Something about convergence for MPFA discretization of Richards' equation

Truls Moholt

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

Contents

Chapter 1

Flow in Porous Media

In this chapter we introduce the basic concepts of flow in porous media, briefly covering the modeling choices and physics that leads to Richards' equation. The theory in this chapter is to a large extent adapted from [1] 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 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, ie. the *representative elementary volume* (REV). An important characterization of a porous medium is the *porosity* ϕ, it is defined as

$$\phi := \frac{volume\ of\ voids\ in\ REV}{volume\ of\ REV}. \tag{1.1}$$

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

$$S_\alpha := \frac{volume\ of\ \alpha\ in\ REV}{volume\ of\ voids\ in\ REV}. \tag{1.2}$$

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*. Then the pressure at height z above datum developed by a water column of height h above datum is given by

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

Where ρ is the density and g is the standard 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*:

$$h = \frac{p}{\rho g} + z. \tag{1.3}$$

A *manometer* is a tube with one end in the reservoir and one in open atmosphere, the water level in this tube is then h. 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:

$$q_d = -\kappa \frac{A(h_2 - h_1)}{L}. \tag{1.4}$$

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

$$\boldsymbol{q} = -\boldsymbol{\kappa} \nabla h. \tag{1.5}$$

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.

The conductivity matrix, $\boldsymbol{\kappa}$, is also *positive definite*, this is because there is never flux in the same direction as the pressure gradient. With further experiments, similar to the one already described, we can understand what makes up $\boldsymbol{\kappa}$.

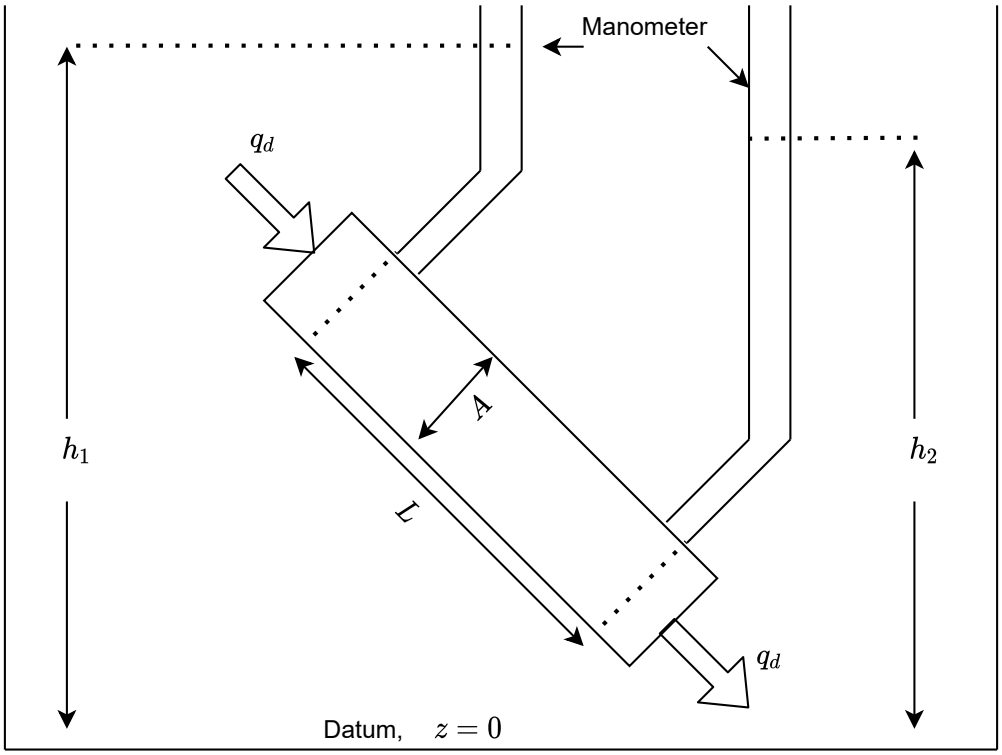


Figure 1.1: The Darcy experiment

Dimensionality analysis shows that it is a function of viscosity μ, density of the fluid ρ, gravity g and *permeability* $\boldsymbol{k}$,

$$\boldsymbol{\kappa} = \frac{\boldsymbol{k}\rho g}{\mu}. \tag{1.6}$$

The *permeability*, which is a property of the soil in the reservoir, is also a rank two tensor which is symmetric positive definite and is in general a function of space, ie. heterogeneous.

If we define the *pressure head* ψ as $\psi := \frac{p}{\rho g}$, we can combine (1.3), (1.5) and (1.6) to get another variant of Darcy's law;

$$\boldsymbol{q} = -\frac{\boldsymbol{k}\rho g}{\mu}\nabla(\psi + z) \tag{1.7}$$

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 (1.5) and appropriate boundary and initial conditions close the system

$$\begin{cases} \boldsymbol{q} = -\boldsymbol{\kappa}\nabla h, & \boldsymbol{x} \in \Omega, \quad t > 0 \\ \dfrac{\partial(\rho\phi)}{\partial t} + \nabla \cdot (\rho\boldsymbol{q}) = f(\boldsymbol{x}, t), & \boldsymbol{x} \in \Omega, \quad t > 0 \\ h(\boldsymbol{x}) = g(\boldsymbol{x}, t), & \boldsymbol{x} \in \partial\Omega, \quad t > 0 \\ h(\boldsymbol{x}) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \quad t = 0 \end{cases} \tag{1.8}$$

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 [1] for a more detailed discussion of (1.8) 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, ie. we assume that Darcy's law (1.7) holds for both phases. Let the subscript α denote the phase, then we have Darcy's law for each phase

$$\boldsymbol{q}_\alpha = \frac{\boldsymbol{k}_{r,\alpha}\boldsymbol{k}\rho g}{\mu}\nabla(\psi_\alpha + z), \tag{1.9}$$

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:

$$\frac{\partial(S_\alpha \rho_\alpha \phi)}{\partial t} + \nabla \cdot (\rho \boldsymbol{q}_\alpha) = f_\alpha. \tag{1.10}$$

Here, we assume that there is no mass transfer between the phases. If we combine equations (1.9) and (1.10), 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

$$p_n - p_w = p_c \tag{1.11}$$

where p_c is called *capillary pressure*. As with the relative permeability, p_c also need 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 (1.11) 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 (1.11) 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 (1.9) and mass balance (1.10) we get **Richards' equation**

$$\frac{\partial \theta(\psi)}{\partial t} - \nabla \cdot (\boldsymbol{\kappa}(\theta(\psi))(\nabla \psi + e_z)) = f \tag{1.12}$$

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 interesting and challenging as we will see. They may also cause the equation to degenerate, ie. the parabolic equation may "collapse" into an elliptic PDE (see figure 1.2) or even an ODE.

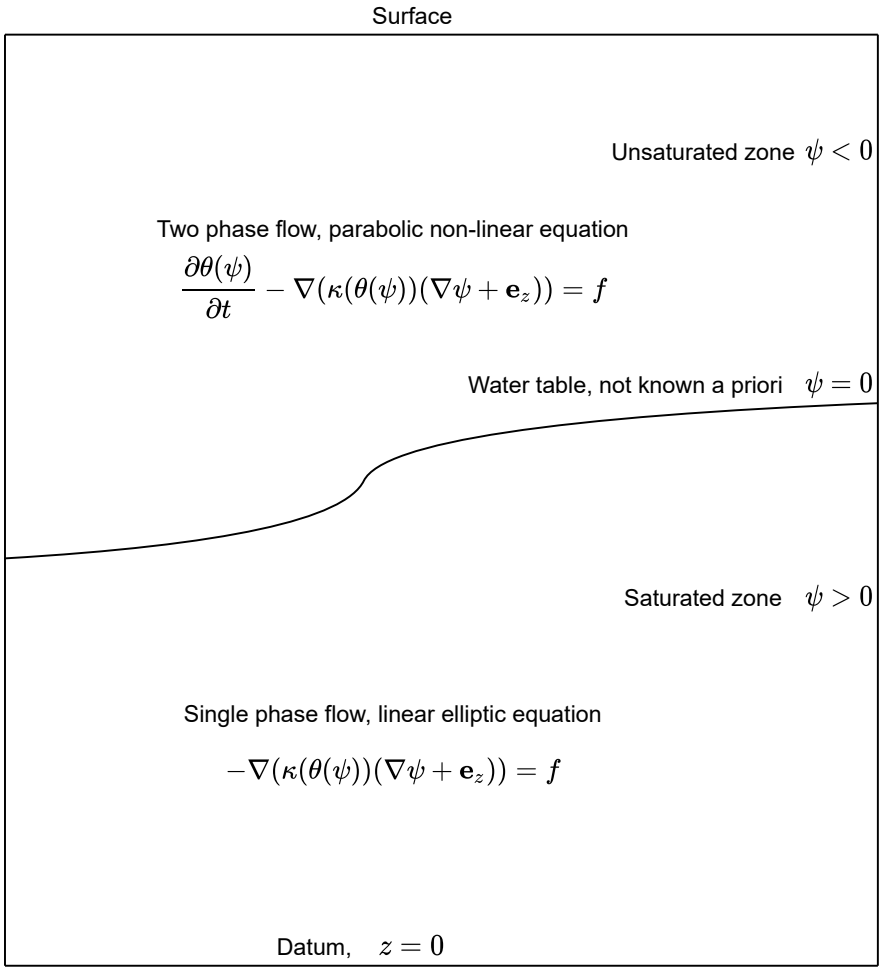


Figure 1.2: A sketch of the degeneracy of Richards' equation

Chapter 2

Numerical approximation techniques

In this chapter, we first discuss two important frameworks for space discretization of PDE's, 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 or 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 became more mainstream [2]. 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 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:

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

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 formulating the Poisson equation in the *variational formulation*. Another motivation is that it allows for an nice framework for computing the solution, as we will soon see. But first, we study some spaces of functions and their properties.

2.1.1 Function spaces

When discussing PDE's 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 for which $\|u\|_p = (\int_\Omega u^p dx)^{1/p} < \infty$*

Remark 1. *Note that an $L^p(\Omega)$ 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* vector spaces or *Banach spaces*.

Theorem 2.1.1 (Riesz-Fischer Theorem [3] 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 $D^{\overline{\alpha}} = (\frac{\partial}{\partial x_1})^{\alpha_1}(\frac{\partial}{\partial x_2})^{\alpha_2}...(\frac{\partial}{\partial x_n})^{\alpha_n}\phi$*

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

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

$$\int_\Omega g\phi dx = (-1)^{|\overline{\alpha}|} \int_\Omega f D^{\overline{\alpha}} \phi dx \quad \forall \phi \in C^\infty \tag{2.2}$$

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*

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

We then define the sovolev 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 $\overline{\alpha},\ |\overline{\alpha}| \leq k$ we have a Cauchy sequence in $L^p(\Omega)$:

$$\begin{aligned} \|u_j - u_i\|_{W^{k,p}} = (\sum_{|\overline{\alpha}|\leq k} \left\| D_w^{\overline{\alpha}} u_j - D_w^{\overline{\alpha}} u_i \right\|_{L^p(\Omega)}^p)^{1/p} < \epsilon \ \forall i,j \geq N \\ \implies \left\| D_w^{\overline{\alpha}} u_j - D_w^{\overline{\alpha}} u_i \right\|_{L^p(\Omega)} < \epsilon. \end{aligned}$$

By (2.1.1), every $L^P(\Omega)$ space is a Banach space. Therefore, for each $|\overline{\alpha}| \leq k$, $D_w^{\overline{\alpha}} u_i$ 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, ie. $D_w^{\overline{\alpha}} u = u_{\overline{\alpha}}$. 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:

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

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

$$\begin{aligned} \int_\Omega (D_w^{\overline{\alpha}} u_i - u_{\overline{\alpha}}) \phi dx &\leq \left\| D_w^{\overline{\alpha}} u_i - u_{\overline{\alpha}} \right\|_{L_p} \|\phi\|_{L_q} \\ \int_\Omega (u_i - u) D^{\overline{\alpha}} \phi dx &\leq \|u_i - u\|_{L_p} \left\| D^{\overline{\alpha}} \phi \right\|_{L_q}. \end{aligned}$$

Taking the limit, the right hand side goes to zero, and we end up with the fact that we can move the limit out of the integral:

$$\begin{aligned} \lim_{i\to\infty} \int_\Omega D_w^{\overline{\alpha}} u_i \phi dx &= \int_\Omega u_{\overline{\alpha}} \phi dx \\ \lim_{i\to\infty} \int_\Omega u_i D^{\overline{\alpha}} \phi dx &= \int_\Omega u D^{\overline{\alpha}} \phi dx \end{aligned}$$

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 follows*

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

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 [4], 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. *$\|Tu\|_{L^p(\partial U)} \le \|u\|_{H^1(U)}$*

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 [4], chapter 5)) Suppose U is as in the previous theorem and $u \in W^{1,p}(U)$, then*

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

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) with multiplying (2.1) by 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

$$\int_\Omega (\nabla v)^T \boldsymbol{K}\nabla u \, dx = \int_\Omega vf \, dx \quad \forall v \in V. \tag{2.3}$$

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 (2.3). But a solution to (2.3) 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

$$a(u, v) = b(v) \quad \forall v \in V, \tag{2.4}$$

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:*

$$\|v\|_{V'} = \sup\left\{|\phi(u)| : \|u\|_V = 1\right\}.$$

In general, a variational formulation can we 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 complicated boundary conditions can be written:

$$\begin{aligned} -\nabla \cdot \boldsymbol{K}\nabla\hat{u}(x) &= f(x) \quad x \in \Omega \\ \hat{u}(x) &= g_D \quad x \in \Gamma_D \\ \boldsymbol{K}\nabla\hat{u}(x) &= g_N \quad x \in \Gamma_N \end{aligned} \tag{2.5}$$

To make a variational formulation of (2.5) 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:

$$a(u, v) := \int_\Omega \nabla u \boldsymbol{K} \nabla v \, dx. \tag{2.6}$$

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 $\hat{u} = u + w$, now we can use integration by parts to as before:

$$a(u+w, v) = \int_\Omega (\nabla u + \nabla w)^T \boldsymbol{K} \nabla v \, dx = \int_\Omega fv \, dx - \int_{\partial\Omega} \boldsymbol{K}\nabla(u+w) \cdot \boldsymbol{n} v \, dx. \tag{2.7}$$

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

$$a(u, v) = b(v) = \int_\Omega fv \, dx - \int_\Omega (\nabla w)^T \boldsymbol{K} \nabla v \, dx - \int_{\Gamma_N} g_N v \, dx. \tag{2.8}$$

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 (2.8) vanishes.

2.1.3 Existence and uniqueness

We still need to show that (2.8) 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* bi linear 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 [3]

Theorem 2.1.5. *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 orthogonal complement. 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* $x \to \langle x, v\rangle$ *for 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 [3] 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*

$$a(u, v) = b(v) \tag{2.9}$$

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, this 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 (2.9):
Find u such that

$$Au = f. \tag{2.10}$$

Now we need to show that (2.10) 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 (2.10) and proved the theorem. We just need to show that T is a contraction.

