenerate. We discretize in time with an implicit method, and get a non-linear elliptic PDE in each time step. This is then linearized with a robust linearization scheme, the L-scheme [12, 17], leading to a sequence of linear elliptic PDEs in each time step. Next, we approximate the solution to these linear elliptic PDEs with a spatial discretization, which will be the main focus of this thesis.

When solving (0.1) we are often additionally interested in transport phenomena, which means that some solute u is transported with the fluid flux, $\boldsymbol{q}$

$$\frac{\partial u}{\partial t} - \nabla \cdot (\boldsymbol{q} u) = 0. \tag{0.2}$$

We therefore want our approximate solution of (0.1) to render an approximate flux field which is *locally mass conservative*, that is important when solving (0.2). Whether or not the approximated flux field is locally mass conservative is a property of the spatial discretization. It turns out that the linear Lagrange finite element method, which we will cover in section 2.1.4, does not in general have this property. There are, however, a class of finite element methods called *mixed finite element methods*, which conserve mass locally, but have the disadvantage of having more unknowns, and are not discussed further in this thesis. We are therefore interested in finite volume methods, as they are designed to be locally mass conservative.

There are also other properties that we want our spatial discretization techniques to have. Among these are *monotonicity*, which means that the discretization does not allow for unphysical oscillations in the approximated solution. Another desirable property is the ability to handle complex geometries, i.e., grids that are not orthogonal, and that consists of general quadrilaterals (rough grids). Moreover, a computationally efficient method is desirable, and therefore, methods with smaller stencils are preferred. In [14], the authors show that it if you have a locally mass conservative method, with a small stencil (nine-point cell stencil in two spatial dimensions), that can handle rough grids, you cannot guarantee unconditional monotonicity. There are therefore always a trade-off between various desirable properties when choosing a spatial discretization method for (0.1). If we, for example, only model groundwater flow in one kind of soil on a easy domain, we could use an orthogonal grid, and our discretization would be efficient and unconditionally monotone.

The MPFA (Multi-Point Flux Approximation) L-method, introduced in [2], is a finite volume method, and will be the main focus of this thesis. It is a compromise between all the properties we want, with a cell stencil usually consisting of only seven points, good monotonicity properties and ability to handle rough grids. We introduce it in section 2.2.3, and provide numerical experiments in chapter 5.

Convergence rate estimates for finite volume methods does not come "out of the box", as they do with finite element methods, at least not for non-orthogonal grids. In [22], A. F. Stephansen shows convergence for the MPFA-L-Method by formulating it as a mimetic finite difference method. Another approach, that was successfully applied in [8] for the MPFA-L-method on a triangular mesh, is to show equivalence with a mixed finite element method. In [10], Klausen and Winther used the same approach for the MPFA-O-method on a quadrilateral grid. In this thesis, we will show equivalence between the MPFA-L-method on a parallelogram mesh, and a modified linear Lagrange finite element method with triangular elements. After equivalence is obtained, we use the finite element framework to prove convergence. Our approach is similar to the one used in [5].

After convergence for the MPFA-L-method is achieved, we will see how it can be applied to obtain a convergence rate estimate for the fully discretized Richards' equation. To achieve this, we use the techniques found in [12] for proving convergence of the linearization scheme, and the techniques in [19, 16] for convergence of the time discretization. In the end, we reach an L^2 error estimate, which we confirm by numerical experiments in section 5.3.

Outline

In chapter one 1, we give a brief introduction to flow in porous media. Highlighting the physical principles that leads to Richards' equation (1.9).

In chapter two 2, we cover some of the numerical approximation techniques one may use to solve Richards' equation, with an emphasis on spatial discretization methods. We start by introducing function spaces, followed by the weak formulation and its well posedness. Then, we introduce the finite element method, how it could be implemented, and the ideas behind proving its convergence. Further, we introduce finite volume methods, then we cover some common finite volume methods; two-point flux approximation, MPFA-O-method and MPFA-L-method. We give a short introduction to time discretization, specifically implicit backward Euler. We end the chapter by discussing iterative methods for solving non-linear problems.

In chapter three 3, we introduce a way of handling boundary conditions for the MPFA-L-method, and show its equivalence with a modified finite element method. Then, we apply standard finite element theory to show convergence for an elliptic problem.

In chapter four 4, we discuss the Kirchhoff transform of Richards' equation, removing the non-linearity in the constitutive law. Next, we prove convergence of a fully discretized and linearized scheme to solve the Kirchhoff transformed Richards' equation.

In chapter five 5, we present some of the code written for this thesis, and do numerical experiments involving elliptic and time-dependent equations on rough grids.

Chapter 1

Flow in Porous Media

In this chapter we introduce the basic concepts of flow in porous media, briefly covering the modelling choices and physics that leads to Richards' equation. The theory in this chapter is to a large extent adapted from [13] and UIB's Porous media course.

1.1 The Representative Elementary Volume

A porous medium consists of a solid matrix and some void filled with fluid of one or more phases. In single phase flow, all the pores are filled with one fluid, in two-phase flow however, we have fluid-fluid interfaces in the void. In porous media research, one has come to the realization that the solid matrix is too complex to model. Instead, one takes averages of variables over a reasonable length scale, i.e., the *representative elementary volume* (REV). An important characterization of a porous medium is the *porosity* ϕ, which is defined as

$$\phi := \frac{\textit{volume of voids in REV}}{\textit{volume of REV}}.$$

Another important quantity is the *saturation* S_α of phase α, this is defined

$$S_\alpha := \frac{\textit{volume of } \alpha \textit{ in REV}}{\textit{volume of voids in REV}}.$$

In single phase flow, the saturation is irrelevant as the saturation is always one. Also note that the volumetric content of phase α in the REV, θ_α, is given by $\theta_\alpha = S_\alpha \phi$.

1.2 Darcy's Law

In 1856, Henri Darcy performed a famous experiment where he studied the flow of water through sand. To understand his experiment we must first define some variables for measuring water content. First, we assume that the external gravitational force on some fluid is balanced by the pressure gradient force, also known as *hydrostatic equilibrium.* This gives us that the pressure at height z above datum developed by a water column of height h above datum is given by

$$p_{abs}(z) = p_{atm} + \rho g(h - z).$$

here ρ is the density and g is the gravitational acceleration. If we define the *gauge pressure* p by $p := p_{abs} - p_{atm}$ we get an expression for p:

$$p = \rho g(h - z).$$

This can be rearranged to give an expression for the height, which we from now on refer to as *hydraulic head*:

$$h = \frac{p}{\rho g} + z. \tag{1.1}$$

A *manometer* is a tube with one end in the reservoir and one in open atmosphere, the water level in this tube is then $\frac{p}{\rho g}$. The volumetric flow of water is denoted by q_d. Darcy's experiment is shown in figure 1.1, where water is poured through a cylinder filled with sand. The cylinder has length L and has cross sectional area A. His observations are given by the equation called Darcy's law:

$$q_d = -\kappa \frac{A(h_2 - h_1)}{L}.$$

Where κ is a positive coefficient of proportionality. Let q denote the volumetric flow-rate per area:

$$q := \frac{q_d}{A} = -\kappa \frac{h_2 - h_1}{L},$$

we will refer to this as the *flux* of hydraulic potential. We can now state the differential version of Darcy's law. Taking the limit as $L \to 0$ we get

$$\boldsymbol{q} = -\boldsymbol{\kappa}\nabla h. \tag{1.2}$$

We call $\boldsymbol{\kappa}$ the *hydraulic conductivity* and note that it in general is a rank two tensor, a matrix. The hydraulic conductivity also has the property that it is *symmetric.* This is because there are, at every point in the reservoir, two orthogonal directions; one with maximum, and one with minimum hydraulic conductivity. Thus, the matrix, $\boldsymbol{\kappa}$, is diagonalizable by a orthogonal matrix.

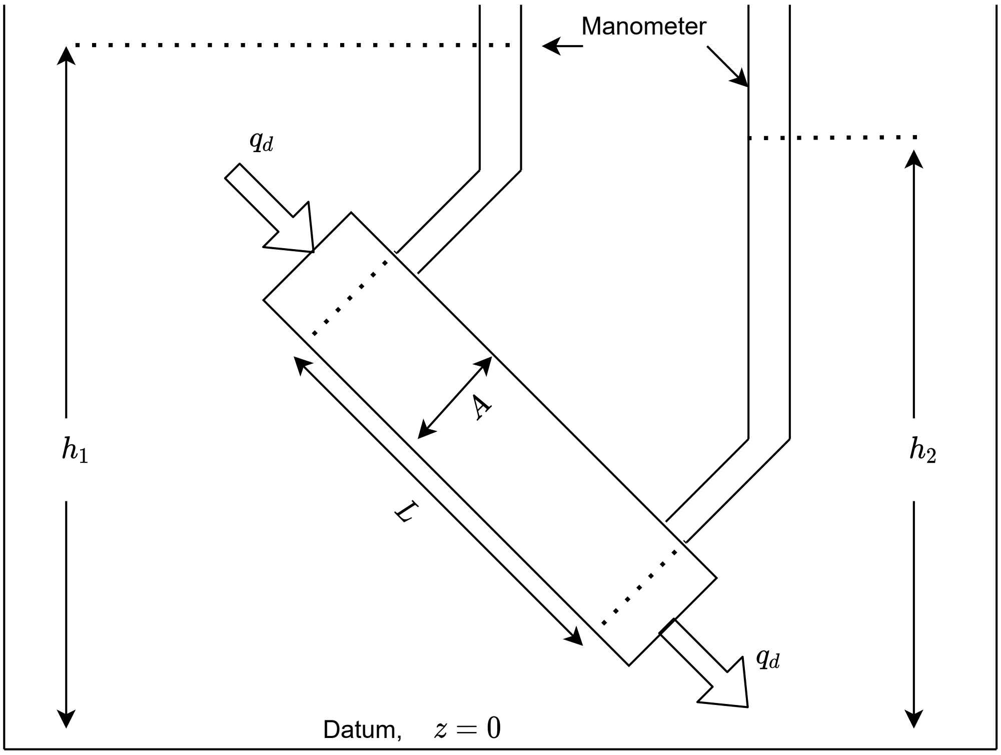


Figure 1.1: The Darcy experiment

The conductivity matrix, $\boldsymbol{\kappa}$, is also *positive definite*, this is because there is never flux towards higher pressure. With further experiments, similar to the one already described, we can understand what makes up $\boldsymbol{\kappa}$. Dimensionality analysis shows that it is a function of viscosity μ, density of the fluid ρ, gravity g and *permeability* $\boldsymbol{k}$,

$$\boldsymbol{\kappa} = \frac{\boldsymbol{k}\rho g}{\mu}. \tag{1.3}$$

The *permeability*, which is a property of the soil in the reservoir, is also a rank two tensor which is symmetric positive definite and it is in general a function of spatial coordinates, i.e., heterogeneous.

If we define the *pressure head* ψ as $\psi := \frac{p}{\rho g}$, we can combine (1.1), (1.2) and (1.3) to get another variant of Darcy's law;

$$\boldsymbol{q} = -\frac{\boldsymbol{k}\rho g}{\mu}\nabla(\psi + z) \tag{1.4}$$

which will be useful later.

1.3 Mass Conservation

Darcy's law is not enough if we want to determine the pressure or flow in a reservoir, but we can use the principle of *mass conservation* to add one more equation. The idea is that for every enclosed region in the reservoir, the change of mass inside the region is balanced by the mass flux into the region and the production of mass inside the region.
We end up with the mass balance equation, let Ω be our domain, then:

$$\int_\omega \frac{\partial(\rho\phi)}{\partial t} dV = -\int_{\partial\omega} \boldsymbol{n}\cdot\rho\boldsymbol{q}\; dS + \int_\omega f dV \quad \forall\omega\subseteq\Omega \;\; \text{with } \omega \text{ being a volume,}$$

where $\boldsymbol{n}$ is an outward pointing normal vector to ω and f corresponds to sources and/or a sinks. We can use the divergence theorem on the surface integral to get

$$\int_\omega \frac{\partial(\rho\phi)}{\partial t} + \nabla\cdot(\rho\boldsymbol{q}) - f dV = 0.$$

Since this is true for all enclosed regions $\omega\subset\Omega$, it also holds for the expressions inside the integral yielding the mass conservation PDE

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

This, together with Darcy's law (1.2) and appropriate boundary and initial conditions close the system

$$\begin{cases} \boldsymbol{q} = -\boldsymbol{\kappa}\nabla h, & \boldsymbol{x}\in\Omega, \quad t>0 \\ \dfrac{\partial(\rho\phi)}{\partial t} + \nabla\cdot(\rho\boldsymbol{q}) = f(\boldsymbol{x},t), & \boldsymbol{x}\in\Omega, \quad t>0 \\ h(\boldsymbol{x},t) = g(\boldsymbol{x},t), & \boldsymbol{x}\in\partial\Omega, \quad t>0 \\ h(\boldsymbol{x},t) = f(\boldsymbol{x}), & \boldsymbol{x}\in\Omega, \quad t=0 \end{cases} \tag{1.5}$$

Now we have a model for single-phase flow. As it is stated now, it is a linear parabolic equation, but for incompressible fluid and matrix it becomes an elliptic equation. One often writes the density as a function of pressure, it then becomes non-linear. See chapter two of [13] for a more detailed discussion of (1.5) and modelling options.

1.4 Two-phase Flow and Richards' Equation

We restrict our discussion to two phases for simplicity, but the theory can be extended to more phases. In two-phase systems one has a *wetting phase* and a

non-wetting phase, denoted by the subscripts w and n respectively.
When we introduce more phases, we continue with the equations we already introduced, i.e., we assume that Darcy's law (1.4) holds for both phases. Let the subscript α denote the phase, then we have Darcy's law for each phase

$$\boldsymbol{q}_\alpha = -\frac{\boldsymbol{k}_{r,\alpha}\boldsymbol{k}\rho g}{\mu}\nabla(\psi_\alpha + z), \tag{1.6}$$

where the coefficient $\boldsymbol{k}_{r,\alpha}$ is known as *relative permeability* and it has to be deduced from experimental observation.

We also assume conservation of mass for each phase:

$$\frac{\partial(S_\alpha\rho_\alpha\phi)}{\partial t} + \nabla\cdot(\rho\boldsymbol{q}_\alpha) = f_\alpha. \tag{1.7}$$

Here, we assume that there is no mass transfer between the phases. If we combine equations (1.6) and (1.7), they give us 2 equations, but we have four unknowns ψ_w, ψ_n, S_w and S_n. We, therefore, introduce the algebraic relation

$$S_w + S_n = 1$$

and the physical relation

$$p_n - p_w = p_c \tag{1.8}$$

where p_c is called *capillary pressure.* As with the relative permeability, p_c also needs to be determined experimentally. With initial and boundary conditions we again have a closed system.

A common simplification is to assume that the capillary pressure and the relative permeability are functions of the saturation, and that the relative permeability is isotropic (a scalar).

Another simplification that is used, especially in groundwater hydrology, is that the non-wetting phase (air) always have $p_n = p_{atm} = 0$. For this assumption to hold it is important that the air always is connected to the surface. Now, equation (1.8) simplifies to

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

Experiments show that the capillary pressure is a monotone decreasing function of saturation, therefore we can invert it. Equation (1.8) now becomes:

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

Finally, we can multiply the above equation by the porosity to get an expression for the *water content* θ_w:

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

Combining this with the two-phase Darcy law (1.6) and mass balance (1.7) we get **Richards' equation**

$$\frac{\partial \theta(\psi)}{\partial t} - \nabla \cdot (\boldsymbol{\kappa}(\theta(\psi))(\nabla \psi + e_z)) = f \tag{1.9}$$

where $\theta = \theta_w$. Note that the density is eliminated, this is because it is assumed to be constant for water. The hydraulic conductivity is parametrized as a function of water content through experiments and can be written $\frac{\boldsymbol{k}_{r,\alpha} \boldsymbol{k} \rho g}{\mu} = \boldsymbol{\kappa}(\theta)$.

Richards' equation contains two non-linearities, θ and $\boldsymbol{\kappa}$, which make the analysis and numerical simulation more challenging as we will see. They may also cause the equation to degenerate, i.e., the parabolic equation may "collapse" into an elliptic PDE (see figure 1.2) or even an ODE when the saturation is so low that there is no flow.

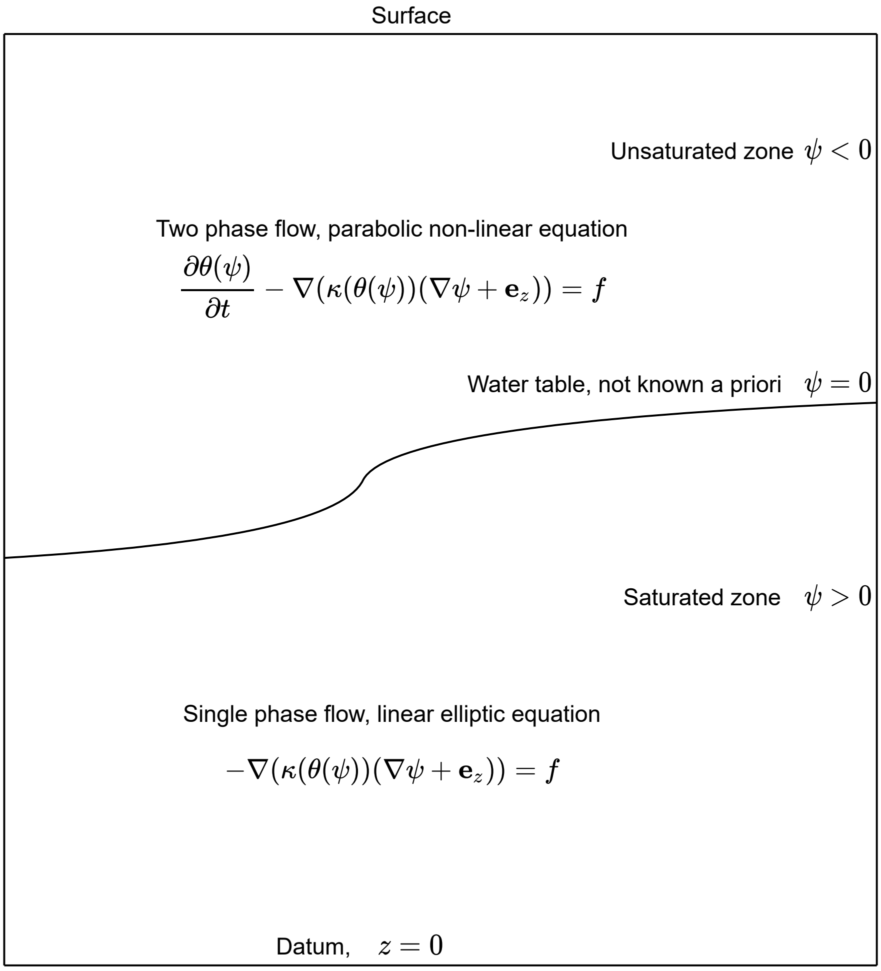

Surface
Unsaturated zone $\psi < 0$
Two phase flow, parabolic non-linear equation
$\frac{\partial\theta(\psi)}{\partial t} - \nabla(\kappa(\theta(\psi))(\nabla\psi + \mathbf{e}_z)) = f$
Water table, not known a priori $\psi = 0$
Saturated zone $\psi > 0$
Single phase flow, linear elliptic equation
$-\nabla(\kappa(\theta(\psi))(\nabla\psi + \mathbf{e}_z)) = f$
Datum, $z = 0$

Figure 1.2: A sketch of the degeneracy of Richards' equation

Chapter 2

Numerical Approximation Techniques

In this chapter, we first discuss two important frameworks for spatial discretization of PDEs, followed by a brief introduction of time discretization, and at the end, an introduction to linearization. The focus will be on two dimensional elliptic and parabolic equations, but the concepts covered can easily be generalized to three dimensions. After reading this chapter, the reader hopefully has some idea of how to implement a few different methods for solving the Poisson equation, the heat equation and maybe even Richards' equation, and some of their properties.

2.1 The Finite Element Method

The finite element method was first developed in the 1940s by Richard Courant for problems in solid mechanics. As computers became better in the 1960s the method increased in popularity [21]. Today there are several general purpose finite element programs being used for a wide range of problems.

In this section we will introduce the finite element method and state the most important results about stability and convergence. We will concentrate on solving the Poisson equation: Let $\Omega \subset \mathbb{R}^n$ be some open and bounded domain, find u such that

$$\begin{aligned} -\nabla \cdot \boldsymbol{K} \nabla u(x) &= f(x), \quad && x \in \Omega, \\ u(x) &= 0, && x \in \partial\Omega. \end{aligned} \tag{2.1}$$

For this equation to be well defined we require that u has double derivatives in Ω, but it is easy to come across physical examples where this does not make sense. This is some of the motivation for the Poisson equation in the *variational formulation.* Another motivation is that it allows for a general framework for computing the solution, as we will soon see. But first, let us introduce some spaces of functions and their properties.

2.1.1 Function Spaces

When discussing PDEs and the numerical schemes to solve them it is important to have a precise notion of what kind of functions we are looking for and their properties. The function spaces discussed here are all normed vector spaces. From now on we assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain.

Definition 1 (Lebesgue spaces, $L^p(\Omega)$)**.** *For $p \in [1, \infty)$ let $L^p(\Omega)$ be the space of functions where $\|u\|_p = (\int_\Omega u^p dx)^{1/p} < \infty$.*

Remark 1. *Note that the $L^p(\Omega)$ norm induces equivalence relations on the set of functions. Two functions in $L^p(\Omega)$ are equal if they only differ on a set of measure zero.*

An important concept when discussing normed vector spaces are that they intuitively do not have any points missing, this is formally defined as spaces where every Cauchy sequence converges. This is known as *complete* normed vector spaces or *Banach spaces.*

Theorem 2.1.1 (Riesz-Fischer Theorem [6] chapter 8)**.** *Each $L^p(\Omega)$ space is a Banach space.*

Remark 2. *The space $L^2(\Omega)$ is a inner product space, with inner product*

$$\langle u, v \rangle_{L^2} = \int_\Omega uv \; dx.$$

Banach spaces with an inner product, that induces the norm

$$\langle u, u \rangle^{\frac{1}{2}} = \|u\|,$$

are called ***Hilbert spaces****.*

Before we continue the study of function spaces we develop some convenient notation for derivatives.

Definition 2 (multi-index notation)**.** *Let $\overline{\alpha}$ be an ordered n-tuple. We call this a multi-index and denote the length $|\overline{\alpha}| = \sum_{i=1}^n \alpha_i$. For $\phi \in C^\infty(\Omega)$ we define $D^{\overline{\alpha}}\phi = (\frac{\partial}{\partial x_1})^{\alpha_1}(\frac{\partial}{\partial x_2})^{\alpha_2}...(\frac{\partial}{\partial x_n})^{\alpha_n}\phi$*

We would also like a more general notion of derivative than the one presented in a basic calculus book.

Definition 3 (weak derivative)**.** *Let $L^1_{loc}(\Omega) = \{\, f \in L^1(K) : \forall K \subset \Omega$ where K is compact $\}$. Let $f \in L^1_{loc}(\Omega)$. If there exists $g \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega g\phi dx = (-1)^{|\overline{\alpha}|} \int_\Omega f D^{\overline{\alpha}} \phi dx \quad \forall \phi \in C^\infty(\Omega)$$

with $\phi = 0$ on $\partial\Omega$ we say that g is the weak derivative of f and denote it by $D^{\overline{\alpha}}_w f$.

We can now define a class of subspaces of the L^p spaces known as the **Sobolev spaces**.

Definition 4 (Sobolev space)**.** *Let k be a non-negative integer, define the Sobolev norm as*

$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\overline{\alpha}| \le k} \left\| D^{\overline{\alpha}}_w u \right\|^p_{L^p(\Omega)} \right)^{1/p} .$$

We then define the Sobolev spaces as

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

Theorem 2.1.2. *The Sobolev spaces $W^{k,p}(\Omega)$ are Banach spaces.*

Proof. Let $\{u_i\}_{i=0}^\infty \subseteq W^{k,p}(\Omega)$ be a Cauchy sequence. This implies that for all $\overline{\alpha}$, $|\overline{\alpha}| \le k$ we have a Cauchy sequence in $L^p(\Omega)$:

$$\|u_j - u_i\|_{W^{k,p}} = (\sum_{|\overline{\alpha}| \le k} \left\| D^{\overline{\alpha}}_w u_j - D^{\overline{\alpha}}_w u_i \right\|^p_{L^p(\Omega)})^{1/p} < \epsilon \ \forall i, j \ge N$$
$$\implies \left\| D^{\overline{\alpha}}_w u_j - D^{\overline{\alpha}}_w u_i \right\|_{L^p(\Omega)} < \epsilon.$$

By theorem (2.1.1), every $L^P(\Omega)$ space is a Banach space. Therefore, for each $|\overline{\alpha}| \le k$, $D^{\overline{\alpha}}_w u_i$ converges to some limit, $u_{\overline{\alpha}} \in L^p(\Omega)$, as $i \to \infty$. In particular $u_i \to u$ in $L^p(\Omega)$, so the limit in the $\|\cdot\|_{W^{k,p}(\Omega)}$ norm, u, is well defined. Now we need to show that $\{u_{\overline{\alpha}}\}_{\overline{\alpha}}$, are in fact the weak derivatives of u, i.e., $D^{\overline{\alpha}}_w u = u_{\overline{\alpha}}$. In other words, that the limit of u_i in the $\|\cdot\|_{W^{k,p}(\Omega)}$ norm, u, is in fact in $W^{k,p}(\Omega)$. By the definition of weak derivative we have:

$$\int_\Omega D^{\overline{\alpha}}_w u_i \phi dx = (-1)^{|\overline{\alpha}|} \int_\Omega u_i D^{\overline{\alpha}} \phi dx.$$

Let $1 = \frac{1}{q} + \frac{1}{p}$, applying Hölder's inequality on both sides we get the two inequalities

$$\int_\Omega (D^{\overline{\alpha}}_w u_i - u_{\overline{\alpha}}) \phi dx \le \left\| D^{\overline{\alpha}}_w u_i - u_{\overline{\alpha}} \right\|_{L_p} \|\phi\|_{L_q} ,$$
$$\int_\Omega (u_i - u) D^{\overline{\alpha}} \phi dx \le \|u_i - u\|_{L_p} \left\| D^{\overline{\alpha}} \phi \right\|_{L_q} .$$

Taking the limit, the right hand side goes to zero, and by the fact that we can move the limit out of the integral

$$\lim_{i\to\infty}\int_\Omega D_w^{\overline{\alpha}} u_i \phi dx = \int_\Omega u_{\overline{\alpha}}\phi dx,$$
$$\lim_{i\to\infty}\int_\Omega u_i D^{\overline{\alpha}}\phi dx = \int_\Omega u D^{\overline{\alpha}}\phi dx.$$

Now, we can put the two equations together with the definition of the weak derivative:

$$\int_\Omega u_{\overline{\alpha}}\phi dx = \lim_{i\to\infty}\int_\Omega D_w^{\overline{\alpha}} u_i\phi dx = \lim_{i\to\infty}(-1)^{|\overline{\alpha}|}\int_\Omega u_i D^{\overline{\alpha}}\phi dx = \int_\Omega u D^{\overline{\alpha}}\phi dx.$$

We have shown $D_w^{\overline{\alpha}} u = u_{\overline{\alpha}}$, and therefore, that $u \in W^{k,p}(\Omega)$. □

Definition 5. *We rename the $L^2(\Omega)$ based Sobolev spaces as*

$$H^k(\Omega) = W^{k,2}(\Omega),$$

with the norm of $H^k(\Omega)$ being written in the more compact form $\|\cdot\|_{\Omega,k}$ or just $\|\cdot\|_k$, and the inner product defined as follows:

$$\langle u, v\rangle_k = \sum_{|\overline{\alpha}|\le k}\int_\Omega D_w^{\overline{\alpha}} u D_w^{\overline{\alpha}} v dx.$$

In Sobolev spaces it is not obvious that a function is well defined on a lower dimensional subset of Ω, because two functions may map elements of this zero measure subset to different values and still be of the same equivalence class. This is important to settle if we want to solve boundary value problems. The following results holds for general $L^p(\Omega)$ based Sobolev spaces, but we will only state them for the Hilbert space $H^1(\Omega)$.

