Something about a comparison of FEM and FVM for 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

Flow in porous media

In this section we start by introducing the physics of single-phase flow in porous media.

The REV

A porous medium consists of a solid matrix and some void in between the matrix, 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 REV or the *representative elementary volume*. An important characterization of a porous medium is the *porosity* ϕ is defined as $\phi = \frac{volume\ of\ voids\ in\ REV}{volume\ of\ REV}$. Another measure is the *saturation* S_α of phase α, $S_\alpha \equiv \frac{volume\ of\ \alpha\ in\ REV}{volume\ of\ voids\ in\ REV}$. In single phase flow, the saturation is irrelevant as the saturation is always one. Also note that the content of the phase α in the REV, θ_α, is given by $\theta_\alpha = S_\alpha \phi$.

Darcy's law

In 1856 Henri Darcy performed a famous experiment where he studied the flow of water trough sand. To understand his experiment we must first define some variables for measuring water content. First note that the pressure at height z above datum developed by a water column of height h above datum is given by

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

If we define the *gauge pressure* p by $p \equiv p_{abs} - p_{atm}$ we get an expression for p:

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

This can be rearranged to give an expression for the *hydraulic head* h:

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

A *manometer* is a tube put into the reservoir with one end in open atmosphere, the water level in this tube is then h. The volumetric flow of water is denoted by $\mathbf{q}_d$. Darcy's experiment is shown in figure 1.1, where water is poured trough 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

$$q_d = \kappa \frac{A(h_2 - h_1)}{L} \tag{1.2}$$

Where κ is a coefficient of proportionality.

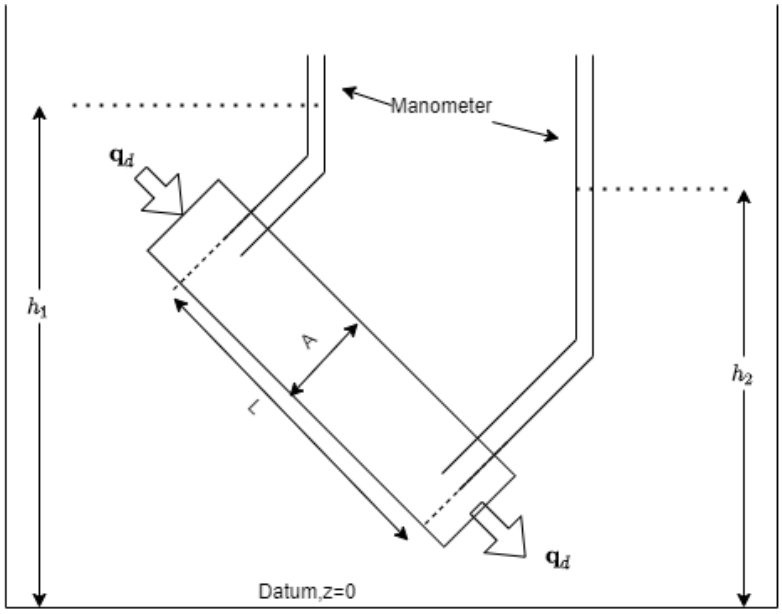


Figure 1.1: The Darcy experiment

Let q denote the volumetric flow-rate per area

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

This is now 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

$$\boldsymbol{q} = \boldsymbol{\kappa} \nabla h \tag{1.3}$$

We call $\boldsymbol{\kappa}$ the *hydraulic conductivity* and note that it is in general a rank two tensor, a matrix. The *hydraulic conductivity* also has the property that it is symmetric positive definite. With further experiments similar to the one described one can understand what makes up $\boldsymbol{\kappa}$, and it turns out that it is a function of viscosity μ, density ρ, gravity g and *permeability* $\boldsymbol{k}$.

$$\boldsymbol{\kappa} = \frac{\boldsymbol{k}\rho g}{\mu} \tag{1.4}$$

The *permability*, which is a property of the soil in the reservoir, is also a second rank tensor which is symmetric positive definite and is in general a function of space.
If we define the *pressure head* ψ as $\psi \equiv \frac{p}{\rho g}$ we can combine (1.1), (1.3) and (1.4) to get another variant of Darcy's law which will be usefull later

$$\boldsymbol{q} = \frac{\boldsymbol{k}\rho g}{\mu} \nabla(\psi + z) \tag{1.5}$$

Governing equations

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

$$\int_\Omega \frac{\partial(\rho\phi)}{\partial t} dV = -\int_{\partial\Omega} \boldsymbol{n} \cdot \rho\boldsymbol{q} \; dS + \int_\Omega f dV$$

Where $\boldsymbol{n}$ is an outward pointing normal vector to Ω and f is a source or a sink. 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 Ω, 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(1.3) and appropriate boundary and initial conditions closes the system

$$\begin{cases} \boldsymbol{q} = \boldsymbol{\kappa}\nabla h \\ \frac{\partial(\rho\phi)}{\partial t} + \nabla\cdot(\rho\boldsymbol{q}) = f \\ h = g(\boldsymbol{x}) \qquad \boldsymbol{x}\in\partial\Omega \\ h = f(\boldsymbol{x}) \qquad \boldsymbol{x}\in\Omega \;\; t=0 \end{cases} \tag{1.6}$$

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 (1.6) and modelling options,

Twophase 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, we write down Darcy's law (1.5) with a modification.

$$\boldsymbol{q}_\alpha = \frac{\boldsymbol{k}_{r,\alpha}\boldsymbol{k}\rho g}{\mu}\nabla(\psi_\alpha + z) \tag{1.7}$$

The scaling in front of the permeability $\boldsymbol{k}_{r,\alpha}$ is known as *relative permeability* and it has to be deduced from experimental observation.
We can also write down a mass balance equation for each phase:

$$\frac{\partial(S_\alpha\rho_\alpha\phi)}{\partial t} + \nabla\cdot(\rho\boldsymbol{q}_\alpha) = f_\alpha \tag{1.8}$$

Here we assume there is no mass transfer between the phases. If we combine equations (1.7) and (1.8), they give us 2 equations, but we have four unknowns ψ_w, ψ_n, S_w and S_n. We therefore introduce a simple algebraic relation

$$S_w + S_n = 1$$

and a not so simple relation

$$p_n - p_w = p_c \tag{1.9}$$

Where p_c is *capillary pressure* and is also 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 always have $p_n = p_{atm}$. For this assumption to hold it is important that the air always has some path to the surface. Now equation (1.9) simplifies to

$$-p_w(S_w) = p_c(S_w)$$

Note that we can divide by ρg to get an expression for ψ. Also experiments show that the capillary pressure is a monotone decreasing function of saturation, we can therefore invert it. Finally, we can multiply by the porosity to get an expression for the *water content* θ_w

$$\theta_w = \theta_w(\psi_w)$$

Combining this with the two-phase Darcy law (1.7) and mass balance (1.8) we get **Richards' equation**

$$\frac{\partial \theta(\psi)}{\partial t} - \nabla \cdot (\boldsymbol{\kappa}(\theta(\psi))(\nabla \psi + e_z)) = F \tag{1.10}$$

Where $\theta = \theta_w$. Note that density is completely eliminated because water is assumed to have constant density. The hydraulic conductivity in (1.7) is simply written $\frac{\boldsymbol{k}_{r,\alpha}\boldsymbol{k}\rho g}{\mu} = \boldsymbol{\kappa}(\theta)$.
Richards' equation contains two non-linearities in θ and $\boldsymbol{\kappa}$, this makes 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.

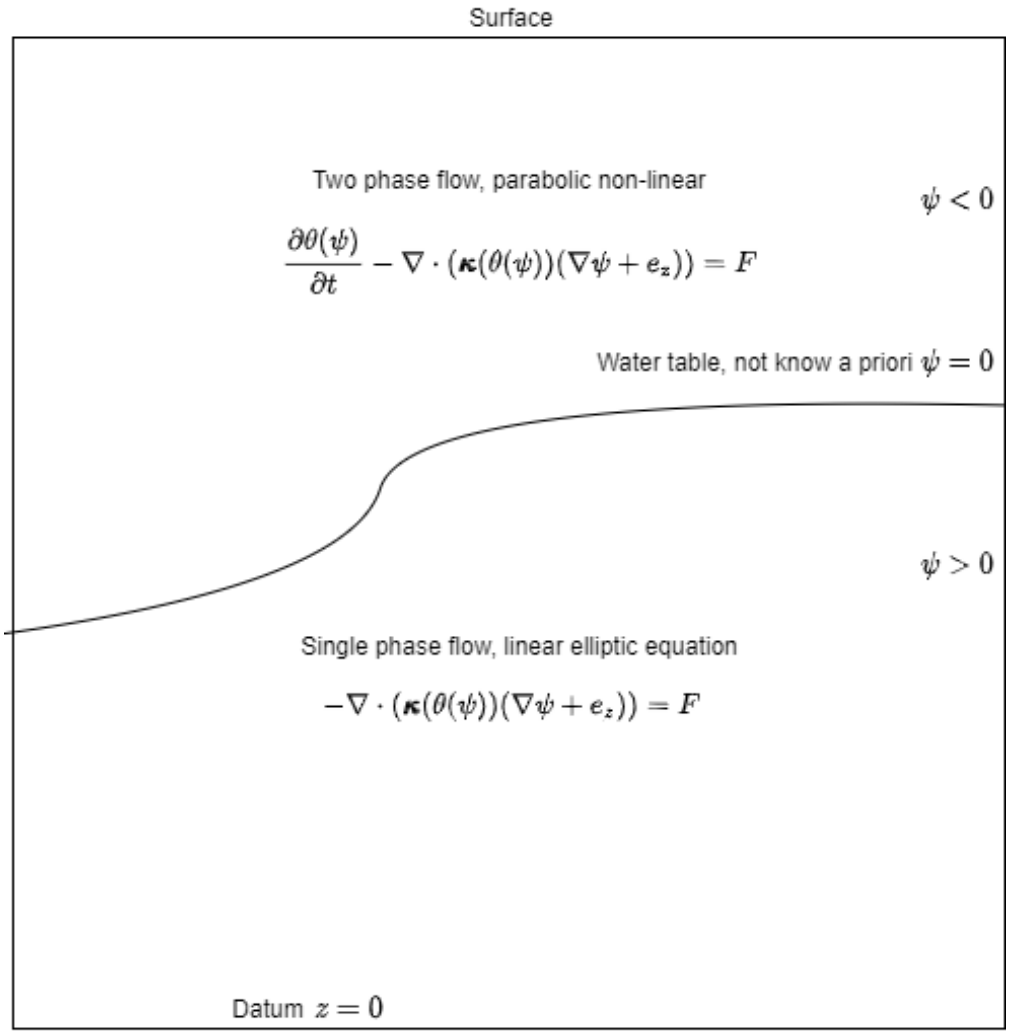


Figure 1.2: A sketch of the degeneracy of Richards' equation

Notes

Chapter 2

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 bacame more mainstream[2] . Today there are several general purpose finite element programs beieng used for a wide range of problems.
In this chapter we will introduce the finite element method and state results about stability and convergence. We will concentrate on solving the poisson equation: let $\Omega \subset \mathbb{R}^n$

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

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

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 (L^p space)**.** *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 a L^p space induces equivalence relations on the set of functions. Two functions in L^p is equal if they only differ on a set of meassure zero.*

These spaces have the property that they are complete.

Theorem 2.0.1 (Riesz-Fischer Theorem [3] chapter 8)**.** *Each L^p 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 are called* ***Hilbert spaces***

Before we continue the study of funtion spaces we develop some convinient 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 $|\overline{\alpha}| = \sum_{i=1}^n \alpha_i$. Let $\phi \in C^\infty(\Omega)$ 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 the basic calculus books.