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

Where $u = u_1 - u_2$, here 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 $\langle Au, Au\rangle = a_u(Au) = a(u, Au)$

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

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

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

So now we have the inequality

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

We can choose ϵ such that T becomes a contraction. $\epsilon < \frac{2C_c}{C_b^2} \Rightarrow (1-2\epsilon+\epsilon^2) < 1$ $\square$

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

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

And note that $b(\cdot)$ is a bounded functional:

$$\Rightarrow \|u\| \leq \frac{b(u)}{C_c \|u\|} \leq \frac{\|b\|_{V'}}{C_c}$$

Now we have proved that (2.4) has an unique solution for suitable a and b. The variational form of Poisson equation (2.3) 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 $\|\cdot\|_{H_0^1}$:*

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

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

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

  *Where we used the **Cauchy Schwarz inequality** on the second line.*

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

$$|b(v)| = |\int_\Omega fv dx| \leq \|f\|_{L^2} \|v\|_{L^2}$$
$$\Rightarrow |\ b\|_{H_0^{1'}} = \sup\left\{\frac{|b(v)|}{\|v\|}\right\} \leq \|f\|_{L^2}$$

Hence, (2.3) *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 (2.4), 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 $u \in V_h$ such that

$$a(u, v_h) = b(v_h) \tag{2.11}$$

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 (2.11) holds for the basis functions of V_h, it holds for all elements in V_h. In the *finite element method*, the finite dimensional subspace are 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 [5])**.** *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 can now define our finite dimensional subspace, V_h.

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

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

Are the space 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*, ie. $V_h \subset V = H_0^1$*

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 is the evaluation on functions in P_K at the nodes of the element.*

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

$$\phi_i(x_j) = \delta_{ij}, \quad \phi_i \in V_h.$$

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

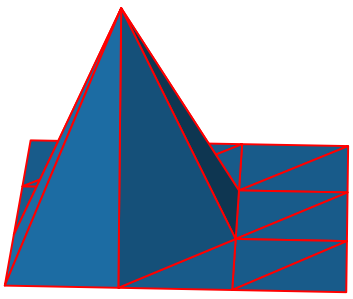


Figure 2.1: A hat function.

Now, we demonstrate how the method works in practice. We seek a solution $u_h \in V_h$. Write this in terms of the basis functions: $u_h = \sum_{i=1}^{n} \hat{u}_i \phi_i$. Now, (2.11) can be written as an equation with a solution vector with real coefficients:

$$\text{Find } \hat{\boldsymbol{u}}_h = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} \in \mathbb{R}^n \text{ such that } \sum_{i=1}^{n} \hat{u}_i a(\phi_i, \phi_j) = b(\phi_j). \tag{2.12}$$

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 (2.3), 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 are not expected to be very regular.

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 (2.3) 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 [5]. 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 (2.13). Also, compute the gradients of the shape

functions.

$$\begin{aligned} N_1(x,y) &= 1 - x - y \\ N_2(x,y) &= x \\ N_3(x,y) &= y \end{aligned} \tag{2.13}$$

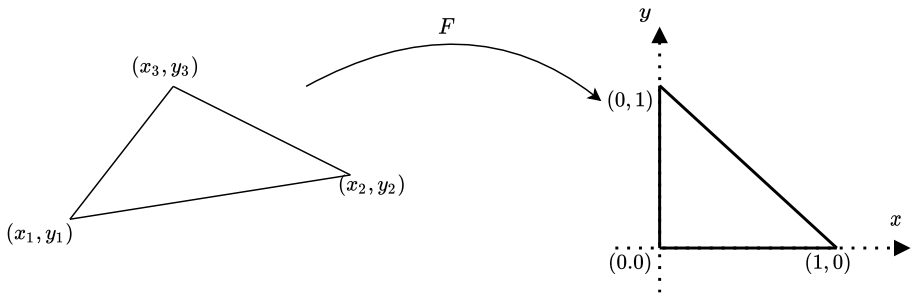


Figure 2.2: The map F from element K to the reference element $\hat{K}$.

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

$$\begin{aligned} F : K &\to \hat{K} \\ x &\mapsto Bx + d. \end{aligned} \tag{2.14}$$

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

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ d_1 & d_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.15}$$

So for each element we solve (2.15) 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:

$$\nabla^T_{\hat{x}} u(F^{-1}(\hat{x})) = \nabla^T_x u(F^{-1}(\hat{x})) \nabla^T_{\hat{x}} F^{-1}(\hat{x}) = \nabla^T_x u(F^{-1}(\hat{x})) B^{-1}. \tag{2.16}$$

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

$$\nabla_x u(F^{-1}(\hat{x})) = B^T \nabla_{\hat{x}} u(F^{-1}(\hat{x})). \tag{2.17}$$

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

$$\begin{aligned} a(\phi_i,\phi_j)|_K &= \int_K (\nabla\phi_i)^T\nabla\phi_j dx \\ &= \int_{\hat{K}} (\nabla_x\phi_i(F^{-1}(\hat{x})))^T\nabla_x\phi_j(F^{-1}(\hat{x}))|\mathrm{Det}(J(F^{-1}))|d\hat{x} \\ &= \int_{\hat{K}} (B^T\nabla_{\hat{x}}\phi_i(F^{-1}(\hat{x})))^T B^T B\nabla_{\hat{x}}\phi_j(F^{-1}(\hat{x}))|\mathrm{Det}(B^{-1})|d\hat{x} \\ &= \int_{\hat{K}} (\nabla_{\hat{x}}N_i(\hat{x}))^T B^T B\nabla_{\hat{x}}N_j(\hat{x})|\mathrm{Det}(B^{-1})|d\hat{x} \\ &= \frac{1}{2}(\nabla_{\hat{x}}N_i(\hat{x}))^T B^T B\nabla_{\hat{x}}N_j(\hat{x})\frac{1}{|\mathrm{Det}(B)|} \end{aligned} \tag{2.18}$$

So for each element we evaluate the last line of (2.18) 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 (2.18), we compute the integral on the reference element:

$$\begin{aligned} b(\phi_i)|_K &= \int_{\hat{K}} f(F^{-1}(\hat{x}))\phi_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})N_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &\approx \frac{1}{\mathrm{Det}(B)}\sum_k \omega_k\hat{f}(\hat{p}_k)N_i(\hat{p}_k) \end{aligned} \tag{2.19}$$

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, ie. 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* (2.8)*, using one-dimensional elements.*

2.1.6 Convergence

In this subsection, we review the most important concepts in studying the convergence fo FEM, for a detailed discussion see [5]. The starting point of convergence 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* (2.4) *and u_h solve the corresponding Galerkin approximation* (2.11)*, where the bilinear form a is bounded and coercive. Then we have:*

$$\|u - u_h\|_V \leq \frac{C_b}{C_c} min \{\|u - v_h\| : \ v_h \in V_h\}. \tag{2.20}$$

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 [3], 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.

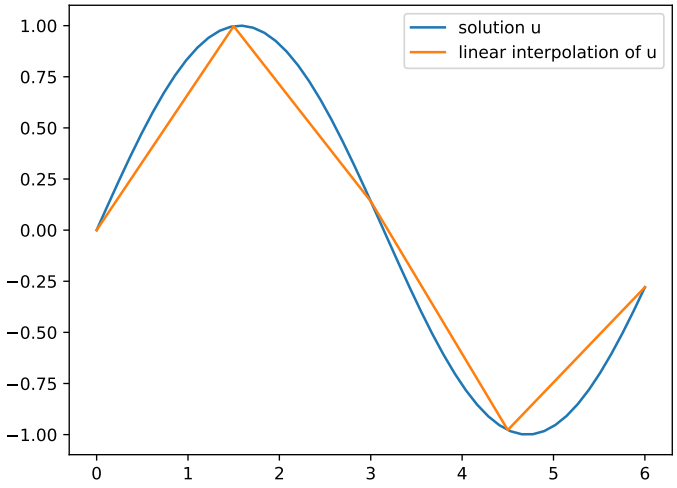


Figure 2.3: The unique linear interpolation of a function in one dimension.

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

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

$$\begin{aligned} I_h : C(\overline{\Omega}) &\to V_h \\ v &\mapsto \sum_i v(n_i)\phi_i \end{aligned} \tag{2.21}$$

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

Remark 10. *The global interpolator operator* (2.21) *maps from continuous functions, so we need to make sure our solution is continuous. By the Sobolev embedding theorem (Evans [4],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* $k \geq 2$*.*

Hence, in the setting of the model problem (2.3), we hope to reach an estimate

$$\|u - u_h\|_1 \leq C \|u - I_h(u)\|_1 \leq C^* h^k |u|_{k+1}, \tag{2.22}$$

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 [5], there is a detailed proof of (2.22).
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)**.** *Consider a finite element discretization as described by* (2.12) *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 10, then*

$$\|u - u_h\|_1 \leq Ch|u|_2. \tag{2.23}$$

Often we are happy with a convergence rate estimate in the $\|\cdot\|_0$ norm, which do not measure an error in the approximation of the derivative. We then expect a better convergence rate, as can be shown by the *duality trick*. We consider the dual problem of our variational problem (2.3): $a(v, u_f) = \langle f, v\rangle_0$, 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 exist an unique solution to the adjoint problem with* $|u_f| \leq C\|f\|_0$*, then there exist a constant* C^* *such that:*

$$\|u - u_h\|_0 \leq C^* h \|u - u_h\|_1 . \tag{2.24}$$

See [5] 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* (2.11)*. 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 for most source terms* 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

$$\begin{aligned} -\nabla \cdot \boldsymbol{K} \nabla u(x) &= f(x) \quad x \in \Omega \\ u(x) &= 0 \quad x \in \partial\Omega. \end{aligned} \tag{2.25}$$

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:

$$\int_{\Omega_i} -\nabla \cdot \boldsymbol{K} \nabla u dx = -\int_{\partial\Omega_i} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} ds = \int_{\Omega_i} f dx. \tag{2.26}$$

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 the edge between cells i and j. Then the main idea is to approximate the flux through $E_{i,j}$, from cell i to cell j,

$$q_{E_{i,j}} = -\int_{E_{i,j}} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} ds \tag{2.27}$$

by a linear combination of u_i at neighbouring cell centers

$$q_{E_{i,j}} \approx \tilde{q}_{E_{i,j}} = \sum_k t_{i,j}^k u^k. \tag{2.28}$$

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

$$\sum_{j\in\mathcal{S}_i} \tilde{q}_{E_{i,j}} = |\Omega_i| f(x_i), \tag{2.29}$$

where $\mathcal{S}_i$ is the set of indexes of neighbouring cells. The system of equations (2.29) ensures local mass conservation. It can also be written in matrix form as:

$$\boldsymbol{A}^V \tilde{\boldsymbol{u}}_h = \boldsymbol{f}. \tag{2.30}$$

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.

When it comes to boundary conditions, it is usually straightforward to implement Neumann boundary conditions. Especially no-flow boundary conditions, where one creates a strip of cells outside the boundary with zero permeability. The same discretization algorithm can the be applied everywhere in the domain.

Dirichlet boundary conditions often require special care. If one, however, knows the solution somewhere outside the domain, one could make a strip of cells outside the boundary where the potential values are known, this is known as ghost Dirichlet boundary conditions. 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 $t_{i,j}^k$ is also the most popular in the industry. As the name suggests, we only use the function value at two points, x_0 and x_1, to compute the numerical flux $\tilde{q}_{E_{0,1}}$.

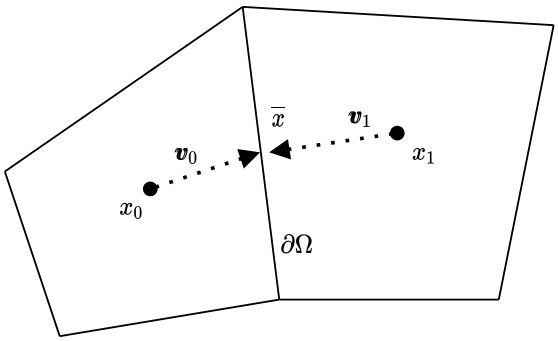


Figure 2.4: The two point flux approximation (TPFA) setup.

Let $\boldsymbol{v}_1$ be the vector from cell center x_0 to the midpoint of the edge between the cells, $\overline{x}$. Then we approximate the flux out of cell x_0 into cell x_1 by:

$$\tilde{q}_{E_{0,1},0} = -\boldsymbol{n}_0^T \boldsymbol{K}_0 \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}(u(\overline{x}) - u(x_0))ds \tag{2.31}$$

or as

$$\tilde{q}_{E_{0,1},1} = -\boldsymbol{n}_\mathbf{1}^T \boldsymbol{K}_1 \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}(u(x_1) - u(\overline{x}))ds \tag{2.32}$$

where $\boldsymbol{n}_i$ is the normal vector pointing out of cell i with length equal to $\partial\Omega$. Because we require flux continuity we have that

$$\tilde{q}_{E_{0,1},0} = \tilde{q}_{E_{0,1},1} = t^0 u(x_0) + t^1 u(x_1) \tag{2.33}$$

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 $T_i := \boldsymbol{n}_i^T \boldsymbol{K}_i \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|}$ to represent the cell transmissivity. So first we solve for $u(\overline{x})$:

$$T_0(u(\overline{x}) - u(x_0)) = T_1(u(x_1) - u(\overline{x})) \Rightarrow u(\overline{x}) = \frac{T_0 u(x_0) + T_1 u(x_1)}{T_0 + T_1}. \tag{2.34}$$

Next we insert this into the expression for $\tilde{q}_0$:

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} =& -T_0(u(\overline{x}) - u(x_0)) \\
=& -T_0\left(\frac{T_0u(x_0) + T_1u(x_1)}{T_0 + T_1} - u(x_0)\right) \\
=& -T_0\left(\frac{T_0u(x_0) + T_1u(x_1) - u(x_0)T_0 - u(x_0)T_1}{T_0 + T_1}\right) \\
=& -T_0\left(\frac{T_1u(x_1) - u(x_0)T_1}{T_0 + T_1}\right) \\
=& \frac{u(x_0) - u(x_1)}{\frac{1}{T_1} + \frac{1}{T_0}}.
\end{aligned} \tag{2.35}$$

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

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} &= t^0u(x_0) + t^1u(x_1) \\
\frac{u(x_0) - u(x_1)}{\frac{1}{T_1} + \frac{1}{T_0}} &= t^0u(x_0) + t^1u(x_1) \\
\Rightarrow t^0 &= \frac{1}{\frac{1}{T_1} + \frac{1}{T_0}}.
\end{aligned} \tag{2.36}$$

Hence, the transmissibility is the *harmonic mean* of the local 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 2.4. 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 (2.31) and (2.32) 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

$$\boldsymbol{n}_2 \cdot \boldsymbol{K}\boldsymbol{n}_1 = 0 \tag{2.37}$$

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 2.4, our grid would not be

K-orthogonal as the control volumes are not a parallelograms. All meshes with orthogonal control volumes are K-orthogonal if the permeability is isotropic.

reference a figure showing the failed convergence of TPFA

2.2.2 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 [6], we only give a brief introduction. Consider the control volumes in 2.5.

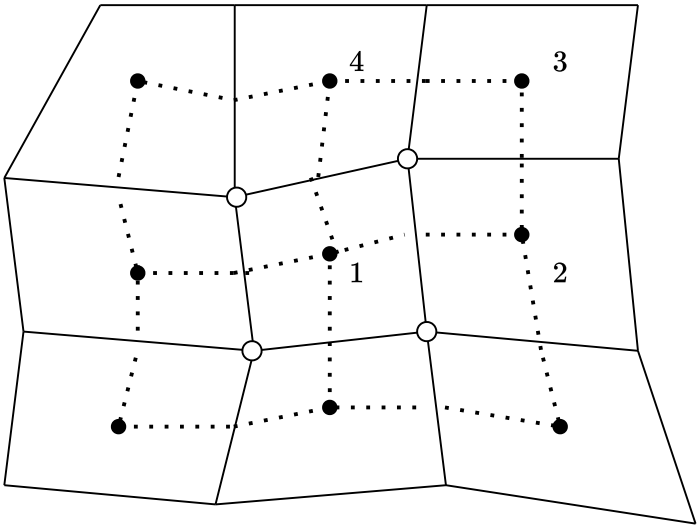


Figure 2.5: The solid lines are the control volumes, the dashed lines are the dual mesh connecting the cell centers, going through the midpoints of each edge. The solid circles are cell centers, the white circles are grid points.