Definition 6. *We denote by $H_0^k(\Omega)$ the closure of $C_c^\infty(\Omega)$ in $H^k(\Omega)$, where $C_c^\infty(\Omega)$ is the space of infinitely differentiable functions with compact support.*

Theorem 2.1.3 (Trace theorem, (Evans [7], chapter 5))**.** *Assume U is bounded and ∂U is C^1. Then there exists a bounded, linear operator*

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

Such that

1. *$Tu = u|_{\partial u}$ if $u \in H^1 \cap C(\overline{U})$*

2. $\|Tu\|_{L^p(\partial U)} \leq \|u\|_{H^1(U)}$

We call Tu the trace of u. Note that the theorem does not state that T is surjective.

Theorem 2.1.4. *(Trace-zero functions in $W^{1,p}$,(Evans [7], chapter 5)) Suppose U is as in the previous theorem and $u \in W^{1,p}(U)$, then*

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

Remark 3. *We often denote the image of T as:*

$$H^{\frac{1}{2}}(\Omega) = T(H^1(\Omega))$$

And define the norm

$$\|f\|_{H^{\frac{1}{2}}(\Omega)} = \inf_{w \in H^1(\Omega),\ Tw=f} \|w\|_1$$

Now we have the theory we need to study elliptic boundary value problems and their weak solutions.

2.1.2 The Variational Problem

We obtain the **variational formulation** of (2.1) by multiplying (2.1) with a function v in a suitable space V called the *test space*, integrating over Ω and using integration by parts/divergence theorem,

$$-\int_\Omega v\nabla \cdot \boldsymbol{K}\nabla u \, dx = -\int_{\partial\Omega} v\boldsymbol{K}\nabla u \cdot \boldsymbol{n} \, dx + \int_\Omega (\nabla v)^T \boldsymbol{K}\nabla u \, dx = \int_\Omega vf \, dx.$$

If we choose v such that $v = 0$ on $\partial\Omega$, then the integral over the boundary vanishes. The new formulation reads: Find u such that

$$\int_\Omega (\nabla v)^T \boldsymbol{K}\nabla u \, dx = \int_\Omega vf \, dx \quad \forall v \in V. \tag{2.2}$$

A good choice of the test space V is $V = H_0^1(\Omega)$. We also choose this as the solution space. We see that if u is a solution to (2.1), it also solves (2.2). But a solution to (2.2) does not necessarily solve (2.1), that is why it is also called the *weak formulation.*

The variational problems that we will look at, will all have the form: Find u such that

$$a(u, v) = b(v) \quad \forall v \in V, \tag{2.3}$$

where $a(\cdot, \cdot)$ is a *bilinear form* on V and $b(\cdot)$ is a *linear functional* on V. To be precise we define a famous concept from functional analysis:

Definition 7 (dual space). *Let V be a normed vector space, then we define it's dual space as the space of functions from V to $\mathbb{R}$ that are linear and continuous, also called linear functionals. We denote it by V'. This is a normed vector space with the norm:*

$$\|v\|_{V'} = \sup_{u \in V} \{|v(u)| : \|u\|_V = 1\}.$$

In general, a variational formulation can be seen as finding the element in a Banach space that is mapped to an element in its dual space by some map.

Boundary Conditions

Let $\partial\Omega = \Gamma_D \bigcup \Gamma_N$ with $\Gamma_D \bigcap \Gamma_N = \emptyset$, then (2.1) with more complex boundary conditions can be written as: Find $\hat{u}(x)$ such that

$$\begin{cases} -\nabla \cdot \boldsymbol{K}\nabla\hat{u}(x) = f(x), & x \in \Omega, \\ \hat{u}(x) = g_D, & x \in \Gamma_D, \\ \boldsymbol{K}\nabla\hat{u}(x) = g_N, & x \in \Gamma_N. \end{cases} \tag{2.4}$$

To make a variational formulation of (2.4) we first define the test space:

$$V = \left\{v \in H^1(\Omega) : T(v) = 0 \text{ on } \Gamma_D\right\}.$$

Next, we define the bilinear form:

$$a(u, v) := \int_\Omega \nabla u \boldsymbol{K} \nabla v \, dx.$$

Further, assume there exists an element w of $H^1(\Omega)$ that are mapped by the trace operator such that Dirichlet boundary conditions are met: $T(w) = g_D$. Let $\hat{u} = u + w$, where $u \in H_0^1(\Omega)$, we can use integration by parts as before:

$$a(u + w, v) = \int_\Omega (\nabla u + \nabla w)^T \boldsymbol{K} \nabla v \, dx = \int_\Omega fv \, dx - \int_{\partial\Omega} \boldsymbol{K}\nabla(u + w) \cdot \boldsymbol{n} v \, dx.$$

Using the linearity of $a(\cdot, \cdot)$ and inserting boundary conditions we get:

$$a(u, v) = b(v) = \int_\Omega fv \, dx - \int_\Omega (\nabla w)^T \boldsymbol{K} \nabla v \, dx - \int_{\Gamma_N} g_N v \, dx. \tag{2.5}$$

Hence both Dirichlet and Neumann boundary conditions are incorporated into the right hand side. For homogeneous Dirichlet boundary conditions, the second term on the right hand side of (2.5) vanishes.

2.1.3 Existence and Uniqueness

We still need to show that (2.5) has an unique solution. First, we define some important properties that a variational problem should have in order to have a unique solution. Let $(V, \|\cdot\|_V)$ be a Hilbert space.

Definition 8. *Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form. We say that:*

- *$a(\cdot,\cdot)$ is **coercive with respect to** V, or **elliptic** if there exists a constant $C_c \in \mathbb{R}$ such that $C_c \|u\|_V^2 \leq a(u,u)\ \forall u \in V$,*
- *$a(\cdot,\cdot)$ is **bounded** or **continuous** if there exists a constant C_B such that $|a(u,v)| \leq C_B \|u\|_V \|v\|_V\ \ \forall u, v \in V$.*

In order to prove existence and uniqueness, we must first state some important results about the underlying space V. The following theory can be found in its entirety in the first four chapters of Cheney [6]

Theorem 2.1.5 ([6] page 64)**.** *If Y is a closed subspace of the Hilbert space X, then*

$$X = Y \otimes Y^{\perp},$$

where $Y^{\perp} = \{x \in X : \langle x, y\rangle = 0\ \forall y \in Y\}$ is the orthogonal complement of Y. In other words, an element in X can always be written as the sum of an element Y and an element in $Y^{\perp}$.

Theorem 2.1.6 (Riesz representation theorem)**.** *Every continuous linear functional, $\phi(x)$, defined on a Hilbert space X can be written $\phi(x) = \langle x, v\rangle$ by a uniquely determined $v \in X$.*

Proof. Let $\phi \in X'$, define $Y = \{x \in X : \phi(x) = 0\}$. Take a non-zero element in the orthogonal complement $u \in Y^{\perp}$ such that $\phi(u) = 1$, (if this does not exist then $X = Y$ and $\phi(x) = \langle x, 0\rangle$, this is ensured by theorem 2.1.5). Now, we can write every vector in X as a linear combination of a vector in Y and the vector u. $x = x - \phi(x)u + \phi(x)u$ for any $x \in X$. Using this, we can find an expression for the inner product of x with a scaled version of u
$\left\langle x, \frac{u}{\|u\|^2}\right\rangle = \left\langle x - \phi(x)u, \frac{u}{\|u\|^2}\right\rangle + \left\langle \phi(x)u, \frac{u}{\|u\|^2}\right\rangle$. The first part of the sum vanishes as $x - \phi(u)x \in Y$. So we end up with
$\left\langle x, \frac{u}{\|u\|}\right\rangle = \phi(x)\frac{\langle u,u\rangle}{\|u\|^2} = \phi(x)$.

□

Theorem 2.1.7 (Banach fixed point theorem)**.** *Let X be a Banach space and $F : X \to X$ an operator where $\|Fx - Fy\|_X \leq \theta \|x - y\|_X$ for some $\theta \in (0,1)$, we call this a **contraction**.*
Then for all $x \in X$ the sequence $[x, Fx, F^2x, ...]$ converges to a point $x^* \in X$ called the fixed point of F.

See page 177 of [6] for a proof.

Theorem 2.1.8 (Lax-Milgram)**.** *Suppose that $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a bilinear, bounded and coercive form and that $b(\cdot) : V \to \mathbb{R}$ is a bounded, linear functional. Then the variational problem has a unique solution $u \in V$, such that*

$$a(u,v) = b(v) \tag{2.6}$$

for all $v \in V$.

Remark 4. *If $a(\cdot,\cdot)$ also is symmetric, it defines an inner product on V giving a complete space. We can then use Riesz representation theorem 2.1.6 to show that it has an unique solution.*

Proof of Lax Milgram theorem 2.1.8.
For each w denote the map $a(w,v) = a_w(v)$, this is a linear continuous functional, and follows from the assumptions on a. By Riesz representation theorem 2.1.6 $a_w(\cdot)$ uniquely determines an element $Aw \in V$ such that $a_w(v) = \langle Aw, v\rangle$. The map

$$\begin{aligned} A :& V \to V \\ & w \mapsto Aw, \end{aligned}$$

- is linear: $\langle A(x+y), v\rangle = a_{x+y}(v) = a(x+y,v) = a_x(v) + a_y(v) = \langle Ax, v\rangle + \langle Ay, v\rangle$. Since this holds for all $v \in V$, we have $A(x+y) = Ax + Ay$.
- is bounded: $\|Ax\| = \|a_x\| = \sup\{a(x,v) : \|v\| = 1\} \leq C_B \|x\|$.

We can also use Riesz representation theorem on the right hand side: $b(\cdot) = \langle f, \cdot\rangle$. Now we have a reformulation of (2.6):
Find u such that

$$Au = f. \tag{2.7}$$

Now we need to show that (2.7) has an unique solution, and for that we need the Banach fixed point theorem. Let $\epsilon > 0$, we define the operator

$$\begin{aligned} T : V &\to V \\ u &\mapsto u - \epsilon(Au - f). \end{aligned}$$

If T has a fixed point u^*, then $u^* - \epsilon(Au^* - f) = u^* \Rightarrow Au^* = f$ and we have solved (2.7) and proved the theorem. We just need to show that T is a contraction. First, let $u_1, u_2 \in V$, and subtract what they are mapped to by T, then we get

$$\|Tu_1 - Tu_2\|^2 = \|u - \epsilon(Au)\|^2,$$

where $u = u_1 - u_2$, we used the linearity of A,

$$= \|u\|^2 - 2\epsilon \langle u, Au \rangle + \epsilon^2 \langle Au, Au \rangle .$$

Now, we can use that $a(u, u) = \langle Au, u \rangle$, and that $\langle Au, Au \rangle = a_u(Au) = a(u, Au)$:

$$\|Tu_1 - Tu_2\|^2 = \|u\|^2 - 2\epsilon a(u, u) + \epsilon^2 a(u, Au).$$

Next, we use the coercivity and boundedness of $a(\cdot, \cdot)$. We also use the boundedness of A

$$\|Tu_1 - Tu_2\|^2 \leq \|u\|^2 - 2\epsilon C_c \|u\|^2 + \epsilon^2 C_B^2 \|u\|^2 .$$

This leads to the inequality the inequality

$$\|Tu_1 - Tu_2\|^2 \leq \|u_1 - u_2\|^2 (1 - 2\epsilon + \epsilon^2).$$

We choose ϵ such that T becomes a contraction: $\epsilon < \frac{2C_c}{C_b^2} \Rightarrow (1 - 2\epsilon + \epsilon^2) < 1$. By the Banach fixed point theorem we have existence and uniqueness of a solution. □

Remark 5. *The solution, u, to our variational problem depends on the data $b(\cdot)$. To see this we use the coercivity:*

$$\|u\|^2 \leq \frac{a(u, u)}{C_c} = \frac{b(u)}{C_c}.$$

Now, we have proved that (2.3) has a unique solution for suitable a and b. The variational form of Poisson equation (2.2) satisfies this:

Example 1 (Well posedness of variational form of Poisson equation)**.** *Let $a(u, v) = \int_\Omega \nabla u \cdot \nabla v \; dx$. Then we have that:*

- *$a(\cdot, \cdot)$ is* ***Coercive*** *with respect to $\|\cdot\|_{H_0^1}$:*

$$\begin{aligned} \|u\|_{H_0^1}^2 &= \|u\|_{L^2}^2 + \sum_{|\overline{\alpha}|=1} \left\|D^{\overline{\alpha}} u\right\|_{L^2}^2 \\ &= \|u\|_{L^2}^2 + a(u, u) \\ &\leq (C_\Omega + 1) a(u, u), \end{aligned}$$

  where we used the ***Poincaré inequality*** *in the last step.*

- *$a(\cdot, \cdot)$ is* ***Bounded*** *with respect to $\|\cdot\|_{H_0^1}$:*

$$|a(u, v)| \leq \left| \int_\Omega \nabla u \cdot \nabla v dx \right| \leq \int_\Omega |\nabla u \cdot \nabla v| dx$$

$$= \int_\Omega |\sum_{|\overline{\alpha}|=1} D^{\overline{\alpha}} u D^{\overline{\alpha}} v| dx = \sum_{|\overline{\alpha}|=1} \left\|D^{\overline{\alpha}} u D^{\overline{\alpha}} v\right\|_{L^1} \leq \sum_{|\overline{\alpha}|=1} \left\|D^{\overline{\alpha}} u\right\|_{L^2} \left\|D^{\overline{\alpha}} v\right\|_{L^2}$$

$$\leq \|u\|_{H_0^1} \|v\|_{H_0^1} ,$$

  where we used the ***Cauchy-Schwarz inequality*** *on the second line.*

- $b(\cdot)$ *is in the dual space of* H_0^1 *if* $f \in L^2(\Omega)$*:*

$$\begin{aligned} |b(v)| &= |\int_\Omega fvdx| \leq \|f\|_{L^2} \|v\|_{L^2} \\ \Rightarrow \|b\|_{H_0^{1'}} &= \sup \left\{ \frac{|b(v)|}{\|v\|} \right\} \leq \|f\|_{L^2} \end{aligned}$$

Hence, (2.2) *is well posed and we get a solution* $u \in H_0^1(\Omega)$*.*

2.1.4 Galerkin FEM

Now we want to discretize the variational equation (2.3), we do this by replacing the test space V by a finite dimensional subspace V_h. This is called the *Galerkin method.* The discretization now reads: Find $u_h \in V_h$ such that

$$a(u_h, v_h) = b(v_h) \tag{2.8}$$

for all v_h in V_h. Since $a(\cdot,\cdot)$ is bilinear and $b(\cdot)$ is linear, it is easy to see that if (2.8) holds for the basis functions of V_h, it holds for all elements in V_h. In the *finite element method* (FEM), the finite dimensional subspace is determined by the *triangulation.* In this thesis, we only consider problems in two spatial dimensions, so let $\Omega \subset \mathbb{R}^2$.

Definition 9 (two dimensional triangulation, page 56 of Knabner [11])**.** *Let* τ_h *be a partition* Ω *into closed triangles* K *including the boundary* $\partial\Omega$*, with the following properties:*

(T1) $\overline{\Omega} = \bigcup_{K \in \tau_h} K$.

(T2) *For* K*,* $K' \in \tau_h$*,* $K \neq K'$

$$int(K) \bigcap int(K') = \emptyset,$$

where $int(K)$ *denotes the interior of* K*.*

(T3) *If* $K \neq K'$*, but* $K \bigcap K' \neq \emptyset$*, then* $K \bigcap K'$ *is either a point or a common edge of* K *and* K'*.*

The above definition sets some rules on how we can divide our domain into triangles, often called elements. Now that we have a triangulation, we now define our finite dimensional subspace, V_h.

Definition 10 (Linear ansatz space)**.** *Let $\mathcal{P}_1(K)$ be the space of linear polynomials in two variables on $K \subset \mathbb{R}^2$, we define the ansatz space*

$$V_h := \left\{ u_h \in C(\overline{\Omega}) : u_{h|K} \in \mathcal{P}_1(K) \ \forall K \in \tau_h, u|_{\Gamma_D} = 0 \right\},$$

of piecewise linear functions on each K.

Remark 6. *Our local ansatz space $P_K = \{v|_K : v \in V_h\}$ is such that $P_K = \mathcal{P}_1(K) \subset H^1(K) \bigcap C(K)$. This together with **(T3)**, which ensures continuity between elements, makes V_h a **conformal finite element method**, i.e., $V_h \subset V = H_0^1(\Omega)$*

Remark 7 (Nodes)**.** *We will refer to the corners of the triangles in τ_h as nodes. For more advanced element types one can have nodes also on the edges or interiors of the triangles.*

Remark 8. *In general, finite elements are defined by an element $K(\in \tau_h)$, the local ansatz space P_K and degrees of freedom Σ_K. In all Lagrange finite element methods Σ_K corresponds the evaluation of functions in P_K at the nodes of the element.*

A choice of basis for V_h could then be the hat functions. Let ϕ_i be the basis function corresponding to the node x_i, it is defined by:

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, \quad \phi_i \in V_h.$$

There are no basis functions defined for the nodes at the Dirichlet boundary.

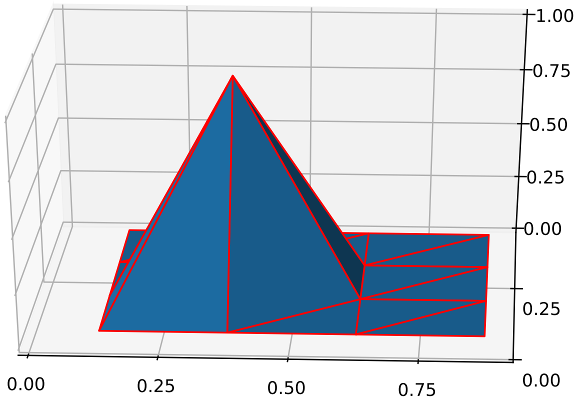


Figure 2.1: A hat function.

To see how the computation works in practice, we write the solution to our Galerkin problem as a linear combination of the basis functions of V_h, i.e., $u_h = \sum_i \hat{u}_i \phi_i$. The Galerkin problem (2.8) can be written

$$\text{Find } \hat{\boldsymbol{u}}_h = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} \in \mathbb{R}^n \text{ such that } \sum_{i=1}^{n} \hat{u}_i a(\phi_i, \phi_j) = b(\phi_j). \tag{2.9}$$

So we get a system of linear equations $\boldsymbol{A}\hat{\boldsymbol{u}}_h = \boldsymbol{b}$, where $\boldsymbol{A}_{i,j} = a(\phi_i, \phi_j)$ and $\boldsymbol{b}_j = b(\phi_j)$. The matrix, $\boldsymbol{A}$, has as many rows and columns as there are nodes (the Dirichlet nodes can be removed, depending on implementation). If we solve (2.2), our variational problem, and also matrix, will be symmetric. The matrix is then often called a *stiffness matrix*. These names originated from mechanics and structural analysis, where the solution represents displacement and the force function represents load. The stiffness matrix is also sparse, which is a very important property when designing algorithms to solve it.

With the setup described in this subsection, the degrees of freedom are the same as the dimension of V_h. If we in definition 10 instead had chosen a space of quadratic polynomials on each element, we had gained three degrees of freedom on each element. In this thesis we focus on linear finite elements because we do not gain anything from increasing regularity, as the solutions to problems in porous media flow are not expected to be very regular. Also, the finite volume methods we will discuss later, in particular the MPFA-L-Method, are not higher order methods.

2.1.5 Implementation

Here we explain the most important parts of the algorithm for discretizing elliptic PDE's with linear Lagrange finite elements. We consider the homogenous elliptic model problem (2.2) in two dimensions, with $\boldsymbol{K} = \boldsymbol{I}$ and zero Dirichlet boundary conditions. The procedure goes as follows:

1. Make a triangulation of the domain. This can be done in a number of different ways, see chapter 4 of Knabner [11]. If we have N nodes, our triangulation would be stored as a $N \times 2$ array of floats, being the coordinates of the nodes. And a $E \times 3$ array of ints being the elements, where each entry is the index of a coordinate in the coordinate matrix, E is the number of elements.

2. Allocate space for the $N \times N$ stiffness matrix $\boldsymbol{A}$ and the $N \times 1$ source vector $\boldsymbol{b}$.

3. Define the basis functions on a reference element, this is also called the shape functions, see figure 2.2 and (2.10). Also, compute the gradients of the shape functions.

$$\begin{aligned} N_1(x,y) &= 1-x-y \\ N_2(x,y) &= x \\ N_3(x,y) &= y \end{aligned} \tag{2.10}$$

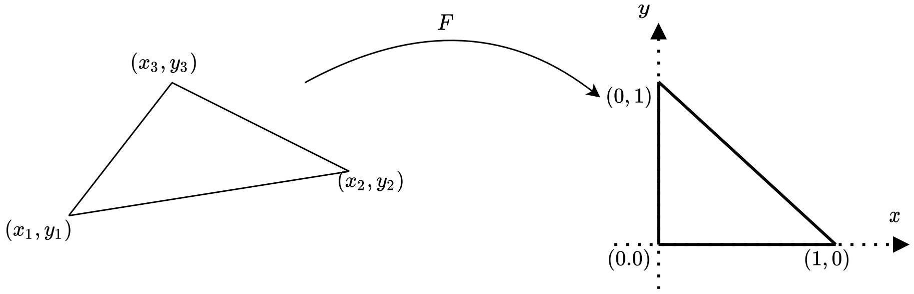


Figure 2.2: The map F from element K to the reference element $\hat{K}$.

4. Loop through the elements. For each element K compute the affine linear map that maps it to the reference element. That means we want to find $B \in \mathbb{R}^{2\times 2}$ and $d \in \mathbb{R}^2$ such that

$$\begin{aligned} F : K &\to \hat{K} \\ x &\mapsto Bx + d. \end{aligned}$$

To achieve this we set up a system of equations inspired by figure 2.2

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ d_1 & d_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.11}$$

So for each element we solve (2.11) for B and d, that means computing an inverse of a three by three matrix and a matrix product. Note that this only needs to be done once per element and could be done in a preprocessing step.

Now that we have T, we do the following on the element:

(a) Use the map and the shape functions to evaluate $a(\phi_i, \phi_j)|_K$ for $1 \leq i, j \leq 3$. Note that for $u : K \to \mathbb{R}$ we get by the chain rule:

$$\nabla_{\hat{x}}^T u(F^{-1}(\hat{x})) = \nabla_x^T u(F^{-1}(\hat{x})) \nabla_{\hat{x}}^T F^{-1}(\hat{x}) = \nabla_x^T u(F^{-1}(\hat{x})) B^{-1}.$$

This gives an expression for the derivative on an element expressed as a derivative in the reference element coordinate system:

$$\nabla_x u(F^{-1}(\hat{x})) = B^T \nabla_{\hat{x}} u(F^{-1}(\hat{x})).$$

Now we can compute the product of the gradients of the basis functions on an element:

$$\begin{aligned}
a(\phi_i, \phi_j)|_K &= \int_K (\nabla \phi_i)^T \nabla \phi_j dx \\
&= \int_{\hat{K}} (\nabla_x \phi_i(F^{-1}(\hat{x})))^T \nabla_x \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(J(F^{-1}))| d\hat{x} \\
&= \int_{\hat{K}} (B^T \nabla_{\hat{x}} \phi_i(F^{-1}(\hat{x})))^T B^T B \nabla_{\hat{x}} \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(B^{-1})| d\hat{x} \\
&= \int_{\hat{K}} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) |\mathrm{Det}(B^{-1})| d\hat{x} \\
&= \frac{1}{2} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) \frac{1}{|\mathrm{Det}(B)|}
\end{aligned} \tag{2.12}$$

So for each element we evaluate the last line of (2.12) for all (9) combinations of i and j on the element and add this to $\boldsymbol{A}_{i,j}$. This approach is called *element-based assembling*, and $\boldsymbol{A}_{i,j} = \sum_{K \in \mathcal{N}(i)} a(\phi_i, \phi_j)|_K$, where $\mathcal{N}(i)$ is the set of all elements that contain node i.

(b) In almost the same way we compute $b(\phi_i)|_K$ and add this to $\boldsymbol{b}_i$. As in (2.12), we compute the integral on the reference element:

$$\begin{aligned}
b(\phi_i)|_K &= \int_{\hat{K}} f(F^{-1}(\hat{x})) \phi_i(F^{-1}(\hat{x})) \frac{1}{\mathrm{Det}(B)} d\hat{x} \\
&= \int_{\hat{K}} \hat{f}(\hat{x}) N_i(F^{-1}(\hat{x})) \frac{1}{\mathrm{Det}(B)} d\hat{x} \\
&\approx \frac{1}{\mathrm{Det}(B)} \sum_k \omega_k \hat{f}(\hat{p}_k) N_i(\hat{p}_k)
\end{aligned}$$

Where $\hat{f} := f(F^{-1}(\hat{x}))$ and $\{(\omega_k, \hat{p}_k)\}_k$ defines a *quadrature rule*. This can be chosen in different ways, for higher order finite elements this may even affect the convergence behaviour. But for linear Lagrange elements, the trapezoidal rule works fine, i.e., using three points per element with appropriate weights.

5. Loop through the Dirichlet boundary nodes x_j at the boundary and set $\boldsymbol{A}_{j,i} = \delta_{ij}$, $b_j = 0$. This fixes the value of u at the Dirichlet boundary to zero.

Remark 9. *If we have inhomogeneous Dirichlet boundary conditions this is in practice done the same way as in the homogenous case, eliminating the degrees of freedom on the boundary. For Neumann conditions one has to evaluate integrals along the boundary as in* (2.5)*, using one-dimensional elements.*

2.1.6 Convergence

In this subsection, we review the most important concepts regarding the convergence of FEM. For a detailed discussion see [11]. The starting point of convergence rate estimates for the finite element method already described are **Cèa's lemma**:

Theorem 2.1.9 (Cèa's lemma)**.** *Let u solve the variational problem* (2.3) *and u_h solve the corresponding Galerkin approximation* (2.8)*, where the bilinear form $a(\cdot,\cdot)$ is bounded and coercive. Then we have:*

$$\|u - u_h\|_V \leq \frac{C_b}{C_c} min\{\|u - v_h\| : \; v_h \in V_h\}\,.$$

Proof. By the coercivity and linearity of $a(\cdot,\cdot)$ we have:

$$C_c \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h).$$

The last term equals zero, since both u and u_h solves the variational problem in V_h: $v_h - u_h = v \in V_h$ and $a(u - u_h, v) = a(u, v) - a(u_h, v) = b(v) - b(v) = 0$, this is called *Galerkin orthogality*. Then, we use the boundedness of $a(\cdot,\cdot)$:

$$C_c \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) \leq C_b \|u - u_h\|_V \|u - v_h\|_V\,.$$

We divide by C_c and $\|u - u_h\|_V$ and take the infimum over $v_h \in V_h$:

$$\|u - u_h\|_V \leq \frac{C_b}{C_c} \inf\{\|u - v_h\|_V : v_h \in V_h\}\,.$$

By (Cheney [6], page 64, theorem 2), as V_h is closed and convex subspace of a Hilbert space, there exist an unique element of V_h closest to u and minimum is attained. □

Hence the solution to Galerkin problem is the best in the subspace V_h up to a constant. We can therefore study convergence rate estimates for a suitable comparison element in V_h. In one dimension it is easy to picture what this comparison element might be, see figure 2.3. A direct proof with techniques from calculus is possible in this case.