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

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

We then define the sovolev spaces as

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

Theorem 2.0.2. *The sobolev spaces $W^{k,p}$ 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 $L^p(\Omega)$.

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

By (2.0.1) we know that $D_w^{\overline{\alpha}} u_i \to u_{\overline{\alpha}}$ as $i \to \infty$. In particular $u_i \to u$, so now we just need to show that $D_w^{\overline{\alpha}} u = u_{\overline{\alpha}}$. By the definition of weak derivative we have:

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

Now applying Hölder's inequality on both sides we get the two inequalities:

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

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:

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

Now we can put the two equations together to obtain $D_w^{\overline{\alpha}} u = u_{\overline{\alpha}}$

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

□

Definition 5. *We rename the L^2 based Sobolev spaces as follows*

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

With the norm of H^k beieng written in the more compact forms $\|\cdot\|_k$ and the inner product defined as follows:

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

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 meassure 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.

Definition 6. *We denote by $W_0^{k,p}(\Omega)$ the closure of $C_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$, where $C_c^\infty(\Omega)$ is the space of infinitely differentiable functions with compact support.*

Theorem 2.0.3 (Trace theorem, (Evans [4], chapter 5))**.** *Assume U is bounded and ∂U is C^1. Then there exists a bounded, linear operator*

$$T : W^{1,p}(U) \to L^p(U)$$

Such that

1. $Tu = u|_{\partial u}$ *if* $u \in W^{1,p} \cap C(\overline{U})$
2. $\|Tu\|_{L^p(\partial U)} \leq \|u\|_{W^{1,p}(U)}$

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

Theorem 2.0.4. *(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*

$$u \in W_0^{1,p} \Leftrightarrow Tu = 0 \text{ on } \partial U \tag{2.2}$$

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

The variational problem

We obtain the **variational formulation** by 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 $\partial\Omega$ the integral over the boundary vanishes. So the new formulation now reads:

$$\begin{gathered} \textit{find } u \textit{ such that} \\ \int_\Omega (\nabla v)^T \boldsymbol{K}\nabla u \, dx = \int_\Omega vF \, dx \\ \forall v \in V \end{gathered} \tag{2.3}$$

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's why it's also called the *weak formulation.*

The variational problems that we will look at, that arises from PDE's, will all have the form

$$\begin{gathered} \textit{find } u \textit{ such that} \\ a(u,v) = b(v) \\ \forall v \in V \end{gathered} \tag{2.4}$$

Where $a(\cdot,\cdot)$ is a *bilinear form* on a V and $b(\cdot)$ is a **linear functional** on V ie. $b \in V'$, where V is a Hilbert space. In general, a variational formulation can we seen as

$$\begin{gathered} \text{find } u \in V \text{ such that} \\ Au = b \text{ where } A : V \to V' \end{gathered} \tag{2.5}$$

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:

$$\begin{aligned} -\nabla \cdot \boldsymbol{K}\nabla u(x) &= F(x) && x \in \Omega \\ u(x) &= g_D && x \in \Gamma_D \\ \boldsymbol{K}\nabla u(x) &= g_N && x \in \Gamma_N \end{aligned} \tag{2.6}$$

To make similar variational formulation of (2.6) we first define the test space:

$$V = \left\{ v \in H^1 : T(v) = 0 \text{ on } \Gamma_D \right\} \tag{2.7}$$

Next assume $\exists w \in H^1(\Omega)$ such that Dirichlet boundary conditions are met: $T(w) = g_D$. Now we use integration by parts to as before:

$$a(u+w,v) = \int_\Omega (\nabla u + \nabla w)^T \boldsymbol{K}\nabla v \; dx = \int_\Omega Fv \; dx - \int_{\partial\Omega} \boldsymbol{K}\nabla(u+w)\cdot \boldsymbol{n} v \; dx \tag{2.8}$$

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

$$a(u,v) = b(v) = \int_\Omega Fv \; dx - \int_\Omega (\nabla w)^T \boldsymbol{K}\nabla v \; dx + \int_{\Gamma_N} g_N v \; dx \tag{2.9}$$

Hence both Dirichlet and Neumann boundary conditions are incorporated into the right hand side.

We still need to show that (2.4) has a unique solution.

Existence and uniqueness

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

- $a(\cdot,\cdot)$ *is* ***coercive*** *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* $|a(u,v)| \leq C_B \|u\| \|v\|\ \ \forall u, v \in V$

To use this to prove existence and uniqueness, we must first state some important results about the underlying space V. The following theory can be found in it's entirety in chapter one-four of Cheney [3]

Theorem 2.0.5. *If* Y *is a closed subspace of a 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 plain english: an element in X *can always be written as the sum of an element* Y *and an element in* $Y^\perp$

Theorem 2.0.6 (Riesz Representation theorem)**.** *Every continuous linear functional defined on a Hilbert space* X *can be written* $x \to \langle x, v\rangle$ *for an uniquely determined* $v \in X$*.*

Proof. Let $\phi \in X'$, define $Y = \{x \in X : \phi(x) = 0\}$ to be the null space of ϕ. Take a non-zero vector 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 2.0.5). 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 2.0.7 (Banach fixed-point theorem)**.** *Let* X *be a complete metric space and* $F : X \to X$ *an operator where* $d(Fx, Fy) \leq \theta d(x,y)$ *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 2.0.8 (Lax Milgram)**.** *Suppose* $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ *is a billinear, bounded and coercive form and that* $b(\cdot) : V \to \mathbb{R}$ *is a bounded, linear functional. Then the variational problem has an unique solution* u*.*

$$a(u,v) = b(v) \tag{2.10}$$

For all $v \in V$

Remark 3. *If $a(\cdot,\cdot)$ also is symmetric, and defines an inner product on V giving a complete space. We can use Riesz Representation theorem 2.0.6 to show that it has an unique solution.*

proof of Lax Milgram theorem 2.0.8. 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 2.0.6 $a_w(\cdot)$ uniquely determines a vector $Aw \in V$ 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 (2.10):
find u such that

$$Au = f \tag{2.11}$$

Now we need to show that (2.11) has an unique solution, and for that we need the Banach fixpoint 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 (2.11) 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. $\epsilon < \frac{2C_c}{C_b^2} \Rightarrow (1-2\epsilon+\epsilon^2) < 1$ □

Remark 4. *u depends on $b(\cdot)$, 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 a 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 **Poincare inequality** in the last step.*

- ***Bounded*** *with respect to $\|\cdot\|_{H_0^1}$*

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

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

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

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

We also see that b is in the dual space of H_0^1 if for example $f \in L^2(\Omega)$:

$$|b(v)| = |\int_\Omega f v dx| \leq \|f\|_{L^2} \|v\|_{L^2}$$

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

Hence (2.3) is well posed and we get a solution $u \in H_0^1(\Omega)$.

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*.

$$\begin{gathered} \textit{find } u \in V_h \textit{ such that} \\ a(u, v_h) = b(v_h) \\ \textit{for all } v_h \in V_h \end{gathered} \tag{2.12}$$

Since a and b both are linear, it's easy to see that if (2.12) 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 8 (two dimensional triangulation, page 56 of Knaber [5])**.** *Let τ_h be a partition Ω into closed trinagles 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 open triangle (without the boundary ∂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 9 (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 5. *Our local ansatz space $P_K = \{v|_K : v \in V_h\}$ is such that $P_K = P_1 \subset H^1(K) \bigcap C(K)$. 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 6. *In general, finite element methods are defined by a domain(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.*

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's defined by the equation:

$$\phi_i(x_j) = \delta_{ij}$$

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

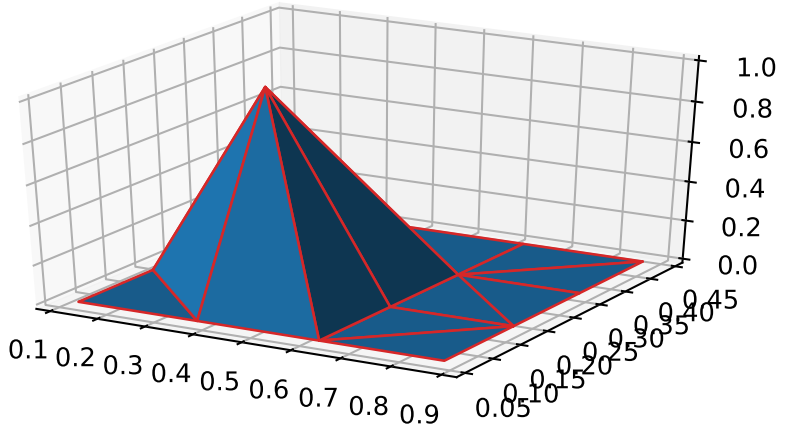


Figure 2.1: A hat function .

Now lets see how the method works in practice. We seek a solution $u_h \in V_h$. Write this out in our basis of hat functions: $u_h = \sum_{i=1}^{n} u_i^* \phi_i$. Now (2.12) becomes

$$\begin{aligned} &\textit{find } \boldsymbol{u}^* \in \mathbb{R}^n \textit{ such that} \\ &\sum_{i=1}^{n} u_i^* a(\phi_i, \phi_j) = b(\phi_j) \end{aligned} \tag{2.13}$$

So we get a system of linear equations $A\boldsymbol{u}^* = \boldsymbol{b}$, where we have one equation for each interior node. If we solve (2.3) our variatonal problem and also matrix will be symmetric, it is then often called a *stiffness matrix*. The system is also sparse, which is a very important property when designing algorithms to solve it.
With this setup described in this section, the degrees of freedom are the same as the dimension of V_h. If we in definition 9 instead had chosen a space of quadratic polynomials on each element, we hade 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 is not expected to be very regular.

Implementation

Implementing linear finite element method on a triangular mesh consists of two main parts.

- Assembling the system described in (2.13). This may be more or less complicated depending on the underlying variational problem.
- Solving the linear system, this is usually done by a sparse iterative solver like GMRES or conjugate gradient descent.

The last part is outside the scope of this chapther, the first part is well illustrated if we choose our variational problem to be the homogenous elliptic model problem (2.3) in two dimensions with $\boldsymbol{K} = \boldsymbol{I}$. 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 beieng 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 (2.14). Also compute the gradients of the shape functions.

$$\begin{aligned} N_1(x, y) &= 1 - x - y \\ N_2(x, y) &= x \\ N_3(x, y) &= y \end{aligned} \tag{2.14}$$

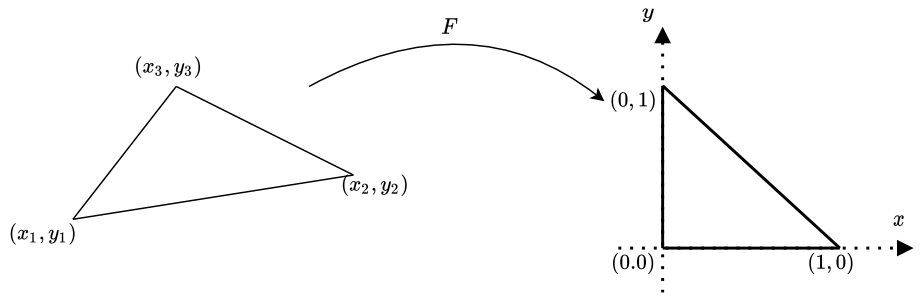


Figure 2.2: The map F from element K to the reference element $\hat{K}$.