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

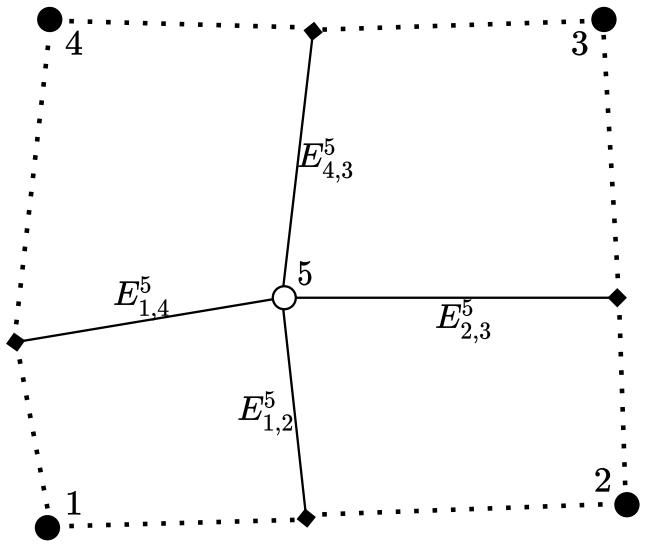


Figure 2.6: The four subcells in the interaction region corresponding to cells $1, 2, 3, 4$ and grid point 5. Here, $\mathcal{R}_5 = \{1, 2, 3, 4\}$.

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

$$\tilde{q}_{E^n_{i,j}} = \sum_{k \in \mathcal{R}_n} t^{k,n}_{i,j} u^k \approx \int_{E_{i,j}} \hat{\boldsymbol{n}}_j^T \boldsymbol{K} \nabla u \, ds \quad i, j \in \mathcal{R}_n \tag{2.38}$$

for the flux through each half edge $E^n_{i,j}$ in the interaction region corresponding to grid point n (figure 2.6). Where $\mathcal{R}_n$ is the index set of the four cells neighbouring grid point n.

We assume for now that the potential is linear in each of the four sub cells in the interaction region, figure 2.6. 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 of the midpoints of the edges.

By these assumptions on flux and potential continuity, the linear potential in each sub cell is well defined given values at the cell center. 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 2.6 and equation (2.38) it would look like

$$\boldsymbol{T}^5 = \begin{bmatrix} t^{1,5}_{1,2} & t^{2,5}_{1,2} & t^{3,5}_{1,2} & t^{4,5}_{1,2} \\ t^{1,5}_{2,3} & t^{2,5}_{2,3} & t^{3,5}_{2,3} & t^{4,5}_{2,3} \\ t^{1,5}_{4,3} & t^{2,5}_{4,3} & t^{3,5}_{4,3} & t^{4,5}_{4,3} \\ t^{1,5}_{1,4} & t^{2,5}_{1,4} & t^{3,5}_{1,4} & t^{4,5}_{1,4} \end{bmatrix}, \tag{2.39}$$

where each row corresponds to the flux over some half edge, and each column corresponds to some cell center. Computing (2.39) involves inverting a four by four matrix with coefficients depending on the mesh and permeability, see [6] for details. Finally, we assemble the system of equations (2.29) 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:

$$\sum_{j \in \mathcal{S}_i} (\tilde{q}_{\hat{E}^1_{i,j}} + \tilde{q}_{\hat{E}^2_{i,j}}) = |\Omega_i| f(x_i) \tag{2.40}$$

Where $\hat{E}^1_{i,j} = E^n_{i,j}$ for some n, see (2.38). 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 we get a six point flux stencil. In other words, for each j in (2.40), 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 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 indexes corresponding to cell i and its eight neighbour cells.

The O-method is consistent for non K-orthogonal grids, and reduces to 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 [7], Nordbotten and Keilegavlen describes a framework of MPFA methods where the O-method is a special case. They consider the problem of finding the four linear potential functions in each interaction region that minimizes the discontinuity across the edges. The discontinuity should be minimized given that the functions respect cell center potential values, that the flux models the constitute law and flux continuity. The O-method is then defined for some special cost function measuring the discontinuity. Other methods, with the potential continuity at other places than the edge midpoint, are also common.

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 my 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 oscillations in the solution and violation of the minimum principle. 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,[5]] page 175. When solving for example the Richards' equation, violating the minimum principle can lead to air bubbles being formed spontaneously in the saturated region. For a discussion on monotonicity see [8].

2.2.3 L-method

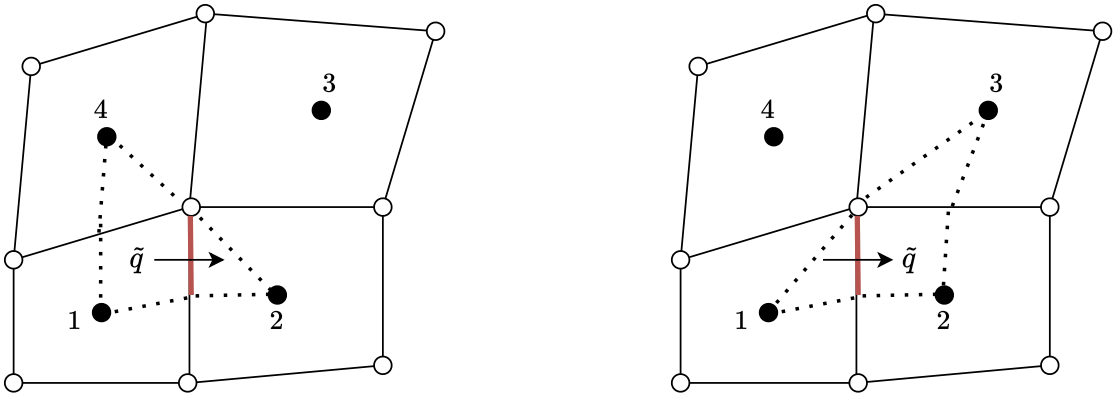
> The L-method is the Ferrari of discretization techniques for porous media flow problems, while conformal finite elements is the Volvo.
>
> *Professor Jan Martin Nordbotten*

Like the O-method, the L-method is also a multi-point flux approximation method. It was introduced in [9], 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 potential continuity at the two edges.

We have two choices of flux stencil for each half edge, see figure 2.7. 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 T_1, then

$$\begin{aligned} &\text{if } |t_1^1| < |t_2^2| \\ &\text{choose } T_1 \text{ else} \\ &\text{choose } T_2. \end{aligned} \tag{2.41}$$



(a) T_1 (b) T_2

Figure 2.7: The two choices of which cell centers to use for computing the flux over the half edge in red. We call the hexagons spanned T_1 and T_2 for L-triangles, as they consists of three cell centers.

A cheap intuition behind (2.41) is that if $|t_1^1| < |t_2^2|$, it is more likely that $sgn(t_1^1) = sgn(t_1^4)$ and if not, $sgn(t_2^2) = sgn(t_2^3)$ is more likely. This is due to the fact that $\sum t^i = 0$. Choosing L-triangle as in (2.41) 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 [10] 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:

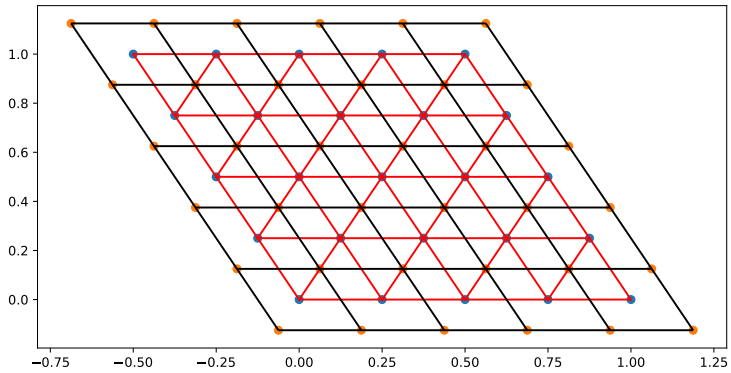
$$\begin{aligned} \sum_{j=1}^{4}(\tilde{q}_{i,j}^1 + \tilde{q}_{i,j}^2) &= |\Omega_i| f(x_i) \\ \sum_{j=1}^{4}(\sum_{k=1}^{3} t_{i,j}^{k,1} u^k + \sum_{k=1}^{3} t_{i,j}^{k,2} u^k) &= |\Omega_i| f(x_i). \end{aligned} \tag{2.42}$$

But the flux stencil across each edge is possibly smaller, often just four points.

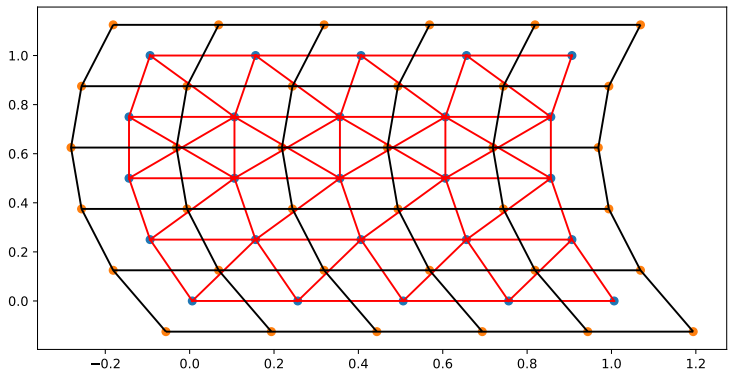
Remark 12. *In the L-method, we need to construct and solve a matrix equation twice for each half edge to compute the transmissibility coefficients, as there are always two choices. In contrast, the O-method only needs this done once for each grid point, and its four half edges.*

In figure 2.8 we see the criterion in practice for a homogenous medium: In figure 2.8a all L-triangles are used by two half-edges, and they are chosen in the

same way throughout the domain. In figure 2.8b 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 chose similarly.



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

Figure 2.8: Examples of L-triangles(in red) in a domain with homogenous permeability tensor.

The observation in figure 2.8a can be stated as a theorem:

Theorem 2.2.1 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[11])**.** *For homogeneous media and uniform parallelogram grids, the MPFA L-method has a seven-point cell stencil for the discretization of each interior cell, ie. the discretization of each cell is a seven point stencil including the center cell and the six closest potential cells, as shown in 2.8a.*

In case of 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 2.9 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 2.9 given by

$$K_{n,m} = (e^{\hat{x}} - 1)^2 \tag{2.43}$$

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 2.9.

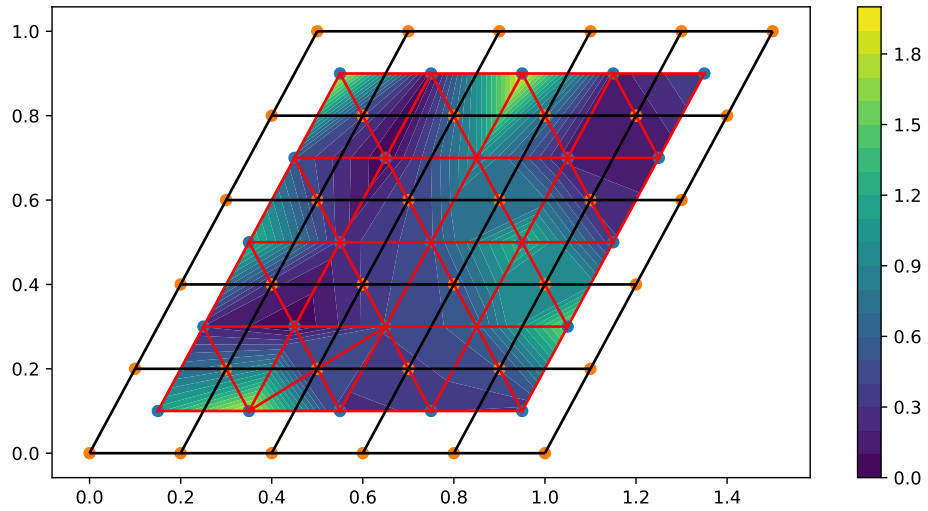


Figure 2.9: L-triangles on a random permeability.

For homogenous media the L-method becomes easier to 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),[10])**.** *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*

$$\tilde{q}_e = -\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_e \tag{2.44}$$

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.

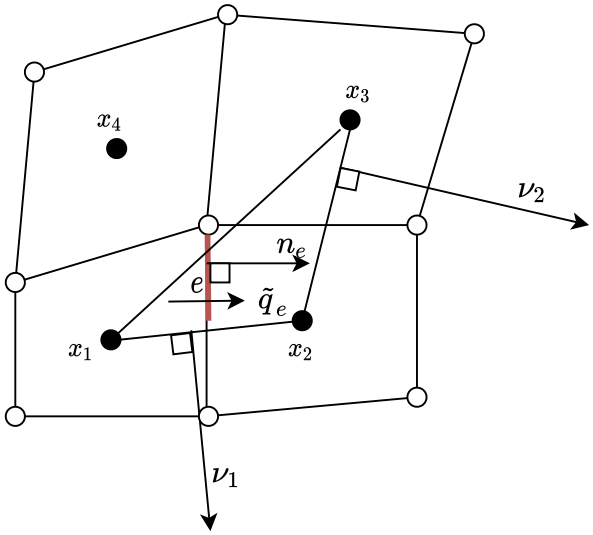


Figure 2.10: Simplified L-triangle, the original L-triangle i shown in figure 2.7b or 2.11. The vector ν_1 is perpendicular to the edge between x_1 and x_2, with the same length as the edge it is perpendicular to. Same for ν_2, with x_2 and x_3.

Moreover, the gradient ∇u, is given by:

$$\nabla u = -\frac{1}{2F}[(u_1 - u_2)\nu_2 + (u_3 - u_2)\nu_1], \tag{2.45}$$

where F is the the area of the simplified L-triangle with corners x_1, x_2 and x_4, see figure 2.10. An expression like (2.45) can be obtained for the other choice of L-triangle as well.

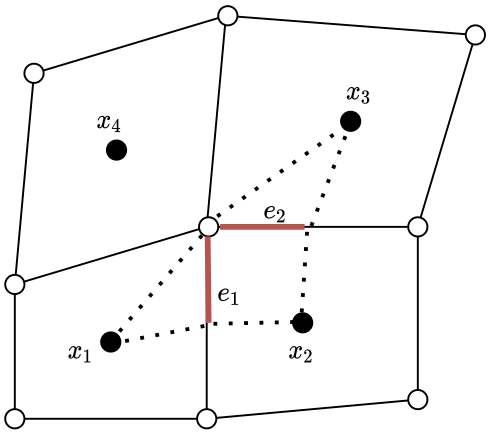


Figure 2.11: Original L-triangle with notations in proof.