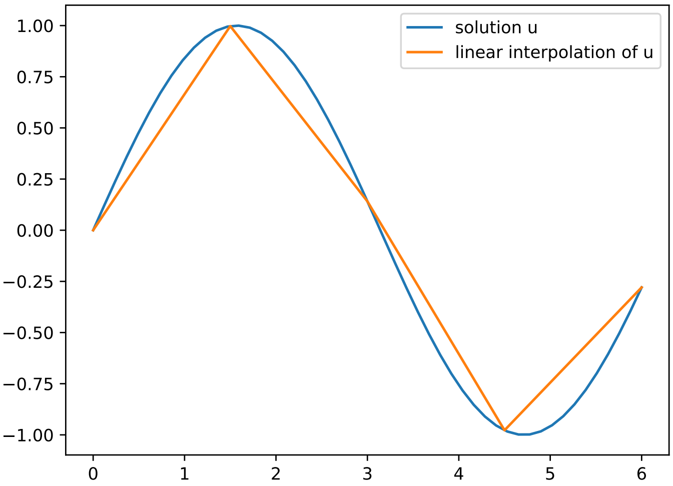


Figure 2.3: The unique linear interpolation of a function in one dimension.

The idea for more dimensions are the same, to be precise we define the interolation operator.

Definition 11 (Global interpolation operator)**.**

$$\begin{aligned} I_h : C(\overline{\Omega}) &\to V_h \\ v &\mapsto \sum_i v(n_i)\phi_i \end{aligned} \tag{2.13}$$

Where $\{n_i\}_i$ *are the nodes and* $\{\phi_i\}_i$ *the corresponding basis functions.*

Remark 10. *The global interpolator operator* (2.13) *maps from continuous functions, so we need to make sure our solution is continuous. By the Sobolev embedding theorem (Evans [7],page 286), we are okay if our space dimension is such that* $\Omega \subset \mathbb{R}^d$ *for* $d \leq 3$*, and* $u \in H^k(\Omega)$ *for* $k \geq 2$*.*

We remind the reader of the notation $\|\cdot\|_1 = \|\cdot\|_{H^1(\Omega)}$, and similarly for the semi-norms. In the setting of the model problem (2.2), we hope to reach an estimate

$$\|u - u_h\|_1 \leq C \|u - I_h(u)\|_1 \leq C^* h^k |u|_{k+1}, \tag{2.14}$$

where h is the maximum diameter of the elements in the triangulation, and k is the polynomial degree on the ansatz space. This bound is indeed attainable if we make sure the triangles in our triangulation have maximum angle less than π. In chapter 3.4 of Knabner [11], there is a detailed proof of (2.14).

Note that this means that our linear finite element method has a linear convergence in the $\|\cdot\|_1$ norm, if our variational problem admits a solution with sufficient regularity. We tie these observations together in a theorem:

Theorem 2.1.10 (energy norm estimate, Knabner [11] page 144)**.** *Consider a finite element discretization as described by* (2.9) *in* $\mathbb{R}^d$ *for* $d \leq 3$ *on a family of triangulations with an uniform upper bound on the maximal angle. Suppose we have a linear ansatz space as in definition 10, then*

$$\|u - u_h\|_1 \leq Ch|u|_2.$$

The above is called an energy norm estimate due to the equivalence of $\|\cdot\|_1^2$ and $a(\cdot,\cdot)$ in case of a symmetric bilinear form, in structural mechanics $a(\cdot,\cdot)$ corresponds to the potential energy.

Often we are happy with a convergence rate estimate in the $\|\cdot\|_0$ norm, which does not measure an error in the approximation of the derivative. We then expect a better convergence rate, as can be shown by the *Aubin-Nitsche technique.* This involves considering the dual problem of our variational problem (2.2): $a(v, u_f) = \langle f, v\rangle_0$, and assume some uniqueness and stability of the solution u_f of this.

Theorem 2.1.11 (L^2 estimate)**.** *Suppose the situation of theorem 2.1.10 and assume there exists a unique solution to the adjoint problem with* $|u_f| \leq C\|f\|_0$, *then there exist a constant* C^* *such that:*

$$\|u - u_h\|_0 \leq C^* h \|u - u_h\|_1 .$$

See [11] for a proof. When it comes to the assumption on the dual problem, this is satisfied for our elliptic model problem 2.1. If we put the last two theorems together we obtain quadratic convergence in the L^2 norm.

Remark 11. *In this chapter we have only discussed the convergence behaviour of the solution to the Galerkin problem* (2.8)*. In practice, one often only solves this approximately. For example the term* $b(v_h) = \int_\Omega f v_h \, dx$ *is impossible to evaluate exactly depending on* f*. We will later see error estimates with this taken into account.*

2.2 The Finite Volume Method

Finite volume methods are designed such that the conservation law we solve hold everywhere in the domain. Consider our elliptic model problem (2.1): Find u such that

$$\begin{aligned} -\nabla \cdot \boldsymbol{K}\nabla u(x) &= f(x) \quad x \in \Omega \\ u(x) &= 0 \quad x \in \partial\Omega. \end{aligned}$$

First we divide our domain Ω into convex quadrilaterals (control volumes, cells), $\{\Omega_i\}_i$. Then we integrate our equation over Ω_i and apply the divergence theorem:

$$\int_{\Omega_i} -\nabla \cdot \boldsymbol{K}\nabla u dx = -\int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} ds = \int_{\Omega_i} f dx. \tag{2.15}$$

The above equation equates the fluxes through the boundary of a control volume, with the source or sinks inside the control volume. The finite volume methods are discrete versions of this. Let $E_{i,j}$ be the edge between cells i and j, we approximate the flux through $E_{i,j}$, from cell i to cell j,

$$q_{E_{i,j}} = -\int_{E_{i,j}} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} ds$$

by a linear combination of u_i at neighbouring cell centers (geometric center of a cell)

$$q_{E_{i,j}} \approx \tilde{q}_{E_{i,j}} = \sum_k t_{i,j}^k u^k.$$

Where the *transmissibility* $t_{i,j}^k$ has the property $\sum_k t_{i,j}^k = 0$. Note that with this notation, we have $q_{E_{i,j}} = -q_{E_{j,i}}$.

We also approximate the integral on the right side, $\int_{\Omega_i} f dx$, with some quadrature rule. In porous media flow, the space discretization used, usually has a truncation error of at most second order. This is because the solution has low regularity due to heterogeneous permeability. The upshot is that we use the midpoint rule for evaluating the right hand side, as this also has a second order truncation error. Hence we evaluate f at the cell center and multiply by the area of Ω_i. We then end up with a system of equations

$$\sum_{j\in\mathcal{S}_i} \tilde{q}_{E_{i,j}} = |\Omega_i| f(x_i), \tag{2.16}$$

where $\mathcal{S}_i$ is the set of indices of neighbouring cells. The system of equations (2.16) ensures local mass conservation. It can also be written in matrix form as

$$\boldsymbol{A}^V \tilde{\boldsymbol{u}}_h = \boldsymbol{f}.$$

We will discuss different ways of constructing the transmissibility coefficients, as they result in very different discretizations.

The motivation for using finite volume methods for problems in porous media, for example Richards' equation, is that the flux appears explicitly in our discretization. If one, for example, wants to simulate the spread of some contaminant by groundwater flow, one can easily obtain a local mass conservative flux field using the finite volume method. This flux field can then be used in the desired transport equation.

We always make sure that our control volumes are inside our domain, with the boundary of our domain aligning the edges of our grid. To set our boundary conditions we thus need to specify the flux across the boundary. If we have Neumann boundary conditions, this is usually straightforward. One way of implementing no-flow boundary conditions is to make a strip of cells outside our domain with zero permeability.

Dirichlet boundary conditions are not always so natural for the equations we consider, that is, knowing the pressure or the saturation on a thin line in a two dimensional domain. This is reflected in the finite volume framework, where a common approach is to make ghost cells outside our domain where the potential is known, see figure 2.4. We can make these ghost cells as small as we like, and this approach is easy to implement. On the other hand, if we insist that we only know the potential at $\partial\Omega$, there exist techniques for determining the flux across the boundary as well. In section 3.1 demonstrate an approach for the MPFA-L-method where this is achieved.

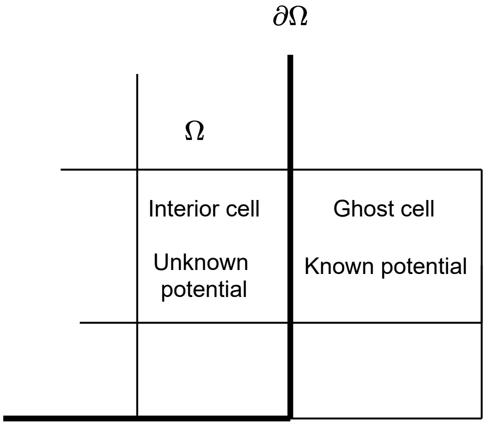


Figure 2.4: Ghost cell outside the boundary.

We will focus on discretizing the interior of the domain in the following sections.

2.2.1 Two-Point flux Approximation

The simplest way of constructing $t^k_{i,j}$ is also the most popular in the industry. As the name suggests, we only use the potential value at two points, x_0 and x_1, which are the cell center of two neighbouring cells, to compute the numerical flux $\tilde{q}_{E_{0,1}}$ between them.

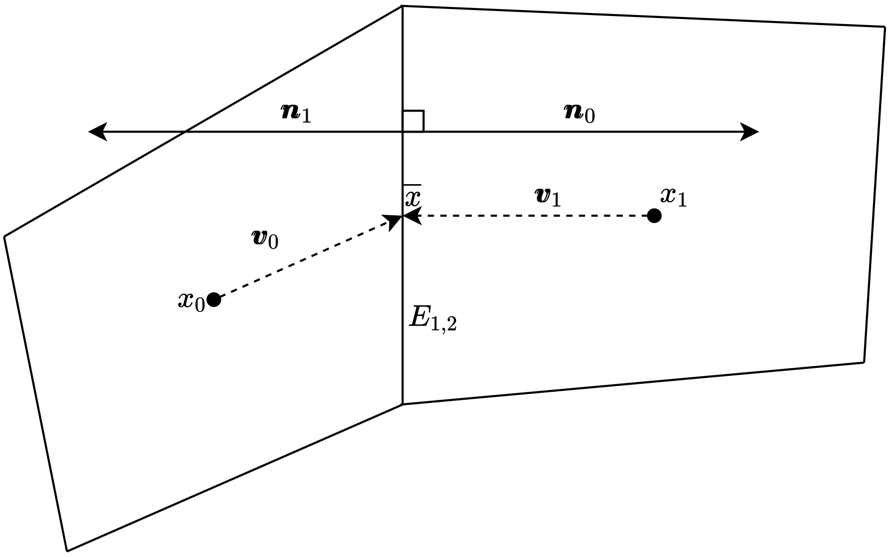


Figure 2.5: The two-point flux approximation (TPFA) setup.

Let $\boldsymbol{v}_1$ be the vector from cell center x_1 to the midpoint of the edge between the cells, $\overline{x}$. Then we approximate the flux out of cell 0 into cell 1 by:

$$\tilde{q}_{E_{0,1},0} = -\boldsymbol{n}_0^T \boldsymbol{K}_0 \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}(u(\overline{x}) - u(x_0)) \tag{2.17}$$

or as

$$\tilde{q}_{E_{0,1},1} = -\boldsymbol{n}_\mathbf{1}^T \boldsymbol{K}_1 \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}(u(x_1) - u(\overline{x})) \tag{2.18}$$

where $\boldsymbol{n}_i$ is the normal vector pointing out of cell i with length equal to $\partial\Omega$. In figure 2.5 we have $\boldsymbol{n}_1 = -\boldsymbol{n}_0$, and they are in general not aligned with $\boldsymbol{v}_1$ or $\boldsymbol{v}_0$. Because we require flux continuity we have that

$$\tilde{q}_{E_{0,1},0} = \tilde{q}_{E_{0,1},1} = t^0 u(x_0) + t^1 u(x_1)$$

where, as before, $t^0 + t^1 = 0 \Rightarrow t^0 = -t^1$, and the subscript on t is dropped for readability. We now have three equations and three unknowns, $u(\overline{x})$, t^0 and t^1. To simplify, we introduce the quantity $\mathcal{T}_i := \boldsymbol{n}_i^T \boldsymbol{K}_i \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|}$ to represent the cell *transmissivity*. So first we solve for $u(\overline{x})$:

$$\mathcal{T}_0(u(\overline{x}) - u(x_0)) = \mathcal{T}_1(u(x_1) - u(\overline{x})) \Rightarrow u(\overline{x}) = \frac{\mathcal{T}_0 u(x_0) + \mathcal{T}_1 u(x_1)}{\mathcal{T}_0 + \mathcal{T}_1}.$$

Next, we insert this into the expression for the numerical flux:

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} =& -\mathcal{T}_0(u(\overline{x}) - u(x_0)) \\
=& -\mathcal{T}_0\left(\frac{\mathcal{T}_0 u(x_0) + \mathcal{T}_1 u(x_1)}{\mathcal{T}_0 + \mathcal{T}_1} - u(x_0)\right) \\
=& -\mathcal{T}_0\left(\frac{\mathcal{T}_0 u(x_0) + \mathcal{T}_1 u(x_1) - u(x_0)\mathcal{T}_0 - u(x_0)\mathcal{T}_1}{\mathcal{T}_0 + \mathcal{T}_1}\right) \\
=& -\mathcal{T}_0\left(\frac{\mathcal{T}_1 u(x_1) - u(x_0)\mathcal{T}_1}{\mathcal{T}_0 + \mathcal{T}_1}\right) \\
=& \frac{u(x_0) - u(x_1)}{\frac{1}{\mathcal{T}_1} + \frac{1}{\mathcal{T}_0}}.
\end{aligned}$$

Now, we have solved the equations for the transmissivity coefficients:

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} &= t^0 u(x_0) + t^1 u(x_1) \\
\frac{u(x_0) - u(x_1)}{\frac{1}{\mathcal{T}_1} + \frac{1}{\mathcal{T}_0}} &= t^0 u(x_0) + t^1 u(x_1) \\
\Rightarrow t^0 &= \frac{1}{\frac{1}{\mathcal{T}_1} + \frac{1}{\mathcal{T}_0}}.
\end{aligned}$$

Hence, the transmissibility is the *harmonic mean* of the transmissivities. This kind of mean appears naturally when one wants to find the permeability of flow through layers of different permeability.

One way of looking at this discretization, is that we assume the potential to be a linear function of one variable, with its gradient pointing in the v_i direction between the cell center and the edge in figure 2.5. So for each edge, we have two linear functions on each side, which gives us four degrees of freedom. Two of them are used to respect the cell center potential values, the other two are used on potential and flux continuity across the edge. With these assumptions, expressions (2.17) and (2.18) are exact. And we only have to solve for the transmissibility coefficients.

Two-point flux approximation has the advantage of being fast to assemble and simple to code. It yields a pleasant five-point stencil for two dimensional problems. However, there is one big disadvantage with two-point flux approximation: Computing the flux with only two points is not consistent when the grid is not aligned with the principal directions of $\boldsymbol{K}$. If our grid is aligned with $\boldsymbol{K}$, we have that

$$\boldsymbol{n}_2 \cdot \boldsymbol{K}\boldsymbol{n}_1 = 0 \tag{2.19}$$

for a uniform parallelogram mesh with the normal vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$. We then call the grid **K-orthogonal**. In the setting of figure 2.5, our grid would not be

K-orthogonal as the control volumes are not parallelograms. All meshes with orthogonal control volumes are K-orthogonal if the permeability is isotropic. In figure 5.6 we observe the failed convergence of the TPFA-method for a parallelogram mesh. Analytical expressions of the error resulting from a grid that is not K-orthogonal, can be found in [25].

2.2.2 MPFA-O-Method

The O-method is a multi-point flux approximation method, these types of methods were developed to make control volume methods converge for grids that are not K-orthogonal. It is described in detail in [1], we only give a brief introduction. Consider the control volumes in 2.6.

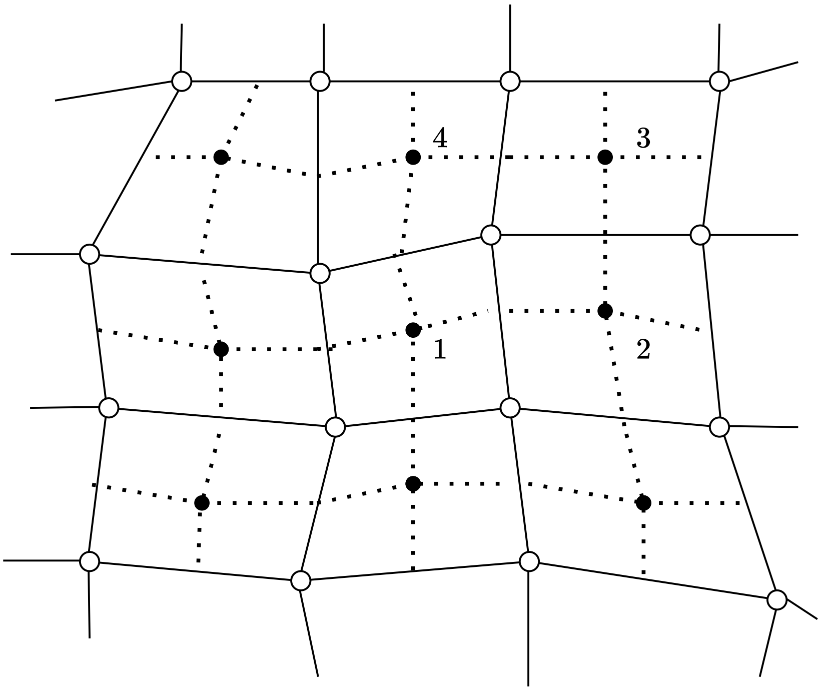


Figure 2.6: The solid lines are the edges of the control volumes, the dashed lines are the dual mesh connecting the cell centers, going through the midpoints of each edge. The solid circles are cell centers, the white circles are interaction points, and the volumes enclosed by the dotted lines are referred to as interaction regions.

For each interaction point, that means where four control volumes intersect, we consider an interaction region. This is the polygon drawn by the dual mesh around the interaction point.

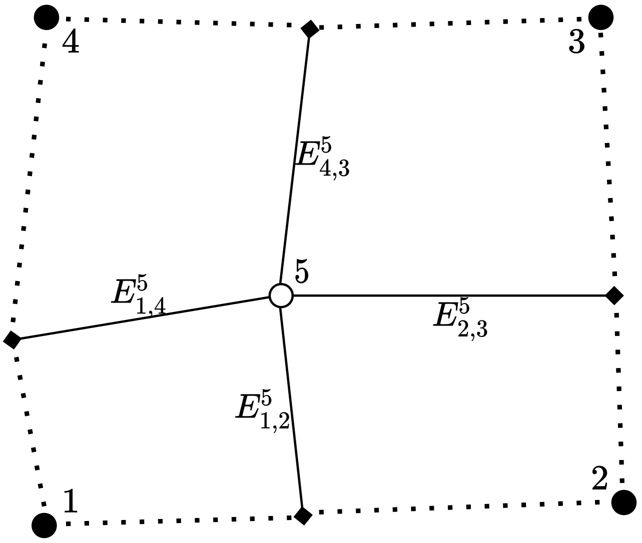


Figure 2.7: The four subcells in the interaction region corresponding to cells $1, 2, 3, 4$ and interaction point 5. Here, $\mathcal{R}_5 = \{1, 2, 3, 4\}$.

In each interaction region there are four half edges. Our goal is to obtain an expression

$$\tilde{q}_{E^n_{i,j}} = \sum_{k \in \mathcal{R}_n} t^{k,n}_{i,j} u^k \approx \int_{E_{i,j}} -\hat{\boldsymbol{n}}^T_j \boldsymbol{K} \nabla u \; ds \quad i, j \in \mathcal{R}_n \tag{2.20}$$

for the flux through each half edge $E^n_{i,j}$ in the interaction region corresponding to interaction point n (figure 2.7). Where $\mathcal{R}_n$ is the index set of the four cells neighbouring interaction point n.

We assume for now that the potential is linear in each of the four sub cells in the interaction region, figure 2.7. This gives $4 \cdot 3 = 12$ degrees of freedom. The linear potential must of course equal the cell center values of the potential in the cell centres, this removes four degrees of freedom. We also require flux continuity on the four half edges in the interaction region, this removes an additional four degrees of freedom. The last four degrees of freedom are spent on potential continuity on the midpoints of the edges.

The linear potential in each sub cell is now well defined given values at the cell center, provided the assumptions on flux and potential continuity. We can now use this to compute the four by four matrix of transmissibility coefficients for each of the four half edges. In the situation of figure 2.7 and equation (2.20) it would look like

$$\boldsymbol{T}^5 = \begin{bmatrix} t^{1,5}_{1,2} & t^{2,5}_{1,2} & t^{3,5}_{1,2} & t^{4,5}_{1,2} \\ t^{1,5}_{2,3} & t^{2,5}_{2,3} & t^{3,5}_{2,3} & t^{4,5}_{2,3} \\ t^{1,5}_{4,3} & t^{2,5}_{4,3} & t^{3,5}_{4,3} & t^{4,5}_{4,3} \\ t^{1,5}_{1,4} & t^{2,5}_{1,4} & t^{3,5}_{1,4} & t^{4,5}_{1,4} \end{bmatrix}, \tag{2.21}$$

where each row corresponds to the flux over some half edge, and each column corresponds to some cell center. Computing (2.21) involves inverting a four by four matrix with coefficients depending on the mesh and permeability, see [1] for details. Finally, we assemble the system of equations (2.16) with the transmissibility coefficients. Note that we write the flux over the jth edge of cell i, $\tilde{q}_{i,j}$ as the flux over the two half edges:

$$\sum_{j\in\mathcal{S}_i}(\tilde{q}_{\hat{E}^1_{i,j}} + \tilde{q}_{\hat{E}^2_{i,j}}) = |\Omega_i|f(x_i), \tag{2.22}$$

where the half edge fluxes $\hat{E}^1_{i,j}$ and $\hat{E}^1_{i,j}$ are related to some interaction point, i.e., $\hat{E}^1_{i,j} = E^n_{i,j}$ for some n, see (2.20). Hence, computing the transmissibility coefficients and assembling them into the discretization matrix, requires two different indexing systems.

Next, we see that the interaction regions of the two half edges sharing same edge overlaps, so that we get a six-point flux stencil. In other words, for each j in (2.22), the union of the two interaction regions used to compute $\tilde{q}_{E^1_{i,j}}$ and $\tilde{q}_{E^2_{i,j}}$ consists of six points. Taking the union of the four flux stencils connected to a cell, we observe that the O-method yields a nine-point potential stencil

$$\sum_{k\in\mathcal{M}_i}\hat{t}^k u^k = |\Omega_i|f(x_i),$$

where $\mathcal{M}_i$ is the set of nine indices corresponding to cell i and its eight neighbouring cells.

The O-method is consistent for non K-orthogonal grids, and reduces to the two-point flux approximation when the grid is K-orthogonal. This happens because the systems of equations to be solved for the transmissibility coefficients in each interaction region, becomes diagonal. This is because $\boldsymbol{n}^T\boldsymbol{K}\nabla u$ can be expressed as two points when u is a linear function given by three points which forms two K-orthogonal vectors.

In [15], Nordbotten and Keilegavlen describes a framework of MPFA methods where the O-method we introduced is a special case. They consider the problem of finding the four linear potential functions in each interaction region that minimizes the potential discontinuity across the edges. This should be minimized given the three constraints:

1. the potentials respect cell center potential values,

2. the flux models the constitute law, that is, consistency of Darcy's law,

3. flux continuity across the half edges.

Other methods, with the potential continuity at other places than the edge midpoint, have also been proposed. An early example is the nine-point finite difference scheme of J.L Yanosik and T.A McCracken [26], which has continuity at the midpoints of the half edges, this is also known as the $O(0.5)$ method.

With our implementation of MPFA-O-method, one needs for each interaction region to assemble four, four by four, matrices. Compute the inverse of one of them, and do two matrix multiplications and one subtraction. All of this could be done in parallel. However, for our implementation, it slows matrix assembly down a lot compared to two-point flux approximation. Another drawback of the O-method is the *monotonicity* properties: One can risk having positive entries off the main diagonal of the discretization matrix for difficult meshes. This may lead to unphysical oscillations in the solution and violation of the maximum principle of elliptic PDEs. For two-point flux approximation we avoid this issue altogether, as the signs of the five-point stencil always are one plus and four negatives. Even for the linear finite element method, this issue is avoided if one imposes some maximum angle condition, see [Knabner,[11]] page 175. When solving for example the Richards' equation, violating the maximum principle can lead to air bubbles being formed spontaneously in the saturated region. For a discussion on monotonicity see [14].

2.2.3 MPFA-L-Method

> The L-method is the Ferrari of discretization techniques for porous media flow problems, while conformal finite elements is the Volvo.
>
> *Professor Jan Martin Nordbotten*

Like the O-method, the L-method is also a multi-point flux approximation method. It was introduced in [2], where the authors demonstrate improved monotonicity properties with numerical experiments. This method is similar to the O-method, in that it goes through the half edges and uses information from the same interaction regions. But instead of using four points for the flux across each half edge, we use three, with two half edges between them.

As in the O-method, we assume linear potential in each cell, this gives us $3 \cdot 3 = 9$ degrees of freedom. Three are eliminated because we respect the cell center value of the potential, this leaves six degrees of freedom. We use two, one at each edge, for flux continuity. The last four are used for full potential continuity at the two edges.