4. Loop trough 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

$$\begin{aligned} F : K &\to \hat{K} \\ x &\mapsto Bx + d \end{aligned} \tag{2.15}$$

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

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ d_1 & d_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2.16}$$

So for each element we solve (2.16) 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 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 $u : K \to \mathbb{R}$

$$\nabla^T_{\hat{x}} u(F^{-1}(\hat{x})) = \nabla^T_x u(F^{-1}(\hat{x})) \nabla^T_{\hat{x}} F^{-1}(\hat{x}) = \nabla^T_x u(F^{-1}(\hat{x})) B^{-1} \tag{2.17}$$

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

$$\nabla_x u(F^{-1}(\hat{x})) = B^T \nabla_{\hat{x}} u(F^{-1}(\hat{x})) \tag{2.18}$$

Now we can compute the product of the gradients of the bassis functions

on an element.

$$\begin{aligned} a(\phi_i,\phi_j)|_K &= \int_K (\nabla\phi_i)^T\nabla\phi_j dx \\ &= \int_{\hat{K}} (\nabla_x\phi_i(F^{-1}(\hat{x})))^T\nabla_x\phi_j(F^{-1}(\hat{x}))|\mathrm{Det}(J(F^{-1}))|d\hat{x} \\ &= \int_{\hat{K}} (B^T\nabla_{\hat{x}}\phi_i(F^{-1}(\hat{x})))^T B^T\nabla_{\hat{x}}\phi_j(F^{-1}(\hat{x}))|\mathrm{Det}(B^{-1})|d\hat{x} \\ &= \int_{\hat{K}} (\nabla_{\hat{x}}N_i(\hat{x}))^T B^T B\nabla_{\hat{x}}N_j(\hat{x})|\mathrm{Det}(B^{-1})|d\hat{x} \\ &= \frac{1}{2}(\nabla_{\hat{x}}N_i(\hat{x}))^T B^T B\nabla_{\hat{x}}N_j(\hat{x})\frac{1}{|\mathrm{Det}(B)|} \end{aligned} \tag{2.19}$$

So for each element we evaluate the last line of (2.19) 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 (2.19) we compute the integral on the reference element:

$$\begin{aligned} b(\phi_i)|_K &= \int_{\hat{K}} f(F^{-1}(\hat{x}))\phi_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})N_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &\approx \frac{1}{\mathrm{Det}(B)}\sum_k \omega_k\hat{f}(\hat{p}_k)N_i(\hat{p}_k) \end{aligned} \tag{2.20}$$

Where $\hat{f} := f(F^{-1}(\hat{x}))$ and $\{(\omega_k,\hat{p}_k)\}_k$ defines a *quadrature rule*. We will see later that this quadrature rule can be chosen in different ways, for higher order finite elements this may even affect the convergenec behauviour.

5. Loop through the nodes x_j at the boundary and set $\boldsymbol{A}_{j,i} = \delta_{ij}$, $b_j = 0$

Remark 7. *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* (2.9)*, using one-dimensional elements.*

Convergence

In this section we review the most important concepts in studying the convergence, for a detailed discussion see [5]. The starting point of convergence estimates for the finite element method already described are **Cèa's lemma**.

Theorem 2.0.9. *Let u solve the variational problem* (2.4) *and u_h solve the corresponding galerkin approximation* (2.12)*, where the bi-linear form a is bounded and coercive. Then we have*

$$\|u - u_h\|_V \leq \frac{C_b}{C_c} min\{\|u - v_h\| : \ v_h \in V_h\} \tag{2.21}$$

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*. Hence we only need to 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.

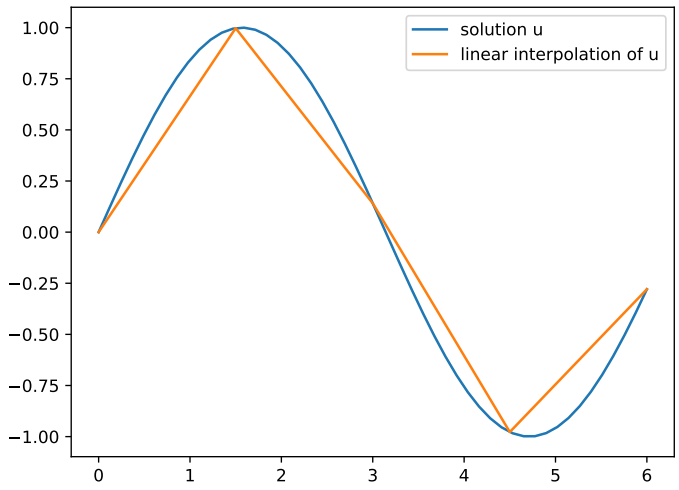


Figure 2.3: The unique linear interpolation of a function in one dimension.

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

Definition 10 (Global interpolation operator)**.**

$$I_h : C(\overline{\Omega}) \to V_h$$
$$v \mapsto \sum_i v(n_i)\phi_i$$

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

Remark 8. *The global interpolator operator 10 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 below three 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 on the form

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

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 varational problem admits a solution with sufficent regularity. We tie these observations together in a theorem

Theorem 2.0.10 (energy norm estimate)**.** *Consider a finite element discretization as described by* (2.13) *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 9, 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 meassure 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 prbolem of our variational problem (2.3): $a(v,u) = \langle f, v\rangle_0$, and assume some uniqueness and stability of the solution $u = u_f$ of this.

Theorem 2.0.11 (L^2 estimate)**.** *Suppose the situation of theorem 2.0.10 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 obtian quadratic convergence in the L^2 norm

Remark 9. *In this chapter we have only discussed the convergence behaviour of the solution to the Galerkin problem* (2.12)*. 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.*

Stability

A stability property for the solution of the Galerkin problem (2.12) follows from remark (4):

$$\|u_h\|_{H_0^1} \leq \frac{1}{C_c} \|b\|_{H_0^{1\prime}}$$

Chapter 3

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

$$\int_{\Omega_i} -\nabla \cdot \boldsymbol{K} \nabla u dx = -\int_{\partial\Omega_i} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} ds = \int_{\Omega_i} F dx \tag{3.1}$$

The above equation equates the fluxes trough the boundary of a control volume with the source or sinks inside the control volume, the finite volume methods are discrete versions of this. The main idea is to approximate the normal flux trough edge j of $\partial\Omega_i$

$$f_{i,j} = -\int_{\partial\Omega_{i,j}} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} ds \tag{3.2}$$

by a linear combination of u_i at neghbouring cell centers.

$$\tilde{f}_{i,j} = \sum_k t_{j,k} u_k \tag{3.3}$$

Where the *transmissibility* $t_{j,k}$ has the property $\sum_k t_{j,k} = 0$. We also approximate the integral on the right side by evaluating F at the cell center and multiply by the area of Ω_i. We then end up with a system of equations

$$\sum_{j=1}^{4} \tilde{f}_{i,j} = |\Omega_i| F(x_i) \tag{3.4}$$

The system of equations (3.4) ensures local mass conservation. We will discuss different ways of constructing the transmissibility coefficients, as they result in very different discretizations.

Two point flux approximation

The simplest way of constructing $t_{j,k}$ is also the most popular in the industry. As the name suggests, we only use the function value at two points, $x_{j,0}$ and $x_{j,1}$, to compute the numerical flux $\tilde{f}_j$.

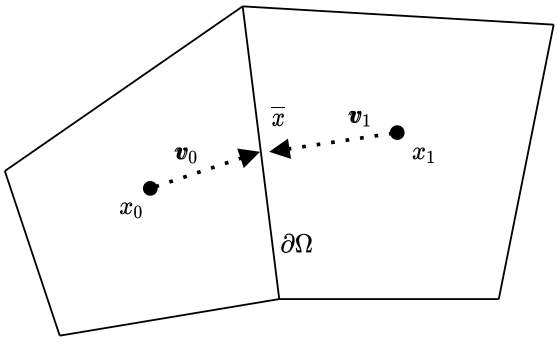


Figure 3.1: The two point flux approximation(TPFA) setup.

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 between the cells by

$$f_0 = -\int_{\partial\Omega} \hat{\boldsymbol{n}}^T \boldsymbol{K}_0 \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}(u(\overline{x}) - u(x_0))ds \tag{3.5}$$

Or as

$$f_1 = -\int_{\partial\Omega} \hat{\boldsymbol{n}}^T \boldsymbol{K}_1 \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}(u(x_1) - u(\overline{x}))ds \tag{3.6}$$

Where $\hat{n}$ is the normal vector to edge between the cells. Because we require flux continuity we have that

$$f_0 = f_1 = t_0 u_0 + t_1 u_1 \tag{3.7}$$

Where, as before, $t_0 + t_1 = 0 \Rightarrow t_0 = -t_1$. We now have three equations and three unknowns, $u(\overline{x})$, t_0 and t_1. To simplify, we introduce the quantity $T_i = \int_{\partial\Omega} \hat{\boldsymbol{n}}^T \boldsymbol{K}_i \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|}$ to represent the cell transisibility. So first we solve for $u(\overline{x})$:

$$T_0(u(\overline{x}) - u(x_0)) = T_1(u(x_1) - u(\overline{x})) \Rightarrow u(\overline{x}) = \frac{T_0 u_0 + T_1 u_1}{T_0 + T_1} \tag{3.8}$$

Next we insert this into the expression for f_0

$$f_0 = -T_0(u(\overline{x}) - u(x_0)) = t_0(u_0 - u_1) \Rightarrow t_0 = \frac{1}{\frac{1}{T_0} + \frac{1}{T_1}} \tag{3.9}$$

Hence, the transmissibility is the harmonic mean of the local transmissibilities. This discretization has the **advantages**

- The matrix in (3.4) is symmetric, this makes the resulting system easier to solve.
- A small stencil, ie. the matrix in (3.4) has a bandwith of five for two dimensional quadrilateral grids.
- The assembly of the matrix in (3.4) is fast, as you have an explicit expression for the transmissibilities.
- It is easy to implement

But there is no such thing as a discretization that has everything, and TPFA has one big **disadvantage**

- It is not convergent when the grid is not alligned with the principal directions of $\boldsymbol{K}$, also called $\boldsymbol{K}$-orthogonality. This is demonstrated in the next chapter, see **??**

O-method

The O-method is a multi-point flux approximation method, these types of methods where developed to make control volume methods converge for grids that are not $\boldsymbol{K}$-orthogonal. It is described in detail in [6], I only give a brief introduction. Consider the control volumes in 3.2.

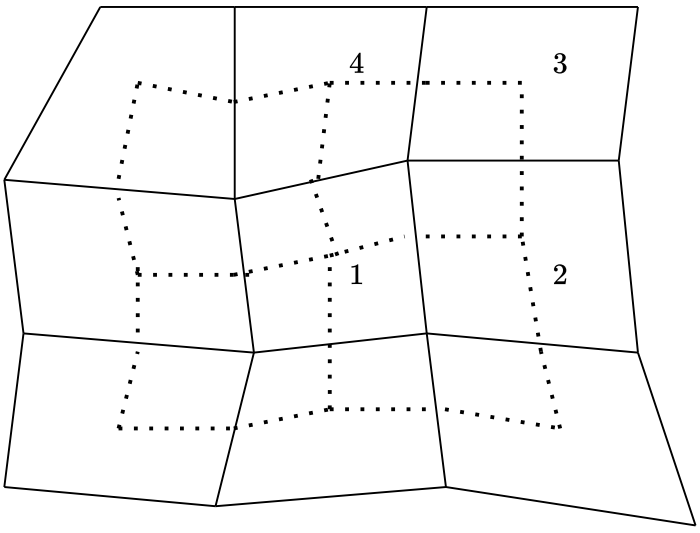


Figure 3.2: The solid lines are the control volumes, the dashed lines are the dual mesh connecting the cell centers, going trough the midpoints of each edge.