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 2.11. 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:

$$[\nabla u \cdot \boldsymbol{t}_{e_1}] = 0. \tag{2.46}$$

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

$$[\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}^T \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}\boldsymbol{n}_{e_1}] = 0. \tag{2.47}$$

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's only possible for homogenous media.*

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

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:

$$\begin{cases} \partial_t u - \nabla \cdot \boldsymbol{K}\nabla u = f, & \text{in } \Omega \times (0, T] \\ u = 0, & \text{on } \partial\Gamma_D \times (0, T] \\ -\boldsymbol{K}\nabla u = g_N, & \text{on } \partial\Gamma_N \times (0, T) \\ u = u_0, & \text{on } \Omega \times \{t = 0\} \end{cases} \tag{2.48}$$

The well-posedness of (2.48) is discussed in chapter seven of [4], it requires a more detailed discussion of Sobolev spaces and Bochner spaces, ie. 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. The two choices we have left is the forward Euler (explicit) and the backward Euler (implicit). The obvious choice is 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 ([5], chapter 7). There, it is shown that 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 (2.48) 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

$$\begin{aligned} u^n - \tau \nabla \cdot \boldsymbol{K} \nabla u^n &= \tau f^n + u^{n-1} && x \in \Omega \\ u^n &= 0 && x \in \partial\Gamma_D \\ \boldsymbol{K}\nabla u &= g_N && x \in \partial\Gamma_N \\ u^0 &= u_0 && x \in \Omega. \end{aligned} \tag{2.49}$$

The above equation shows that this time discretization is implicit, ie. we cannot solve (2.49) for u^n with simple algebraic manipulation. Now, we have an elliptic problem (2.49) 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 (2.49) 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

$$\langle u^n, v\rangle_0 + \tau \langle \boldsymbol{K}\nabla u^n, \nabla v\rangle_0 = \tau \langle f^n, v\rangle_0 + \left\langle u^{n-1}, v\right\rangle_0 \tag{2.50}$$

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

$$(\boldsymbol{B} + \tau \boldsymbol{A})\hat{\boldsymbol{u}}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}\hat{\boldsymbol{u}}^{n-1}, \tag{2.51}$$

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

Finite volume approach

As before, we divide our domain Ω into d control volumes $\{\Omega_i\}_i$. Either, one can write the heat equation(2.48) in conservation form on each control volume

$$\partial_t \int_{\Omega_i} u \; dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} \; dx = \int_{\Omega_i} f \; dx, \tag{2.52}$$

and discretize the first term with backward Euler, or one can make sure the semi-discrete heat equation (2.48) holds for each control volume and use the divergence theorem. Both ways, we end up with

$$\int_{\Omega_i} u^n \, dx - \tau \int_{\partial\Omega_i} \boldsymbol{K}\nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \tau \int_{\Omega_i} f^n \, dx + \int_{\Omega_i} u^{n-1} \, dx, \tag{2.53}$$

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

$$\int_{\Omega_i} u^n(x_i) \, dx - \tau \int_{\partial\Omega_i} \boldsymbol{K}\nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \tau \int_{\Omega_i} f^n(x_i) \, dx + \int_{\Omega_i} u^{n-1}(x_i) \, dx. \tag{2.54}$$

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

$$(\boldsymbol{B}^V + \tau \boldsymbol{A}^V)\tilde{\boldsymbol{u}}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}^V \tilde{\boldsymbol{u}}^{n-1} \tag{2.55}$$

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.

If $\boldsymbol{A} = \boldsymbol{A}^V$, then the discretization of the constitutive law is the same for both the finite volume and the finite element method. As we will see later, this is challenging.

2.3 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 (1.12) leads to a system of non-linear equations. We start by discussing this in a general setting: Find $x \in U$ such that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \text{ where } f : U \subset \mathbb{R}^n \to \mathbb{R}^n. \tag{2.56}$$

The solution of (2.56) is called a *root*, it is almost always found using an iterative method.

A common iterative scheme to solve (2.56) 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:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} - D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{j-1}). \tag{2.57}$$

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

Theorem 2.3.1. *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*

$$|x_{j+1} - \overline{x}| \leq \frac{1}{\delta}|x_j - \overline{x}|^2 < |x_j - \overline{x}|. \tag{2.58}$$

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

$$0 = f(\overline{x}) = f(x_j - e_j) = f(x_j) - f'(x_j)e_j + \frac{f''(\psi)e_j^2}{2}. \tag{2.59}$$

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

$$\begin{aligned} e_{j+1} = x_{j+1} - \overline{x} &= x_j - \frac{f(x_j)}{f'(x_j)} - \overline{x} \\ &= e_j - \frac{f(x_j)}{f'(x_j)} \\ &= \frac{e_j f'(x_j) - f(x_j)}{f'(x_j)} \end{aligned} \tag{2.60}$$

By the Taylor expansion around x_j, (2.59), we get

$$f'(x_j) = \frac{f(x_j)}{e_j} + \frac{f''(\psi)e_j}{2}. \tag{2.61}$$

Inserting this into (2.60), we get the equality

$$e_{j+1} = \frac{e_j^2 f''(\psi)}{2f'(x_j)}. \tag{2.62}$$

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

$$|e_1| \leq \frac{2}{2\delta}|e_0|^2 < |e_0| \tag{2.63}$$

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

$$|e_{j+1}| < |e_j|, \tag{2.64}$$

and the quadratic convergence

$$|e_{j+1}| \leq \frac{1}{\delta}|e_j|^2 \tag{2.65}$$

□

For a similar result in more dimensions see (Knabner [5], chapter 8). One apparent drawback of this method is that it is only locally convergent, ie. 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

$$D\boldsymbol{f}(\boldsymbol{x}_{j-1})\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}), \tag{2.66}$$

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

$$L\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}). \tag{2.67}$$

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 2.3.2. *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

$$e_{j+1} = x_j - \frac{f(x_j)}{L} - \overline{x} = e_j - \frac{f(x_j)}{L}. \tag{2.68}$$

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

$$0 = f(\overline{x}) = f(x_j - e_j) = f(e_j) - f'(\psi)e_j \Rightarrow e_j = \frac{f(x_j)}{f'(\psi)}. \tag{2.69}$$

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

$$|e_{j+1}| = |e_j(1 - \frac{f'(\psi)f(x_j)}{f(x_j)L})| \leq |e_j||1 - \frac{f'(\psi)}{L}| < |e_j|. \tag{2.70}$$

□

Chapter 3

Convergence for 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, such as (2.49). That is, we prove equivalence for 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 as before. Once equivalence is obtained, we prove convergence for the finite element method using standard techniques in the spirit of 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. Moreover, that the L-method can be used as a locally mass conservative flux recovery algorithm on the finite element solution. See figure for a comparison of the methods. todo

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),[11]), there, convergence is proved for the Poisson equation, ie. without the u term.

3.1 Modified MPFA-L method

First of all, we assume that we have a uniform parallelogram grid, as in 2.8a. As we saw in the previous chapter, one gets with the finite volume method the following

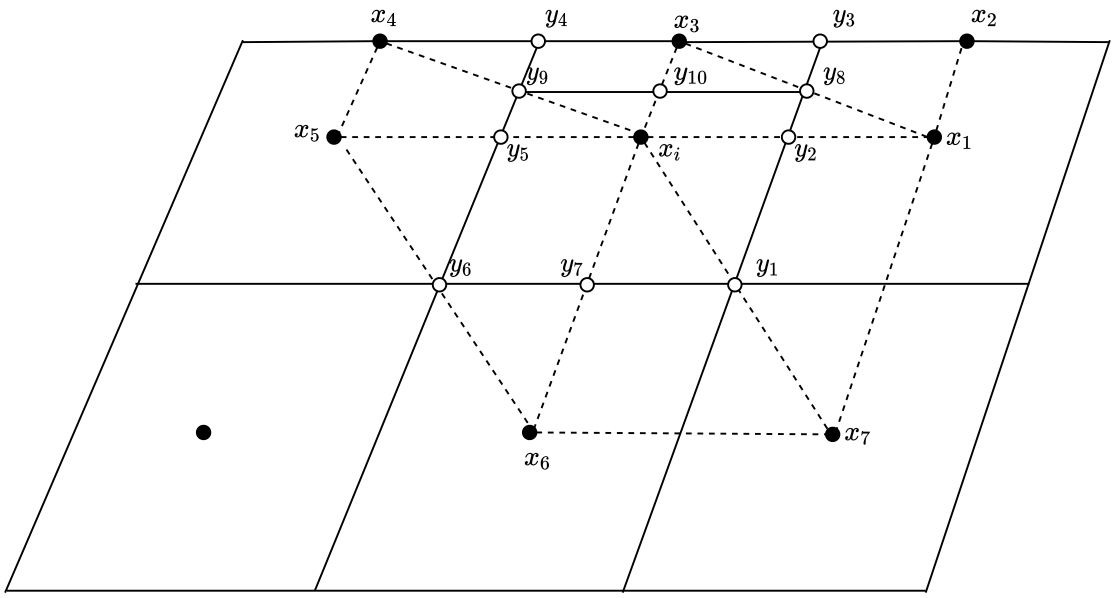


Figure 3.1: Control volumes in solid lines and interaction regions in dashed lines at the boundary.

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).

We need to modify the Neumann boundaries, this is to be expected as finite element methods have degrees of freedom on the boundaries as opposed to finite volume methods. We will also see how we could enforce Dirichlet boundary conditions in a way that is equivalent to the finite element method. On the **interior** control volumes, we use the original MPFA-L method already covered.

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 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

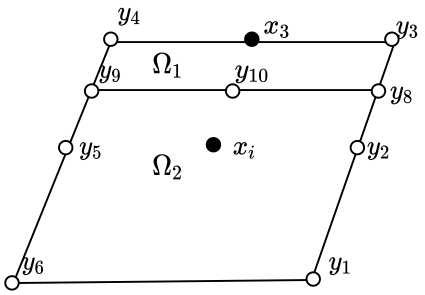


Figure 3.2: Control volume along top boundary.

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 [11] for details.

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 dot products that make up the mass matrix and load vector, become mass conservative in each control volume.

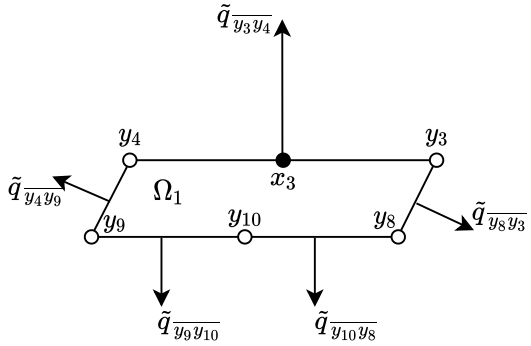


Figure 3.3: The fluxes on the Dirichlet boundary.

We need some notation so that we can distinguish between nodes in the interior, at cell centers along the boundary and at the boundary. In addition, corner cells introduce edge cases. Let $\mathcal{N}_h^*$ be a set of indexes 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. $\mathcal{N}_h^b$ 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.

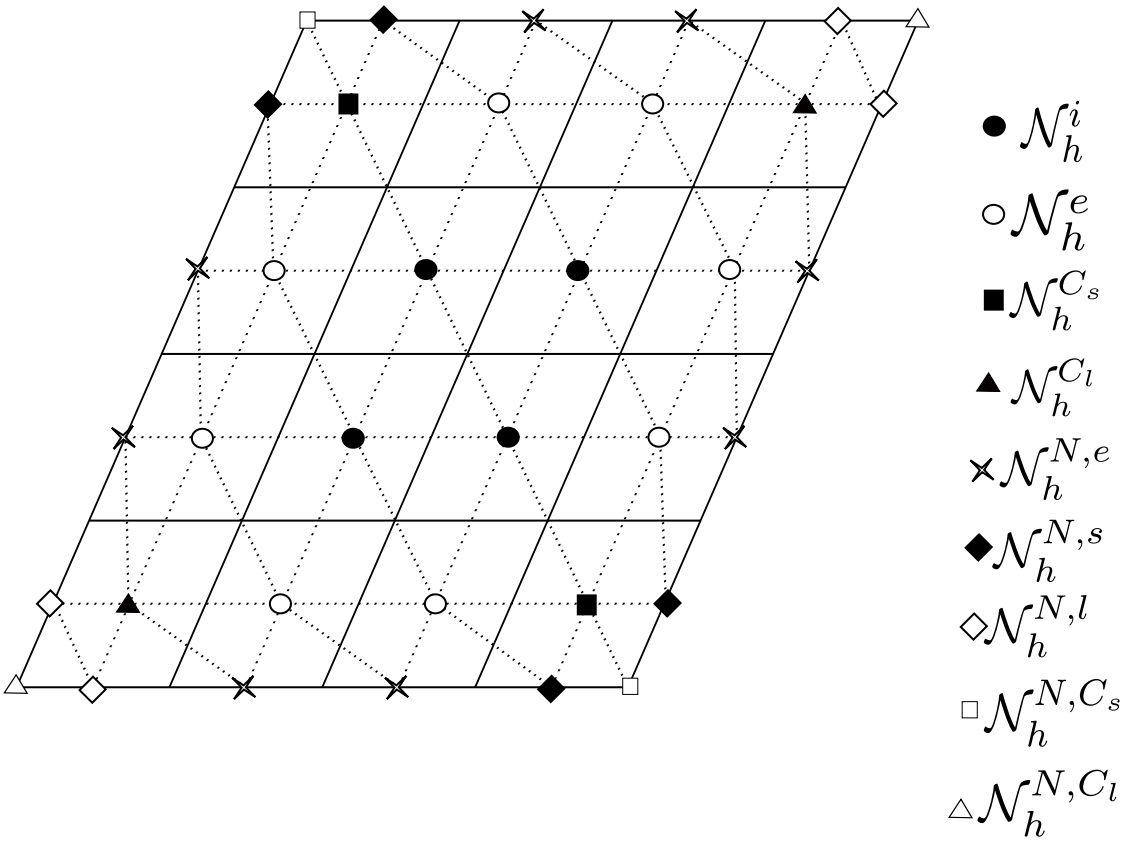


Figure 3.4: A paralellogram mesh with finite element triangles in dotted lines and control volumes in solid lines. In this case we have a pure Neumann problem.

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:

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.*

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

And

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

Where

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

In interior cells, $i \in \mathcal{N}_h^i$, we have $D_i = \Omega_i$. 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 paralellogram closest to the boundary, ie. $D_i = \Omega_2$ 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, ie. $D_i = \Omega_1$ in figure 3.2*
- *On the corners there are special definitions, see (Cao Wolmuth [11],2009)*

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 is often referred to as *mass lumping*, see for [12] for examples.