We have two choices of flux stencil for each half edge, see figure 2.8. We compute the transmissibility coefficients for both, then we choose the one "best" aligned with the flow: Let t_1^i be the ith transmissibility coefficient of $\triangle_1$, then

$$\begin{aligned}&\text{if } |t_1^1| < |t_2^2|\\ &\text{choose } \triangle_1 \text{ else}\\ &\text{choose } \triangle_2.\end{aligned} \tag{2.23}$$

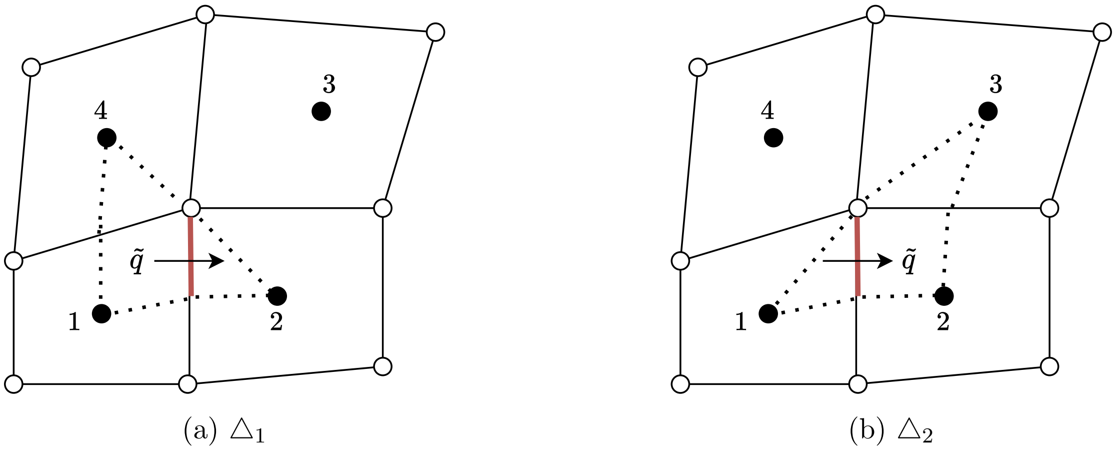


Figure 2.8: The two choices of which cell centers to use for computing the flux over the half edge in red. We call the interaction regions (the dotted lines) for L-triangles, as they related to three cell centers.

A cheap intuition behind (2.23) is that if $|t_1^1| < |t_2^2|$, it is more likely that $sgn(t_1^1) = sgn(t_1^4)$ than $sgn(t_2^2) = sgn(t_2^3)$ due to the fact that $\sum t^i = 0$. Choosing L-triangle as in (2.23) increases the chances that we get the same sign of t^i on the same side of the half edge, thus increasing the chance that we get a monotone discretization. See [4] for a more detailed geometric intuition of choosing L-triangle in the case of homogenous permeability.

To compute transmissibility coefficients in a given L-triangle, we use the assumptions on flux and potential continuity, to construct a linear system. The coefficients depends on mesh and permeability in the three cells. As with the O-method, we end up with a system assembled from the fluxes over the half edges:

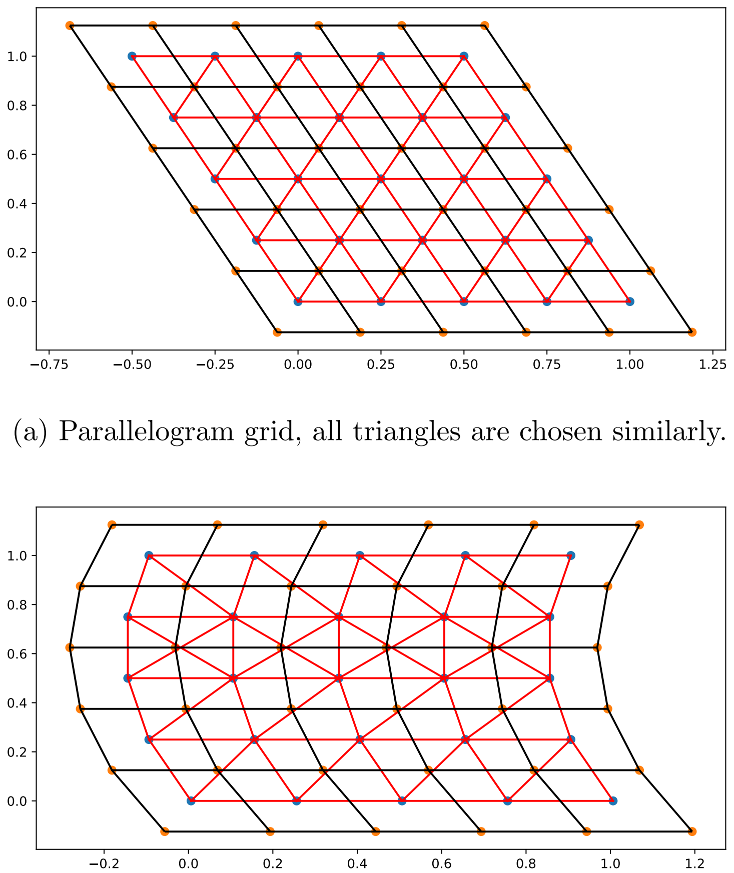
$$\sum_{j\in\mathcal{S}_i} (\tilde{q}_{\hat{E}^1_{i,j}} + \tilde{q}_{\hat{E}^2_{i,j}}) = |\Omega_i| f(x_i)$$

$$\sum_{j=1}^{4} (\sum_{k=1}^{3} t_{i,j}^{k,a} u^k + \sum_{k=1}^{3} t_{i,j}^{k,b} u^k) = |\Omega_i| f(x_i),$$

where a and b corresponds to the two interaction points sharing edge $E_{i,j}$. Thus, the flux stencil across each edge always consists of four points.

Remark 12. *In the L-method, we need to construct and solve a matrix equation twice for each half edge to compute the transmissibility coefficients, as there are always two choices. In contrast, the O-method only needs this done once for each interaction point, and its four half edges.*

In figure 2.9, we see the criterion in practice for a homogenous medium: In figure 2.9a all L-triangles are used by two half edges, and they are chosen in the same way throughout the domain. In figure 2.9b there are some triangles that overlap, this is due to the fact that some L-triangles are used by only one half edge.



(a) Parallelogram grid, all triangles are chosen similarly.

(b) Complicated grid, note that some of the L-triangles overlap.

Figure 2.9: Examples of L-triangles (in red) in a domain with homogenous permeability tensor.

The observation in figure 2.9a can be stated as a theorem:

Theorem 2.2.1 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[5])**.** *For homogeneous media and uniform parallelogram grids, the MPFA L-method has a seven-point cell stencil for the discretization of each interior cell, i.e., the discretization of each cell is a seven-point stencil including the center cell and the six closest potential cells, as shown in 2.9a.*

In the case of a parallelogram grid with heterogeneous permeability, it may also happen that one gets overlapping L-triangles. This is the case even if the permeability only changes as a scalar in the domain. In figure 2.10 the L-triangles are shown for a random, scalar permeability. Let $K_{m,n}$ be the permeability of the mth cell in y direction and nth cell in x direction. Then the random permeability used in figure 2.10 given by

$$K_{n,m} = (e^{\hat{x}} - 1)^2, \tag{2.24}$$

where $\hat{x}$ is a random sample drawn from a uniform distribution over $[0,1)$. We see that two of the L-triangles overlap. This is due to some combination of permeability at four neighbouring cells. Also note that the permeability is not so low that it causes numerical rounding errors, as $\min_{m,n} K_{m,n} = 0.0017$ in figure 2.10.

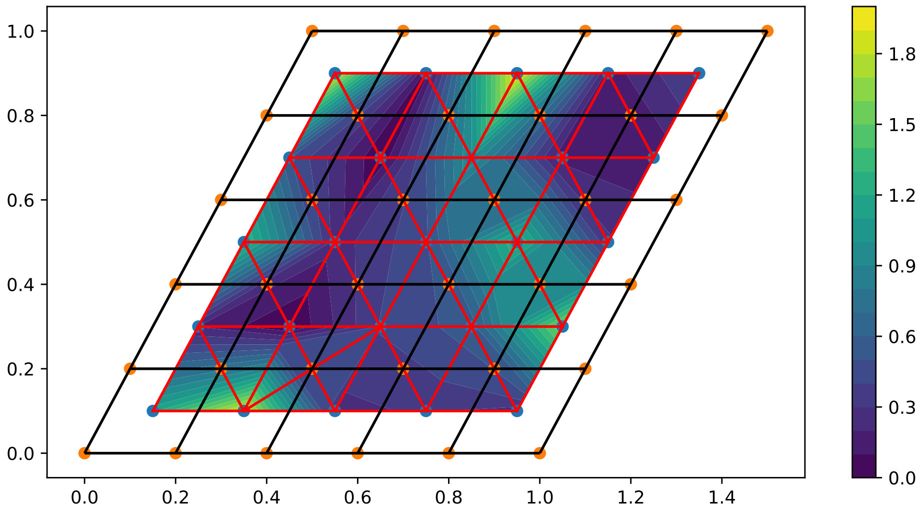


Figure 2.10: L-triangles on a random permeability.

For homogenous media the L-method becomes simpler. We continue with a useful theorem which we will use later:

Lemma 2.2.2 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[4])**.** *Assume that the permeability $\boldsymbol{K}$ is homogeneous on Ω, then the flux through each half edge e, computed by the L-method, can be written as*

$$\tilde{q}_e = -\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_e, \tag{2.25}$$

where $\boldsymbol{n}_e$ is the scaled normal vector to the half edge e, having the same length as e. u is a linear scalar field uniquely given by the potential values at the three cell centers chosen by the L-method.

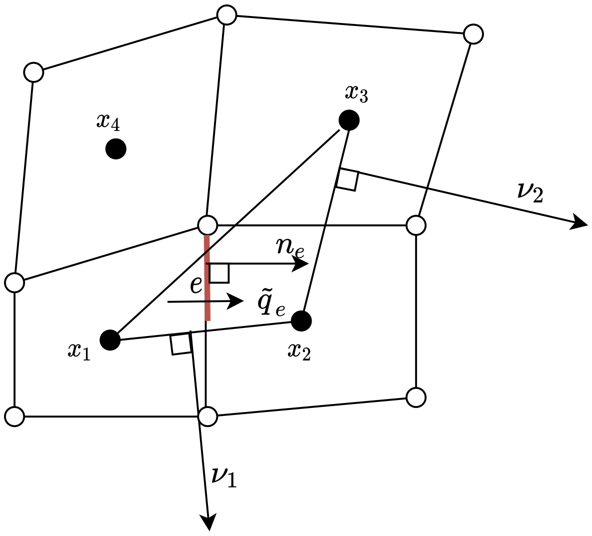


Figure 2.11: Simplified L-triangle, the original L-triangle i shown in figure 2.8b or 2.12. The vector ν_1 is perpendicular to the edge between x_1 and x_2, with the same length as the edge it is perpendicular to. Same for ν_2, with x_2 and x_3.

Moreover, the gradient ∇u, is given by:

$$\nabla u = -\frac{1}{2F}[(u_1 - u_2)\nu_2 + (u_3 - u_2)\nu_1], \tag{2.26}$$

where F is the area of the simplified L-triangle with corners x_1, x_2 and x_4, see figure 2.11. An expression like (2.26) *can be obtained for the other choice of L-triangle as well.*

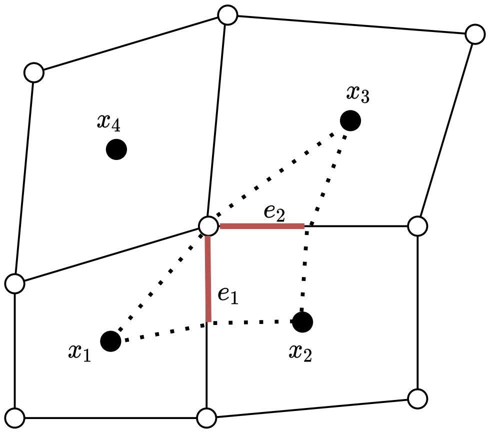


Figure 2.12: Original L-triangle with notations in proof.

Proof. It is enough to check that the jump $[\nabla u]$ is zero on e_1 and e_2 on the original L-triangle in figure 2.12. Let $\boldsymbol{t}_{e_1}$ and $\boldsymbol{n}_{e_1}$be the tangent and normal vector to e_1. Since we require potential continuity on each half edge, we get:

$$[\nabla u \cdot \boldsymbol{t}_{e_1}] = 0. \tag{2.27}$$

Using the fact that $\boldsymbol{K}$ is symmetric and homogenous, we obtain:

$$[\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}^T \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}\boldsymbol{n}_{e_1}] = 0. \tag{2.28}$$

Where we used flux continuity across each half edge in the last equality. Since $\boldsymbol{K}$ is positive definite, we have that $\boldsymbol{K}\boldsymbol{n}_{e_1}$ and $\boldsymbol{t}_{e_1}$ are independent, thus $[\nabla u] = 0$ on e_1. Same arguments holds for e_2. Hence ∇u is constant on the original L-triangle and the desired result follows. □

Remark 13. *The above lemma suggests that we can obtain the transmissibility coefficients without solving a system of equations for each half edge. This simplifies implementation, but it is only possible for homogenous media.*

To conclude; the L-method is the most sophisticated method in that it adjusts to the anisotropy locally. It has the best monotonicity properties, it is consistent for non K-orthogonal grids, but it is more complicated to implement than the O-method.

2.3 Time Discretization

We start by considering the most famous parabolic equation, namely the heat equation. Let $u = u(x, t)$, given appropriate boundary and initial conditions, find

u such that

$$\begin{cases} \partial_t u - \nabla \cdot \boldsymbol{K} \nabla u = f, & \text{in } \Omega \times (0,T], \\ u = 0, & \text{on } \partial\Gamma_D \times (0,T], \\ -\boldsymbol{K}\nabla u = g_N, & \text{on } \partial\Gamma_N \times (0,T), \\ u = u_0, & \text{on } \Omega \times \{t=0\}. \end{cases} \tag{2.29}$$

The well-posedness of (2.29) is discussed in chapter seven of [7], it requires a more detailed discussion of Sobolev spaces and Bochner spaces, i.e., spaces containing functions from the real numbers to some Sobolev space.

We expect low regularity in time, so there is not much to be gained by using a higher order discretization in time. Next, we need to decide whether we should use an explicit or an implicit scheme. The obvious choice is the implicit backward Euler, as it is stable for large time step sizes. This can be understood intuitively by considering the parabolic nature of the equation, the signals propagate through the domain instantaneously. A careful analysis of time discretizations of parabolic equations is done in ([11], chapter 7). There, it is shown that fully explicit schemes only are stable for time step sizes proportional to the square of the diameter of the space discretization, whereas fully implicit schemes are stable for all time step sizes.

Let $\{t_n\}_n$ be a sequence of $N+1$ uniformly distributed numbers from 0 to T and let $\tau = \frac{T}{N}$ be the time step size. Then we state the semi-discrete version of (2.29) by exchanging the time derivative by a difference quotient $\partial_t u = \frac{u^n - u^{n-1}}{\tau}$. We end up with: Given u^{n-1} and f^n, find u^n such that

$$\begin{aligned} u^n - \tau\nabla \cdot \boldsymbol{K}\nabla u^n &= \tau f^n + u^{n-1}, & x &\in \Omega, \\ u^n &= 0, & x &\in \partial\Gamma_D, \\ \boldsymbol{K}\nabla u &= g_N, & x &\in \partial\Gamma_N, \\ u^0 &= u_0, & x &\in \Omega. \end{aligned} \tag{2.30}$$

The above equation shows that this time discretization is implicit, i.e., we cannot solve (2.30) for u^n with simple algebraic manipulation, instead, we have an elliptic problem (2.30) for each time step. This has almost the same structure as the elliptic model problem (2.1) we solved in the previous chapters, the difference being the u^n term.

Finite element approach

We are now ready to fit this problem into our finite element framework from chapter 2. The variational formulation of (2.30) is achieved as before by multiplying by test functions in $H_0^1(\Omega)$: Given $u^{n-1} \in V$, $f^n \in V'$, find $u^n \in V$ such that

$$\langle u^n, v\rangle_0 + \tau \langle \boldsymbol{K}\nabla u^n, \nabla v\rangle_0 = \tau \langle f^n, v\rangle_0 + \langle u^{n-1}, v\rangle_0$$

for all v in V. If we exchange V with a finite dimensional subspace V_h, and write $u_h^n = \sum_{i=1}^{d} \hat{u}_i^n \phi_i$, as in the Galerkin FEM section 2.1.4, we end up with the system: Find $\hat{\boldsymbol{u}}^n \in \mathbb{R}^d$ such that

$$(\boldsymbol{B} + \tau \boldsymbol{A})\hat{\boldsymbol{u}}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}\hat{\boldsymbol{u}}^{n-1}, \tag{2.31}$$

where the *stiffness matrix*, $\boldsymbol{A}$, is as before, that is $\boldsymbol{A}_{j,i} := \int_\Omega (\nabla \phi_i)^T \nabla \phi_j dx$. The matrix $\boldsymbol{B}$ is often called the *mass matrix* and is defined as $\boldsymbol{B}_{j,i} := \int_\Omega \phi_i \phi_j dx$.

Finite volume approach

As before, we divide our domain Ω into d control volumes $\{\Omega_i\}_i$. Either, one can write the heat equation (2.29) in conservation form on each control volume

$$\partial_t \int_{\Omega_i} u \, dx - \int_{\partial \Omega_i} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} f \, dx, \tag{2.32}$$

and discretize the first term with backward Euler, or one can make sure the semi-discrete heat equation (2.29) holds for each control volume and use the divergence theorem. Both ways, we end up with

$$\int_{\Omega_i} u^n \, dx - \tau \int_{\partial \Omega_i} \boldsymbol{K} \nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \tau \int_{\Omega_i} f^n \, dx + \int_{\Omega_i} u^{n-1} \, dx,$$

if we, as discussed earlier, use the midpoint rule to evaluate the integrals, we get

$$\int_{\Omega_i} u^n(x_i) \, dx - \tau \int_{\partial \Omega_i} \boldsymbol{K} \nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \tau \int_{\Omega_i} f^n(x_i) \, dx + \int_{\Omega_i} u^{n-1}(x_i) \, dx.$$

As in the previous section we end up with a system of equations, where superscript V is just to distinct between FVM and FEM. Find $\tilde{\boldsymbol{u}} \in \mathbb{R}^d$, such that

$$(\boldsymbol{B}^V + \tau \boldsymbol{A}^V)\tilde{\boldsymbol{u}}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}^V \tilde{\boldsymbol{u}}^{n-1}$$

The matrix $\boldsymbol{A}^V$ is as in chapter 3, with the fluxes through the edges of cell i described by the jth row of $\boldsymbol{A}^V$. The matrix $\boldsymbol{B}^V$ is diagonal with the entry i being the volumes of the volume of cell i. That is, for two dimensional problems, the entries of $\boldsymbol{B}^V$ are the areas of the control volumes. In contrast, the $\boldsymbol{B}$ matrix from the finite element method, is in general not diagonal.

2.4 Linearization

We have seen that the heat equation leads to a sequence of linear systems. In the same way, we expect that our non-linear Richards' equation (1.9) leads to a system of non-linear equations. We start by discussing this in a general setting: Find $x \in U$ such that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \text{ where } f : U \subset \mathbb{R}^n \to \mathbb{R}^n. \tag{2.33}$$

The solution of (2.33) is called a *root*, it is almost always found using an iterative method.

A common iterative scheme to solve (2.33) is the *Newton's method.* Let $D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ be the Jacobian of $\boldsymbol{f}(\boldsymbol{x}_{j-1})$, then the newton iteration is given by:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} - D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{j-1}).$$

In one dimension a convergence proof i easily obtained by techniques from calculus, the following theorem is found in slightly more detail in (Cheney[6], chapter 3):

Theorem 2.4.1. *Let $f'' < 2$ with $f(\overline{x}) = 0$ and $f'(x) > \delta \ \forall x \in B_\epsilon(\overline{x})$, then the Newton method is locally quadratic convergent: For $x_0 \in B_\epsilon(\overline{x})$ we have*

$$|x_{j+1} - \overline{x}| \leq \frac{1}{\delta}|x_j - \overline{x}|^2 < |x_j - \overline{x}|.$$

Proof. Define $e_j := x_j - \overline{x}$. Then we have by Taylor expansion

$$0 = f(\overline{x}) = f(x_j - e_j) = f(x_j) - f'(x_j)e_j + \frac{f''(\psi)e_j^2}{2}. \tag{2.34}$$

For some ψ between x_j and $\overline{x}$. Further, we get by the definition of the newton method:

$$\begin{aligned} e_{j+1} = x_{j+1} - \overline{x} = x_j - \frac{f(x_j)}{f'(x_j)} - \overline{x} \\ = e_j - \frac{f(x_j)}{f'(x_j)} \\ = \frac{e_j f'(x_j) - f(x_j)}{f'(x_j)} \end{aligned} \tag{2.35}$$

By the Taylor expansion around x_j, (2.34), we get

$$f'(x_j) = \frac{f(x_j)}{e_j} + \frac{f''(\psi)e_j}{2}.$$

Inserting this into (2.35), we get the equality

$$e_{j+1} = \frac{e_j^2 f''(\psi)}{2f'(x_j)}.$$

The assumptions on f' and f'' combined with $|e_0| < \delta$ give us the estimate

$$|e_1| \leq \frac{2}{2\delta}|e_0|^2 < |e_0|$$

The above equation implies $x_1 \in B_\epsilon(\overline{x})$, and by induction we get

$$|e_{j+1}| < |e_j|,$$

and the quadratic convergence

$$|e_{j+1}| \leq \frac{1}{\delta}|e_j|^2$$

□

For a similar result in more dimensions see (Knabner [11], chapter 8). One apparent drawback of this method is that it is only locally convergent, i.e., one needs to start the iteration in a neighbourhood of the root where the Jacobian is well defined. In practice one often solves the system

$$D\boldsymbol{f}(\boldsymbol{x}_{j-1})\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}),$$

and then update the current iterate: $\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + \boldsymbol{\delta}_j$. Typically, the matrix $D\boldsymbol{f}(\boldsymbol{x}_{j-1})$, needs to be computed and assembled for every iteration. This may be computationally expensive. So Newton's method may be slow despite its quadratic convergence, if it even converges.

A simpler approach is to exchange the Jacobian with a diagonal matrix $L\boldsymbol{I}$ such that

$$L\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}). \tag{2.36}$$

This is called the *L-scheme*, and will be the method we use for linearization in this thesis. In one dimension it is easy to prove convergence:

Theorem 2.4.2. *Let $f \in C(\mathbb{R})$ and $L > \sup_{x\in\mathbb{R}} f'(x)$, then the L-scheme converges linearly for all $x_0 \in \mathbb{R}$.*

Proof. Define $e_j := e_j - \overline{x}$, then we get

$$e_{j+1} = x_j - \frac{f(x_j)}{L} - \overline{x} = e_j - \frac{f(x_j)}{L}.$$

We use the same trick as before with the Taylor expansion around the root

$$0 = f(\overline{x}) = f(x_j - e_j) = f(e_j) - f'(\psi)e_j \Rightarrow e_j = \frac{f(x_j)}{f'(\psi)}.$$

Using this and the assumption on L, we get the estimate:

$$|e_{j+1}| = \left|e_j\left(1 - \frac{f'(\psi)f(x_j)}{f(x_j)L}\right)\right| \leq |e_j|\left|1 - \frac{f'(\psi)}{L}\right| < |e_j|.$$

□

In practice we need to stop the linearization scheme at some point, and we decide on a **stopping criterion**. A common choice is, and the one we will use, is

$$|x_j - x_{j-1}| < TOL_1 + TOL_2|x_{j-1}|, \tag{2.37}$$

Where TOL_1 and TOL_2 is some constants chosen to be smaller than the error expected from spatial discretization, in our numerical experiments, we set both to be 10^{-8}. See (Storvik, [23]) for a discussion of the L-scheme and how to choose the L parameter in a smart way. One can also study other linearization approaches with different properties. In (List and Radu, [12]) the authors compare different iterative linearization methods for the Richards equation and propose a method that combines the Newton method and the L-scheme with desirable convergence rate and robustness.

Chapter 3

Convergence of the MPFA-L-Method

In this chapter we show equivalence between a modified MPFA-L method and a modified Lagrange finite element method, for linear time dependent problems discretized in time with backward Euler (2.30). That is, we prove equivalence between the two discretizations of the equation: Let $x \in \Omega \subset \mathbb{R}^2$, find $u(x)$ such that

$$\begin{cases} u - \nabla \cdot \boldsymbol{K} \nabla u = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Gamma_D \\ -\boldsymbol{K}\nabla u = g_N, & \text{on } \partial\Gamma_N, \end{cases} \tag{3.1}$$

where $\boldsymbol{K}$ is homogeneous, in addition to being symmetric positive definite. Once equivalence is obtained, we prove convergence for the finite element method using techniques from section 2.1.6.

After reading this chapter, the reader should be convinced that the finite element method covered in section 2.1 is almost the same as the L-method for homogeneous media. Moreover, that the L-method can be used as a locally mass conservative flux recovery algorithm on the modified finite element solution. See section 5.2 for a comparison of the MPFA-L method and normal linear Lagrange finite element method.

We saw in the section about the MPFA-L method that the interaction regions (L-triangles) may form a triangulation of our domain. With this observation in mind, modifications are made to both methods so that we obtain equivalence. This entire chapter is adapted from (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[5]), where, convergence is proved for the Poisson equation, i.e., without the first term u.

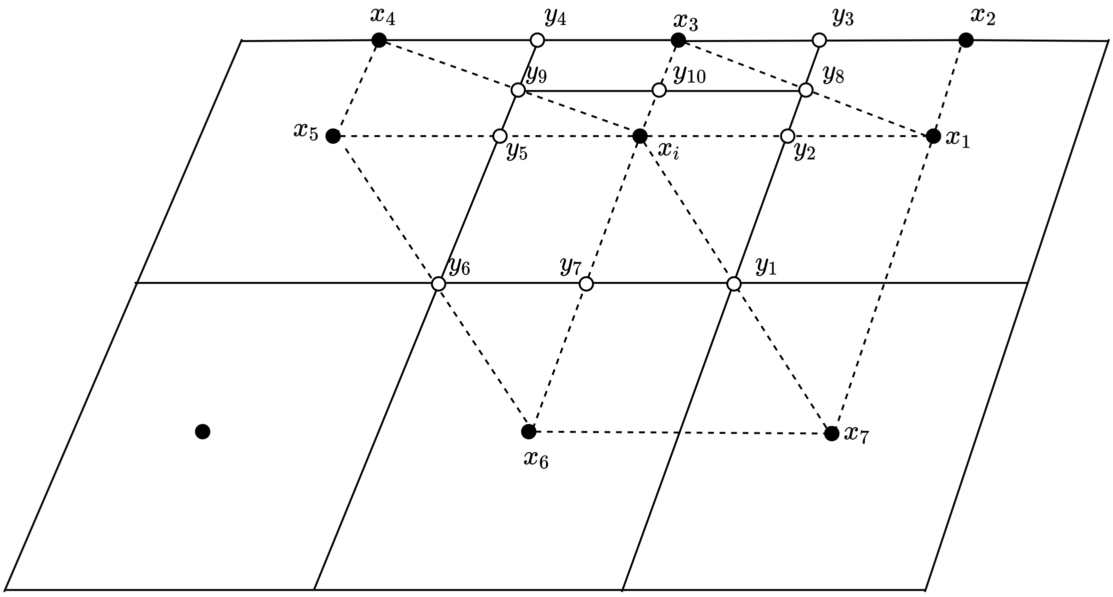


Figure 3.1: Control volumes in solid lines and interaction regions in dashed lines at the boundary.

3.1 Boundary Conditions for the MPFA-L-Method

First of all, we assume that we have a uniform parallelogram grid, as in 2.9a. As we saw in the previous chapter, one gets with the finite volume method the following relation for all control volumes Ω_i:

$$\int_{\Omega_i} u \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} f \, dx. \tag{3.2}$$

The MPFA-L method deals with the second term, approximating the constitutive law. The other two terms are common to all control volume methods solving time dependent problems or (3.1).