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

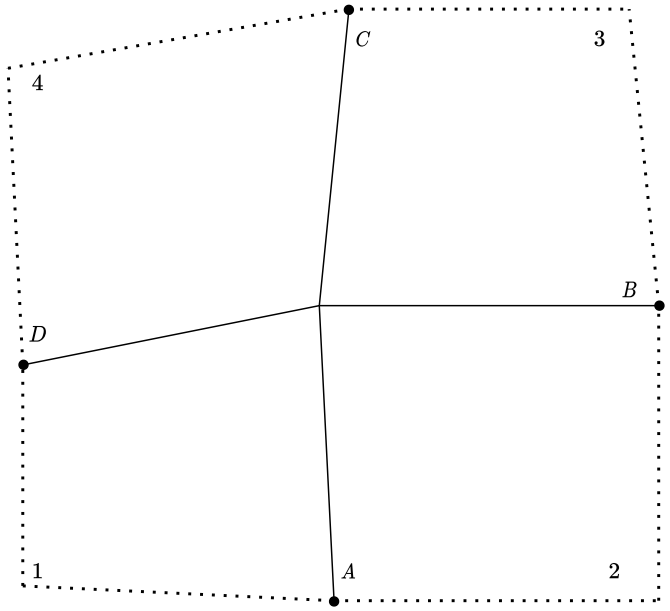


Figure 3.3: The interaction region corresponding to cells $1, 2, 3, 4$.

In each interaction region there are four half edges, and the flux trough each of the half edges are determined by a linear combination of the four cell center potential values.

$$f_i = \int_{\partial\Omega_i} \hat{\boldsymbol{n}}_i{}^T \boldsymbol{K} \nabla u ds = \sum_{j=1}^{4} t^j u^j \quad i, j \in \{1, 2, 3, 4\}$$

Where subscript i denotes the half edges of the interaction region and j denotes cell center potential values.

We assume for now that the potential is linear in each of the four subcells in the interaction region, figure 3.3. This gives $4 \cdot 3 = 12$ degrees of freedom. The linear potential must ofcourse equal the cellcenter values of the potential in the cellcenters, this removes four degrees of freedom. We also require flux continuity on the four half edges in the interaction region, this removes an aditional four degrees of freedom. The last four degrees of fredom are spent on potential continuity of the midpoints of the edges.

By theese assumptions of flux and potential continuity we can calculate the transmissibility coefficients. This involves equations for flux and potential derived

with properties from the mesh, see [6] for the details. What we end up with is a four by four transmissibility matrix for each interaction region. Finally we assemble the system of equations (3.4) with the transmissibility coefficients. Note that we write the flux over the jth edge of cell i, $\tilde{f}_{i,j}$ as the flux over the two half edges.

$$\sum_{j=1}^{4}(\tilde{f}_{i,j}^{1} + \tilde{f}_{i,j}^{2}) = |\Omega_i|F(x_i)$$

$$\sum_{j=1}^{4}(\sum_{k=1}^{4} t_{i,j}^{k,1}u^k + \sum_{k=1}^{4} t_{i,j}^{k,2}u^k) = |\Omega_i|F(x_i)$$

Next, we see that the interaction regions of the two half edges sharing same edge overlaps, so we get a six point flux stencil.

$$\sum_{j=1}^{4}\sum_{k=1}^{6} \bar{t}_{i,j}^{k}u^k = |\Omega_i|F(x_i)$$

We can simplify this further and see that we get a nine point stencil.

$$\sum_{k=1}^{9} \hat{t}_{i,j}^{k}u^k = |\Omega_i|F(x_i)$$

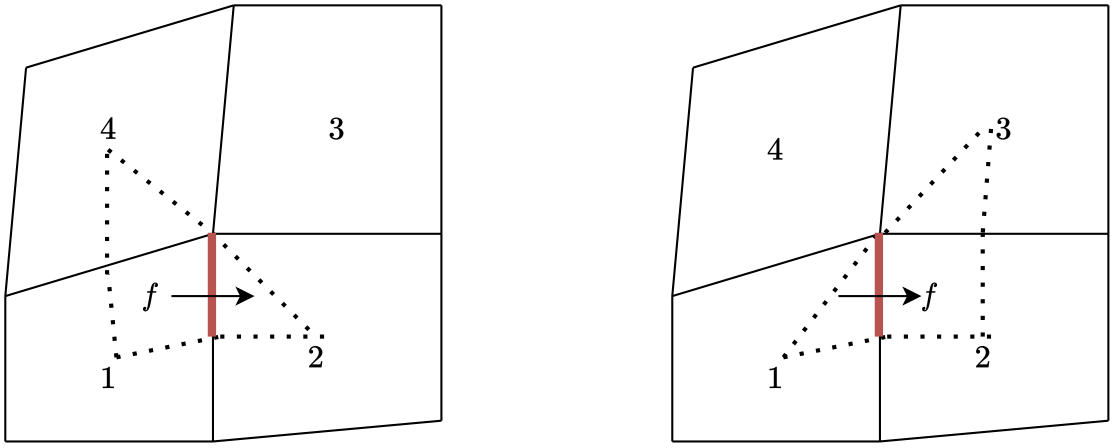
L-method

As the O-method, the L-method is also a multipoint flux approximation method. It was introduced in [7]. This method is similar to the O-method in that it goes trough 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 cellcenter 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 3.4. 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

$$\begin{aligned} &\text{if } |t_1^1| < |t_2^2| \\ &\text{choose } T_1 \text{ else} \\ &\text{choose } T_2 \end{aligned} \tag{3.10}$$

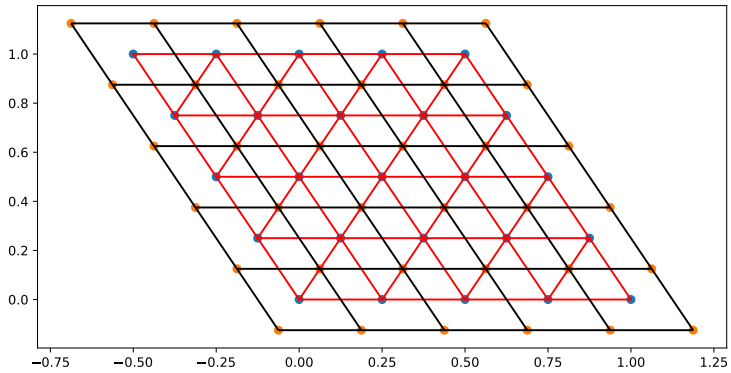


(a) T_1 (b) T_2

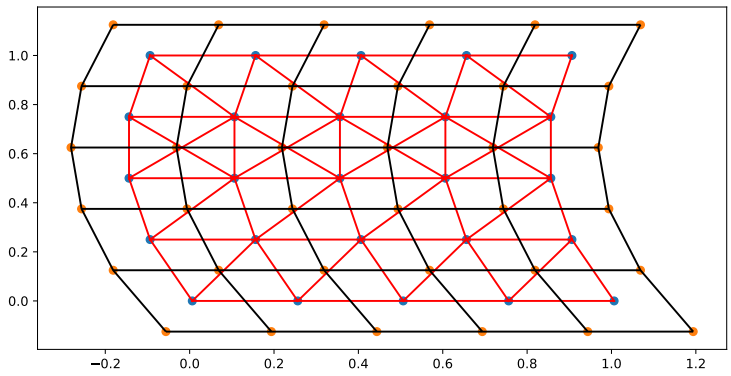
Figure 3.4: The two choices of which cellcenters to use for computing the flux over the halfedge in red.

A cheap intuition behind (3.10) 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 increases the chances that we get the same sign of t^i on the same side of the half edge which is more stable. See [8] for a more detailed geometric intuition in the case of homegenous permability tensor. In figure 3.5 we see the criterion in practice for a homogenous medium. In figure 3.5a all L-triangles are used by two half-edges, and they are chosen in the same way troughout the domain. In figure 3.5b there are some triangles that overlap, this is due to the fact that some L-triangles are used by only one half edge.

explain some more



(a) Paralellogram grid, all triangles are chose similarily.



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

Figure 3.5: Examples of L-triangles(in red) in a domain with homogenous permability tensor.

As with the O-method, we end up with a system assembled from the fluxes over the half edges.

$$\begin{aligned} \sum_{j=1}^{4}(\tilde{f}_{i,j}^{1} + \tilde{f}_{i,j}^{2}) &= |\Omega_i| F(x_i) \\ \sum_{j=1}^{4}(\sum_{k=1}^{3} t_{i,j}^{k,1} u^k + \sum_{k=1}^{3} t_{i,j}^{k,2} u^k) &= |\Omega_i| F(x_i) \end{aligned} \tag{3.11}$$

But the flux stencil across each edge is possibly smaller, often just four points as would be the case in figure 3.5a.

Lemma 3.0.1 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[8]). *Assume that the permability* $\boldsymbol{K}$ *is homogenous on* Ω*, then the flux trough each half edge* e*, computed by the L-method, can be written as*

$$f_e = -\boldsymbol{K}\nabla p \cdot \boldsymbol{n}_e \tag{3.12}$$

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

Chapter 4

Richards' 2

In this chapter we discuss numerical solving algorithms for parabolic equations with non-linearities, such as Richards' equation (1.10).

Time discretization

We start by considering the most famous parabolic equation, namely the heat equation:

$$\begin{aligned} \partial_t u - \nabla \cdot \boldsymbol{K} \nabla u &= F & x &\in \Omega & t &\in (0,T] \\ u &= 0 & x &\in \partial\Gamma_D & t &\in (0,T] \\ \boldsymbol{K}\nabla u &= g_N & x &\in \partial\Gamma_N & t &\in (0,T] \\ u &= u_0 & x &\in \Omega & t &= 0 \end{aligned} \tag{4.1}$$

We expect low regularity in time, so there is not much 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 long timesteps. This can be understood intuitively by considering the parabolic nature of the equation, the signals spread trough the domain instantaneously. A careful analysis time discretization of parabolic equations is done in ([5], chapter 7). Here it is shown that explicit schemes only are stable for time-step proportional to the square of the space step, whereas fully implicit schemes are stable for all time-steps.

Let $\{t_n\}_n$ be a sequence of $N+1$ evenly spaced numbers from 0 to T and let $\Delta t = \frac{T}{N}$ be the time-step Then we state the semi-discrete version of (4.1) by exchanging the time derivative by a difference quotient $(\partial_t u)^n = \frac{u^n - u^{n-1}}{\Delta t}$. Note that this difference quotient is implicit because u^n is not explicitly given by terms

of the previous time-step

$$\begin{aligned} u^n - \Delta t \nabla \cdot \boldsymbol{K} \nabla u^n &= \Delta t F^n + u^{n-1} && x \in \Omega \\ u^n &= 0 && x \in \partial\Gamma_D \\ \boldsymbol{K}\nabla u &= g_N && x \in \partial\Gamma_N \\ u^0 &= u_0 && x \in \Omega \end{aligned} \tag{4.2}$$

Now we have an elliptic problem (4.2) 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 that we have a 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 (4.2) is achieved as before by multiplying by test functions in $H_0^1(\Omega)$:

$$\begin{gathered} \text{find } u^n \in V \text{ such that} \\ \langle u^n, v\rangle_0 + \tau \langle \boldsymbol{K}\nabla u^n, \nabla v\rangle_0 = \tau \langle F^n, v\rangle_0 + \langle u^{n-1}, v\rangle_0 \\ \text{for all } v \in V \end{gathered} \tag{4.3}$$

If we swap V with a finite dimensional subspace V_h, and write $u_h^n = \sum_{i=1}^d (u_i^*)^n \phi_i$, as in the Galerkin FEM section, we end up with the system.

$$\begin{gathered} \text{find } (\boldsymbol{u}^*)^n \in \mathbb{R}^d \text{ such that} \\ \boldsymbol{B}(\boldsymbol{u}^*)^n + \tau \boldsymbol{A}(\boldsymbol{u}^*)^n = \tau \boldsymbol{F}^n + \boldsymbol{B}(\boldsymbol{u}^*)^{n-1} \end{gathered} \tag{4.4}$$

Where the *stiffness matrix*, $\boldsymbol{A}$, is as before. 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 control volumes $\{\Omega_i\}_i$. One could write the heat equation(4.1) in conservation form on each control volume

$$\partial_t \int_{\Omega_i} u \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} F \, dx \tag{4.5}$$

And discretize the first term with backward Euler. Or one could make sure the semi-discrete heat equation (4.1) holds for each control volume and use the divergence theorem. Both ways, we end up with

$$\int_{\Omega_i} u^n \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} F^n \, dx + \int_{\Omega_i} u^{n-1} \, dx \tag{4.6}$$

As in chapter 3 we end up with a system of equations, where superscript V is just to distinct between FVM and FEM.

$$(\boldsymbol{B}^V + \tau \boldsymbol{A}^V)\boldsymbol{u}^n = \tau \boldsymbol{F}^n + \boldsymbol{B}^V \boldsymbol{u}^{n-1} \tag{4.7}$$

The matrix $\boldsymbol{A}^V$ is as in chapter 3 with $\boldsymbol{A}^V_{i,j} = \tilde{f}_{i,j}$ is the flux between cell i and j. The matrix $\boldsymbol{B}^V$ is diagonal with the entry i being the volumes of the volume of cell i.