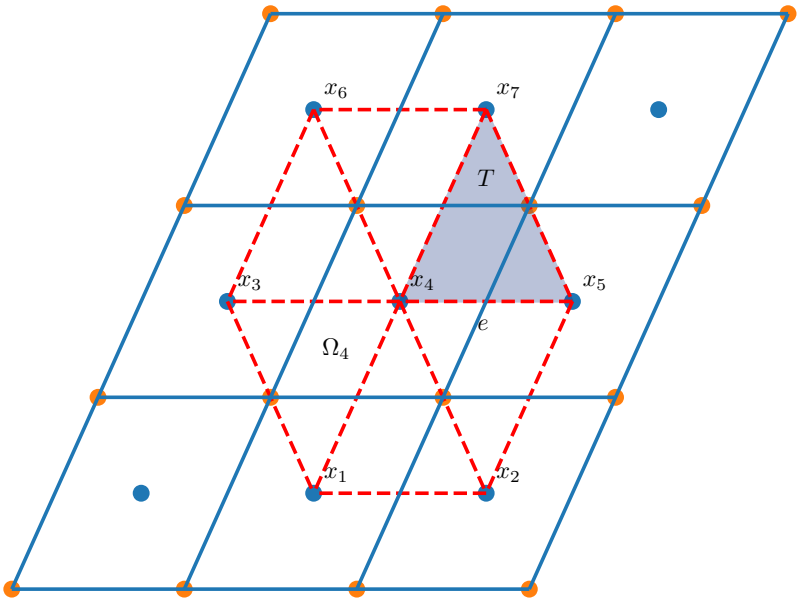


Figure 3.5: The support of ϕ_4

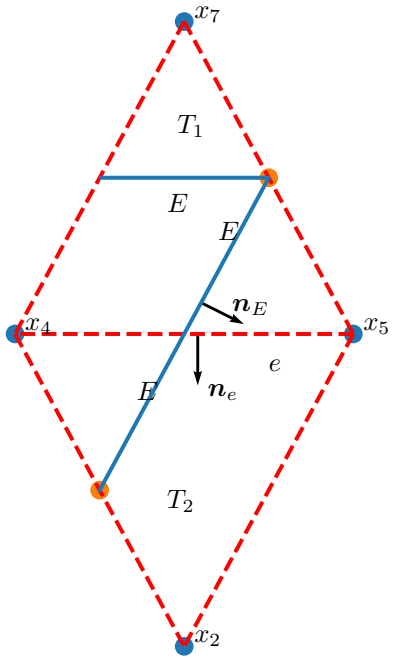


Figure 3.6: Notation in the proof

Now we can state the equivalence theorem:

Theorem 3.2.1. *The modified finite element* (3.5) *method and the modified MPFA-L method are equivalent on uniform parallelogram grid for the time discretized heat equation, ie.* (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 $T \in \tau_h \bigcap \text{supp}(\phi_i)$ 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:

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

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

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

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:

$$\int_{\Omega_i} u_h(x_i) \; dx - \sum_{E\in\partial\Omega_i} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_i} f \; dx. \tag{3.10}$$

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 (2.44). Hence, if $\hat{u}_h$ is the solution to (3.1) with the original L-method in the interior, then $\hat{u}_h(x_i) = u_h(x_i)$ 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:

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_i \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla \phi_i \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_i \right\rangle_{0,\Omega}. \tag{3.11}$$

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

$$\langle \boldsymbol{K} \nabla u_h, \nabla \phi_i \rangle_{0,\Omega} = - \sum_{E \in \partial \Omega_{i,2}} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n}_E |E|, \tag{3.12}$$

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 (3.11) is equivalent to:

$$\int_{\Omega_{i,2}} u_h(x_i)\ dx - \sum_{E \in \partial \Omega_{i,2}} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n}_E |E| = \int_{\Omega_{i,2}} f(x_i)\ dx. \tag{3.13}$$

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

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla \phi_j \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \left\langle g, \hat{I}_{\Gamma_N} \phi_j \right\rangle_{0,\Gamma_N}. \tag{3.14}$$

Similarly as in (3.7) we have

$$\begin{aligned}
\langle \boldsymbol{K} \nabla u_h, \nabla \phi_j \rangle_0 &= \int_{\text{supp}(\phi_j)} (\nabla u_h)^T \boldsymbol{K} \nabla \phi_j\ dx = \sum_{T \in \text{supp}(\phi_j)} \int_T (\nabla u_h)^T \boldsymbol{K} \nabla \phi_j\ dx \\
&= \sum_{T \in \text{supp}(\phi_j)} \left(\int_{\partial T} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n} \phi_j\ ds - \int_T \nabla \cdot \boldsymbol{K} \nabla u_h \phi_j\ dx \right) \\
&= \sum_{T \in \text{supp}(\phi_j)} \int_{\partial T} (\boldsymbol{K} \nabla u_h)^T \boldsymbol{n} \phi_j\ ds.
\end{aligned} \tag{3.15}$$

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

$$\begin{aligned}\sum_{T\in\text{supp}(\phi_j)} \int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_j \, ds &= \sum_{e\in\text{supp}(\phi_j)} \int_e \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}}\right) \phi_j \, ds \\ &+ \int_{\Gamma_N \bigcap \text{supp}(\phi_j)} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_j \, ds \\ &= \sum_{e\in\text{supp}(\phi_j)} \left((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,0}} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{e,1}}\right) \frac{|e|}{2} \, ds \\ &+ (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}|E_{\Gamma_N}| \, ds \\ &= - \sum_{E\in\partial\Omega_j\setminus\Gamma_N} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E|E|\end{aligned} \tag{3.16}$$

Combining (3.15) and (3.16) and using the definition of $\hat{I}_h$, definition 11, we get that (3.14) is equivalent to:

$$\int_{\Omega_{j,1}} u_h(x_i) \, dx - \sum_{E\in\partial\Omega_{j,1}} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E|E| = \int_{\Omega_{j,1}} f(x_i) \, dx. \tag{3.17}$$

Where $\Omega_{j,1}$ is as Ω_1 in figure 3.2. Now, (3.13) and (3.17) 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 $i \in \mathcal{N}_h^e$, ie. the cell center. Then, our modified finite element method

$$\left\langle \hat{I}_h u_h, \hat{I}_h \phi_j \right\rangle_{0,\Omega} + \langle \boldsymbol{K}\nabla u_h, \nabla\phi_j \rangle_{0,\Omega} = \left\langle f, \hat{I}_h \phi_j \right\rangle_{0,\Omega} \tag{3.18}$$

is equivalent to

$$\int_{\Omega_{i,2}} u_h(x_i) \, dx - \sum_{E\in\partial\Omega_{i,2}} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E|E| = \int_{\Omega_{i,2}} f \, dx, \tag{3.19}$$

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:

$$- \sum_{E\in\Omega_i\setminus\Gamma_D} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E|E| + \sum_{E\in\Omega_{i,1}\setminus\Gamma_D} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E|E| + \int_{\Omega_{i,1}} f \, dx = \int_{\Omega_i} f \, dx. \tag{3.20}$$

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. Corner cells!

□

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 [5])**.** *Suppose there exists some $\alpha > 0$ such that for all $h > 0$ and $v_h \in V_h$*

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

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

$$\|u - u_h\|_1 \leq C \left\{ \inf_{v_h \in V_h} \left\{ \|u - v_h\|_1 + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_1} + \sup_{w_h \in V_h} \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_1} \right\} \right\} \tag{3.21}$$

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

$$a_h(u_h, v_h) = \left\langle \hat{I}_h u_h, \hat{I}_h v_h \right\rangle_{0,\Omega} + \langle \boldsymbol{K} \nabla u_h, \nabla v_h \rangle_{0,\Omega} . \tag{3.22}$$

And the linear form:

$$b_h(v_h) = \left\langle F, \hat{I}_h v_h \right\rangle_{0,\Omega} + \left\langle g, \hat{I}_{\Gamma_N} v_h \right\rangle_{0.\Gamma_N} . \tag{3.23}$$

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

$$\begin{aligned} \|u_h\|_1^2 &= \|\partial_{x_1} u_h\|_0^2 + \|\partial_{x_2} u_h\|_0^2 + \|u_h\|_0^2 \\ &= \langle \nabla u_h, \nabla u_h \rangle_0 + \|u_h\|_0^2 . \end{aligned} \tag{3.24}$$

Using the Poincaré inequality on the second term:

$$\begin{aligned} \|u_h\|_1^2 &\leq \langle \nabla u_h, \nabla u_h \rangle_0 + C_\Omega \langle \nabla u_h, \nabla u_h \rangle_0 \\ &\leq \left(\frac{1 + C_\Omega}{\tau \|\boldsymbol{K}\|} \right) \tau \langle \boldsymbol{K} \nabla u_h, \nabla u_h \rangle_0 \\ &\leq \left(\frac{1 + C_\Omega}{\tau \|\boldsymbol{K}\|} \right) \left(\tau \langle \boldsymbol{K} \nabla u_h, \nabla u_h \rangle_0 + \left\langle \hat{I}_h u_h, \hat{I}_h u_h \right\rangle_0 \right) \\ &\leq \frac{1}{\alpha} a_h(u_h, u_h), \end{aligned} \tag{3.25}$$

we obtain coercivity with $\alpha = \tau \|\boldsymbol{K}\| / (1 + C_\Omega)$.

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:*

$$\left\| \hat{I}_h u_h - u_h \right\|_{0,\Omega} \leq ch|u_h|_{1,\Omega} \ \forall u_h \in V_h \tag{3.26}$$

The proof is trivial as our test space, $V_h \subset H^1(\Omega) \bigcap C(\Omega)$, consists of continuous functions. 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 such that*

$$\|u - u_h\|_1 \leq Ch(\|u\|_2 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}). \tag{3.27}$$

Proof. The hypothesis in Strang's lemma 3.3.1 on continuity and coercivity are fulfilled. We start by controlling the second term on the right hand side in (3.21), the truncation error in the bi-linear form:

$$\begin{aligned}
&\sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_1} \\
= &\sup_{w_h \in V_h} \frac{|\langle v_h, w_h \rangle + \tau \langle \boldsymbol{K} \nabla v_h, \nabla w_h \rangle - \left\langle \hat{I}_h v_h, \hat{I}_h w_h \right\rangle - \tau \langle \boldsymbol{K} \nabla v_h, \nabla w_h \rangle|}{\|w_h\|_1} \\
= &\sup_{w_h \in V_h} \frac{|\langle v_h, w_h \rangle - \left\langle \hat{I}_h v_h, w_h \right\rangle + \left\langle \hat{I}_h v_h, w_h \right\rangle - \left\langle \hat{I}_h v_h, \hat{I}_h w_h \right\rangle|}{\|w_h\|_1} \\
= &\sup_{w_h \in V_h} \frac{|\left\langle \hat{I}_h v_h - v_h, w_h \right\rangle + \left\langle \hat{I}_h v_h, \hat{I}_h w_h - w_h \right\rangle|}{\|w_h\|_1}.
\end{aligned} \tag{3.28}$$

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:

$$\begin{aligned}
&\leq \sup_{w_h \in V_h} \frac{Ch \|v_h\|_0 |w_h|_1 + \left\| \hat{I}_h v_h \right\|_0 Ch|w_h|_1}{\|w_h\|_1} \\
\leq &\sup_{w_h \in V_h} \frac{Ch \|v_h\|_0 |w_h|_1 + \left\| \hat{I}_h v_h \right\|_0 Ch|w_h|_1}{\|w_h\|_1} + \frac{ch \|v_h\|_0 \|w_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 \|w_h\|_0}{\|w_h\|_1} \\
&\leq Ch \left(\|v_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 \right).
\end{aligned} \tag{3.29}$$

The third term in (3.21), the linear form, can be controlled similarly:

$$\sup_{w_h \in V_h} \frac{l(w_h) - l_h(w_h)}{\|w_h\|_1} = \sup_{w_h \in V_h} \frac{\left\langle f, w_h - \hat{I}_h w_h \right\rangle_{0,\Omega} + \left\langle g, w_h - \hat{I}_{\Gamma_N} w_h \right\rangle_{0,\Gamma_N}}{\|w_h\|_1} \leq \sup_{w_h \in V_h} \frac{\|f\|_0 Ch \|w_h\|_1 + \|g\|_{\frac{1}{2},\Gamma_N} \left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N}}{\|w_h\|_1}. \tag{3.30}$$

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 piecewise constant functions on the boundary in each Neumann boundary triangle. Then we have:

$$\begin{aligned} \left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N} &= \sup_{0 \neq v \in H^{\frac{1}{2}}(\Omega)} \frac{\left\langle w_h - \hat{I}_{\Gamma_N} w_h, v \right\rangle_{\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}} \\ &= \sup_{0 \neq v \in H^{\frac{1}{2}}(\Omega)} \frac{\left\langle w_h - \hat{I}_{\Gamma_N} w_h, v - v_h \right\rangle_{\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}}, \end{aligned} \tag{3.31}$$

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

$$\left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N} \leq \sup_{0 \neq v \in H^{\frac{1}{2}}(\Omega)} \frac{\left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{0,\Gamma_N} \|v - v_h\|_{0,\Gamma_N}}{\|v\|_{\frac{1}{2},\Gamma_N}}. \tag{3.32}$$

What is this inner product?

By the inequality

$$\|v - v_h\|_{0,\Gamma_N} \leq Ch^{\frac{1}{2}} \|v\|_{\frac{1}{2},\Gamma_N}, \tag{3.33}$$

Does this make sense?

we can bound the right hand side of (3.32):

$$\left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N} \leq Ch^{\frac{1}{2}} \left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{0,\Gamma_N}. \tag{3.34}$$

Using (3.33) again, we get

$$\left\| w_h - \hat{I}_{\Gamma_N} w_h \right\|_{-\frac{1}{2},\Gamma_N} \leq Ch \|w_h\|_{\frac{1}{2},\Gamma_N} \leq Ch \|w_h\|_1. \tag{3.35}$$

Where the last inequality is due to the definition of the $H^{\frac{1}{2}}$ norm. Inserting this into (3.30), gives us a bound on the truncation error of our linear form:

$$\sup_{w_h \in V_h} \frac{l(w_h) - l_h(w_h)}{\|w_h\|_1} \leq Ch(\|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}). \tag{3.36}$$

Hence, from (3.21), we have the error estimate:

$$\|u-u_h\|_1 \leq \inf_{v_h \in V_h} \left\{ \|u-v_h\|_1 + Ch\left(\|v_h\|_0 + \left\| \hat{I}_h v_h \right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N} \right) \right\} \tag{3.37}$$

If we let $v_h = I_h u \in V_h$ in (3.37), we get the inequality:

$$\|u-u_h\|_1 \leq \|u-I_h u\|_1 + Ch\left(\|I_h u\|_0 + \left\| \hat{I}_h I_h u \right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N} \right) \tag{3.38}$$

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

$$\|u-I_h u\|_1 \leq Ch|u|_2. \tag{3.39}$$

If we insert this into (3.38), we get:

$$\|u-u_h\|_1 \leq Ch\left(|u|_2 + \|I_h u\|_0 + \left\| \hat{I}_h I_h u \right\|_0 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N} \right). \tag{3.40}$$

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:

$$\|u-u_h\|_1 \leq Ch\left(\|u\|_2 + \|f\|_0 + \|g\|_{\frac{1}{2},\Gamma_N} \right). \tag{3.41}$$

□

Chapter 4

Convergence of Richards' equation

In this chapter we use the results from chapter three to prove convergence of the Richards equation discretized in space with MPFA-L method, in time with backward Euler and linearized with the L-scheme. 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. The hydraulic conductivity, $\kappa(\theta(\psi))$, depends on our solution which is heterogeneous, and is thus itself heterogeneous. To remedy this we use the Kirchhoff transform

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\to \mathbb{R}^+ \\ \psi &\mapsto \int_0^\psi \kappa(\theta(\phi))\, d\phi = u. \end{aligned} \tag{4.2}$$

As previously discussed, the functions $\theta(\cdot)$ and $\kappa(\cdot)$ are continuous, monotone increasing functions, the Kirchhoff transform, $\mathcal{K}$, therefore has an inverse, $\mathcal{K}^{-1}$. We define

$$\begin{aligned} b(u) &:= \theta(\mathcal{K}^{-1}(u)) \\ k(u) &:= \kappa(\theta(\mathcal{K}^{-1}(u))). \end{aligned} \tag{4.3}$$

make assumptions ensuring that b is nicely behaved

by the chain rule, we get

$$\nabla u = \kappa(\theta(\psi))\nabla\psi. \tag{4.4}$$

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

$$\begin{cases} \partial_t b(u) - \nabla\cdot\nabla u = f, & \text{in } \Omega\times(0,T] \\ u = 0, & \text{on } \partial\Gamma_D\times(0,T] \\ -\nabla u = g_N, & \text{on } \partial\Gamma_N\times(0,T) \\ u = u_0, & \text{on } \Omega\times\{t=0\} \end{cases} \tag{4.5}$$

We start by discretizing (4.5) with the MPFA-L method, we divide our domain into d quadrilaterals (control volumes). Writing (2.52) in vector form we find $\tilde{u}_h \in \mathbb{R}^d$ such that:

$$\partial_t \boldsymbol{B}^V b(\tilde{u}_h) + \boldsymbol{A}^V \tilde{u}_h = \boldsymbol{q}^V. \tag{4.6}$$

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:

$$\boldsymbol{B}^V b(\tilde{u}_h)^n + \tau \boldsymbol{A}^V \tilde{u}_h^n = \tau \boldsymbol{q}^{Vn} + \boldsymbol{B}^V b(\tilde{u}_h)^{n-1}. \tag{4.7}$$

Now we need to linearize (4.7) with the L-scheme. We see from (2.67) that the applying this linearization leads to the equation:

$$L\boldsymbol{B}^V(\tilde{u}_h^{n,j} - \tilde{u}_h^{n,j-1}) + \tau \boldsymbol{A}\tilde{u}_h^{n,j} = -\boldsymbol{B}^V\theta(\tilde{u}_h^{n,j-1}) + \tau\boldsymbol{q}^{Vn} + \boldsymbol{B}^V\theta(\tilde{u}_h^{n-1}). \tag{4.8}$$

The above equation can be solved for $\tilde{u}_h^{n,j}$, and is in principal the same as solving a time discretized heat equation (3.1), which we have proved equivalence with the modified finite element method for.