On the interior control volumes, we use the original MPFA-L method already covered. On the Neumann boundaries we need a modification. This is to be expected, as control volume methods handle flux at the boundary in a very simple way; specifying it the same way we deal with with the source term, adding it to the load vector. In finite element methods however, we have degrees of freedom on the boundary, one dimensional elements. We will also make a special treatment of the Dirichlet boundary, in a way that is equivalent to the finite element method. In [5] they claim that this is a very natural way of dealing with the Dirichlet boundary conditions, and a good practical alternative to other ways of enforcing Dirichlet boundaries in the MPFA-L method.

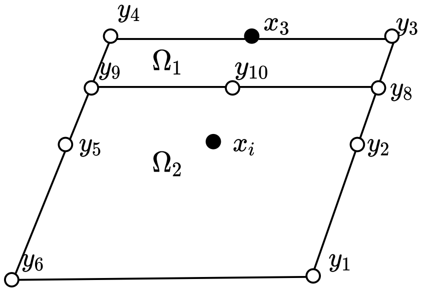


Figure 3.2: Control volume along top boundary.

Consider the control volume $y_1y_6y_4y_3$. For the **Neumann** boundary conditions, we split the control volume into two, $y_1y_6y_9y_8$ as Ω_2 and $y_8y_9y_4y_3$ as Ω_1, see figure 3.2 or 3.1. We therefore get one equation each for u_3 and u_i as the potential at x_3 and x_i. For the fluxes on Ω_2 we have six interaction triangles and a normal seven-point stencil. For the Ω_1 we compute the flux through $\overline{y_3y_8}$ using $\triangle x_1x_3x_2$, the flux through $\overline{y_8y_{10}}$ using $\triangle x_1x_ix_3$, for $\overline{y_{10}y_9}$ and $\overline{y_9y_4}$ the L triangle $\triangle x_ix_4x_3$ is used. Finally the Neumann boundary condition is used at the edge $\overline{y_4x_3}$ and $\overline{x_3y_3}$. We are not able to eliminate the unknown value at x_3 and it remains a degree of freedom, which makes sense if we want equivalence with finite element method.

In the case of **Dirichlet** boundary conditions, we compute the fluxes into $y_1y_6y_4y_3$ using seven L-triangles, as can be seen in figure 3.1. The flux over the edge $\overline{y_3y_1}$ are computed as the sum of the flux over $\overline{y_3y_8}$, $\overline{y_8y_2}$ and $\overline{y_2y_1}$ using the L-triangles $\triangle x_1x_3x_2$, $\triangle x_1x_ix_3$ and $\triangle x_1x_7x_i$ respectively. Similarly for the edge $\overline{y_6y_4}$. For $\overline{y_1y_6}$ we only use the two big L-triangles at the bottom, $\triangle x_ix_7x_6$ and $\triangle x_ix_6x_5$.

The flux over $\overline{y_4y_3}$, at the boundary, we compute by balancing with the other fluxes out of the small control volume Ω_1, see figure 3.3. Let $\tilde{q}_{\overline{y_iy_j}}$ be the flux through edge $\overline{y_iy_j}$, out of the volume Ω_1. Then we get the expression for the flux through the Dirichlet boundary:

$$\tilde{q}_{\overline{y_3y_4}} = -(\tilde{q}_{\overline{y_3y_8}} + \tilde{q}_{\overline{y_{10}y_8}} + \tilde{q}_{\overline{y_9y_{10}}} + \tilde{q}_{\overline{y_4y_9}}) + \int_{\Omega_1} f \; dx. \tag{3.3}$$

The fluxes on the right hand side of (3.3) are computed as for the Neumann case.

On the **corners**, special treatment is needed. Our modified MPFA-L method is modified to become equivalent to the finite element method here. This is done by splitting the corner control volume into four smaller cells, where mass conservation does not necessarily hold, see [5] for details.

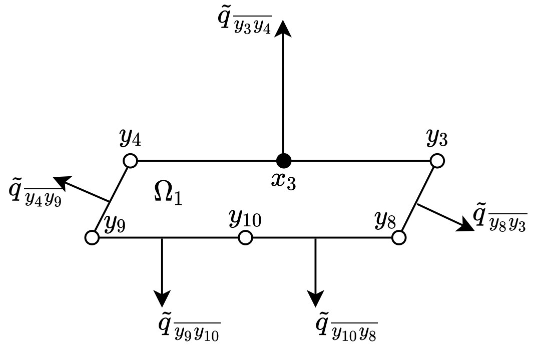

$\tilde{q}_{\overline{y_3y_4}}$
y_4
y_3
x_3
Ω_1
$\tilde{q}_{\overline{y_4y_9}}$
y_9
y_{10}
y_8
$\tilde{q}_{\overline{y_8y_3}}$
$\tilde{q}_{\overline{y_9y_{10}}}$
$\tilde{q}_{\overline{y_{10}y_8}}$

Figure 3.3: The fluxes on the Dirichlet boundary.

3.2 Modified Finite Element Method

In this section we introduce a finite element method for solving (3.1). By theorem 2.2.1 the L-triangles form a triangulation $\{\tau_h\}$, we will use linear Lagrange elements on this triangulation. The only modifications we need to make are to the mass matrix and the load vector, we let the stiffness matrix stay the same as before. That is, we do not touch the discretization of the constitutive law. We do want however, to define an interpolation operator such that the inner products that make up the mass matrix and load vector, become mass conservative in each control volume.

We need some notation so that we can distinguish between the cell centers in the interior, at cell centers along the boundary and the nodes at the boundary. In addition, corner cells need special treatment. Let $\mathcal{N}_h^*$ be a set of indices corresponding to all interior nodes of $\{\tau_h\}$, which are also the cell centers of the control volume mesh. This index set contains two disjoint sets $\mathcal{N}_h^* = \mathcal{N}_h^b \bigcup \mathcal{N}_h^i$, where superscript i denotes the cell centers of the interior cells and b the boundary cells. The index set $\mathcal{N}_h^b$ are further subdivided as we see in figure 3.4. The nodes at the boundary is indexed by the set $\mathcal{N}_h^N \bigcup \mathcal{N}_h^D$, where N and D represent Neumann and Dirichlet boundary nodes, these are further subdivided as illustrated in figure 3.4.

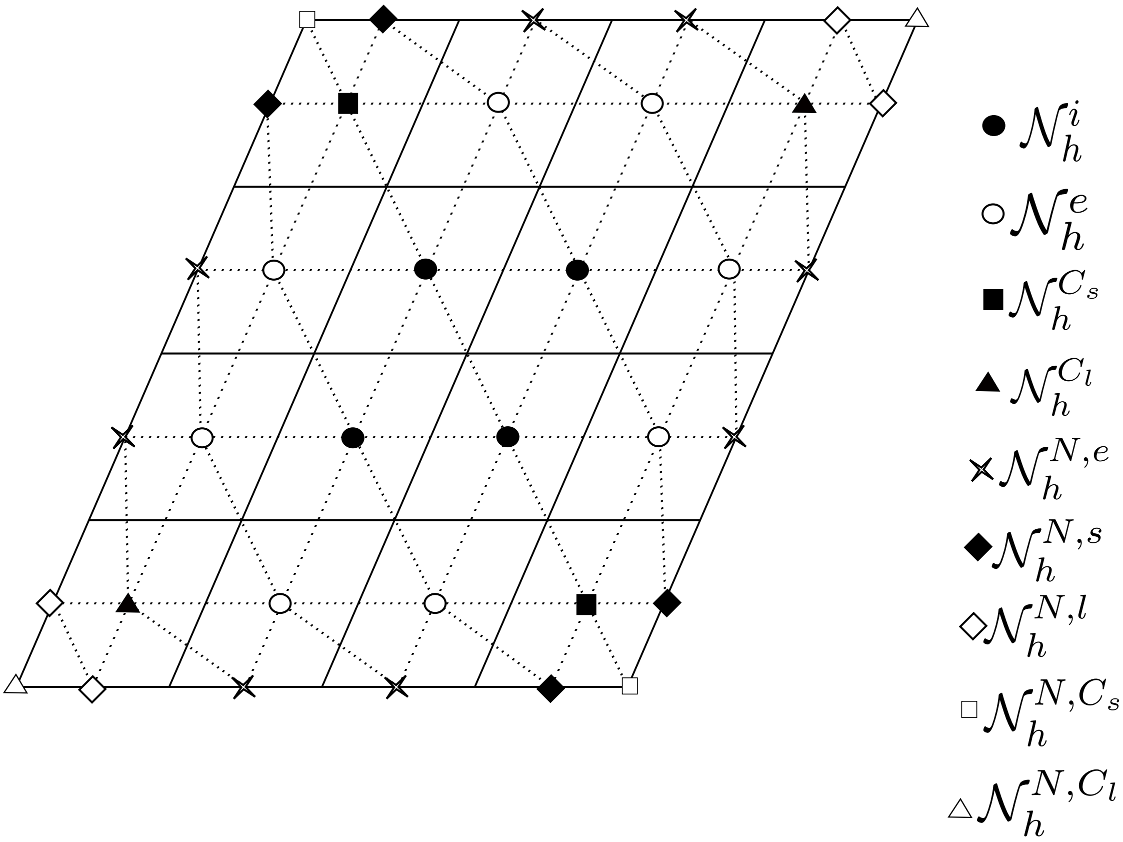


Figure 3.4: A paralellogram mesh with finite element triangles in dotted lines and control volumes in solid lines. In this case we have a pure Neumann problem. The special notation at the nodes is used in [5].

As before we denote by V_h the linear ansatz space as in definition 10:

$$V_h = \left\{u_h \in C(\overline{\Omega}) : u_{h|K} \in \mathcal{P}_1(K)\ \forall K \in \tau_h, u|_{\Gamma_D} = 0\right\}$$

similarly ϕ_i is the standard nodal basis function, where $i \in \mathcal{N}_h \setminus \mathcal{N}_h^D$. In addition to our global interpolation operator, definition 11, we define an operator that maps functions $v_h \in V_h$ to functions that are piecewise constant on the control volumes. This piecewise function are equal to v_h at the nodes of the triangulation. This is an example of *mass lumping*, see [3] for more examples.

Definition 12 (Piecewise global interpolator)**.** *Let $\hat{I}_h$ be an operator that maps from the test space to functions that are piecewise constant on control volumes.*

$$\hat{I}_h : V_h \to L^2(\Omega)$$

And

$$\hat{I}_h v_h = \sum_{i \in \mathcal{N}_h \backslash \mathcal{N}_h^d} v_h(x_i) \hat{I}_h \phi_i(x)$$

Where

$$\hat{I}_h \phi_i(x) = \begin{cases} 1 & \textit{if } x \in D_i \\ 0 & \textit{otherwise} \end{cases} \tag{3.4}$$

In interior cells, $i \in \mathcal{N}_h^i$, we have $D_i = \Omega_i$, i.e., the support of (3.4) *is the control volume corresponding to ϕ_i. If we are close or on the boundary the situation is more complicated:*

- *$i \in \mathcal{N}_h^e$: In this case the function vanishes for the quarter of the parallelogram closest to the boundary,i.e., $D_i = \Omega_2$ from figure 3.2*

- *$i \in \mathcal{N}_h^{N,e}$ In this case of the Neumann boundary node $\hat{I}_h \phi_i(x)$ vanishes outside the quarter of the control volume closest to the edge,i.e., $D_i = \Omega_1$ in figure 3.2*

- *On the corners there are special definitions, see [5].*

Let $\hat{I}_{\Gamma_N} = \hat{I}_h|_{\Gamma_N}$ be the trace of the piecewise interpolation operator on the Neumann boundary. The finite element method we end up with reads as follows: Find $u_h \in V_h$ such that

$$\left\langle \hat{I}_h u_h, \hat{I}_h v_h \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla v_h \rangle_{0,\Omega} = \left\langle f, \hat{I}_h v_h \right\rangle_{0,\Omega} + \left\langle g, \hat{I}_{\Gamma_N} v_h \right\rangle_{0,\Gamma_N}, \tag{3.5}$$

for all $v_h \in V_h$. The key takeaway here is the local support of the inner products, this will make the mass matrix diagonal.

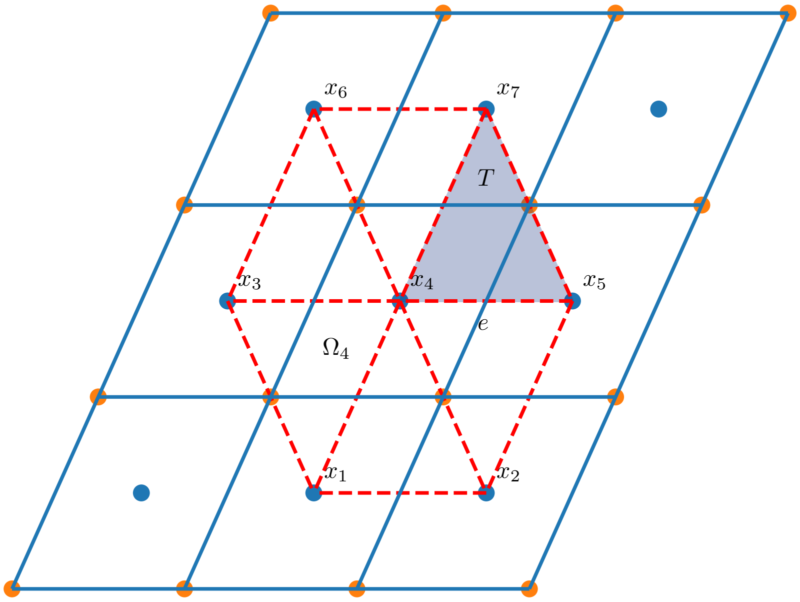


Figure 3.5: The support of ϕ_4, the coloured area corresponds to one triangle (element) in the support of ϕ_4.

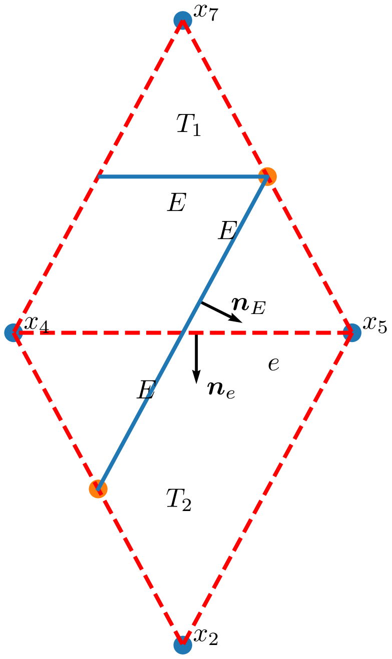


Figure 3.6: Notation in the proof

Now we can state the equivalence theorem:

Theorem 3.2.1. *The modified finite element method* (3.5) *and the modified MPFA-L method are equivalent on uniform parallelogram grid for the time discretized heat equation,i.e.,* (3.1)*, on homogeneous media.*

Proof. We do the proof in four steps:

1. First, we show the equivalence for the interior, so let Ω_i be an interior control volume and ϕ_i be the corresponding basis function evaluating to one at the centre of Ω_i, where $i \in \mathcal{N}_h^i$. We test (3.5) with $v_h = \phi_i$:

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_i \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla \phi_i \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_i \right\rangle_{0,\Omega}. \tag{3.6}$$

   Let $T \in \tau_h \bigcap \mathrm{supp}(\phi_i)$ be one of the elements in the triangulation that makes up the support of ϕ_i. $S = T \bigcap \Omega_i$ is a part of the control volume that lies in some element, and $E \subset S \bigcap \partial\Omega_i$ are the half edges of Ω_i. e are the interior edges of τ_h inside the support of ϕ_i, see fig 3.6 and 3.5. $\boldsymbol{n}_e$ is the unit normal on e with fixed and arbitrary orientation, and $\boldsymbol{n}_E$ is the unit normal on E pointing out of Ω_i. Let $T_{e,0}$ and $T_{e,1}$ be the two elements having e as a common edge, with the numbering corresponding to the orientation of $\boldsymbol{n}_e$. Since u_h and ϕ_i are piecewise linear and $\boldsymbol{K}$ is constant on each triangle T, we have:

$$\begin{aligned}
\langle \boldsymbol{K}\nabla u_h, \nabla\phi_i\rangle_0 &= \int_{\mathrm{supp}(\phi_i)} (\boldsymbol{K}\nabla u_h)^T \nabla\phi_i \; dx = \sum_{T\in\mathrm{supp}(\phi_i)} \int_T (\boldsymbol{K}\nabla u_h)^T \nabla\phi_i \; dx \\
&= \sum_{T\in\mathrm{supp}(\phi_i)} \left(\int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_i \; ds - \int_T \nabla\cdot\boldsymbol{K}\nabla u_h \phi_i \; dx \right) \\
&= \sum_{T\in\mathrm{supp}(\phi_i)} \int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_i \; ds \\
&= \sum_{e\in\mathrm{supp}(\phi_i)} \int_e \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}} \right) \phi_i \; ds \\
&= \sum_{e\in\mathrm{supp}(\phi_i)} \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}} \right) \frac{|e|}{2} \\
&= \sum_{S\in\mathrm{supp}(\phi)} \int_{\partial S} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n} \; ds - \sum_{E\in\partial\Omega_i} \int_E (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E \; ds \\
&= \sum_{S\in\mathrm{supp}(\phi)} \int_S \nabla\cdot\boldsymbol{K}\nabla u_h \; ds - \sum_{E\in\partial\Omega_i} \int_E (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E \; ds \\
&= -\sum_{E\in\partial\Omega_i} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E|.
\end{aligned} \tag{3.7}$$

Note that this last sum is a sum of integrals over the half edges of Ω_i. Further, we have that

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_i \right\rangle_0 = \int_\Omega \hat{I}_h u_h \hat{I}_h \phi_i \; dx = \int_{\Omega_i} u_h(x_i) \; dx \tag{3.8}$$

and

$$\left\langle f, \hat{I}_h \phi_i \right\rangle_0 = \int_{\Omega_i} f \; dx. \tag{3.9}$$

Combining equation (3.7), (3.8) and (3.9) we get that (3.6) is equivalent to:

$$\int_{\Omega_i} u_h(x_i) \; dx - \sum_{E\in\partial\Omega_i} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_i} f \; dx.$$

We know from theorem 2.2.2 that the flux over each half edge in the L-method is given uniquely by the potential values of the three cell centers in the L-triangle. Since the L-triangles and the elements are the same, ∇u_h corresponds to the gradient used in the L-method, see equation (2.25). Hence, if $\tilde{u}_h$ is the solution to (3.1) with the original L-method in the interior, then $\tilde{u}_h(x_i) = u_h(x_i)$ for $x_i \in \mathcal{N}_h^i$.

2. For a control volume bordering the **Neumann** boundary, first let $i \in \mathcal{N}_h^e$, we have:

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_i \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla \phi_i \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_i \right\rangle_{0,\Omega}. \tag{3.10}$$

With similar computations and reasoning as for (3.7) we get:

$$\langle \boldsymbol{K} \nabla u_h, \nabla \phi_i \rangle_{0,\Omega} = - \sum_{E \in \partial \Omega_{i,2}} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n}_E |E|,$$

where $\Omega_{i,2}$ is as Ω_2 in figure 3.2. As $\hat{I}_h$ is carefully defined close to the Neumann boundary, we get that (3.10) is equivalent to:

$$\int_{\Omega_{i,2}} u_h(x_i)\ dx - \sum_{E \in \partial \Omega_{i,2}} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_{i,2}} f(x_i)\ dx. \tag{3.11}$$

Next, let $j \in \mathcal{N}_h^{N,e}$,i.e., the index of a node on the boundary. Then we have

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla \phi_j \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \left\langle g, \hat{I}_{\Gamma_N} \phi_j \right\rangle_{0,\Gamma_N}. \tag{3.12}$$

Similarly as in (3.7) we have

$$\begin{aligned}
\langle \boldsymbol{K} \nabla u_h, \nabla \phi_j \rangle_0 &= \int_{\text{supp}(\phi_j)} (\boldsymbol{K} \nabla u_h)^T \nabla \phi_j\ dx = \sum_{T \in \text{supp}(\phi_j)} \int_T (\boldsymbol{K} \nabla u_h)^T \nabla \phi_j\ dx \\
&= \sum_{T \in \text{supp}(\phi_j)} \left(\int_{\partial T} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n} \phi_j\ ds - \int_T \nabla \cdot \boldsymbol{K} \nabla u_h \phi_j\ dx \right) \\
&= \sum_{T \in \text{supp}(\phi_j)} \int_{\partial T} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n} \phi_j\ ds.
\end{aligned} \tag{3.13}$$

But because $\phi_j \neq 0$ on $\text{supp}(\phi_j) \bigcap \Gamma_N$, we get

$$\begin{aligned}
\sum_{T \in \text{supp}(\phi_j)} \int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_j \; ds = & \sum_{e \in \text{supp}(\phi_j)} \int_e \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}} \right) \phi_j \; ds \\
& + \int_{\Gamma_N \bigcap \text{supp}(\phi_j)} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_j \; ds \\
& = \sum_{e \in \text{supp}(\phi_j)} \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}} \right) \frac{|e|}{2} \; ds \\
& + (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}|E_{\Gamma_N}| \; ds \\
& = - \sum_{E \in \partial\Omega_j \setminus \Gamma_N} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E|
\end{aligned} \tag{3.14}$$

Combining (3.13) and (3.14) and using the definition of $\hat{I}_h$, definition 11, we get that (3.12) is equivalent to:

$$\int_{\Omega_{j,1}} u_h(x_i) \; dx - \sum_{E \in \partial\Omega_{j,1}} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_{j,1}} f(x_i) \; dx. \tag{3.15}$$

Where $\Omega_{j,1}$ is as Ω_1 in figure 3.2. Now, (3.11) and (3.15) are exactly the L-method for the Neumann boundary, as described earlier, see figure 3.1.

3. For a control volume near the **Dirichlet** boundary, let first $i \in \mathcal{N}_h^e$,i.e., the cell center. Then, our modified finite element method

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \langle \boldsymbol{K}\nabla u_h, \nabla \phi_j \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_j \right\rangle_{0,\Omega}$$

is equivalent to

$$\int_{\Omega_{i,2}} u_h(x_i) \; dx - \sum_{E \in \partial\Omega_{i,2}} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_{i,2}} f \; dx,$$

with the same reasoning as in (3.7), (3.8) and (3.9). As $\Omega_i = \Omega_{i,1} \bigcup \Omega_{i,2}$ and $\Omega_{i,1} \bigcap \Omega_{i,2} = \emptyset$, see figure 3.2, we have:

$$- \sum_{E \in \Omega_i \setminus \Gamma_D} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| + \sum_{E \in \Omega_{i,1} \setminus \Gamma_D} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| + \int_{\Omega_{i,1}} f \; dx = \int_{\Omega_i} f \; dx.$$

We recognize the second and third terms in the above equation as the flux across the Dirchlet boundary in the modified L method, see (3.3).

4. See [5] for equivalence on the corner cells.

□

Remark 14. *There may be ways of extending the theorem above to inhomogeneous permeabilities, as long as the L-triangles form a triangulation. This includes using a special quadrature rule for integrating the bilinear form, or proving equivalence with a nonconforming finite element method with discontinuous basis functions.*

3.3 Convergence Rate Estimates

Our modified finite element method only approximates the bi-linear and linear form, and we need to take this into account when proving a convergence rate estimate. The following lemma is an extension of Cèa's lemma 2.1.9, it is useful for estimating the error when our bi-linear and linear form is not exact.

Lemma 3.3.1 (First Lemma of Strang, page 155 [11])**.** *Suppose there exists some $\alpha > 0$ such that for all $h > 0$ and $v_h \in V_h$*

$$\alpha \left\| v_h \right\|_1^2 \leq a_h(v_h, v_h)$$

and let a be continuous in $V \times V$. Then there exist some constant C independent of V_h such that

$$\begin{aligned} \left\| u - u_h \right\|_1 \leq C \left\{ \inf_{v_h \in V_h} \left\{ \left\| u - v_h \right\|_1 + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\left\| w_h \right\|_1} \right\} \right. \\ \left. + \sup_{w_h \in V_h} \frac{|l(w_h) - l_h(w_h)|}{\left\| w_h \right\|_1} \right\} \end{aligned} \tag{3.16}$$

From (3.5) we see that we have a bi-linear form

$$a_h(u_h, v_h) = \left\langle \hat{I}_h u_h, \hat{I}_h v_h \right\rangle_{0,\Omega} + \left\langle \boldsymbol{K} \nabla u_h, \nabla v_h \right\rangle_{0,\Omega}.$$

And the linear form:

$$b_h(v_h) = \left\langle F, \hat{I}_h v_h \right\rangle_{0,\Omega} + \left\langle g, \hat{I}_{\Gamma_N} v_h \right\rangle_{0.\Gamma_N}.$$

To apply the first Lemma of Strang 3.3.1, we first show that $a_h(\cdot, \cdot)$ is coercive. We write out the Sobolev norm

$$\left\| u_h \right\|_1^2 = \left\langle \nabla u_h, \nabla u_h \right\rangle_0 + \left\| u_h \right\|_0^2.$$

Using the Poincaré inequality on the second term:

$$\begin{aligned}\|u_h\|_1^2 &\leq \langle \nabla u_h, \nabla u_h \rangle_0 + C_\Omega \langle \nabla u_h, \nabla u_h \rangle_0 \\ &\leq \left(\frac{1 + C_\Omega}{\tau \|\boldsymbol{K}\|} \right) \tau \langle \boldsymbol{K} \nabla u_h, \nabla u_h \rangle_0 \\ &\leq \left(\frac{1 + C_\Omega}{\tau \|\boldsymbol{K}\|} \right) \left(\tau \langle \boldsymbol{K} \nabla u_h, \nabla u_h \rangle_0 + \left\langle \hat{I}_h u_h, \hat{I}_h u_h \right\rangle_0 \right) \\ &= \frac{1}{\alpha} a_h(u_h, u_h),\end{aligned}$$

we obtain coercivity with $\alpha = \tau \|\boldsymbol{K}\| / (1 + C_\Omega)$, where C_Ω is some constant depending on the domain and the boundary conditions.

Another important piece that must be in place for a convergence proof is the piecewise interpolation error:

Lemma 3.3.2. *For the previously defined piecewise global interpolator $\hat{I}_h$, definition 12, we have the estimate:*

$$\left\| \hat{I}_h u_h - u_h \right\|_{0,\Omega} \leq Ch |u_h|_{1,\Omega} \ \forall u_h \in V_h,$$

for some constant C independent of the mesh diameter.

Proof.

$$\begin{aligned}\left\| \hat{I}_h u_h - u_h \right\|_0^2 &= \sum_{i \in \mathcal{N}_h^*} \left\| \hat{I}_h u_h - u_h \right\|_{0,\Omega_i}^2 \\ &= \sum_{i \in \mathcal{N}_h^*} \int_{\Omega_i} (u_h(x_i) - u_h(x))^2 \ dx \\ &= \sum_{i \in \mathcal{N}_h^*} \int_{\Omega_i} h^2 \left(\frac{u_h(x_i) - u_h(x)}{h} \right)^2 \ dx \\ &\leq \sum_{i \in \mathcal{N}_h^*} \int_{\Omega_i} h^2 (\nabla u_h)^T \nabla u_h \ dx \\ &= Ch^2 |\nabla u_h|_1^2.\end{aligned}$$

□

We are now ready to state the H^1 error estimate for the modified finite element method and thus the MPFA-L method.