If $\boldsymbol{A} = \boldsymbol{A}^V$, ie. that the discretization of the constitutive law is the same for both finite volume and finite element method. We can modify the mass matrix and the load vector of the finite element method so that they become equivalent. To achieve this we define an interpolation operator as in (Cao and Wolmuth [8],2009):

Definition 11 (Piecewise global interpolator)**.** *Let $\hat{I}_h$ be an operator that maps from the test space to functions that are piecewise continuous on each control volume.*

$$\hat{I}_h : C(\Omega) \to \left\{ v_h \in L^2(\Omega) : v_h|_{\Omega_i} = K \right\}$$

And

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

Where

$$\hat{I}_h \phi_i(x) = \begin{cases} 1 & \text{if } x \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

Lemma 4.0.1. *Suppose the stiffness matrix $\boldsymbol{A}$ arising from the finite element discretization* (4.4) *is equivalent to the corresponding matrix in the finite volume discretization* (4.7) *for some finite volume method. Then the Galerkin discretization*

$$\begin{gathered} \text{find } u_h^n \in V_h \\ \left\langle \hat{I}_h u_h^n, \hat{I}_h v_h \right\rangle_0 + \tau \left\langle \boldsymbol{K} \nabla u_h^n, \nabla v_h \right\rangle_0 = \tau \left\langle F^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h u_h^{n-1}, \hat{I}_h v_h \right\rangle_0 \\ \forall v_h \in V_h \\ \text{where } V_h = \left\{ u_h \in C(\overline{\Omega}) : u_{h|K} \in P_1(K) \ \forall K \in \tau_h, u|_{\partial\Omega} = 0 \right\} \end{gathered} \tag{4.9}$$

Is equivalent to the finite volume method (4.7)

Remark 10. *Note that the use of the $\hat{I}_h$ operator is the same as using the trapezoidal quadrature rule to compute the inner product $\langle \phi_i, \phi_j \rangle_0$ instead of evaluating it exactly. This procedure is known as* ***mass lumping*** *and is shown in detail in (Baranger [9],1995).*

Proof. Let $u_h = \sum_i u_i^* \phi_i$, and let the test functions in (4.9) be the basis functions $\{\phi_j\}_j$, then we have

$$\sum_{i\in\mathcal{N}}(u^*)_i^n(\left\langle \hat{I}_h\phi_i, \hat{I}_h\phi_j \right\rangle_0 + \tau \left\langle \boldsymbol{K}\nabla\phi_i, \nabla\phi_j \right\rangle_0) = \tau \left\langle F^n, \hat{I}_h\phi_j \right\rangle_0 + \sum_{i\in\mathcal{N}}(u^*)_i^{n-1} \left\langle \hat{I}_h\phi_i, \hat{I}_h\phi_j \right\rangle_0 \tag{4.10}$$

Now we see by the definition of the piecewise interpolator that we get:

$$\left\langle \hat{I}_h\phi_i, \hat{I}_h\phi_j \right\rangle_0 = \int_{\text{supp}(\phi_i)\bigcap\text{supp}(\phi_j)} 1 \; dx = |\Omega_i| \tag{4.11}$$

This means that the mass matrix is a diagonal matrix as in the finite volume method. Similarly we have

$$\left\langle F^n, \hat{I}_h\phi_j \right\rangle_0 = \int_{\Omega_j} F^n \; dx \tag{4.12}$$

□

Equivalence between modified FEM and modified MPFA-L

Linearization

Now 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 (1.10) leads to a system of non-linear equations. We start by discussing this in a general setting

$$\text{find } x \in U \text{ such that } \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \text{ where } f : U \subset \mathbb{R}^n \to \mathbb{R}^n \tag{4.13}$$

The solution in (4.13) is called a *root*, it is almost always found using an iterative method.

A common iterative scheme to solve (4.13) is the *Newton method*, let $D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ be the Jacobian of $\boldsymbol{f}(\boldsymbol{x}_{j-1})$.

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} - D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{4.14}$$

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 4.0.2. *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*

$$|x_{j+1} - \overline{x}| \leq \frac{1}{\delta}|x_j - \overline{x}|^2 < |x_j - \overline{x}| \tag{4.15}$$

Proof. Define $e_n = x_n - \overline{x}$. Then we have by Taylor expansion

$$0 = f(\overline{x}) = f(x_j - e_j) = f(x_j) - f'(x_j)e_j + \frac{f''(\psi)e_j^2}{2} \tag{4.16}$$

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

$$\begin{aligned} e_{j+1} = x_{j+1} - \overline{x} &= x_n - \frac{f(x_j)}{f'(x_j)} - \overline{x} \\ &= e_j - \frac{f(x_j)}{f'(x_j)} \\ &= \frac{e_j f'(x_j) - f(x_j)}{f'(x_j)} \end{aligned} \tag{4.17}$$

By the Taylor expansion around x_j, (4.16), we get

$$e_{j+1} = \frac{e_j^2 f''(\psi)}{2f'(x_j)} \tag{4.18}$$

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

$$|e_1| \leq \frac{2}{2\delta}|e_0|^2 < |e_0| \tag{4.19}$$

By the same reasoning we get convergence

$$|e_{j+1}| < |e_j| \tag{4.20}$$

And the quadratic convergence

$$|e_{j+1}| \leq \frac{1}{\delta}|e_j|^2 \tag{4.21}$$

□

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

$$D\boldsymbol{f}(\boldsymbol{x}_{j-1})\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{4.22}$$

And then update the current iterate with $\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + \boldsymbol{\delta}_j$. One often end end up with a situation where the matrix $D\boldsymbol{f}(\boldsymbol{x}_{j-1})$ needs to be computed and assembled for every iteration. This may be computationally expensive. So Newtons method may be slow despite it's quadratic convergence, if it even converges.

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

$$L\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{4.23}$$

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

Theorem 4.0.3. *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

$$e_{j+1} = x_j - \frac{f(x_j)}{L} - \overline{x} = e_j - \frac{f(x_j)}{L} \tag{4.24}$$

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

$$0 = f(\overline{x}) = f(x_j - e_j) = f(e_j) - f'(\psi)e_j \Rightarrow e_j = \frac{f(x_j)}{f'(\psi)} \tag{4.25}$$

Using this and the assumption on L we get the estimate

$$|e_{j+1}| = |e_j(1 - \frac{f'(\psi)f(x_j)}{f(x_j)L})| \leq |e_j||1 - \frac{f'(\psi)}{L}| < |e_j| \tag{4.26}$$

□

To see how this could be applied to parabolic PDE's we consider the equation: We will first consider the equation

$$\begin{aligned} \partial_t\theta(u) - \nabla\cdot\kappa(u)\nabla u = F \quad & x \in \Omega \quad & t \in (0,T] \\ u = g \quad & x \in \partial\Omega \quad & t \in (0,T] \\ u = u_0 \quad & x \in \Omega \quad & t = 0 \end{aligned} \tag{4.27}$$

Which is similar to Richards' equation (1.10), only without the gravity. We proceed as before by backward euler in time and we then discretize the elliptic PDE with the Galerkin discretization with mass lumping (4.9). We then get:

$$\begin{gathered} \text{find } u_h^n \in V_h \\ \left\langle \hat{I}_h\theta(u_h^n), \hat{I}_h v_h \right\rangle_0 + \tau \left\langle \kappa(u_h^n)\nabla u_h^n, \nabla v_h \right\rangle_0 = \tau \left\langle F^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h u_h^{n-1}, \hat{I}_h v_h \right\rangle_0 \\ \text{for all } v_h \in V_h \end{gathered} \tag{4.28}$$

We can then linearize $\theta(u_h^n)$ with the L-method, such that we en up with: Given u_h^{n-1} and $u_h^{j-1,n}$ find $u_h^{j,n}$ such that:

$$\begin{aligned}\left\langle \hat{I}_h\theta(u_h^{n,j-1}), \hat{I}_h v_h \right\rangle_0 + L\left\langle \hat{I}_h u_h^{j,n} - \hat{I}_h u_h^{j-1,n}, \hat{I}_h v_h \right\rangle_0 + \tau\left\langle \kappa(u_h^{j-1,n})\nabla u_h^{j,n}, \nabla v_h \right\rangle \\ = \tau\left\langle F^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h u_h^{n-1}, \hat{I}_h v_h \right\rangle_0\end{aligned} \tag{4.29}$$

Theorem 4.0.4. *Assume θ and κ monotone ...*

Equivalence between MPFA-L method and modified finite element method

In this section we prove the equivalence for the elliptic model problem (4.30) on inhomogeneous media discretized by a parallelogram grid.

$$\begin{aligned} -\nabla\cdot\boldsymbol{K}\nabla u(x) &= F(x) \quad && x\in\Omega \\ u(x) &= 0 && x\in\Gamma_D \\ \boldsymbol{K}\nabla u(x) &= g_N && x\in\Gamma_N \end{aligned} \tag{4.30}$$

We

Chapter 5

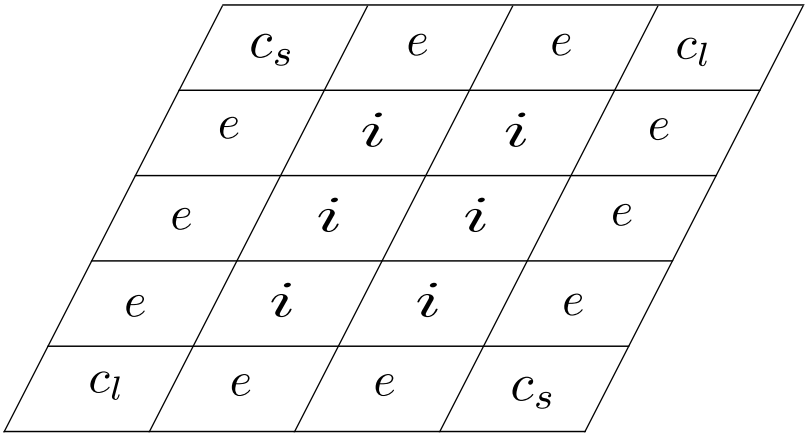
Equivalence between MPFA-L and FEM

In this chapter we show equivalence between a modified MPFA-L method and a modified finite element method for time dependent problems such as (4.2). We saw in the section about the MPFA-L method that the interaction regions(L-triangles) may form a triangulation of our domain. Modifications are made to both methods so that we can exploit this fact and obtain equivalence. This section is adapted from

add citation

Modified MPFA-L method

First of all we assume that we have a parallelogram grid.