Let us see how it would look if we discretized (4.5) with the modified finite element method. Let $u_h, v_h \in V_h$, where V_h is defined as piecewise linear functions as in definition 10. Then our semidisrectization would be to find $u_h \in V_h$ such that

$$\left\langle \partial_t \hat{I}_h b(u_h^n), \hat{I}_h v_h \right\rangle_0 + \langle \nabla u_h^n, \nabla v_h\rangle_0 = \langle q, v_h\rangle_0 \quad \forall v_h \in V_h. \tag{4.9}$$

Next we would discretize in time, our discretization now reads; given $u_h^{n-1} \in V_h$ find $u_h \in V_h$

$$\left\langle \hat{I}_h b(u_h^n), \hat{I}_h v_h \right\rangle_0 + \tau \langle \nabla u_h^n, \nabla v_h\rangle_0 = \tau \left\langle q^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h b(u_h^{n-1}), \hat{I}_h v_h \right\rangle_0 \tag{4.10}$$

for all $v_h \in V_h$. Finally we also linearize this with the L-scheme. Following the same procedure as for (4.8) we get; given $u_h^{n-1}, u_h^{n,j-1} \in V_h$ find $u_h^{n,j} \in V_h$ such that

$$\begin{aligned} &L\left\langle \hat{I}_h u_h^{n,j} - \hat{I}_h u_h^{n,j-1}, \hat{I}_h v_h \right\rangle_0 + \tau \left\langle \nabla u_h^{n,j}, \nabla v_h \right\rangle \\ &= \tau \left\langle q^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h u_h^{n-1}, \hat{I}_h v_h \right\rangle_0 - \left\langle \hat{I}_h b(u_h^{n,j-1}), \hat{I}_h v_h \right\rangle_0 \end{aligned} \tag{4.11}$$

for all $v_h \in V_h$. Since the mass matrix is diagonal with our modified finite element method, we can express (4.11) as a matrix-vector equation similar to the MPFA-L method (4.8). Let $\hat{u}_h^{n,j} \in \mathbb{R}^d$ be the coefficients that together with the basis functions that represents $u_h^{n,j} \in V_h$. Then our (4.5) discretized with modified finite elements in space, backward euler in time and L-scheme linearization reads; given $\hat{u}_h^{n,j-1}, \hat{u}_h^{n-1} \in \mathbb{R}^d$ find $\hat{u}_h^{n,j} \in \mathbb{R}^d$ such that

$$L\boldsymbol{B}(\hat{u}_h^{n,j} - \hat{u}_h^{n,j-1}) + \tau \boldsymbol{A}\hat{u}_h^{n,j} = -\boldsymbol{B}\theta(\hat{u}_h^{n,j-1}) + \tau \boldsymbol{q}^n + \boldsymbol{B}b(\hat{u}_h^{n-1}). \tag{4.12}$$

We can now state an important result for obtaining a convergence estimate for Richards' equation.

Lemma 4.0.1. *The system* (4.8) *obtain from MPFA-L method for discretizing the simplified Richards' equation* (4.5)*, is equivalent to the one obtained with the modified finite element method* (4.12)*. Moreover we have the equalities:*

$$\boldsymbol{A}^V = \boldsymbol{A}, \quad \boldsymbol{B}^V = \boldsymbol{B}, \quad \boldsymbol{q}^V = \boldsymbol{q}, \quad \hat{u}_h = \tilde{u}_h \tag{4.13}$$

Proof.

todo

follows from equivalence chapter. □

How this fit in with the convergence proof for Richards' equation with MPFA can be seen in the figure below.

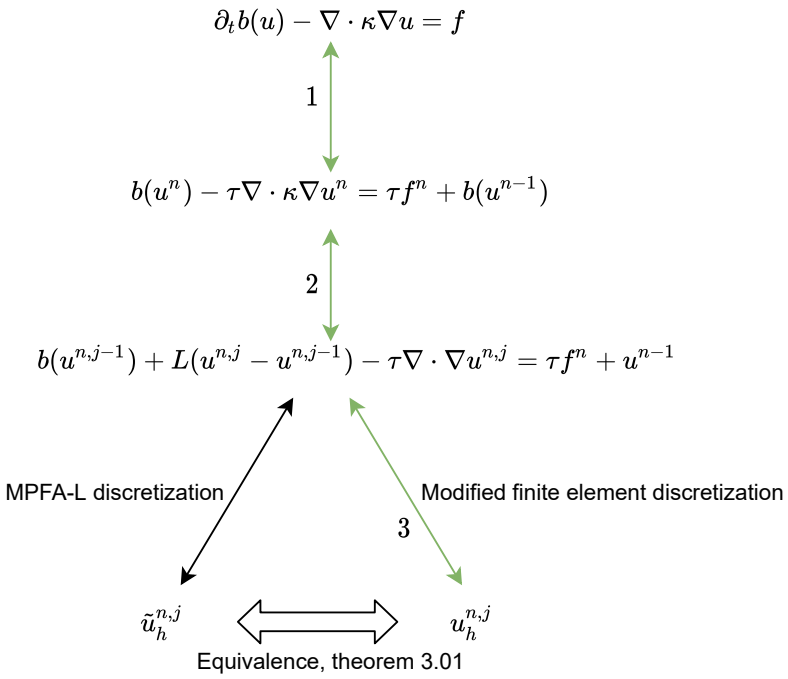


Figure 4.1

Theorem 4.0.2. *Richards' equation after Kirchoff transform* (4.5) *discretized with backward Euler in time, L-scheme linearization and MPFA-L discretization in space* (4.8) *converges,*

$$\tilde{u}_h^{n,j} \to u, \tag{4.14}$$

as $j \to \infty$, $h \to 0$ *and* $\tau \to 0$.

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

$$\left\|u_h^{n,j} - u\right\|_1 \leq \left\|u_h^{n,j} - u^{n,j}\right\|_1 + \left\|u^{n,j} - u^n\right\|_1 + \|u^n - u\|_1 . \tag{4.15}$$

1. The first term is the error of solving an elliptic problem

$$u^{n,j} - \frac{\tau}{L}\nabla \cdot \nabla u^{n,j} = \frac{Lu^{n,j-1} + \tau f^n - b(u^{n,j-1}) + u^{n-1}}{L}. \tag{4.16}$$

By theorem 3.3.3 we have the error estimate

$$\left\|u_h^{n,j} - u^{n,j}\right\|_1 \leq Ch(\left\|u^{n,j-1}\right\|_2 + \left\|\frac{Lu^{n,j-1} + \tau f^n - b(u^{n,j-1}) + u^{n-1}}{L}\right\|_0 + \|g\|_{\frac{1}{2},\Gamma_N}). \tag{4.17}$$

2. The second term, $\|u^{n,j} - u^n\|_1$,...

□

Convergence of L-scheme for modified FEM This is done (List, Radu, 2016,[14]) for the normal finite element method, but we do the proof here for our modified finite element method. First we need assumptions on the parametrizations κ and θ:

A 1 The water content parametrization $\theta(\cdot)$ is monotonically increasing with $sup|\theta'| = L_\theta$ and Lipschitz continuos.

A 2 The permeability κ is positive.

Theorem 4.0.3. *Assume **A 1** and **A 2** above and that the constant L is chosen such that $L \geq L_\theta$. Then the L-scheme (??) converges linearly.*

Proof. Let as before $e^{n,j} = u^{n,j} - u^n$ be the iteration error. We start by subtracting (??) from (??) and obtain:

$$\begin{aligned}\left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h v_h \right\rangle_0 + L\left\langle \hat{I}_h e_h^{n,j} - \hat{I}_h e_h^{n,j-1}, \hat{I}_h v_h \right\rangle_0 \\ +\tau\left\langle \kappa(\nabla u_h^{n,j} - \nabla u_h^n), \nabla v_h \right\rangle_0 = 0\end{aligned} \tag{4.18}$$

Now we test with $v_h = e^{n,j}$:

$$\begin{aligned}\left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h e^{n,j} \right\rangle_0 + L\left\langle \hat{I}_h e_h^{n,j} - \hat{I}_h e_h^{n,j-1}, \hat{I}_h e^{n,j} \right\rangle_0 \\ +\tau\left\langle \kappa\nabla e^{n,j}, \nabla e^{n,j} \right\rangle_0 = 0\end{aligned} \tag{4.19}$$

We use the identity $\langle x - y, x\rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|x-y\|^2 - \frac{1}{2}\|y\|^2$ and some algebraic manipulation to obtain:

$$\begin{aligned}\left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h e^{n,j-1} \right\rangle + \left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h e^{n,j} - \hat{I}_h e^{n,j-1} \right\rangle \\ +\frac{L}{2}\left\|\hat{I}e^{n,j}\right\|^2 + \frac{L}{2}\left\|\hat{I}e^{n,j} - \hat{I}e^{n,j-1}\right\|^2 - \frac{L}{2}\left\|\hat{I}e^{n,j-1}\right\|^2 \\ +\tau\left\langle \kappa\nabla e^{n,j}, \nabla e^{n,j} \right\rangle_0 = 0\end{aligned} \tag{4.20}$$

Then we put some terms on the right hand side:

$$\begin{aligned}
&\left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h e^{n,j-1} \right\rangle + \frac{L}{2}\left\|\hat{I}e^{n,j}\right\|^2 \\
&+\frac{L}{2}\left\|\hat{I}e^{n,j} - \hat{I}e^{n,j-1}\right\|^2 + \tau\left\langle \kappa\nabla e^{n,j}, \nabla e^{n,j}\right\rangle_0 = \\
-&\left\langle \hat{I}_h\theta(u_h^{n,j-1}) - \hat{I}_h\theta(u_h^n), \hat{I}_h e^{n,j} - \hat{I}_h e^{n,j-1} \right\rangle + \frac{L}{2}\left\|\hat{I}e^{n,j-1}\right\|^2
\end{aligned} \tag{4.21}$$

Now we use the Cauchy-Schwarz inequality, and the monotonicity **A 1** on the first term. Similarly we use Cauchy-Schwarz and **A 2** on the second term. Finally we use Young's inequality on the first term on the right hand side.

$$\begin{aligned}
&\frac{1}{L_\theta}\left\|\hat{I}_h(\theta(u^{n,j-1} - \theta(u^n)))\right\|^2 + \frac{L}{2}\left\|\hat{I}e^{n,j}\right\|^2 \\
&+\frac{L}{2}\left\|\hat{I}e^{n,j} - \hat{I}e^{n,j-1}\right\|^2 + \tau\kappa_m\left\|\nabla e^{n,j}\right\|^2 \\
\leq &\frac{1}{2L}\left\|\hat{I}_h(\theta(u^{n,j-1}) - \theta(u^n))\right\|^2 + \frac{L}{2}\left\|\hat{I}_h(e^{n,j} - e^{n,j-1})\right\|^2 + \frac{L}{2}\left\|\hat{I}e^{n,j-1}\right\|^2
\end{aligned} \tag{4.22}$$

Next we can use Poincaré inequality:

$$\frac{L}{2}\left\|\hat{I}_h e^{n,j}\right\|^2 + \frac{\tau\kappa_m}{C_\Omega}\left\|e^{n,j}\right\|^2 \leq (\frac{1}{2L} - \frac{1}{L_\theta})\left\|\hat{I}_h(\theta(u^{n,j-1} - \theta(u^n)))\right\|^2 + \frac{L}{2}\left\|\hat{I}e^{n,j-1}\right\|^2 \tag{4.23}$$

Since $L_\theta \leq L$ we can remove the first term on the right side from the inequality. We can also use the lemma 3.3.2 about the interpolation error:

$$\left\|\hat{I}_h e^{n,j}\right\|^2 = \left\|\hat{I}_h e^{n,j} - e^{n,j} + e^{n,j}\right\|^2 \leq (ch\left\|e^{n,j}\right\| + \left\|e^{n,j}\right\|)^2 = (1+ch)^2\left\|e^{n,j}\right\|^2 \tag{4.24}$$

Now (4.23) becomes:

$$(\frac{L}{2} + \frac{\tau\kappa_m}{C_\Omega(1+ch)^2})\left\|\hat{I}_h e^{n,j}\right\|^2 \leq \frac{L}{2}\left\|\hat{I}_h e^{n,j-1}\right\|^2 \tag{4.25}$$

□

Chapter 5

Numerical results

convergence for homogenous elliptic model problem

The convergence tests in this section are similar to some of the tests done in chapter three of [9]. We consider the elliptic model problem.

$$\begin{aligned} \nabla \cdot \boldsymbol{q} &= f \\ \boldsymbol{q} &= -\boldsymbol{K}\nabla u \end{aligned} \tag{5.1}$$

We set the solution

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

And set $\boldsymbol{K}$ to be the identity matrix. We call $u : \mathbb{R}^2 \to \mathbb{R}$ for the potential and $\boldsymbol{q} : \mathbb{R}^2 \to \mathbb{R}^2$ for the flux. And both values are of importance wehen solving (5.1). The flux term, $\boldsymbol{q}$, could for example be use to compute the transport of some contaminant in a porous medium.
As in [9] page 1340 we define the normalized discrete L_2 norms:

$$\|u - u_h\| = \left(\frac{1}{V} \sum_i V_i (u_{h,i} - u_i)^2 \right)^{\frac{1}{2}} \tag{5.3}$$

$$\|q - q_h\| = \left(\frac{1}{Q} \sum_a Q_a (q_{h,a} - q_a)^2 \right)^{\frac{1}{2}} \tag{5.4}$$

Where $q_a = -\hat{\boldsymbol{n}} \cdot \boldsymbol{q}$ is the normal flow density over edge a, with $\hat{\boldsymbol{n}}$ beieng unit normal to the edge. $q_{h,a}$ is the discrete flux over a, $u_{h,i}$ is the discrete potential at cell i, and u_i is the potential evaluated at the cell-center. Q_a is the volume

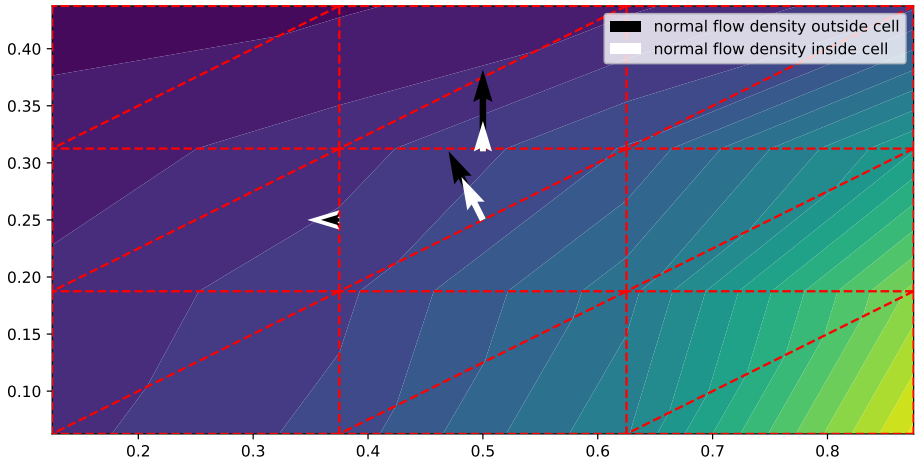


Figure 5.1: An illustration of why it's a bad idea to use the piecewise constant gradient on each element for computing normal flow. The normal flow is discontinuous across the edges, and flow into the cell is not equal to flow out.

associated with edge a, ie. the sum of the two volumes sharing edge a. $V = \sum_i V_i$ and $Q = \sum_a Q_a$.