Theorem 3.3.3. *Let u solve* (3.1) *and u_h be the solution resulting from MPFA-L , then there exists a positive constant C independent of the mesh diameter, h, such that*

$$\|u - u_h\|_1 \leq Ch(\|u\|_2 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}). \tag{3.17}$$

Proof. The hypothesis in Strang's lemma 3.3.1 on continuity and coercivity are fulfilled. Let C be a generic positive constant. We start by controlling the second term on the right hand side in (3.16), the truncation error in the bi-linear form:

$$\begin{aligned}
&\sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_1} \\
&= \sup_{w_h \in V_h} \frac{| \langle v_h, w_h \rangle + \tau \langle \boldsymbol{K} \nabla v_h, \nabla w_h \rangle - \left\langle \hat{I}_h v_h, \hat{I}_h w_h \right\rangle - \tau \langle \boldsymbol{K} \nabla v_h, \nabla w_h \rangle |}{\|w_h\|_1} \\
&= \sup_{w_h \in V_h} \frac{| \langle v_h, w_h \rangle - \left\langle \hat{I}_h v_h, w_h \right\rangle + \left\langle \hat{I}_h v_h, w_h \right\rangle - \left\langle \hat{I}_h v_h, \hat{I}_h w_h \right\rangle |}{\|w_h\|_1} \\
&= \sup_{w_h \in V_h} \frac{| \left\langle \hat{I}_h v_h - v_h, w_h \right\rangle + \left\langle \hat{I}_h v_h, \hat{I}_h w_h - w_h \right\rangle |}{\|w_h\|_1}.
\end{aligned}$$

We see from the above computations, that the truncation error in the bi-linear form, only has a contribution from the *mass lumping*. By Cauchy Schwarz inequality and lemma 3.3.2 we get:

$$\begin{aligned}
&\leq \sup_{w_h \in V_h} \frac{Ch|v_h|_1 \|w_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 Ch|w_h|_1}{\|w_h\|_1} \\
&\leq \sup_{w_h \in V_h} \frac{Ch|v_h|_1 \|w_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 Ch|w_h|_1}{\|w_h\|_1} + \frac{Ch \|v_h\|_0 \|w_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 Ch \|w_h\|_0}{\|w_h\|_1} \\
&\leq Ch \left(\|v_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 \right).
\end{aligned}$$

The third term in (3.16), the linear form, can be controlled similarly:

$$\begin{aligned}
\sup_{w_h \in V_h} \frac{l(w_h) - l_h(w_h)}{\|w_h\|_1} &= \sup_{w_h \in V_h} \frac{\left\langle f, w_h - \hat{I}_h w_h \right\rangle_{0,\Omega} + \left\langle g, w_h - \hat{I}_{\Gamma_N} w_h \right\rangle_{0,\Gamma_N}}{\|w_h\|_1} \\
&\leq \sup_{w_h \in V_h} \frac{\|f\|_0 Ch \|w_h\|_1 + \|g\|_{\frac{1}{2},\Gamma_N} \left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N}}{\|w_h\|_1}.
\end{aligned} \tag{3.18}$$

Now, we want to bound $\left\| w - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N}$ by $\|w_h\|_1$. Let v_h be a piecewise constant function on the boundary in each Neumann boundary triangle. Then we have:

$$\begin{aligned}\left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{-\frac{1}{2},\Gamma_N} &= \sup_{0\neq v\in H^{\frac{1}{2}}(\Omega)} \frac{\left\langle w_h - \hat{I}_{\Gamma_N} w_h, v\right\rangle_{\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}} \\ &= \sup_{0\neq v\in H^{\frac{1}{2}}(\Omega)} \frac{\left\langle w_h - \hat{I}_{\Gamma_N} w_h, v - v_h\right\rangle_{\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}},\end{aligned}$$

as $\int_{\Gamma_N}(w_h - \hat{I}_{\Gamma_N} w_h)\ dx = 0$. Now, we can use Cauchy Schwarz inequality:

$$\left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{-\frac{1}{2},\Gamma_N} \leq \sup_{0\neq v\in H^{\frac{1}{2}}(\Omega)} \frac{\left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{0,\Gamma_N} \|v - v_h\|_{0,\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}}. \tag{3.19}$$

By the inequality

$$\|v - v_h\|_{0,\Gamma_N} \leq Ch^{\frac{1}{2}} \|v\|_{\frac{1}{2},\Gamma_N}, \tag{3.20}$$

we can bound the right hand side of (3.19):

$$\left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{-\frac{1}{2},\Gamma_N} \leq Ch^{\frac{1}{2}} \left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{0,\Gamma_N}.$$

Using (3.20) again, we get

$$\left\|w_h - \hat{I}_{\Gamma_N} w_h\right\|_{-\frac{1}{2},\Gamma_N} \leq Ch \|w_h\|_{\frac{1}{2},\Gamma_N} \leq Ch \|w_h\|_1.$$

Where the last inequality is due to the definition of the $H^{\frac{1}{2}}$ norm. Inserting this into (3.18), gives us a bound on the truncation error of our linear form:

$$\sup_{w_h\in V_h} \frac{l(w_h) - l_h(w_h)}{\|w_h\|_1} \leq Ch(\|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}).$$

Hence, from (3.16), we have the error estimate:

$$\|u - u_h\|_1 \leq \inf_{v_h\in V_h} \left\{\|u - v_h\|_1 + Ch\left(\|v_h\|_0 + \left\|\hat{I}_h v_h\right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}\right)\right\}. \tag{3.21}$$

If we let $v_h = I_h u \in V_h$, in (3.21), where $I_h : C(\Omega) \to V_h$ is the global interpolation operator, we get the inequality:

$$\|u - u_h\|_1 \leq \|u - I_h u\|_1 + Ch\left(\|I_h u\|_0 + \left\|\hat{I}_h I_h u\right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}\right). \tag{3.22}$$

As discussed earlier, in section 2.1.6 about convergence of finite element method, we have the estimate:

$$\|u - I_h u\|_1 \leq Ch|u|_2.$$

If we insert this into (3.22), we get:

$$\|u-u_h\|_1 \leq Ch\left(|u|_2 + \|I_h u\|_0 + \left\|\hat{I}_h I_h u\right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}\right).$$

As $I_h u, \hat{I}_h u \to u$ as $h \to 0$, we can control the first three terms inside the parenthesis by the H^2 norm of u, and we get the desired result:

$$\|u-u_h\|_1 \leq Ch\left(\|u\|_2 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}\right). \tag{3.23}$$

□

In this chapter we have introduced a way to handle Dirichlet and Neumann boundary conditions for the MPFA-L method, and showed convergence when applied to (3.1). The convergence was obtained with showing equivalence to a modified linear Lagrange finite element method.

Remark 15. *In our convergence rate estimate, we showed that* $\|u-u_h\|_1$ *decreases proportional to the mesh diameter, h. In [5], the authors show, using the Aubin-Nitsche technique, an estimate where* $\|u-u_h\|_0$ *decreases proportional to the square of the mesh diameter. We expect similar results can be shown here, as the equations are similar.*

Chapter 4

Convergence of the MPFA-L-Method for Richards' Equation

In this chapter, we use the results from chapter three to prove a convergence rate estimate of the backward Euler, L-scheme and MPFA-L-method applied to Richards' equation. We start by considering the Richards' equation without the gravity term in an isotropic medium: Find $\psi = \psi(x,t)$ such that

$$\begin{cases} \partial_t \theta(\psi) - \nabla \cdot (\kappa(\theta(\psi))\nabla\psi) = f, & \text{in } \Omega \times (0,T] \\ \psi = 0, & \text{on } \partial\Gamma_D \times (0,T] \\ -\kappa(\theta(\psi))\nabla\psi = g_N, & \text{on } \partial\Gamma_N \times (0,T) \\ \psi = u_0, & \text{on } \Omega \times \{t=0\} \end{cases} \tag{4.1}$$

This equation has a non linearity in the flux, $\boldsymbol{q} = -\kappa(\theta(\psi))\nabla\psi$, which makes it hard to apply our results, as they require a homogeneous medium. To remedy this we use the Kirchhoff transform

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\to \mathbb{R}^+ \\ \psi &\mapsto \int_0^\psi \kappa(\theta(\phi))\, d\phi = u. \end{aligned}$$

We assume that the functions $\theta(\cdot)$ and $\kappa(\cdot)$ are Lipschitz continuous, monotone increasing functions, see Van Genuchten [24] for an example. The Kirchhoff transform, $\mathcal{K}$, therefore has an inverse, $\mathcal{K}^{-1}$. We define

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

Further, assume that the hydraulic conductivity is bounded below and above

$$0 < \kappa_m \leq \kappa(\theta) \leq \kappa_M,$$

and that the water content is a Lipschitz continuous function of pressure

$$sup_\psi |\theta(\psi)| \leq L_\theta.$$

Then $b(\cdot)$ is also Lipschitz continuous

$$\sup_u |b'(u)| = \left|\theta'(\mathcal{K}^{-1}(u))\frac{1}{\mathcal{K}'(\mathcal{K}^{-1}(u))}\right| = \left|\theta'(\mathcal{K}^{-1}(u))\frac{1}{\kappa(\mathcal{K}^{-1}(u))}\right| \leq L_B.$$

We note that we in fact remove the non linearity in the constitutive law; by the chain rule, we get

$$\nabla u = \kappa(\theta(\psi))\nabla\psi.$$

We can write the Richards' equation (4.1) in the transformed variable u to get: Find $u = u(x,t)$ such that

$$\begin{cases} \partial_t b(u) - \nabla \cdot \nabla u = f, & \text{in } \Omega \times (0,T] \\ u = 0, & \text{on } \partial\Gamma_D \times (0,T] \\ -\nabla u = g_N, & \text{on } \partial\Gamma_N \times (0,T) \\ u = u_0, & \text{on } \Omega \times \{t = 0\}. \end{cases} \tag{4.2}$$

We start by discretizing (4.2) with the MPFA-L method, we divide our domain into d quadrilaterals (control volumes). Writing (2.32) in vector form we find $\tilde{u}_h \in \mathbb{R}^d$ such that:

$$\partial_t \boldsymbol{B}^V b(\tilde{u}_h) + \boldsymbol{A}^V \tilde{u}_h = \boldsymbol{q}^V.$$

We can then discretize in time using backward Euler. Given $\tilde{u}_h^{n-1}, \boldsymbol{q}^n \in \mathbb{R}^d$ we should then find $\tilde{u}_h^n \in \mathbb{R}^d$ such that:

$$\boldsymbol{B}^V b(\tilde{u}_h)^n + \tau \boldsymbol{A}^V \tilde{u}_h^n = \tau \boldsymbol{q}^{Vn} + \boldsymbol{B}^V b(\tilde{u}_h)^{n-1}. \tag{4.3}$$

Now, we to linearize (4.3) with the L-scheme. We see from (2.36) that the applying this linearization leads to the equation: Given $\tilde{u}_h^{n,j-1}, \tilde{u}_h^{n-1} \in \mathbb{R}^d$ find $\tilde{u}_h^{n,j} \in \mathbb{R}^d$ such that

$$L\boldsymbol{B}^V(\tilde{u}_h^{n,j} - \tilde{u}_h^{n,j-1}) + \tau \boldsymbol{A}\tilde{u}_h^{n,j} = -\boldsymbol{B}^V\theta(\tilde{u}_h^{n,j-1}) + \tau\boldsymbol{q}^{Vn} + \boldsymbol{B}^V\theta(\tilde{u}_h^{n-1}), \tag{4.4}$$

We set $\tilde{u}_h^{n,0} = \tilde{u}_h^{n-1}$ and solve the above equation until we reach the stopping criterion (2.37) each time step. See listing 5.2 for the code. Note that (4.4) is equivalent to a finite element discretization, this is what we will use to prove convergence.

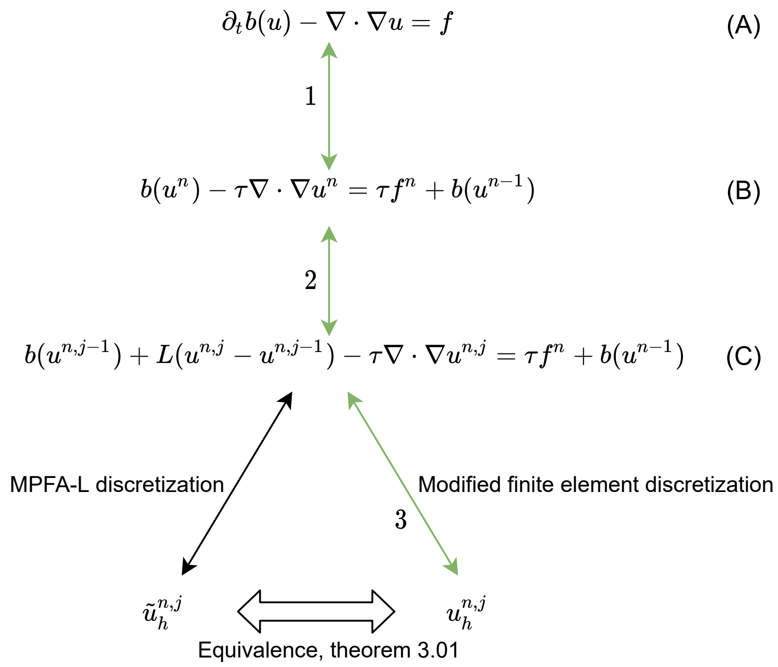

$\partial_t b(u) - \nabla \cdot \nabla u = f$
(A)
1
$b(u^n) - \tau \nabla \cdot \nabla u^n = \tau f^n + b(u^{n-1})$
(B)
2
$b(u^{n,j-1}) + L(u^{n,j} - u^{n,j-1}) - \tau \nabla \cdot \nabla u^{n,j} = \tau f^n + b(u^{n-1})$
(C)
MPFA-L discretization
Modified finite element discretization
3
$\tilde{u}_h^{n,j}$
$u_h^{n,j}$
Equivalence, theorem 3.01

Figure 4.1

To prove convergence we need some assumptions:

A-1 The domain $\Omega \subset \mathbb{R}^2$ is bounded and have a Lipschitz continuos boundary, i.e., it is locally the graph of a Lipschitz continuous function.

A-2 $b \in C^1$ is non-decreasing and Lipschitz continuous, with Lipschitz constant L_B.

A-3 $b(u_0)$ is essentially bounded in Ω and $u_0 \in L^2(\Omega)$.

Theorem 4.0.1. *Assume **A 1-3**, let h denote the diameter of the control volumes, τ denote the time step length and j the linearization iterate. Then for* (4.2) *discretized in space with MPFA-L-method on a parallelogram grid, backward Euler in time and L-scheme linearization, we have the estimate*

$$\left\|\tilde{u}_h^{n,j} - u(t^n)\right\|_0 \leq C\left(h + \tau + \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j\right). \tag{4.5}$$

From the above, it is clear that the approximated solution at t^n converges as $h, \tau \to 0$ and $j \to \infty$.

Proof. As the MPFA-L method and the modified finite element method are equivalent, we prove the convergence of our finite element solution, $u_h^{n,j}$. The proof will be done in three steps, see figure 4.1. We have by the triangle inequality

$$\left\|u(t_n) - u_h^{n,j}\right\|_0 \leq \left\|u(t_n) - u^n\right\|_0 + \left\|u^n - u^{n,j}\right\|_0 + \left\|u^{n,j} - u_h^{n,j}\right\|_0 ,$$

where j denotes the linearization iterate.

- The third term is the error of solving the elliptic problem (C) in figure 4.1, or

$$u^{n,j} - \frac{\tau}{L}\nabla \cdot \nabla u^{n,j} = \frac{Lu^{n,j-1} + \tau f^n - b(u^{n,j-1}) + b(u^{n-1})}{L}. \tag{4.6}$$

  By theorem 3.3.3 we have the error estimate

$$\left\|u_h^{n,j} - u^{n,j}\right\|_1 \leq C\left(\left\|u^{n,j}\right\|_2 + \left\|\frac{Lu^{n,j-1} + \tau f^n - b(u^{n,j-1}) + b(u^{n-1})}{L}\right\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}\right).$$

  The terms on the right hand side involving $u^{n,j}$ u^{n-1} and $u^{n,j-1}$ are bounded independently of the mesh. They are also bounded independently of τ, given sufficient regularity assumptions, see [16] theorem 1 for a stability estimate, hence $\left\|u_h^{n,j} - u^{n,j}\right\|_0 \leq C_3 h$.

- To bound the second term, $\|u^{n,j} - u^n\|_0$, we will use techniques found in (Radu, List, [12]). First, we subtract the variational form of (B) from the variational form of (C) in figure 4.1: For any $v \in H_0^1$ and $j > 1$

$$\left\langle b(u^{n,j-1}) - b(u^n), v\right\rangle_0 + \tau\left\langle \nabla(u^{n,j} - u^n), \nabla v\right\rangle_0 + L\left\langle u^{n,j} - u^{n,j-1}, v\right\rangle_0 = 0. \tag{4.7}$$

  Let $e^{n,j} := u^{n,j} - u^n$, then test (4.7) with $e^{n,j}$:

$$\left\langle b(u^{n,j-1}) - b(u^n), e^{n,j}\right\rangle_0 + \tau\left\|\nabla e^{n,j}\right\|_0 + L\left\langle u^{n,j} - u^{n,j-1}, e^{n,j}\right\rangle_0 = 0.$$

  Now, we use the relation $\langle b-a, b\rangle_0 = \frac{1}{2}\|b\|^2 + \frac{1}{2}\|b-a\|^2 - \frac{1}{2}\|a\|^2$ and some simple algebraic manipulation to obtain

$$\begin{aligned}\left\langle b(u^{n,j-1}) - b(u^n), e^{n,j-1}\right\rangle_0 + \tau\left\|\nabla e^{n,j}\right\|_0 + \frac{L}{2}\left\|e^{n,j}\right\|_0^2 + \frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|_0^2 \\ \leq \frac{L}{2}\left\|e^{n,j-1}\right\|_0^2 - \left\langle b(u^{n,j-1}) - b(u^n), e^{n,j} - e^{n,j-1}\right\rangle_0 .\end{aligned}$$

  Next, we use Cauchy Schwarz inequality on the first term, and Young's inequality on the last term

$$\begin{aligned}&\left\|b(u^{n,j-1}) - b(u^n)\right\|\left\|e^{n,j-1}\right\|_0 + \tau\left\|\nabla e^{n,j}\right\|_0 + \frac{L}{2}\left\|e^{n,j}\right\|_0^2 + \frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|_0^2 \\ &\leq \frac{L}{2}\left\|e^{n,j-1}\right\|_0^2 + \frac{1}{2L}\left\|b(u^{n,j-1}) - b(u^n)\right\|_0^2 + \frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|_0^2 .\end{aligned}$$

We cancel the last term on the right side against the last term on the left side. Since $b(\cdot)$ is Lipschitz continuous with $\|b(x) - b(y)\| \leq L_B \|x - y\|$, we have

$$\begin{aligned}&\frac{1}{L_B}\left\|b(u^{n,j-1}) - b(u^n)\right\|_0^2 + \tau\left\|\nabla e^{n,j}\right\|_0 + \frac{L}{2}\left\|e^{n,j}\right\|_0^2\\ &\leq \frac{L}{2}\left\|e^{n,j-1}\right\|_0^2 + \frac{1}{2L}\left\|b(u^{n,j-1}) - b(u^n)\right\|_0^2.\end{aligned}$$

Using the Poincaré inequality we obtain

$$\left(\frac{L}{2} + \frac{\tau}{C_\Omega}\right)\left\|e^{n,j}\right\|_0^2 \leq \frac{L}{2}\left\|e^{n,j-1}\right\|_0^2 + \left(\frac{1}{2L} - \frac{1}{L_B}\right)\left\|b(u^{n,j-1}) - b(u^n)\right\|_0^2.$$

Since $L_B < 2L$ we reach the convergence estimate

$$\left\|e^{n,j}\right\|_0^2 \leq \frac{L}{L + \frac{2\tau}{C_\Omega}}\left\|e^{n,j-1}\right\|_0^2.$$

We can use recursion to obtain the estimate

$$\left\|e^{n,j}\right\|_0 \leq \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j \left\|e^{n,1}\right\|_0. \tag{4.8}$$

To bound $\|e^{n,1}\|_0 = \|u^{n,1} - u^{n-1}\|_0$, we subtract the variational form of (B), in figure 4.1, at time step $n - 1$

$$\left\langle b(u^{n-1}), v\right\rangle_0 + \tau\left\langle \nabla u^{n-1}, \nabla v\right\rangle_0 = \tau\left\langle f^{n-1}, v\right\rangle_0 + \left\langle b(u^{n-2}), v\right\rangle_0,$$

from the variational form of (C), in figure 4.1, with $j = 1$

$$L\left\langle u^{n,1} - u^{n-1}, v\right\rangle_0 + \tau\left\langle \nabla u^{n,1}, \nabla v\right\rangle_0 = \tau\left\langle f^n, v\right\rangle_0.$$

Note that $u^{n,0} = u^{n-1}$, so two terms cancel in the above equation. Next, we test with $v = e^{n,1}$

$$\begin{aligned}L\left\langle u^{n,1} - u^{n-1}, u^{n,1} - u^{n-1}\right\rangle_0 + \tau\left\langle \nabla e^{n,1}, \nabla e^{n,1}\right\rangle_0 - \left\langle b(u^{n-1}), e^{n,1}\right\rangle_0\\ = \tau\left\langle f^n - f^{n-1}, e^{n,1}\right\rangle_0 - \left\langle b(u^{n-2}), e^{n,1}\right\rangle_0.\end{aligned}$$

Using the Poincarè inequality, we get

$$(L + C_\Omega)\left\|e^{n,1}\right\|_0^2 \leq \tau\left\langle_0 f^n - f^{n-1}, e^{n,1}\right\rangle_0 + \left\langle b(u^{n-1}) - b(u^{n-2}), e^{n,1}\right\rangle_0.$$

Next, we apply the Cauchy Schwarz inequality on the right hand side and divide by $\|e^{n,1}\|$ to obtain

$$(L + C_\Omega)\left\|e^{n,1}\right\|_0 \leq \tau\left\|f^n - f^{n-1}\right\| + \left\|b(u^{n-1}) - b(u^{n-2})\right\|_0.$$

The last term on the right hand side is bounded by using the Lipschitz continuity of $b(\cdot)$

$$\left\|e^{n,1}\right\|_0 \leq \frac{\tau \left\|f^n - f^{n-1}\right\|_0 + L_B \left\|u^{n-1} - u^{n-2}\right\|_0}{(L + C_\Omega)}.$$

By theorem 1 in [16], the terms involving u^{n-1} and u^{n-2} are bounded independent of τ. We can now rewrite (4.8) with a constant, C_2, that is independent of j and not increasing as τ decreases

$$\left\|e^{n,j}\right\|_0 \leq \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j C_2.$$

- The first term $\|u(t^n) - u^n\|$ can be bounded by the techniques used in (Radu, Pop and Knabner, [19]). We reach the estimate

$$\left\|u^n - u(t^n)\right\|_0 \leq C_1\tau.$$

Using all of the above, we get

$$\left\|\tilde{u}_h^{n,j} - u(t^n)\right\|_0 \leq C_3 h + C_2 \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j + C_1\tau \leq C\left(h + \tau + \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j\right),$$

where $C = C_1 + C_2 + C_3$. □

We have showed convergence for Richards' equation after Kirchhoff transform (4.2), discretized in space by MPFA-L method, in time by backward Euler and L-scheme for linearization. In section 5.3 we do numerical tests to confirm this.

Remark 16. *We expect that a better convergence rate estimate*

$$\left\|\tilde{u}_h^n - u(t^n)\right\|_0 \leq C\left(h^2 + \tau + \left(\sqrt{\frac{L}{L + \frac{2\tau}{C_\Omega}}}\right)^j\right),$$

with the square of the mesh diameter, is possible. This is because we use the $\|\cdot\|_1$ estimate for the spatial discretization, but according to remark 15, there exists a better $\|\cdot\|_0$ estimate we could use instead in the above proof.

Chapter 5

Numerical Results

In this chapter we do several numerical experiments with the algorithms covered in this thesis. We focus on convergence properties of the spatial discretizations, and confirm the error estimates we have discussed so far. We also briefly discuss the code used to do the experiments.

5.1 Computer Code

Most of the code used in this thesis can be found on https://github.com/trulsmoholt/masterthesis, and is written in python and numpy. It is primarily intended for educational purposes and for comparing convergence rates of different spatial approximation techniques. An example of a very simple use case can be seen in listing 5.1.

```
1
2  from discretization.mesh import Mesh
3  from discretization.FVML import compute_matrix,compute_vector
4  import numpy as np
5  import math
6
7  #Function to perturb mesh from unit square. Takes a 2d numpy vector and returns a 2d numpy vector. This particular choice makes a paralellogram mesh.
8  perturbation=lambda p: np.array([p[0],0.5*p[0]+p[1]])
9
10 #Number of grid points in x and y direction
11 nx=ny=10
12
13 mesh = Mesh(nx,ny,perturbation,ghostboundary=True)
14 source = lambda x,y:math.sin(y)*math.cos(x)
15 boundary_condition = lambda x,y:0
16 tensor = np.eye(2)
17 permeability = np.ones((mesh.num_unknowns,mesh.num_unknowns))
18
19 A = np.zeros((mesh.num_unknowns,mesh.num_unknowns))#stiffness matrix
20 f = np.zeros(mesh.num_unknowns)#load vector
21
22 compute_matrix(mesh,A,tensor,permeability)
23 compute_vector(mesh,f,source,boundary_condition)
24
25 u = np.linalg.solve(A,f)
26 mesh.plot_vector(u)
27
```

Listing 5.1: Solving simple Poisson equation.

The code is centred around the mesh class, which contains information about how the domain is discretized into quadrilaterals and its properties. This class also contains public functions to make plots of different kinds and compute errors. The spatial numerical methods implemented are: TPFA, MPFA-L, MPFA-O and linear Lagrange finite elements. They all have the same call signature, as in 5.1 line 22 and 23. The control volume methods also has the option of taking a matrix to store the flux stencils. One can use sparse matrices instead of dense numpy matrices in 5.1, as long as the indexing signature is the same as in numpy, for example scipy has a compatible sparse matrix library.

The code also has implementations of mass matrix and the gravitation term. Also included in the github are an example of how to solve Richards' equation using L-scheme linearization and backward Euler, see 5.2 for some of the code.

```
u_l = u.copy() #linearization/L-scheme iterate
u_t = u.copy() #timestep iterate
F = u.copy() #source vector
A = np.zeros((mesh.num_unknowns,mesh.num_unknowns)) #stiffness matrix
B = mass_matrix(mesh)

#time iteration
for t in time_partition[1:]:
  #empty source vector
  F.fill(0)
  #compute source vector
  compute_vector(mesh,F,lambda x,y: f(x,y,t),lambda x,y:u_exact(x,y,t))
  #L-scheme iteration
  while True:
    #compute the heterogeneous hydraulic conductivity, kappa
    conductivity = kappa(np.reshape(u_l, (mesh.cell_centers.shape[0],mesh
   .cell_centers.shape[1]),order='F'))
    A.fill(0)#empty the stiffness matrix
    compute_matrix(mesh, A, K,conductivity)#compute stiffnes matrix
    lhs = L*B+tau*A
    rhs = L*B@u_l + B@theta(u_t) - B@theta(u_l) + tau*F
    u = np.linalg.solve(lhs,rhs)
    #check if L-scheme linearization has acceptable error
    if np.linalg.norm(u-u_l)<=TOL+TOL*np.linalg.norm(u_l):
      #quit linearization and do another time step
      break
    else:
      #update linearization iterate
      u_l = u
  #update time step iterate
  u_t = u
  #update linearization iterate
  u_l = u
```

Listing 5.2: Linearization and time stepping of Richards' equation.