To obtain a finite volume discretization one could write the heat equation(4.1) in conservation form on each control volume

$$\partial_t \int_{\Omega_i} u \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} F \, dx \tag{5.1}$$

And discretize the first term with backward Euler. Or one could make sure the semi-discrete heat equation (4.2) holds for each control volume and use the divergence theorem. Both ways, we end up with

$$\int_{\Omega_i} u^n \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} F^n \, dx + \int_{\Omega_i} u^{n-1} \, dx \tag{5.2}$$

The MPFA-L method deals with the second term, approximating the constitutive law. The other three terms are common to all control volume methods solving (4.2) and we will not make modifications or discuss them further.

We will 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 $y_1 y_6 y_4 y_3$.

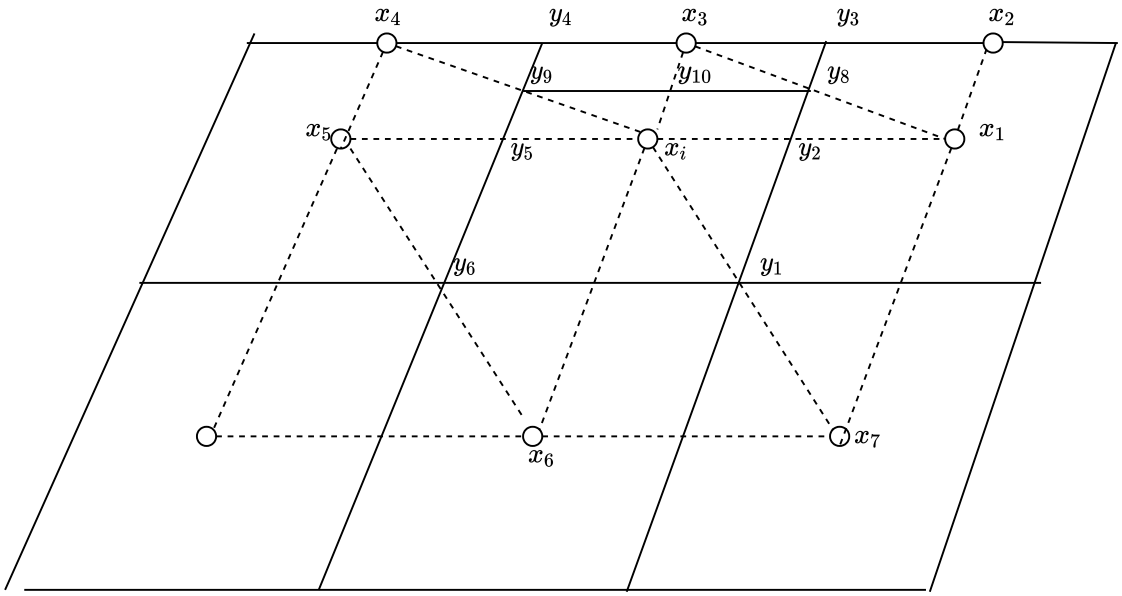


Figure 5.1: Control volumes in solid lines and interaction regions in dashed lines at the boundary.

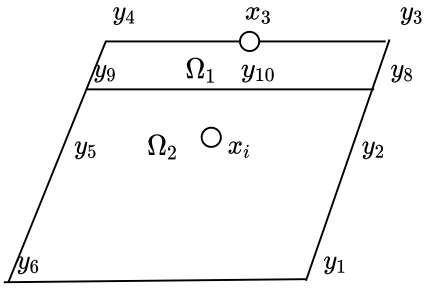


Figure 5.2: Control volume along top boundary.

For the **Neumann** boundary conditions, we split the cell into two, $y_1 y_6 y_9 y_8$ as Ω_2 and $y_8 y_9 y_4 y_3$ as Ω_1. 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 trough $\overline{y_3 y_8}$ using $\triangle x_1 x_3 x_2$, the flux trough $\overline{y_8 y_{10}}$ using $\triangle x_1 x_i x_3$, for $\overline{y_{10} y_9}$ and $\overline{y_9 y_4}$ the L triangle $\triangle x_i x_4 x_3$ is used. Finally the Neumann boundary condition is used at the the edge $\overline{y_4 x_3}$ and $\overline{x_3 y_3}$. We are not able to eliminate the unknown value at x_3 and it remains a degree of freedom, which makes sense if we want equivalence with finite element method.

In the case of **Dirichlet** boundary conditions, we compute the fluxes into $y_1y_6y_4y_3$ using seven L-triangles. 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.

The flux over $\overline{y_4y_3}$, at the boundary, we compute by balancing it by the other fluxes into the small control volume Ω_1. Let $f_{\overline{y_iy_j}} := \int_{\overline{y_iy_j}} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}}\, dx$

$$f_{\overline{y_4y_3}} = f_{\overline{y_3y_8}} + f_{\overline{y_8y_{10}}} + f_{\overline{y_{10}y_9}} + f_{\overline{y_9y_4}} \tag{5.3}$$

The fluxes on the right hand side of (5.3) are computed as for the Neumann case.

Modified finite element method

In this section we introduce a finite element method for solving (4.2). We start by observing that by theorem the L-triangles form a triangulation. The only modifications we need to make are to the linear form, we let the bi-linear form stay the same as before. We want to define an interpolation operator such that the dot products that make up the linear form become mass conservative in each control volume. We need some notation so that we can distinguish between nodes in the interior, at at cell centers along the boundary and at the boundary. In addition, corner cells introduce edge cases.

add theorem in L-method chapter

Let $\mathcal{N}_h^*$ be a set of indexes corresponding to all interior nodes of τ_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. The boundary nodes consists of the set $\mathcal{N}_h^N \bigcup \mathcal{N}_h^D$, where N and D represent neumann and dirichlet boundary nodes. The rest of the notation are explained in figure **??**

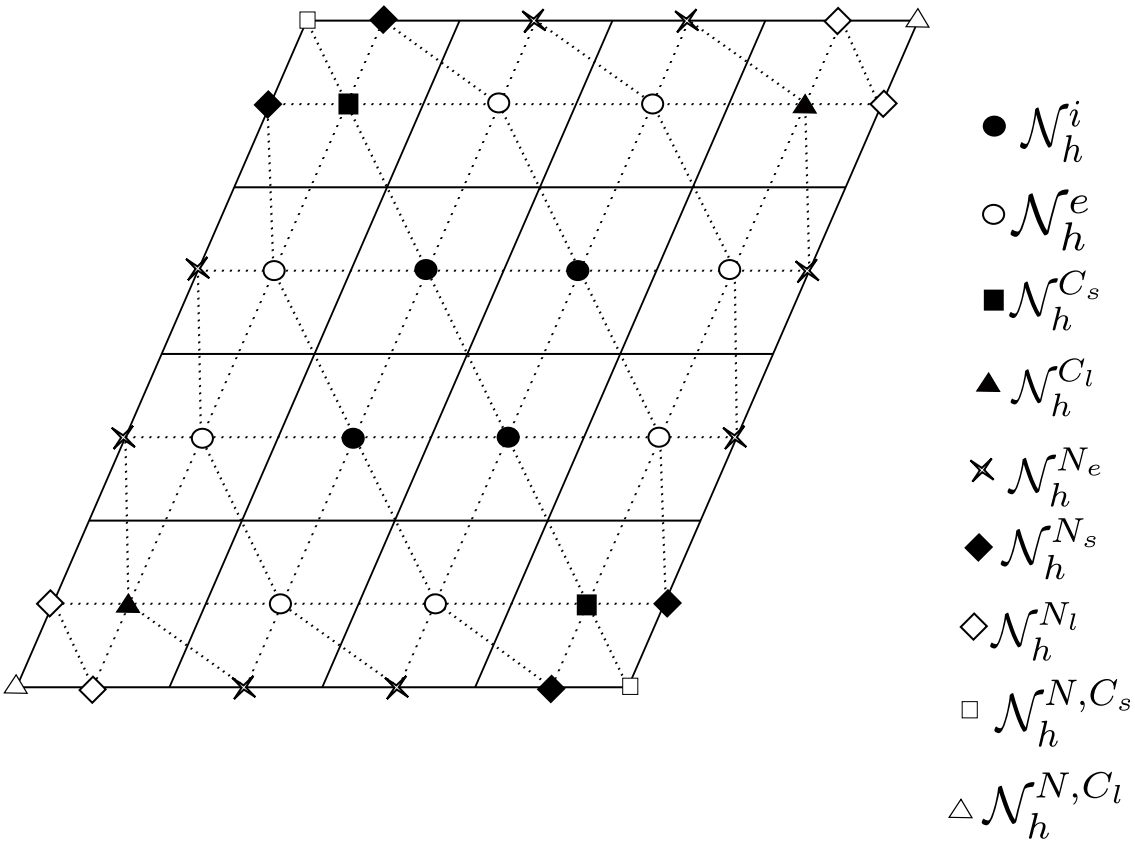


Figure 5.3: A paralellogram mesh with finite element triangles in dotted lines and control volumes in solid lines.

As before one denotes by V_h the linear ansatz space, see definition 9. 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 10, 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 each control volume.*

$$\hat{I}_h : C(\Omega) \to \left\{ v_h \in L^2(\Omega) : v_h|_{\Omega_i} = K \right\}$$

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

$$\hat{I}_h \phi_i(x) = \begin{cases} 1 & \text{if } x \in D_i \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

$D_i = \Omega_i$, ie. the corresponding control volume, if $i \in \mathcal{N}_h^i$. If we are close or on the boundary the situation is more complicated:

- *$i \in N_h^e$: In this case the function vanishes for the quarter of the paralellogram closest to the boundary, ie. $D_i = \Omega_2$ from figure* **??**
- *$i \in \mathcal{N}_h^{N_e}$ 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* **??**
- *On the corners there are special definitions, see (Cao Wolmuth [10],2009)*

The finite element method we end up with reads as follows:

$$\text{find } u_h^n \in V_h \text{ such that}$$

$$\left\langle \hat{I}_h u_h^n, \hat{I}_h v_h \right\rangle_0 + \tau \left\langle \boldsymbol{K}\nabla u_h^n, \nabla v_h \right\rangle_0 = \tau \left\langle F^n, \hat{I}_h v_h \right\rangle_0 + \left\langle \hat{I}_h u_h^{n-1}, \hat{I}_h v_h \right\rangle_0 + \left\langle g, \hat{I}_h v_h \right\rangle_0 \tag{5.5}$$

The key takeaway here is that the inner products become conservative in the control volumes.