The normal flux density (5.4) is easily obtained when working with finite volume methods, it is implicitly computed when assembling the matrix. For the finite element method however we use the transmissiblity coefficients from the L-method. As we see in (Cao, Y., Helmig, R. and Wohlmuth, [11]) chapter three, the bi-linear form of the linear lagrange finite element method on triangular grid is equivalent to the flux integral of the L-method for uniform paralellogram grids. When the grid is perturbed as in figure 5.9, this way of computing the normal flow density is not justified and is only approximate. There are other choices of flux recovery from the finite element method. The most obvious one would be to use the piecewise constant gradients on each triangle, with a triangle-centered finite volume method. This would however not be a conservative method, and one would get numerical diffusion when solving the corresponding transport equation.

In the first setup with uniform rectangular mesh, all the methods are identical and we get a quadratic convergence for normal flow density and potential as we see in figures 5.4 and 5.5.

In the second setup with uniform trapezoidal mesh illustrated in figure 5.6 we get quadratic convergence for normal flow density and potential, except for TPFA. This method is not convergent for grids that are not K-orthogonal, see figures 5.7 and 5.8.

In the perturbed mesh setup 5.9, the convergence rate for the normal flow density drops to about $O(h)$ in the L^2 norm and even worse for the max norm, see figures 5.10 and 5.11.
The next setup is with perturbed mesh and an aspect ratio of 0.1, see figure 5.12 for an illustration of aspect ratio 0.5. Here we clearly see that the O-method performs worse than the other two. We also see that the finite element method is the only method to achieve quadratic covergence for the potential.

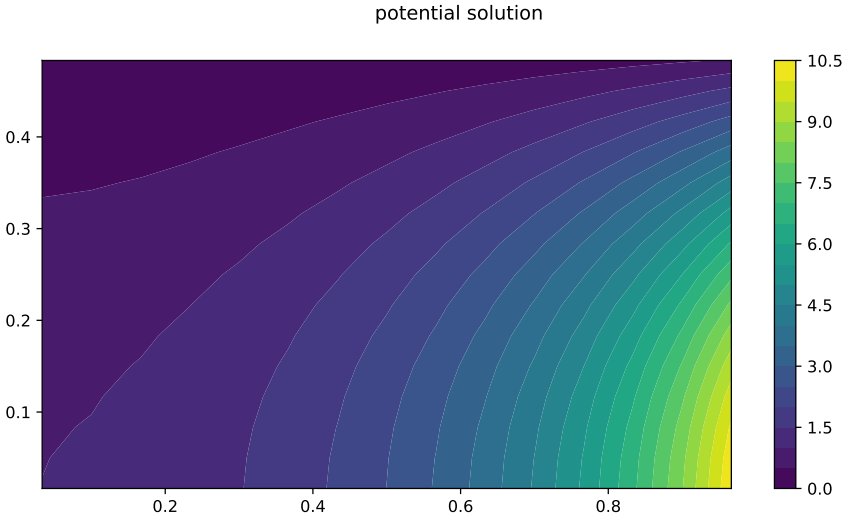


Figure 5.2: The solution (5.2) on half the unit square

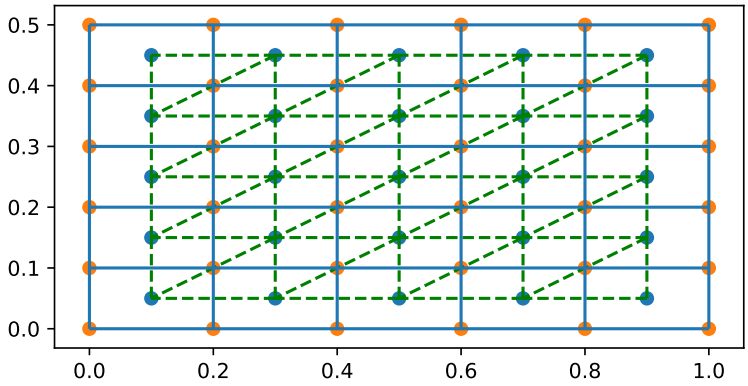


Figure 5.3: Unifrom rectangular mesh on half the unit square. The triangles are used for the finite element solution and are spanned between the nodes of the cell centers of the finite volume methods.

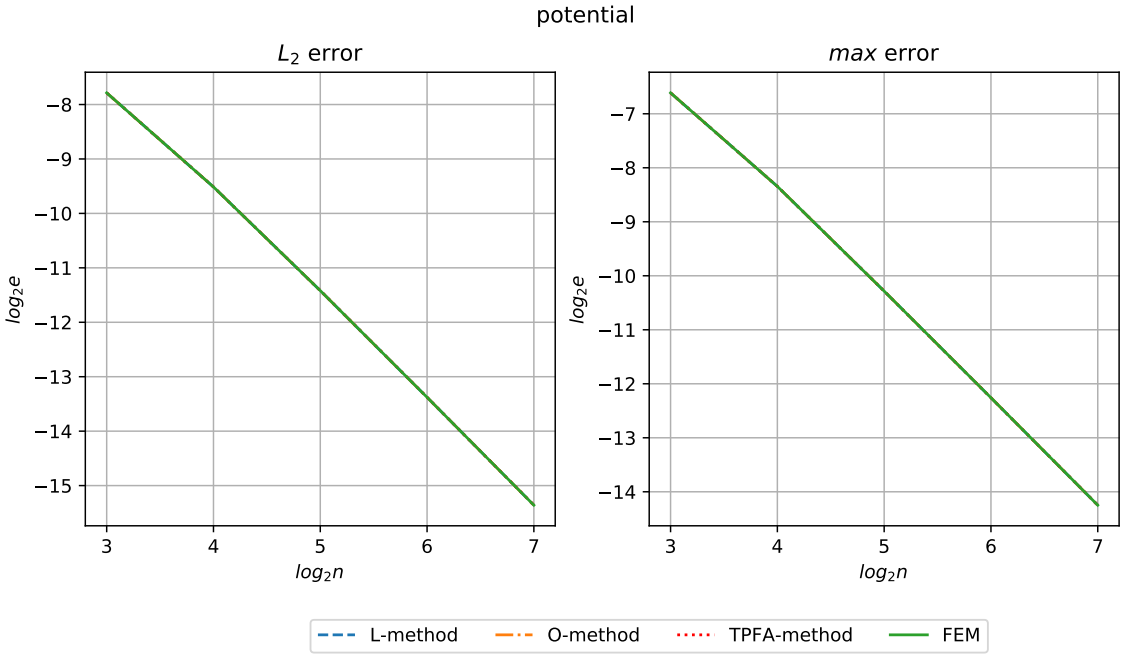


Figure 5.4: Potential error on refinements of the uniform rectangular mesh 5.3

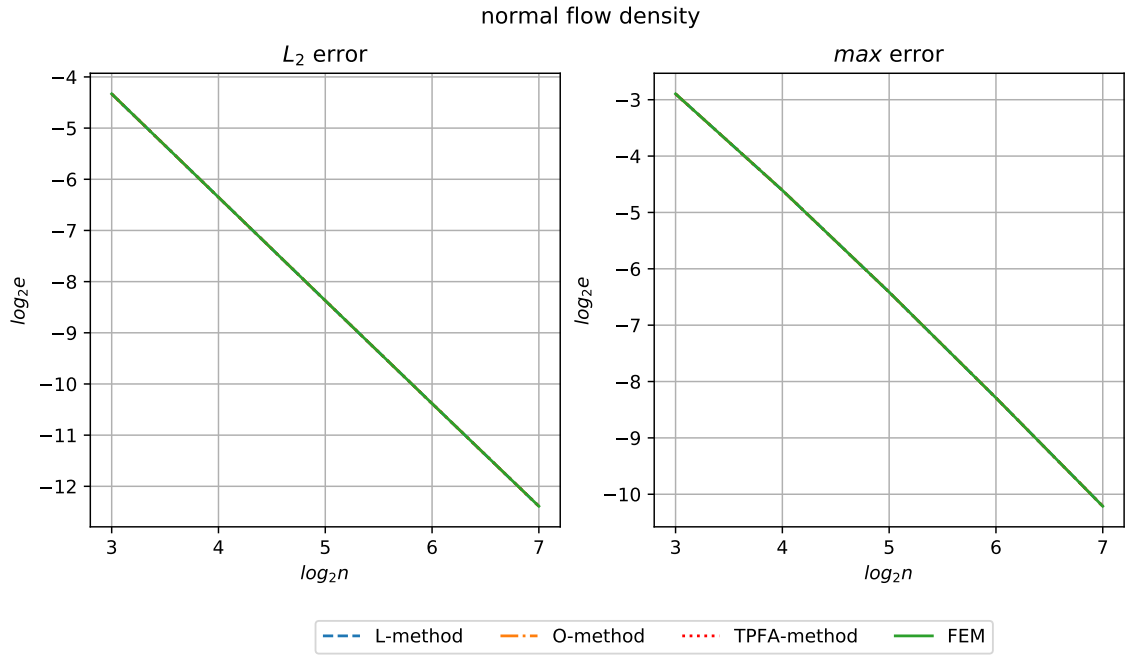


Figure 5.5: Normal flow density error on refinements of the uniform rectangular mesh 5.3

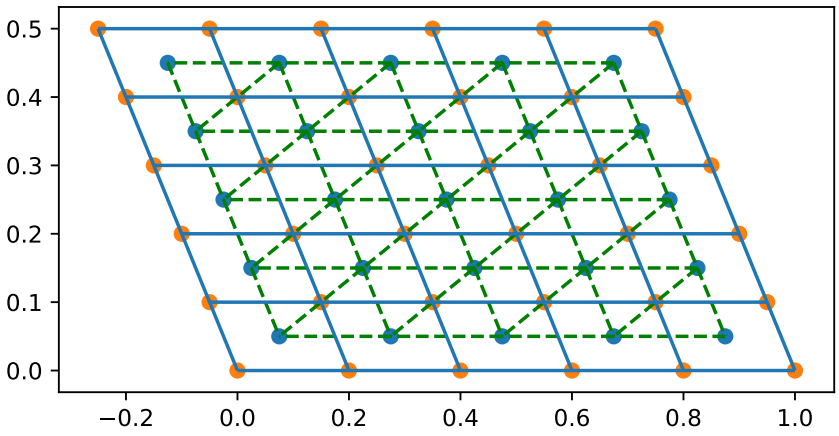


Figure 5.6: Trapezoidal mesh, now every point is transformed by $(x, y) \mapsto (x - 0.5y, y)$