5.2 Elliptic Equation

The convergence tests in this section are similar to some of the tests done in chapter three of [2]. We consider the elliptic model problem (2.1), find $u = u(x)$ such that

$$\begin{cases} \nabla \cdot \boldsymbol{q} = f, & \text{in } \Omega \\ \boldsymbol{q} = -\boldsymbol{K}\nabla u, & \text{in } \Omega \\ u = u_{\partial\Omega}, & \text{on } \partial\Omega \end{cases} \tag{5.1}$$

We set the solution

$$u = cosh(\pi x)cos(\pi y), \tag{5.2}$$

set $\boldsymbol{K}$ to be the identity matrix and compute the Dirichlet boundary condition, $u_{\partial\Omega}$, and the source term f, from the equation above. The domain Ω will be variations of half the unit square through the rest of this section.

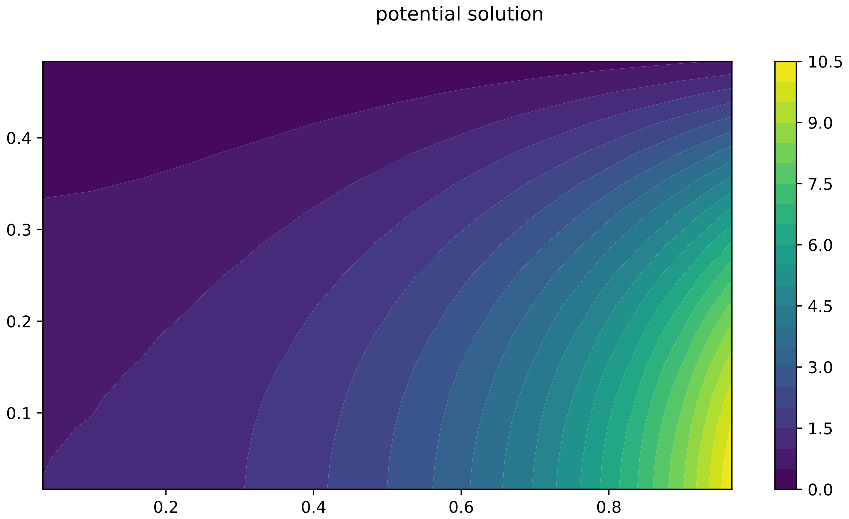


Figure 5.1: The solution (5.2) on half the unit square

As in [2] page 1340 we define the normalized discrete L_2 norms:

$$\|u - u_h\|_{0,h} := \left(\frac{1}{V} \sum_i V_i (u_{h,i} - u_i)^2 \right)^{\frac{1}{2}}$$

$$\|q - q_h\|_{0,h} := \left(\frac{1}{Q} \sum_E Q_E (q_{h,E} - q_E)^2 \right)^{\frac{1}{2}},$$

where $q_E := -\hat{\boldsymbol{n}} \cdot \boldsymbol{q}$ is the normal flow density over the edge E, with $\hat{\boldsymbol{n}}$ being unit normal to the edge and $\boldsymbol{q}$ evaluated at the midpoint of the edge. $q_{h,E}$ is the normal discrete flux over E defined similarly, with $\boldsymbol{q}_h$ being the discrete normal flow density. For a finite volume method, $q_{h,E}$ would be the flux across some edge divided by the edge length. For the finite element method, we use the MPFA-L flux stencil to recover the flux in the experiments where it is present. Let $u_{h,i}$ denote the discrete potential at cell i, and u_i is the potential evaluated at cell i. For the finite element method, this would be the function value at the grid points/nodes. Q_E is the volume associated with edge E, i.e., the sum of the two volumes sharing edge E. $V = \sum_i V_i$ and $Q = \sum_a Q_E$.

We also define the discrete max norms

$$\|u - u_h\|_{h,\infty} := \max_i |u_{h,i} - u_i|$$

and

$$\|q - q_h\|_{h,\infty} := \max_E |q_{h,E} - q_E|$$

In the figures (5.12-5.15) in this section we see the potential and normal flow density error in the $\|\cdot\|_{0,h}$ and $\|\cdot\|_{0,\infty}$ norms plotted against mesh diameter for MPFA-L, MPFA-O, TPFA and the linear Lagrange finite element method without mass lumping. More specifically, the y-axis is the log_2 of the error, and the x-axis is $log_2 n$, where n is the number of points in the x direction, and thus proportional with the inverse of the mesh diameter. The slope of the graph in the plots are the convergence rate.

We use ghost cell Dirichlet boundary conditions for the finite volume methods. For the finite element method, we fix the potential at the cell centers of the ghost cells, see figure 5.2. Therefore, the boundary modifications introduced in chapter 3 are not applied in these tests. We instead treat our entire domain as the interior.

5.2.1 Setup 1: Orthogonal Grid

We let Ω be half the unit square, $[0, 1] \times [0, \frac{1}{2}]$, and consider refinements of the grid in figure 5.2. In figures 5.3 and 5.3 we see that all the methods converge with the same quadratic rate. This fits well with the fact that all the methods covered in this thesis are equivalent to the TPFA method for uniform grids.

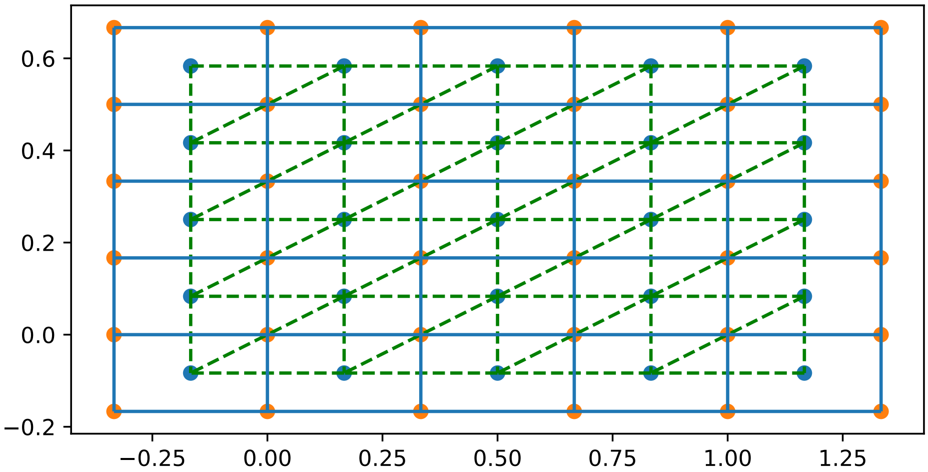


Figure 5.2: Unifrom rectangular mesh on half the unit square. The triangles (dotted green lines) are used for the finite element solution and are spanned between the cell centers of the control volumes (solid blue line). The ghost cell boundary is included, so this mesh has nine degrees of freedom, i.e., the interior cells.

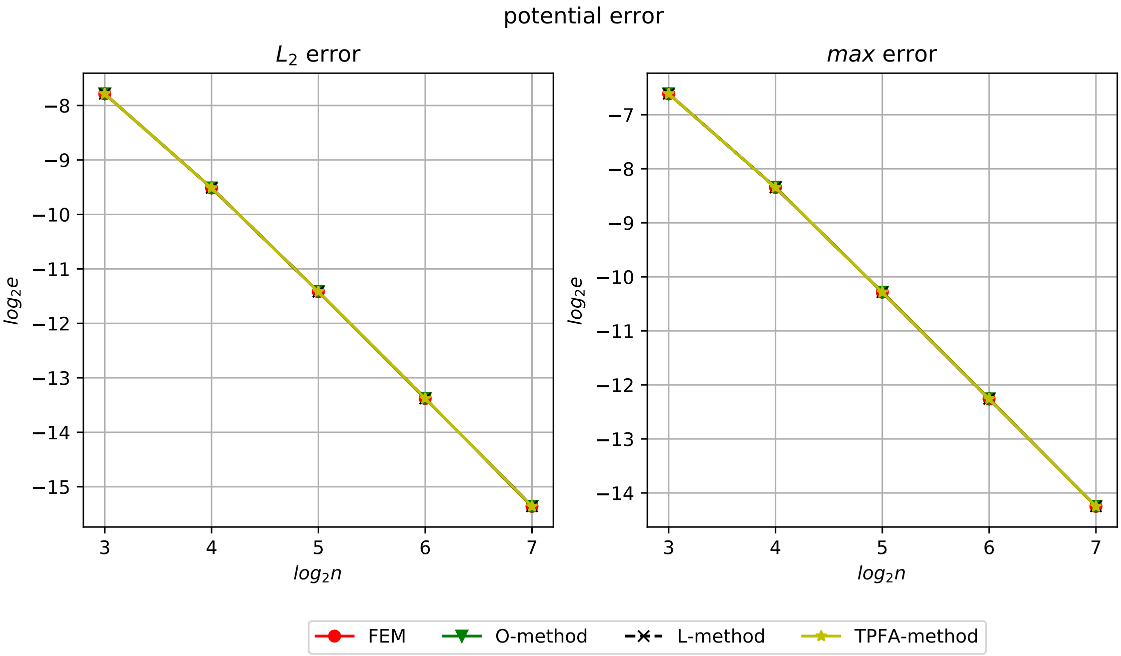


Figure 5.3: Potential error on refinements of the uniform rectangular mesh 5.2. Left: $e = \|u - u_h\|_{h,0}$, right: $e = \|u - u_h\|_{h,\infty}$

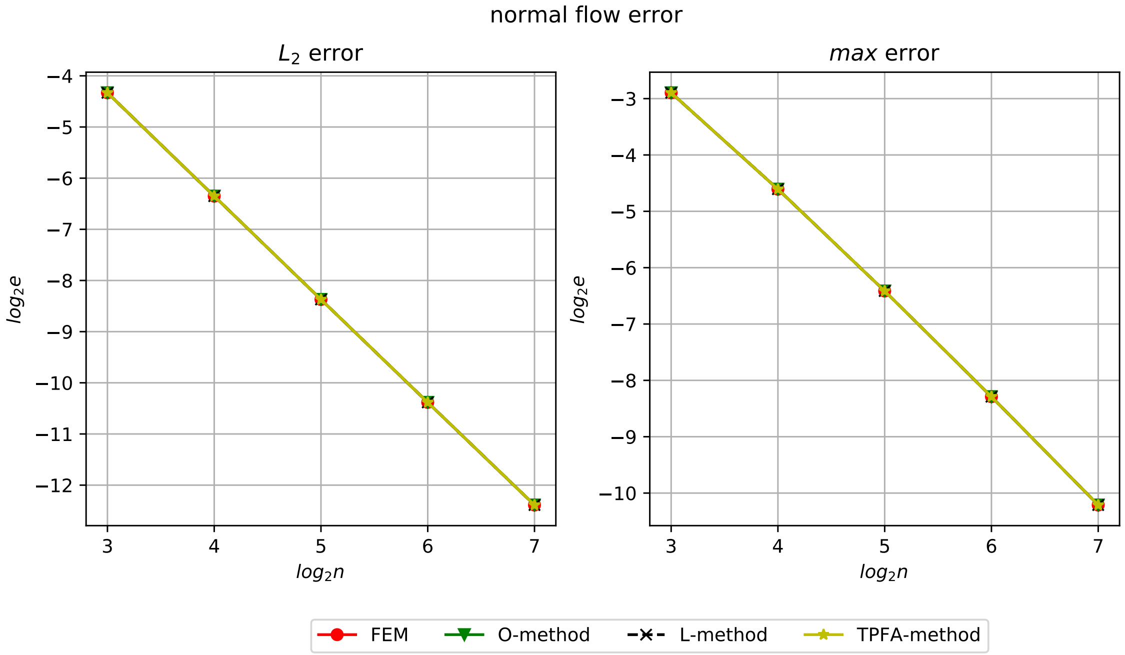


Figure 5.4: Normal flow density error on refinements of the uniform rectangular mesh 5.2. Left: $e = \|q - q_h\|_{h,0}$, right: $e = \|q - q_h\|_{h,\infty}$

5.2.2 Setup 2: Parallelogram Grid

In this setup, we perturb the half unit square by $(x, y) \mapsto (x-0.5y, y)$, and consider a parallelogram grid as in figure 5.5. In figures 5.6 and 5.7 we observe that the TPFA method does not converge, this makes sense as the grid is not K-orthogonal. The other methods still have quadratic convergence for potential and flow density.

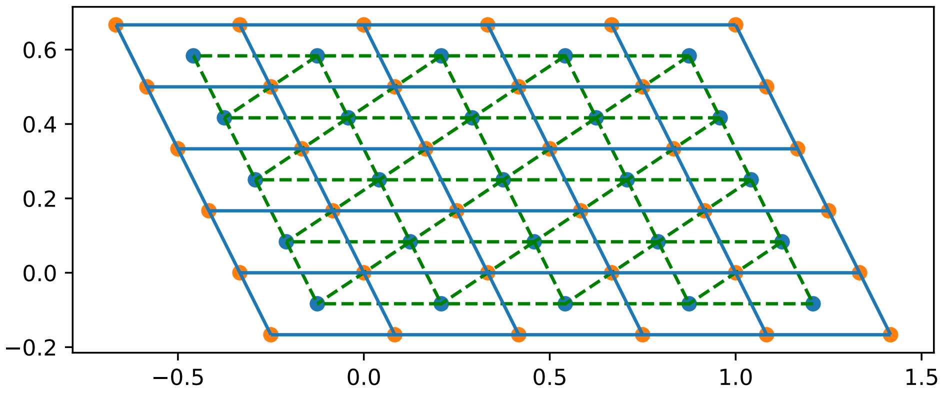


Figure 5.5: Trapezoidal mesh, now every point is transformed by $(x, y) \mapsto (x - 0.5y, y)$

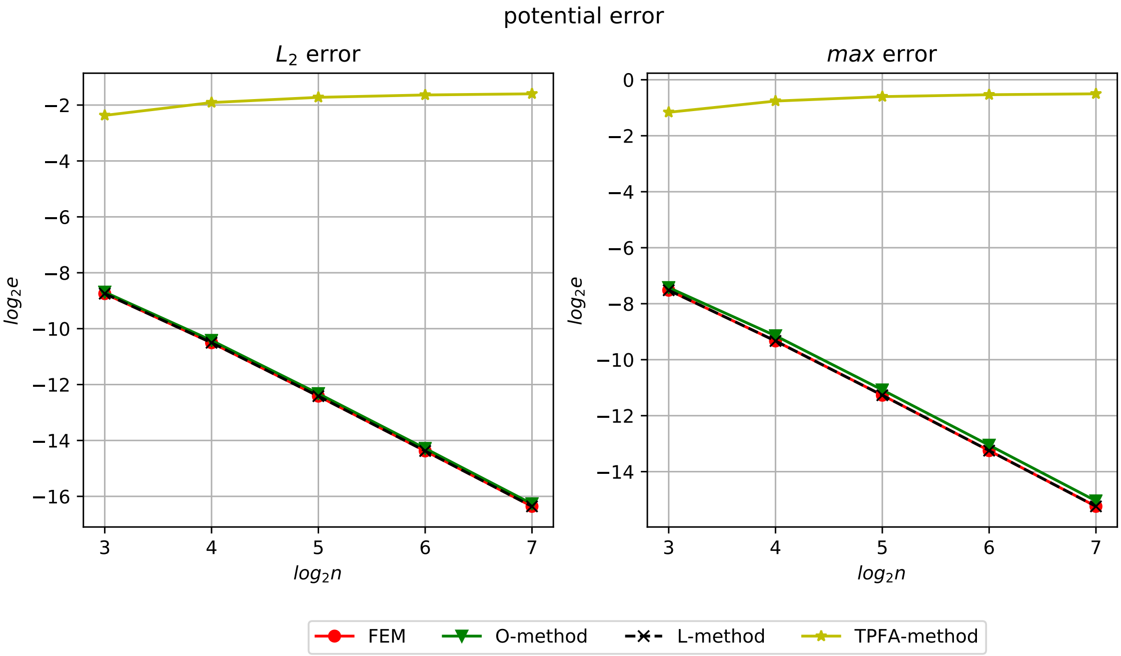


Figure 5.6: Pressure error on refinements of the mesh 5.5. Left: $e = \|u - u_h\|_{h,0}$, right: $e = \|u - u_h\|_{h,\infty}$

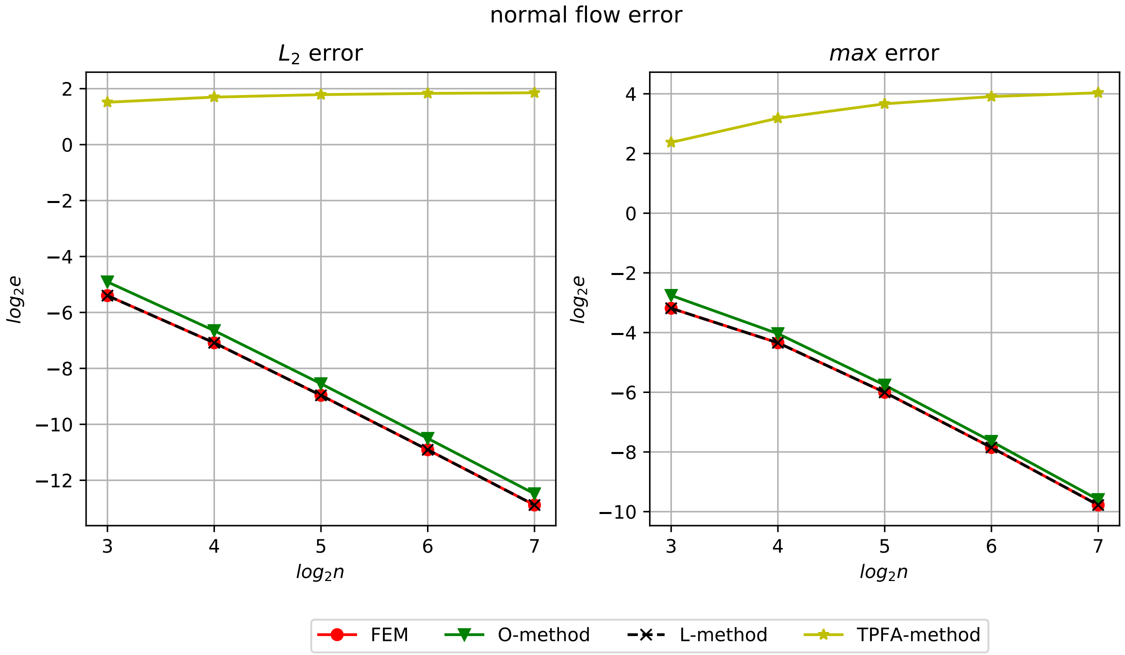


Figure 5.7: Normal flow density error on refinements of the mesh 5.5. Left: $e = \|q - q_h\|_{h,0}$, right: $e = \|q - q_h\|_{h,\infty}$

5.2.3 Setup 3: Rough Paralellogram Grid

Here, we perturb every interaction point in the parallelogram grid, that is, we perturb the x and y coordinate by a number no bigger than $\frac{h}{5}$, where $h = \frac{1}{n}$, with n being the number of interaction points in y direction. See figure 5.8 for an example. Remark that our finite volume methods do not handle control volumes that are not convex, so there are limits to how much we can perturb.

In figures 5.9 and 5.10 the MPFA-L, MPFA-O and FEM still converges quadratically when we refine the rough grids. For the normal flow density however, the convergence rate drops to about one. Also note that results for the MPFA-L-method and FEM are different, this suggests that integrating the source term in way that is not mass conservative, yields a slightly better potential error and a slightly worse normal flow density error.

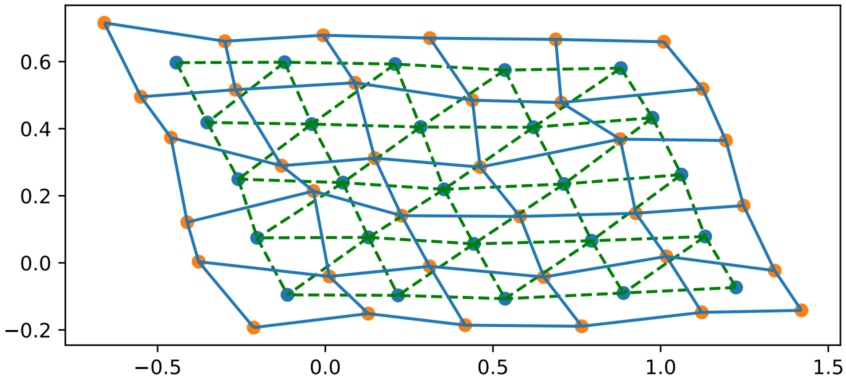


Figure 5.8: Perturbed mesh, every point in the mesh is perturbed by a random number which is $O(\frac{h}{5})$, in both x and y direction.

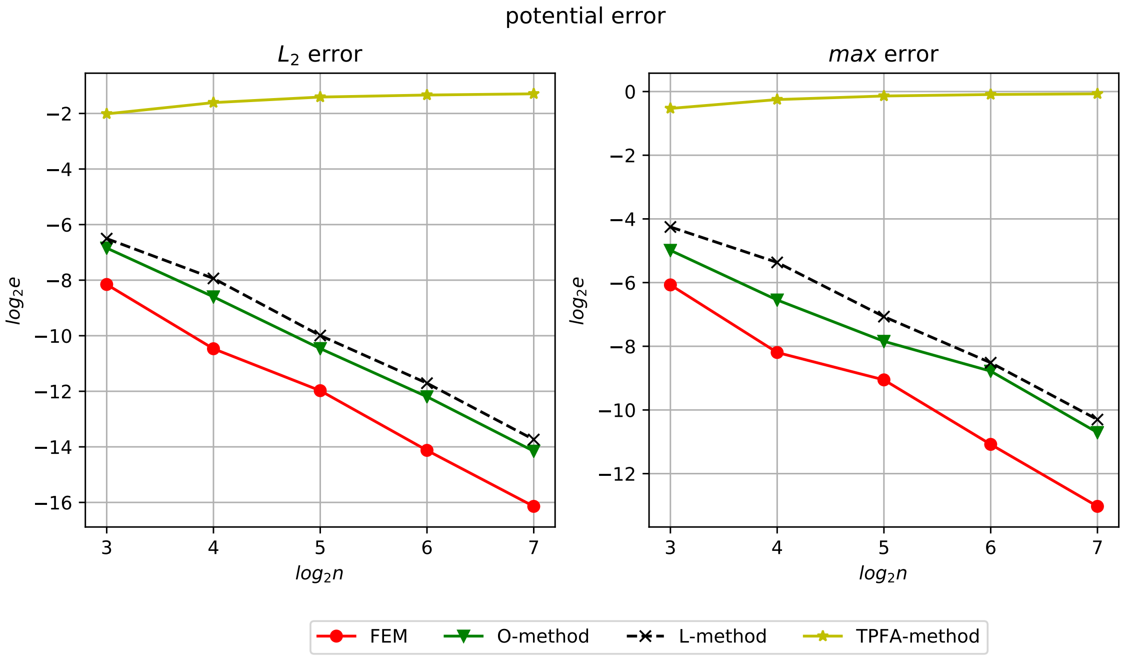


Figure 5.9: The pressure error of perturbed mesh. Left: $e = \|u - u_h\|_{h,0}$, right: $e = \|u - u_h\|_{h,\infty}$

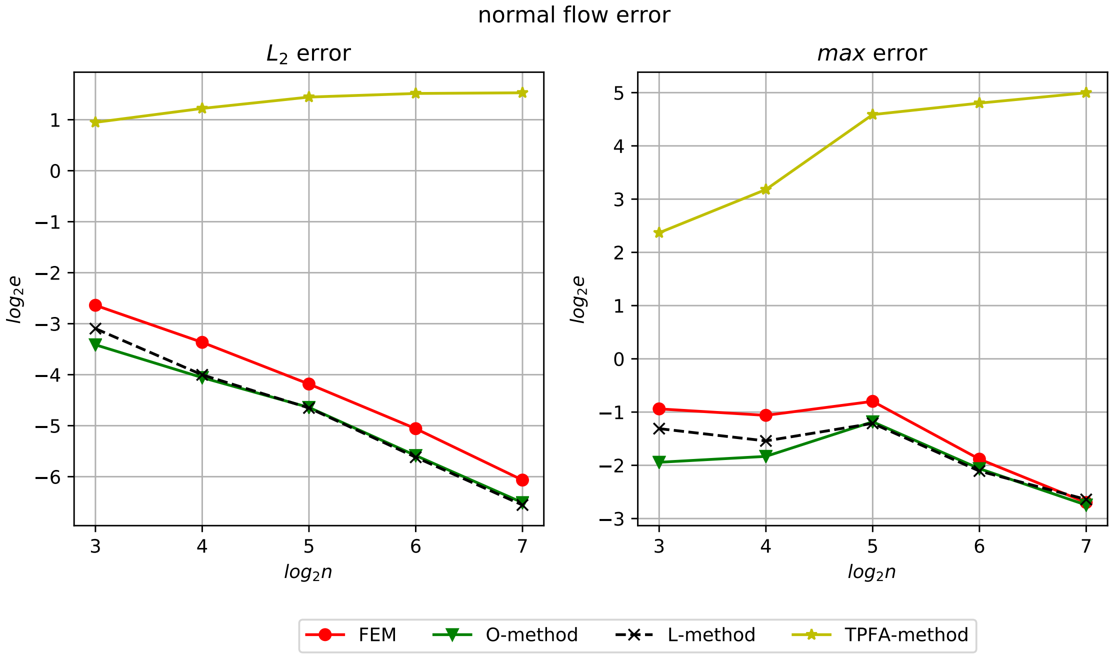


Figure 5.10: The normal flow density error of perturbed mesh. Left: $e = \|q - q_h\|_{h,0}$, right: $e = \|q - q_h\|_{h,\infty}$

5.2.4 Setup 4: Aspect Ratio

In figure 5.11 we introduce grids with aspect ratio, i.e., grids with more points in the y direction than the x direction, see figure 5.11 for a graphical explanation. We still perturb the interaction points as before.

In figure 5.12 we observe that MPFA-L, MPFA-O and FEM has a convergence rates for the potential of about 1.5 for the grid with aspect ratio 0.1. In figure 5.14 we see that the MPFA-O method fail to converge for the grid with aspect ratio 0.01. Thus, the MPFA-L method performs best in this round of numerical experiments.

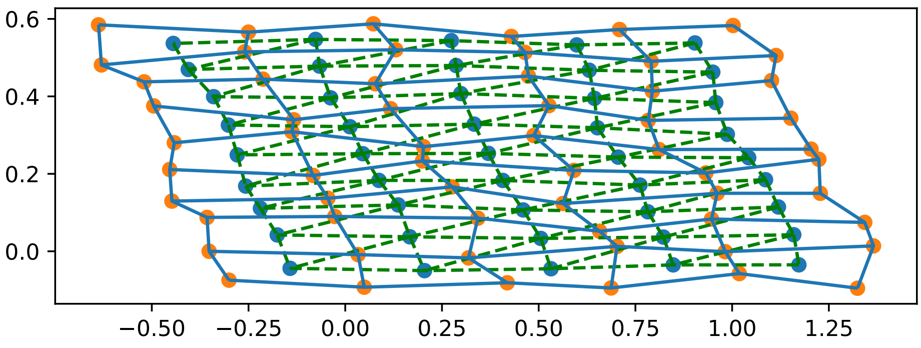


Figure 5.11: Perturbed mesh with aspect ratio 0.5, there are half as many points in the x-direction as in the y-direction.