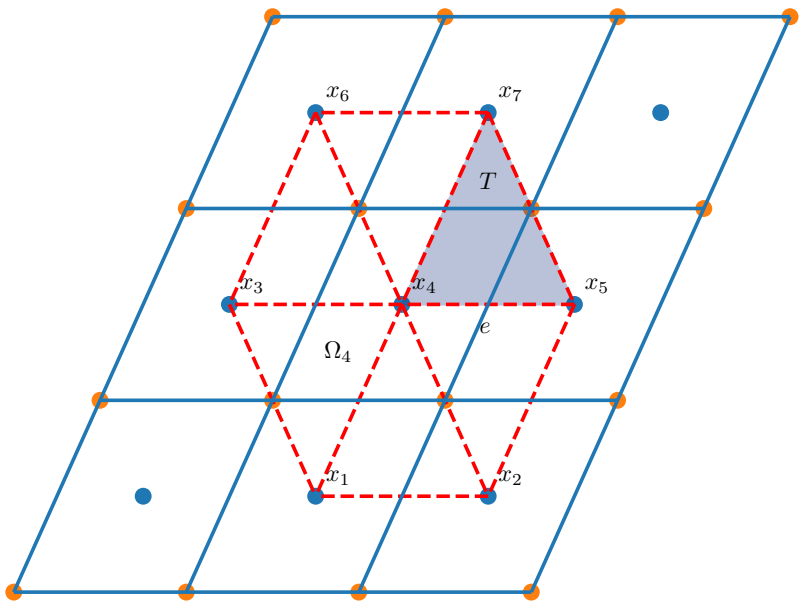


Figure 5.4: The support of ϕ_4

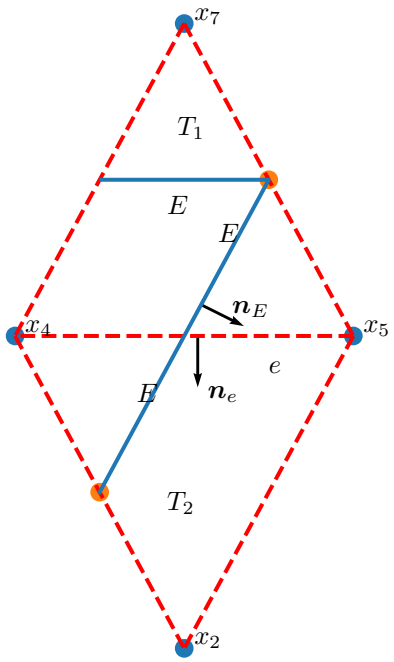


Figure 5.5: Notation in the proof

Now we can state the equivalence theorem:

Theorem 5.0.1. *The modified finite element* (5.5) *method and the modified MPFA-L method are equivalent on uniform parallelogram grid for the heat equation* (4.1) *on homogenous media.*

Proof.

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, ie. $i \in \mathcal{N}_h^i$. $T \in \tau_h \bigcap \text{supp}(\phi_i)$ is an element of the triangulation. $S = T \bigcap \Omega_i$ 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 **??** and **??**. Since u_h and ϕ

is piecewise linear and $\hat{\boldsymbol{K}}$ is constant on each triangle T we have:

$$\begin{aligned}
\langle \boldsymbol{K}\nabla u_h, \nabla\phi_i\rangle_0 &= \int_{\mathrm{supp}(\phi_i)} (\nabla u_h)^T \boldsymbol{K}\nabla\phi_i \; dx = \sum_{T\in\mathrm{supp}(\phi_i)} \int_T (\nabla u_h)^T \boldsymbol{K}\nabla\phi_i \; dx \\
&= \sum_{T\in\mathrm{supp}(\phi_i)} \left(\int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_i \; ds - \int_T \nabla\cdot\boldsymbol{K}\nabla u_h\phi_i \; dx \right) \\
&= \sum_{T\in\mathrm{supp}(\phi_i)} \int_{\partial T} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}\phi_i \; ds \\
&= \sum_{e\in\mathrm{supp}(\phi_i)} \int_e ((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_j} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{j+1}})\phi_i \; ds \\
&= \sum_{e\in\mathrm{supp}(\phi_i)} ((\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_j} - (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_e|_{T_{j+1}}) \frac{|e|}{2} \\
&= \sum_{S\in\mathrm{supp}(\phi)} \int_{\partial S} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n} \; ds - \sum_{E\in\partial\Omega_i} \int_E (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E \; ds \\
&= \sum_{S\in\mathrm{supp}(\phi)} \int_S \nabla\cdot\boldsymbol{K}\nabla u_h \; ds - \sum_{E\in\partial\Omega_i} \int_E (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E \; ds \\
&= -\sum_{E\in\partial\Omega_i} (\boldsymbol{K}\nabla u_h)^T \boldsymbol{n}_E |E|
\end{aligned} \tag{5.6}$$

Hence we have showed that the bi-linear form from the finite element method is equivalent to the flux integral in the finite volume L method. For the source terms and the mass matrix the equivalence is obvious:

$$\left\langle \hat{I}_h u_h, \hat{I}\phi_i \right\rangle_0 = \int_\Omega \hat{I}_h u_h \hat{I}_h \phi_i \; dx = u_h(x_i) \int_{\Omega_i} dx \tag{5.7}$$

□

This chapter is adapted from (Cao Wolmuth [10],2009). We prove the equivalence for the elliptic model problem (4.30) on homogeneous media discretized by a parallelogram grid.

$$\begin{aligned}
-\nabla\cdot\boldsymbol{K}\nabla u(x) &= F(x) \quad && x\in\Omega \\
u(x) &= 0 && x\in\Gamma_D \\
\boldsymbol{K}\nabla u(x) &= g_N && x\in\Gamma_N
\end{aligned} \tag{5.8}$$

Chapter 6

Richards' equation

In this chapter we discuss numerical solving algorithms for the Richards' equation (1.10). We start by discretizing in time, then handling the non-linearities and finally using the elliptic discretizations we discussed in chapter 2 and 3. We consider a simplified version of Richards' equation where the gravity and the non-linearity in the permability have been removed.

$$\begin{aligned} \partial_t \theta(\psi) - \nabla \cdot \boldsymbol{K} \nabla p &= F \quad x \in \Omega \quad t \in (0,T] \\ \psi &= g \quad x \in \partial\Omega \quad t \in (0,T] \\ \psi &= u_0 \quad x \in \Omega \quad t = 0 \end{aligned} \tag{6.1}$$

We expect low regularity in time, so there is not much 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 long timesteps. Let $\{t_n\}_n$ be a sequence of $N+1$ evenly spaced numbers from 0 to T and let $\tau = \frac{T}{N}$ be the timestep. Then we state the semidiscrete version of (6.1) by exchanging the time derivative by a difference quotient $\partial_t \theta(\psi^n) = \frac{\theta(\psi^n) - \theta(\psi^{n-1})}{\tau}$

show why?

$$\theta(\psi^n) - \tau \nabla \cdot \boldsymbol{K} \nabla \psi^n = \tau F^n + \theta(\psi^{n-1}) \tag{6.2}$$

Definition 13. *content...*

explain and possibly prove convergence of L-scheme

The modified finite element method

Let V_h be the finite dimensional test space as defined in the Galerkin FEM section, definition 9, with $\{\phi_i\}_i$ beieng the standard nodal basis. To make our finite element method conservative, we define as in (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2011) [8]), another interpolation operator.

Definition 14 (Piecewise global interpolator)**.** *Let $\hat{I}_h$ be an operator that maps from the test space to functions that are piecewise continuous on each control volume.*

$$\hat{I}_h : C(\Omega) \to \left\{ v_h \in L^2(\Omega) : v_h|_{\Omega_i} = K \right\}$$

And

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

Where

$$\hat{I}_h \phi_i(x) = \begin{cases} 1 & \text{if } x \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

Extend this definition with some figures and for the boundary

The linearized system arising from the simplyfied Richards equation (6.1) now becomes:

$$\begin{gathered} \text{find } \psi^{n,j} \in V_n \text{ such that} \\ \left\langle \hat{I}_h \theta(\psi^{n,j-1}), \hat{I}_h v_h \right\rangle + L \left\langle \psi^{n,j} - \psi^{n,j-1}, \hat{I}_h v_h \right\rangle \\ + \tau \left\langle \boldsymbol{K} \nabla \psi^{n,j}, \nabla v_h \right\rangle = \tau \left\langle F^n, \hat{I}_h v_h \right\rangle + \left\langle \theta(\psi^{n-1}), \hat{I}_h v_h \right\rangle \\ \text{for all } v_h \in V_h \end{gathered} \tag{6.4}$$

Theorem 6.0.1. *Assume a homogenous domain discretized with paralellograms. Then the MPFA-L method with L-scheme linearization for Richards' equation gives a system that is equivalent to the modified Finite element method* (6.4)

Proof. If we in equation (6.4) test with the basis functions ϕ_i and express the solution u_h as $u_h = \sum u_j^* \phi_j$ we end up with the system

$$\begin{aligned} A_{i,j} &= L \left\langle \phi_i, \hat{I}_h \phi_j \right\rangle + \tau \left\langle \boldsymbol{K} \nabla \phi_i, \nabla \phi_j \right\rangle \\ B_i &= \left\langle \tau F^n + \theta(\psi^{n-1}) + L\psi^{n,j-1} - \theta(\psi^{n,j-1}), \hat{I}_h \phi_i \right\rangle \end{aligned} \tag{6.5}$$

1. $L \left\langle \phi_i, \hat{I}_h \phi_j \right\rangle$ is equivalent to $L \int_{\Omega_i} dx \delta_{ij}$
2.
3.

□

Lemma 6.0.2. *The Bi-linear form in the modified finite element method* (6.4)

$$a_h(v,w) = L\left\langle v, \hat{I}_h w\right\rangle + \tau\left\langle \boldsymbol{K}\nabla v, \nabla w\right\rangle \tag{6.6}$$

Is coercive

Proof. content.. □

Lemma 6.0.3 (First Lemma of Strang, page 155 [5])**.** *Suppose there exists some* $\alpha > 0$ *such that for all* $h > 0$ *and* $v \in V_h$

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

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

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

Chapter 7

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 [7]. We consider the elliptic model problem.

$$\begin{aligned} \nabla \cdot \boldsymbol{q} &= f \\ \boldsymbol{q} &= -\boldsymbol{K}\nabla u \end{aligned} \tag{7.1}$$

We set the solution

$$u = cosh(\pi x)cos(\pi y) \tag{7.2}$$

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 (7.1). The flux term, $\boldsymbol{q}$, could for example be use to compute the transport of some contaminant in a porous medium.
As in [7] page 1340 we define the normalized discrete L_2 norms:

$$\|u - u_h\| = \left(\frac{1}{V} \sum_i V_i (u_{h,i} - u_i)^2 \right)^{\frac{1}{2}} \tag{7.3}$$

$$\|q - q_h\| = \left(\frac{1}{Q} \sum_a Q_a (q_{h,a} - q_a)^2 \right)^{\frac{1}{2}} \tag{7.4}$$

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

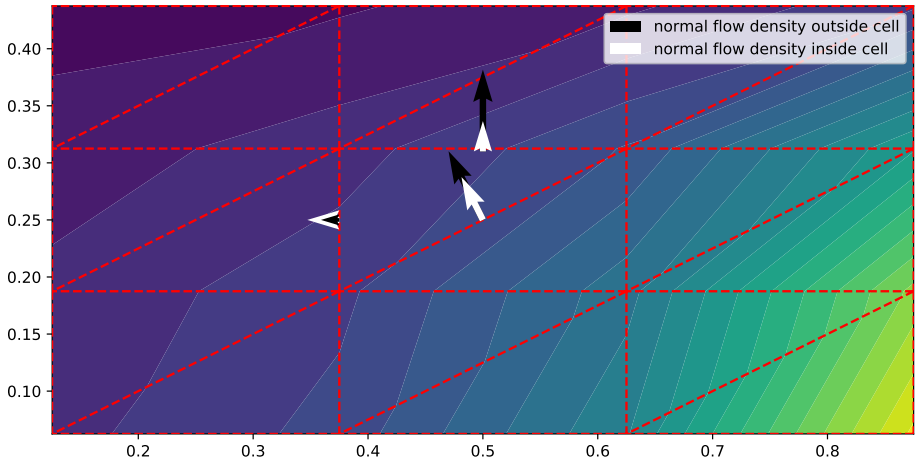


Figure 7.1: An illustration of why it's a bad idea to use the piecewise constant gradient on each element for computing normal flow. The normal flow is discontinuous across the edges, and flow into the cell is not equal to flow out.

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 (7.4) 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, [10]) 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 7.9, 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 7.4 and 7.5.