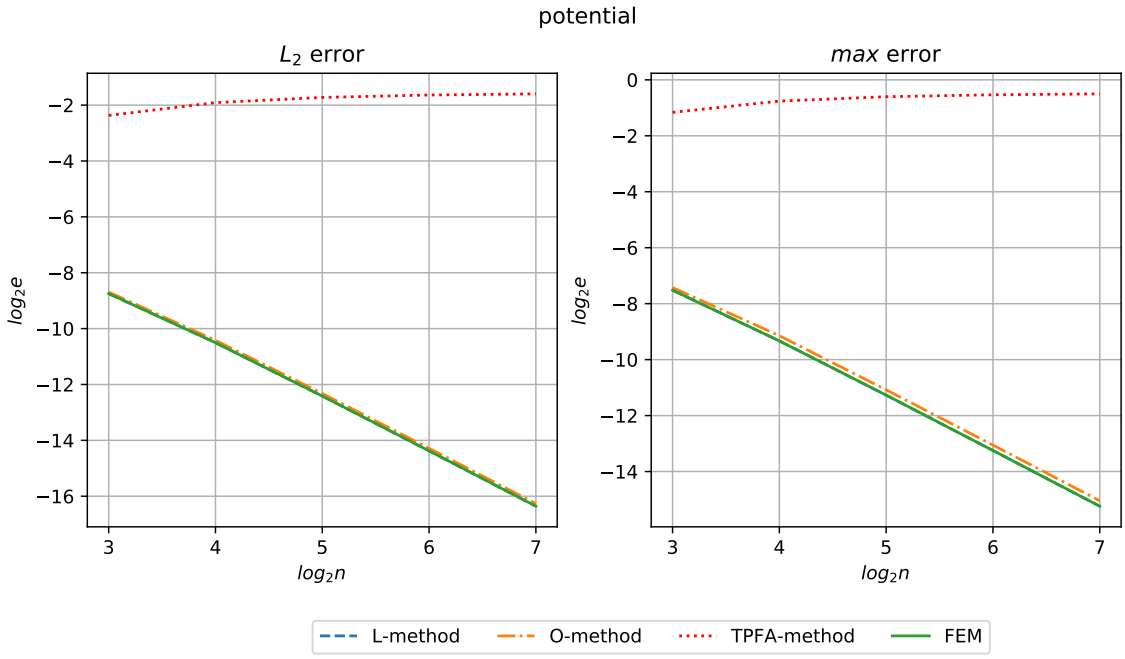


Figure 5.7: Pressure error on refinements of the mesh 5.6

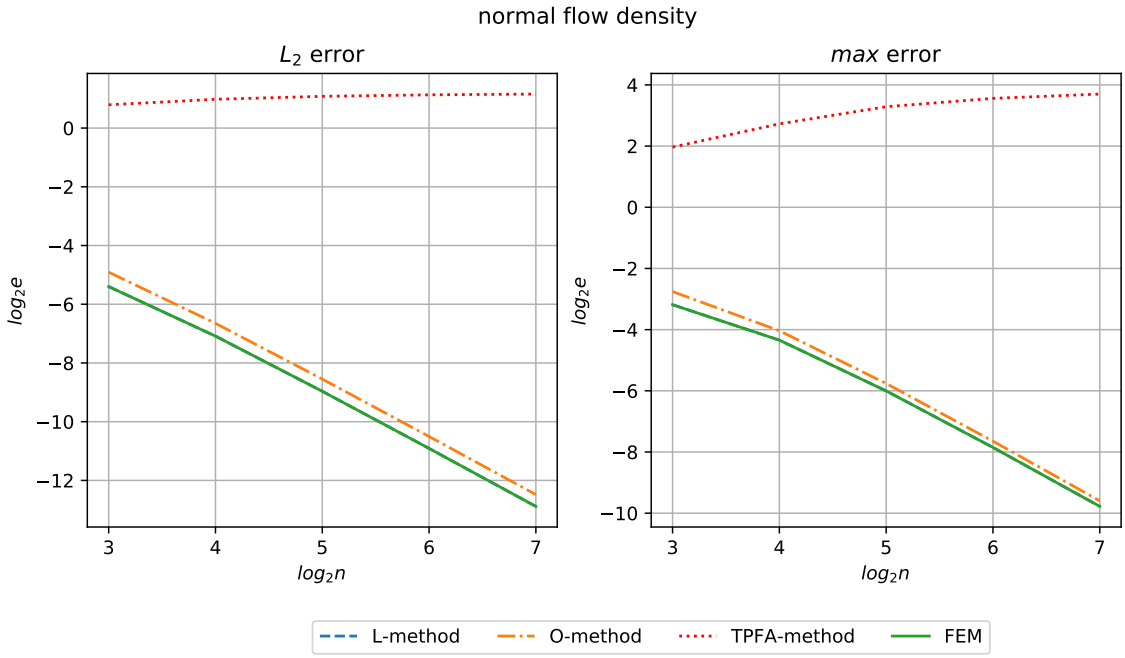


Figure 5.8: Normal flow density error on refinements of the mesh 5.6

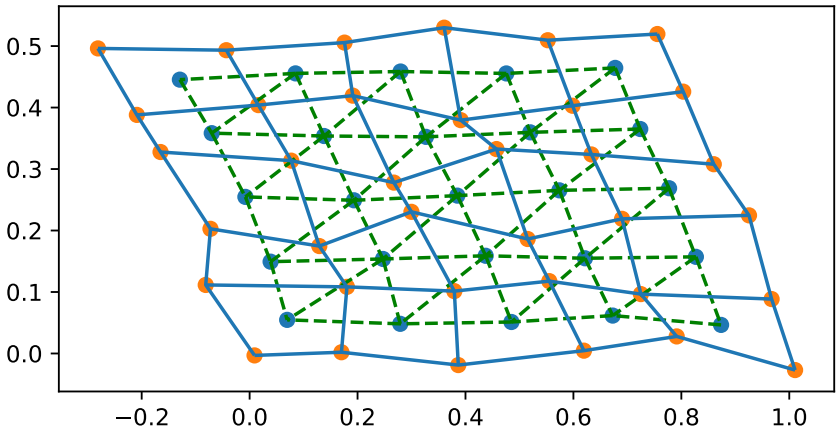


Figure 5.9: Perturbed mesh, every point in the mesh is perturbed by a random number which is $O(\frac{h}{5})$, in both x and y direction.

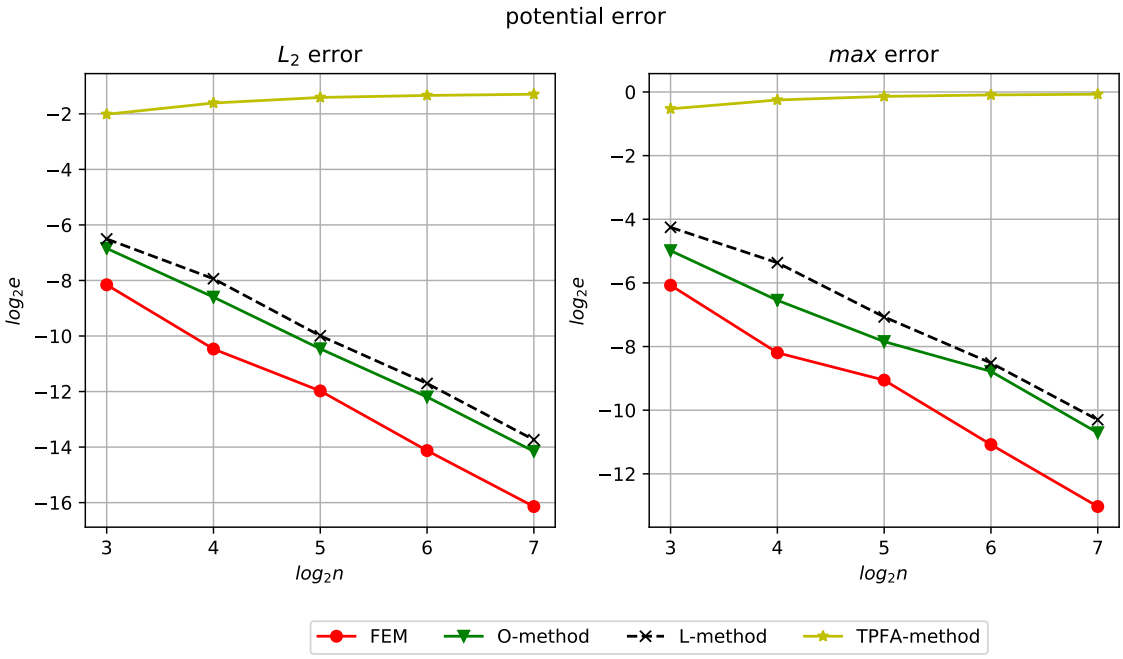


Figure 5.10: The pressure error of perturbed mesh.

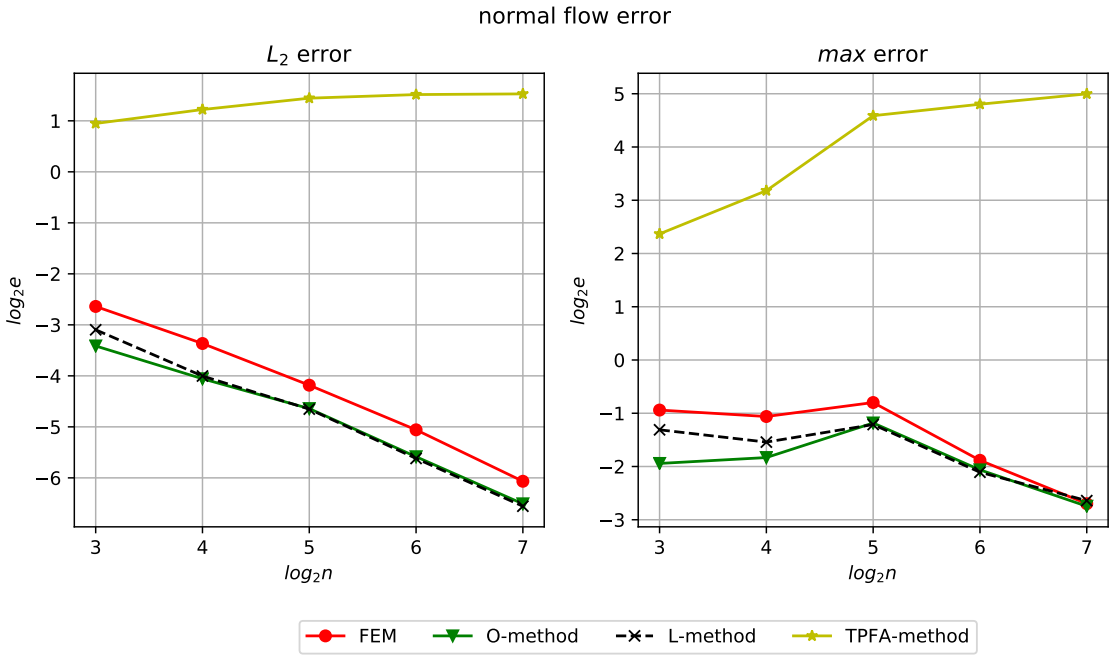


Figure 5.11: The normal flow density error of perturbed mesh

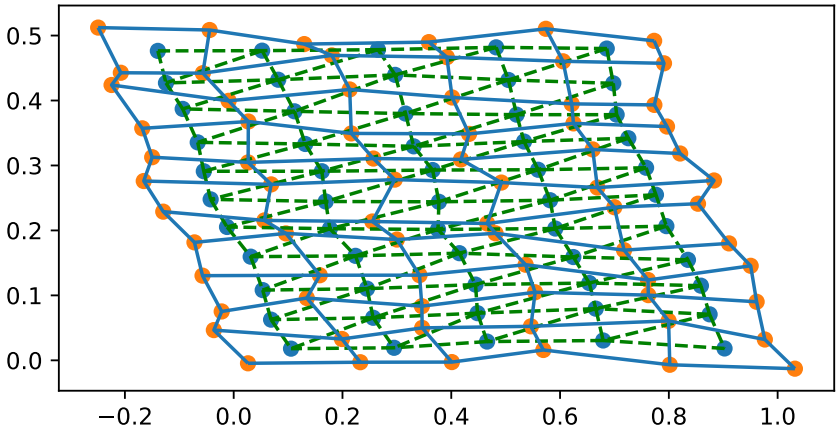


Figure 5.12: Perturbed mesh with aspect ratio 0.5, there are half as many points in the x-direction as in the y-direction.

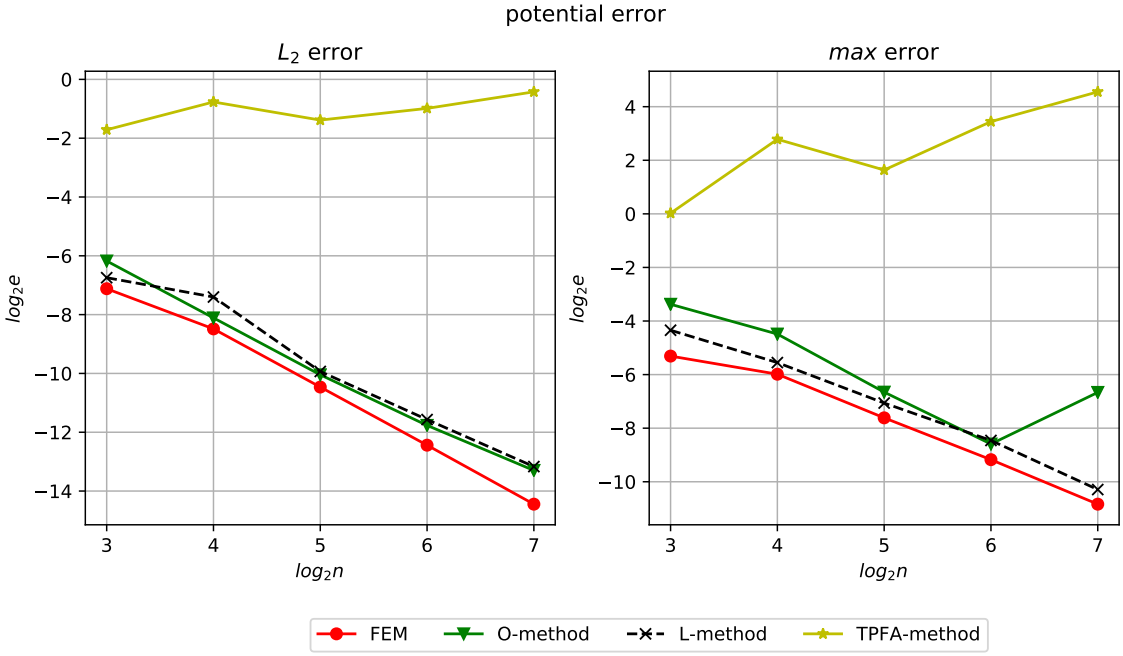


Figure 5.13: The pressure error of perturbed mesh with aspect ratio 0.1.

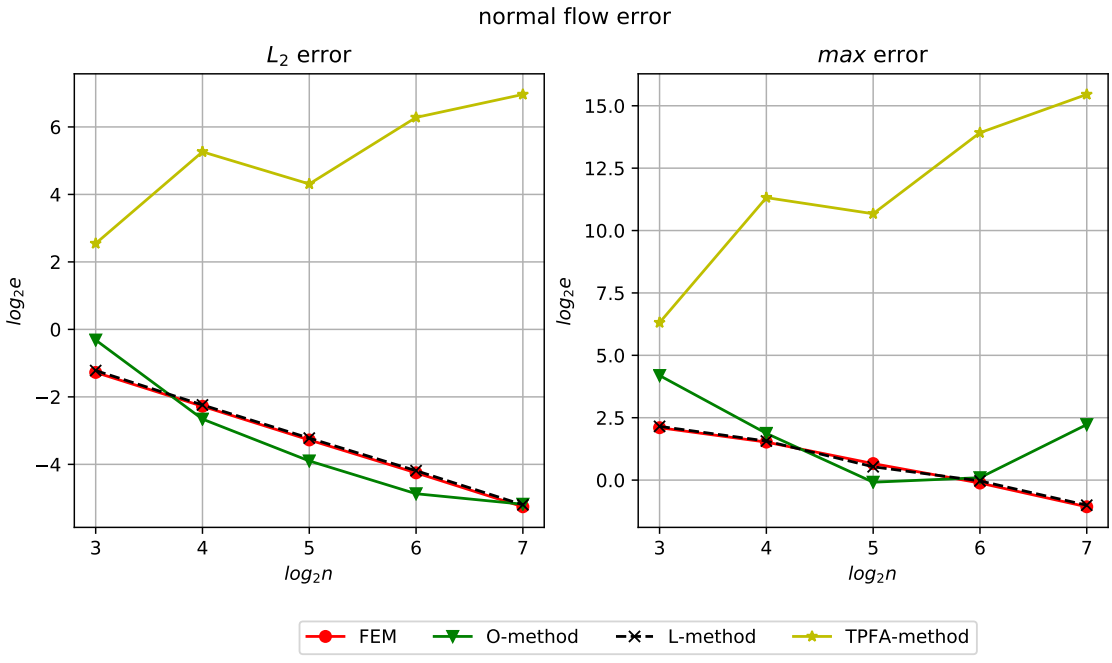


Figure 5.14: The normal flow density error of perturbed mesh with aspect ratio 0.1.

Chapter 6

Computer Code

Bibliography

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

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

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

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

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

[6] Ivar Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Computational Geosciences*, 6(3):405–432, Sep 2002.

[7] Jan Martin Nordbotten and Eirik Keilegavlen. An introduction to multi-point flux (mpfa) and stress (mpsa) finite volume methods for thermo-poroelasticity, 2020.

[8] J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad. Monotonicity of control volume methods. *Numer. Math.*, 106(2):255–288, March 2007.

[9] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, and J.M. Nordbotten. A compact multipoint flux approximation method with improved robustness. *Numerical Methods for Partial Differential Equations*, 24(5):1329–1360, 2008.

[10] Yufei Cao, Rainer Helmig, and Barbara I. Wohlmuth. Geometrical interpretation of the multi-point flux approximation l-method. *International Journal for Numerical Methods in Fluids*, 60(11):1173–1199, 2009.

[11] Yufei Cao, Rainer Helmig, and Barbara I. Wohlmuth. Convergence of the multipoint flux approximation l-method for homogeneous media on uniform grids. *Numerical Methods for Partial Differential Equations*, 27(2):329–350, 2011.

[12] Jacques Baranger, Jean-François Maitre, and Fabienne Oudin. Connection between finite volume and mixed finite element methods. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 30(4):445–465, 1996.

[13] L. E. Payne and H. F. Weinberger. An optimal poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, Jan 1960.

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