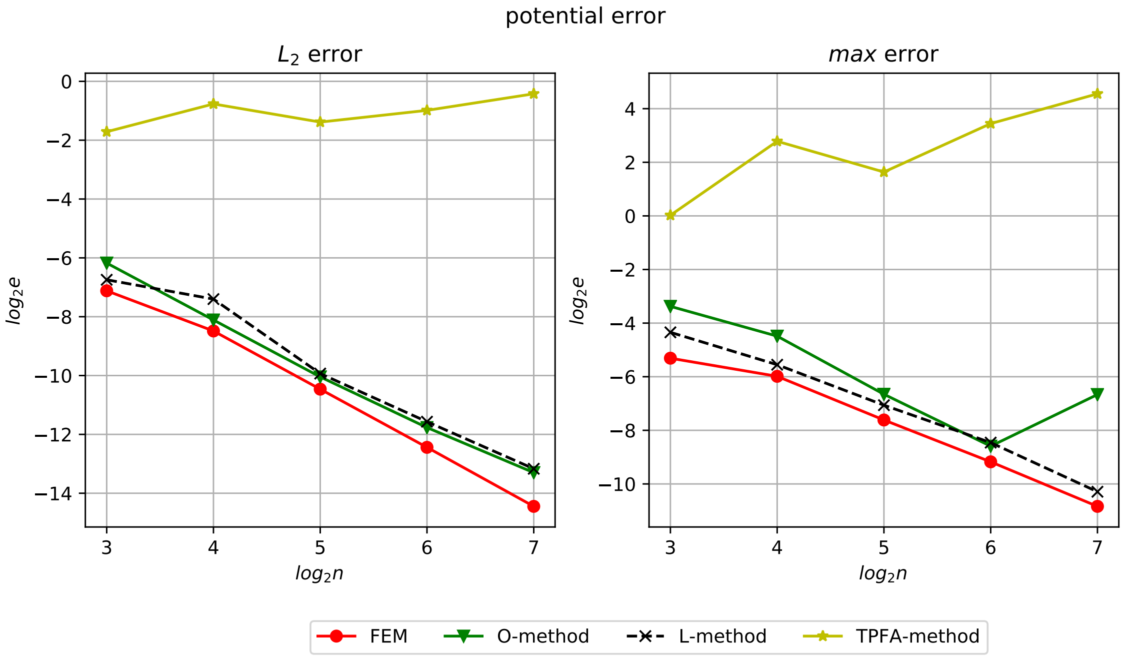


Figure 5.12: The potential error of perturbed mesh with aspect ratio 0.1. Left: $e = \|u - u_h\|_{h,0}$, right: $e = \|u - u_h\|_{h,\infty}$

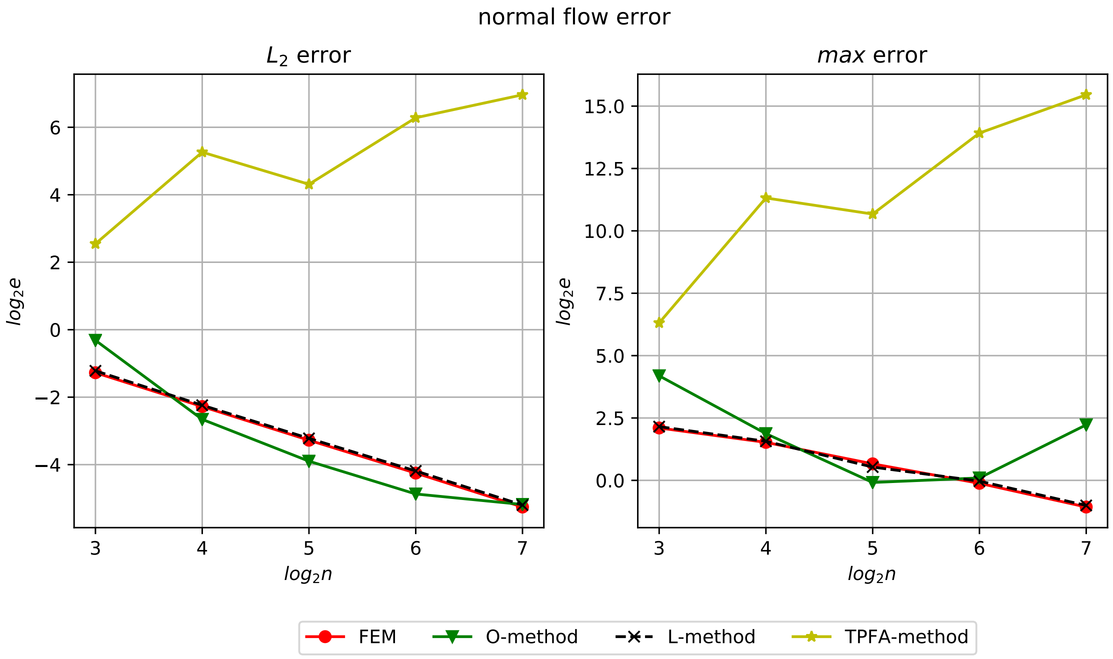


Figure 5.13: The normal flow density error of perturbed mesh with aspect ratio 0.1. Left: $e = \|q - q_h\|_{h,0}$, right: $e = \|q - q_h\|_{h,\infty}$

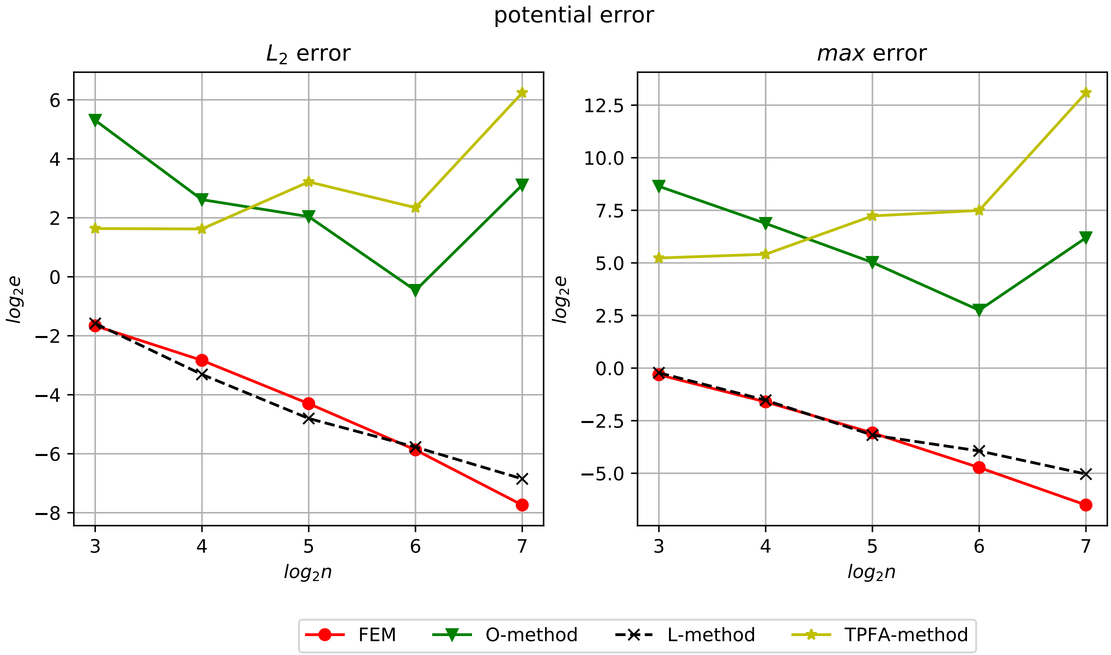


Figure 5.14: The potential error of perturbed mesh with aspect ratio 0.01. Left: $e = \|u - u_h\|_{h,0}$, right: $e = \|u - u_h\|_{h,\infty}$

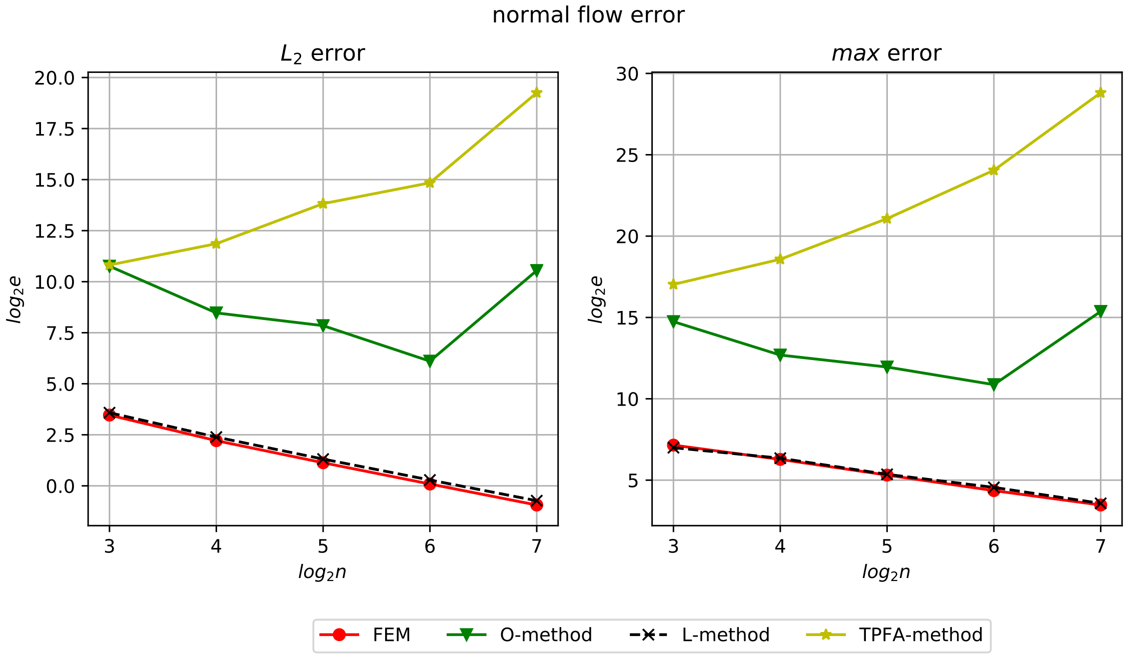


Figure 5.15: The normal flow density error of perturbed mesh with aspect ratio 0.01. Left: $e = \|q - q_h\|_{h,0}$, right: $e = \|q - q_h\|_{h,\infty}$

Remark 17. *Comparing the the linear Lagrange finite element method, with the control volume methods on rough grids is a somewhat strange comparison. The cell centres in 5.8 does not get as affected by the perturbations of the control volumes.*

Remark 18. *We have so far only considered homogeneous permeability, and have therefore not demonstrated the ability of the control volume methods in handling discontinuous permeability. For a comparison between MPFA-methods and the linear Lagrange finite element method on a heterogeneous domain, we refer to [9], where the authors demonstrate the importance of locally mass conservative methods.*

5.3 Richards' Equation

In this section we test the convergence of backward Euler, L-scheme and MPFA-L-method applied Richards' equation. We are interested in how the error in the numerical approximation changes with respect to time step length and mesh diameter.

5.3.1 Constant Hydraulic Conductivity

Here, we consider numerical experiments for (4.2), with Dirichlet boundary conditions: Find $u = u(x,t)$ such that

$$\begin{cases} \partial_t b(u) - \nabla \cdot \nabla u = f, & \text{in } \Omega \times (0,T] \\ u = u|_{\Gamma_D}, & \text{on } \partial\Gamma_D \times (0,T] \\ u = u_0, & \text{on } \Omega \times \{t = 0\} \end{cases} \tag{5.3}$$

We define

$$b(u) := \frac{1}{1-u}, \tag{5.4}$$

and compute the source term f, such that the solution becomes

$$u = -tx(1-x)y(1-y) - 1, \tag{5.5}$$

which is the same equation and solution as they use in [18]. We let Ω be the unit square transformed by $(x,y) \mapsto (x - 0.5y, y)$. The L-scheme linearization has the parameters $L := 1.5$ and error tolerance $TOL = 5e^{-8}$. We use a parallelogram grid as in figure (5.16). In table 5.1 we observe quadratic convergence when the time step length is set equal to the square of the mesh diameter, $\tau = h^2$. We observe the same convergence rate in table 5.2, when $\tau = h$. One would not expect this, as the the time discretization has linear convergence rate. An explanation could be that the solution (5.4) is linear in time, and even with the non-linearity $b(\cdot)$, is approximated exactly by the backward Euler time discretization.

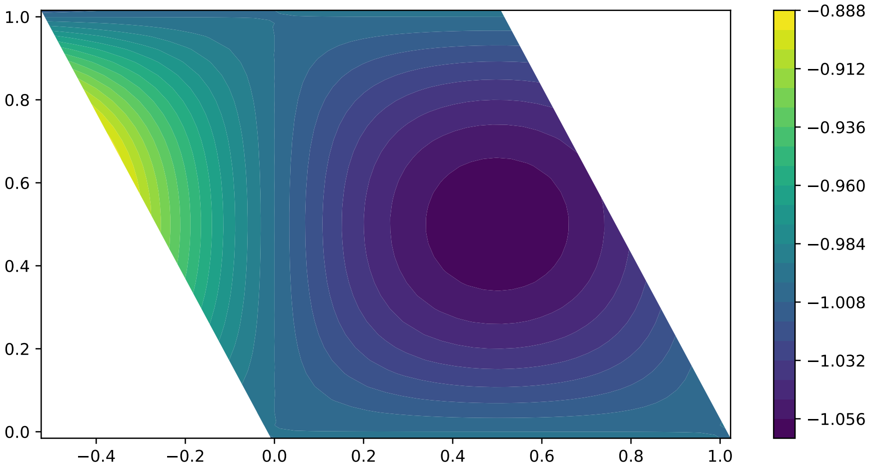


Figure 5.17: The solution of (5.3) at $T = 1$, with the ghost Dirichlet boundary.

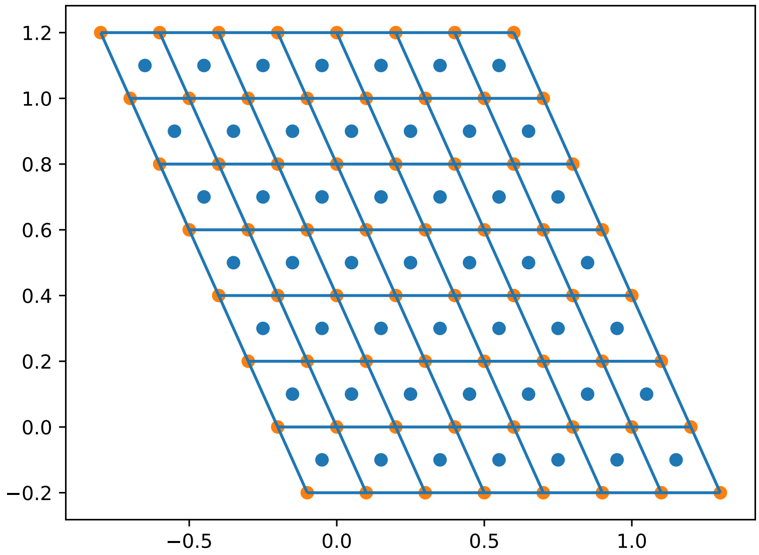


Figure 5.16: Parallelogram grid, with ghost Dirichlet boundary cells.

mesh diameter, h	time step length, τ	discrete $L_2(\Omega)$ error	improvement
0.45069	0.20312	0.001695	-
0.22535	0.05078	0.000375	4.51623
0.11267	0.01270	0.000087	4.31530
0.05634	0.00317	0.000021	4.20040

Table 5.1: Convergence table for (5.3),(5.4) and (5.5). The time step length, τ, is set proportional to the square of the mesh diameter, that is $\tau = h^2$.

mesh diameter, h	time step length, τ	discrete $L_2(\Omega)$ error	improvement
0.45069	0.45069	0.001694	-
0.22535	0.22535	0.000374	4.52868
0.11267	0.11267	0.000086	4.33993
0.05634	0.05634	0.000020	4.24067
0.02817	0.02817	0.000005	4.19727

Table 5.2: Convergence table for (5.3),(5.4) and (5.5). The time step length, τ, is set proportional to the mesh diameter, that is $\tau = h$.

If we however construct a solution which is not linear in time

$$u = -t^2 x(1-x)y(1-y) - 1. \tag{5.6}$$

And do the same experiment as described above, with the time step length $\tau = h$, we do not observe quadratic convergence, see table 5.3.

mesh diameter, h	time step length, τ	discrete $L_2(\Omega)$ error	improvement
0.45069	0.45069	0.001922	-
0.22535	0.22535	0.000471	4.08322
0.11267	0.11267	0.000125	3.76197
0.05634	0.05634	0.000036	3.43254
0.02817	0.02817	0.000012	2.97651

Table 5.3: Convergence table for (5.3),(5.4) and (5.6). The time step length, τ, is set proportional to the mesh diameter, that is $\tau = h$.

5.3.2 Non-Linear Hydraulic Conductivity

Here, we consider Richards' equation (1.9) in pressure variable, find $p = p(x,t)$ such that

$$\begin{cases} \partial_t \theta(p) - \nabla \cdot \kappa(\theta(p)) \nabla p = f, & \text{in } \Omega \times (0,T] \\ p = p|_{\Gamma_D}, & \text{on } \partial\Gamma_D \times (0,T] \\ p = p_0, & \text{on } \Omega \times \{t = 0\} \end{cases} \tag{5.7}$$

With Ω being the perturbed unit square as before (see figure 5.16), and $T = 1$. We consider the Van Genuchten-Mualem parametrizations

$$\theta(p) := \begin{cases} (1 + (-\alpha_{vG} p)^{n_{vG}})^{-\frac{n_{vG}-1}{n_{vG}}}, & p \le 0 \\ 1, & p > 0 \end{cases}, \tag{5.8}$$

and

$$\kappa(\theta) := \frac{\kappa_{abs}}{\mu} \sqrt{\theta} \left(1 - \left(1 - \theta^{\frac{n_{vG}}{n_{vG}-1}} \right)^{\frac{n_{vG}-1}{n_{vG}}} \right)^2, \tag{5.9}$$

where the inverse of air suction $\alpha_{vG} = 0.1844$, the pore size distribution $n_{vG} = 3$, absolute permeability $\kappa_{abs} = 0.03$ and the viscosity $\mu = 1$ are model parameters. The source term f is computed such that the solution becomes

$$p = -3tx(1-x)y(1-y) - 1. \tag{5.10}$$

We include the factor of three in the solution above, to make sure that we capture more of the non-linearity, i.e., so low pressure that it affects the saturation. The L-scheme linearization has the parameters $L = 0.3$ and $TOL = 10^{-8}$.

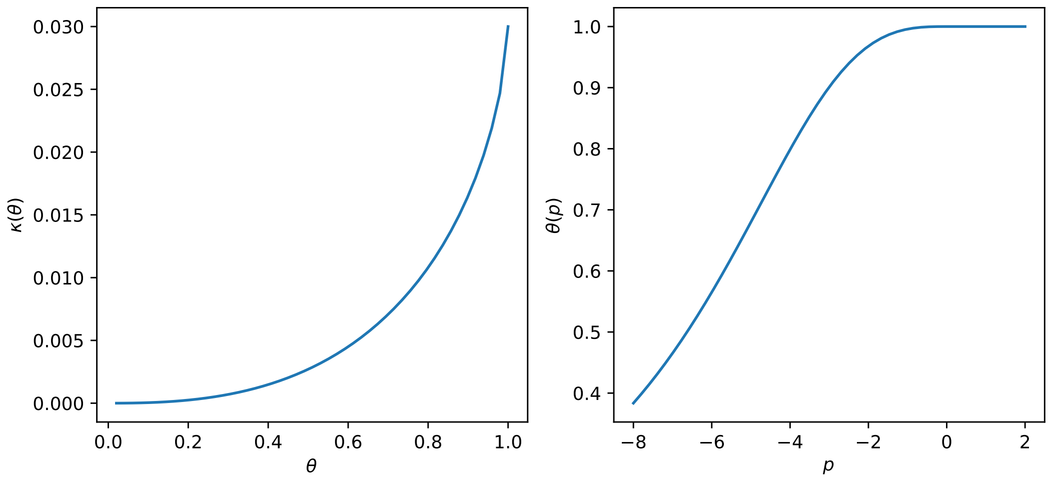


Figure 5.18: The Van Genuchten-Mualem non-linearities, (5.8) and (5.9).

In table 5.4, we see a quadratic convergence when $\tau = h^2$, as expected. In table 5.5, we see a slowly decreasing convergence rate, but still well above linear, when $\tau = h$. We expect the trend towards lower convergence to continue for finer refinements.

mesh diameter, h	time step length, τ	discrete $L_2(\Omega)$ error	improvement
0.45069	0.20312	0.005779	-
0.22535	0.05078	0.001443	4.00516
0.11267	0.01270	0.000350	4.12657
0.05634	0.00317	0.000086	4.06126

Table 5.4: Convergence table for (5.7),(5.10), (5.8) and (5.9). The time step length, τ, is set proportional to the square of the mesh diameter, that is $\tau = h^2$.

mesh diameter, h	time step length, τ	discrete $L_2(\Omega)$ error	improvement
0.45069	0.45069	0.005802	-
0.22535	0.22535	0.001484	3.90847
0.11267	0.11267	0.000378	3.93106
0.05634	0.05634	0.000099	3.81519

Table 5.5: Convergence table for (5.7),(5.10), (5.8) and (5.9). The time step length, τ, is set proportional to the mesh diameter, that is $\tau = h$.

Referring to listing 5.2, we observe that we need to assemble the stiffness matrix and solve the elliptic problem for each linearization iteration when working with non-linear hydraulic conductivity. This requires a lot more computational effort than for the constant hydraulic conductivity in the previous section, where we only invert the mass and stiffness matrix once.

Chapter 6

Conclusions

In this thesis, we have given an introduction to finite volume methods and finite element methods and seen how they can be applied to the Richards' equation. Moreover, we introduced the techniques in [5] to prove convergence for the MPFA-L-method by showing equivalence with a finite element method on a parallelogram grid. Next, we used this technique to prove a convergence rate estimate for the MPFA-L-method, backward Euler and L-scheme applied to Richards' equation after Kirchhoff transform. Unfortunately, the theory in [5] cannot be applied to elliptic PDEs with heterogeneous permeability. This limitation makes it unsuitable for proving convergence in the case of non-linear hydraulic conductivity.

We have also implemented the numerical methods covered in this thesis, and done numerical experiments with our code. In section 5.2 we compared convergence rates for the spatial discretization methods applied on a homogeneous elliptic problem. We conclude that the MPFA-L-Method is the only control volume method we covered, that handles rough grids with small aspect ratio (thin control volumes), this was also observed in [2]. Moreover, we see that it has the same convergence rate as the linear Lagrange finite element method for rough grids with small anisotropy. This also confirms the link between the two methods, even when the grid does not consist of parallelograms.

In section 5.3, we did numerical experiments regarding the convergence rate of the MPFA-L-method, backward Euler and L-scheme applied to Richards' equation with, and without, non-linear hydraulic conductivity. We conclude that for sufficiently many linearization iterations in each time step, the L^2 convergence rate in either case, is $(h^2 + \tau)$. This aligns with our theoretical findings in chapter 4.

Bibliography

[1] I. Aavatsmark, *An introduction to multipoint flux approximations for quadrilateral grids*, Computational Geosciences, 6 (2002), pp. 405–432.

[2] I. Aavatsmark, G. Eigestad, B. Mallison, and J. Nordbotten, *A compact multipoint flux approximation method with improved robustness*, Numerical Methods for Partial Differential Equations, 24 (2008), pp. 1329–1360.

[3] J. Baranger, J.-F. Maitre, and F. Oudin, *Connection between finite volume and mixed finite element methods*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique, 30 (1996), pp. 445–465.

[4] Y. Cao, R. Helmig, and B. I. Wohlmuth, *Geometrical interpretation of the multi-point flux approximation L-method*, International Journal for Numerical Methods in Fluids, 60 (2009), pp. 1173–1199.

[5] ——, *Convergence of the multipoint flux approximation L-method for homogeneous media on uniform grids*, Numerical Methods for Partial Differential Equations, 27 (2011), pp. 329–350.

[6] W. Cheney, *Analysis for Applied Mathematics*, Springer-Verlag New York Inc., 2001.

[7] L. C. Evans, *Partial differential equations*, American Mathematical Society, Providence, R.I., 2010.

[8] R. A. Klausen, F. A. Radu, and G. T. Eigestad, *Convergence of MPFA on triangulations and for Richards' equation*, International Journal for Numerical Methods in Fluids, 58 (2008), pp. 1327–1351.

[9] R. A. Klausen and T. F. Russell, *Relationships among some locally conservative discretization methods which handle discontinuous coefficients*, Computational Geosciences, 8 (2004), pp. 341–377.

[10] R. A. Klausen and R. Winther, *Robust convergence of multi point flux approximation on rough grids*, Numerische Mathematik, 104 (2006), pp. 317–337.

[11] P. Knabner and L. Angerman, *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*, vol. 44 of Texts in Applied Mathematics, Springer-Verlag New York, 2003.

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

[13] J. Nordbotten and M. Celia, *Geological storage of CO2: Modeling Approaches for Large-Scale Simulation*, "John Wiley & Sons", 2011.

[14] J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad, *Monotonicity of control volume methods*, Numer. Math., 106 (2007), p. 255–288.

[15] J. M. Nordbotten and E. Keilegavlen, *An introduction to multi-point flux (MPFA) and stress (MPSA) finite volume methods for thermo-poroelasticity*, 2020.

[16] I. S. Pop, *Error estimates for a time discretization method for the Richards' equation*, Computational Geosciences, 6 (2002), pp. 141–160.

[17] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: Linearization procedure*, J. Comput. Appl. Math., 168 (2004), p. 365–373.

[18] F. Radu and W. Wang, *Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media*, Nonlinear Analysis: Real World Applications, 15 (2011).

[19] F. A. Radu, I. S. Pop, and P. Knabner, *Order of convergence estimates for an euler implicit, mixed finite element discretization of Richards' equation*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 1452–1478.

[20] L. A. Richards, *Capillary conduction of liquids through porous mediums*, Physics, 1 (1931), pp. 318–333.

[21] E. Stein, *History of the Finite Element Method – Mathematics Meets Mechanics – Part I: Engineering Developments*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 399–442.

[22] A. F. Stephansen, *Convergence of the multipoint flux approximation L-method on general grids*, SIAM J. Numer. Anal., 50 (2012), pp. 3163–3187.

[23] E. Storvik, *On the optimization of iterative schemes for solving non-linear and/or coupledPDEs*, Master's thesis, University of Bergen, 2018.

[24] M. Van Genuchten, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Science Society of America Journal, 44 (1980).

[25] X.-H. Wu and R. Parashkevov, *Effect of grid deviation on flow solutions*, SPE journal (Society of Petroleum Engineers (U.S.) : 1996), 14 (2009), pp. 67–77.

[26] J. Yanosik and T. McCracken, *A Nine-Point, Finite-Difference Reservoir Simulator for Realistic Prediction of Adverse Mobility Ratio Displacements*, Society of Petroleum Engineers Journal, 19 (1979), pp. 253–262.