In the second setup with uniform trapezoidal mesh illustrated in figure 7.6 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 7.7 and 7.8.

In the perturbed mesh setup 7.9, 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 7.10 and 7.11.
The next setup is with perturbed mesh and an aspect ratio of 0.1, see figure 7.12 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.

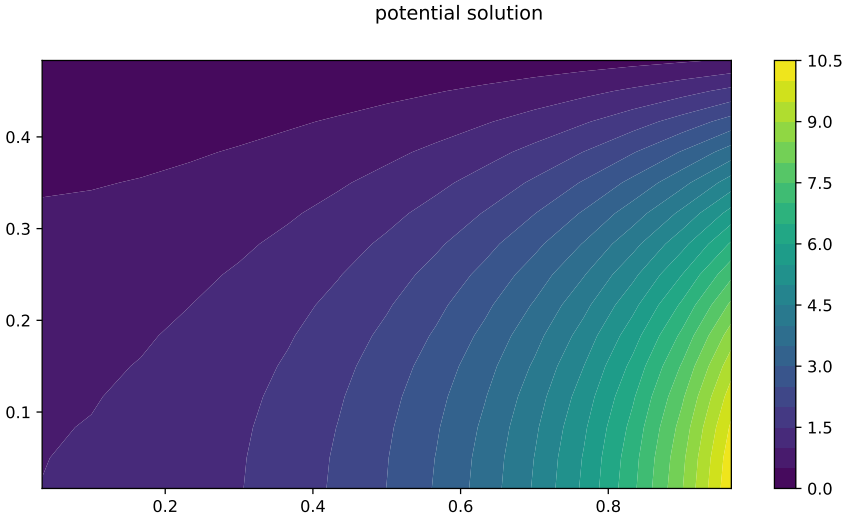


Figure 7.2: The solution (7.2) on half the unit square

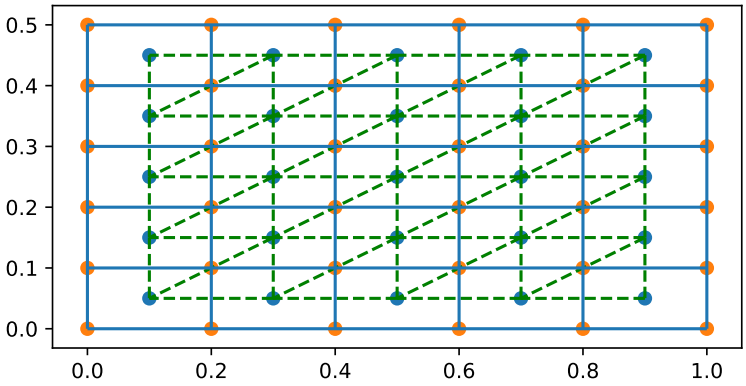


Figure 7.3: Unifrom rectangular mesh on half the unit square. The triangles are used for the finite element solution and are spanned between the nodes of the cell centers of the finite volume methods.

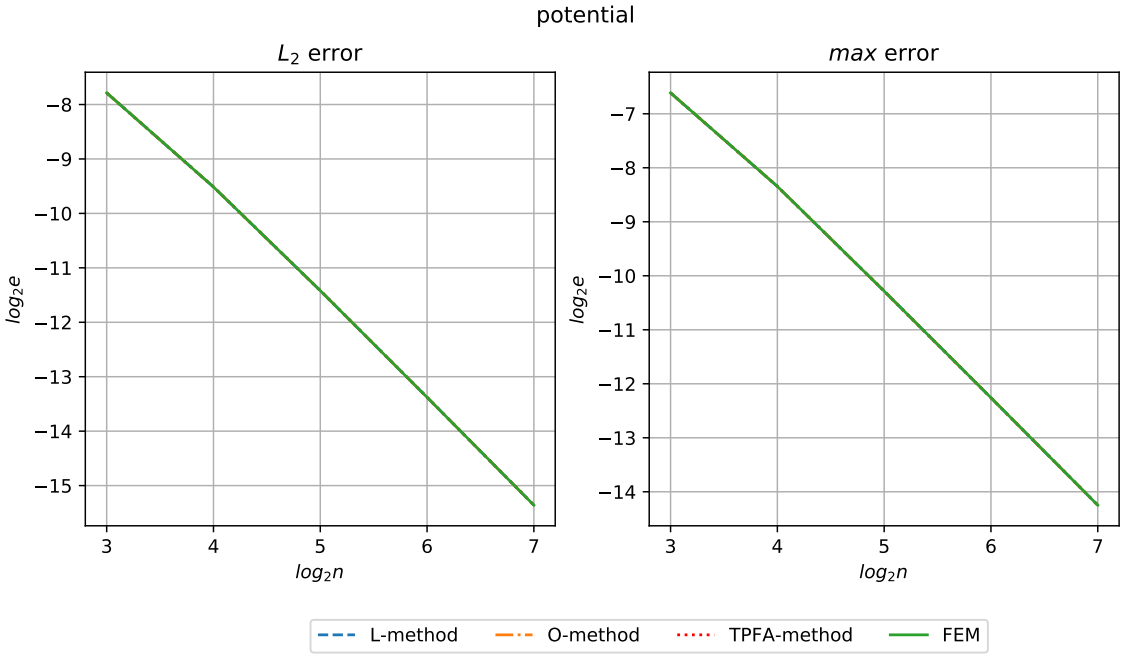


Figure 7.4: Potential error on refinements of the uniform rectangular mesh 7.3

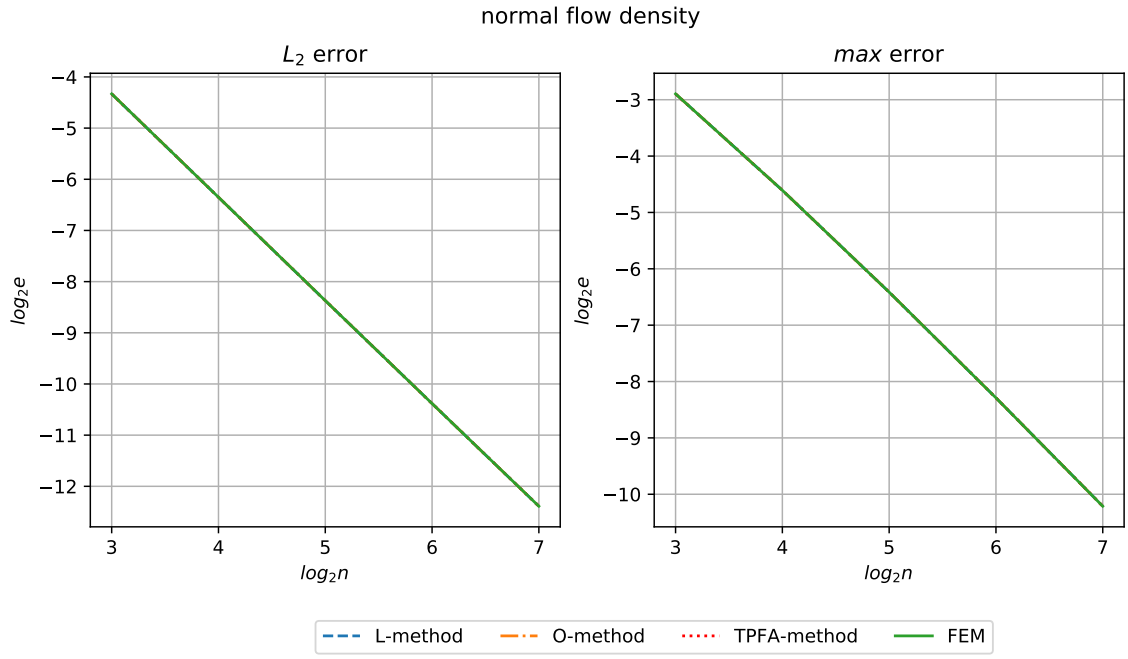


Figure 7.5: Normal flow density error on refinements of the uniform rectangular mesh 7.3

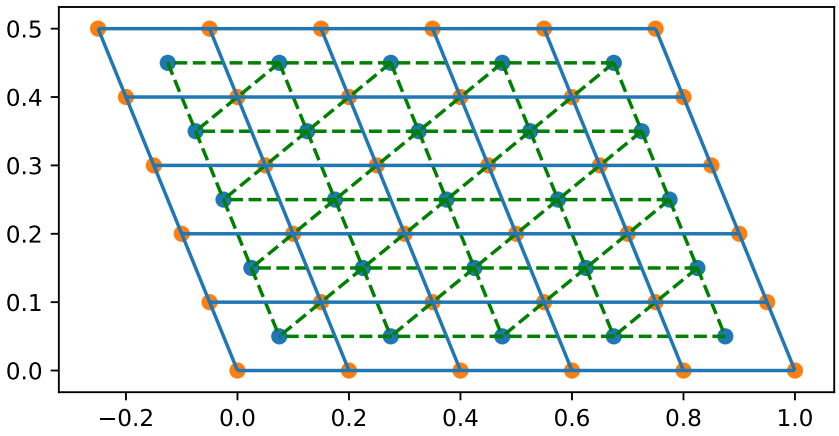


Figure 7.6: Trapezoidal mesh, now every point is transformed by $(x, y) \mapsto (x - 0.5y, y)$

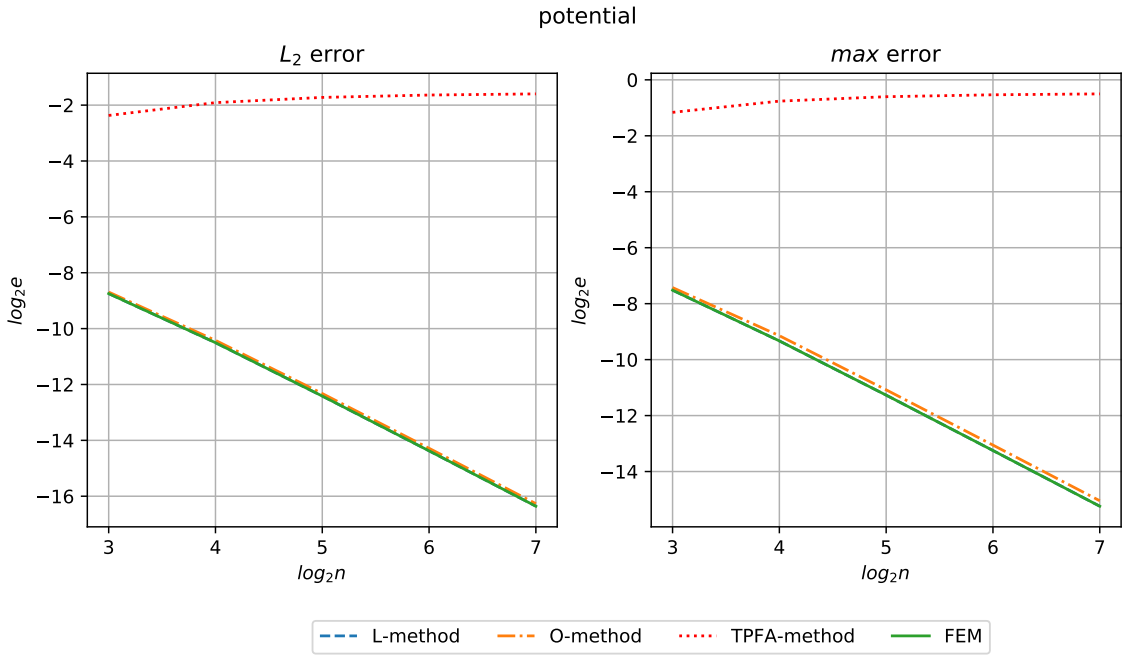


Figure 7.7: Pressure error on refinements of the mesh 7.6

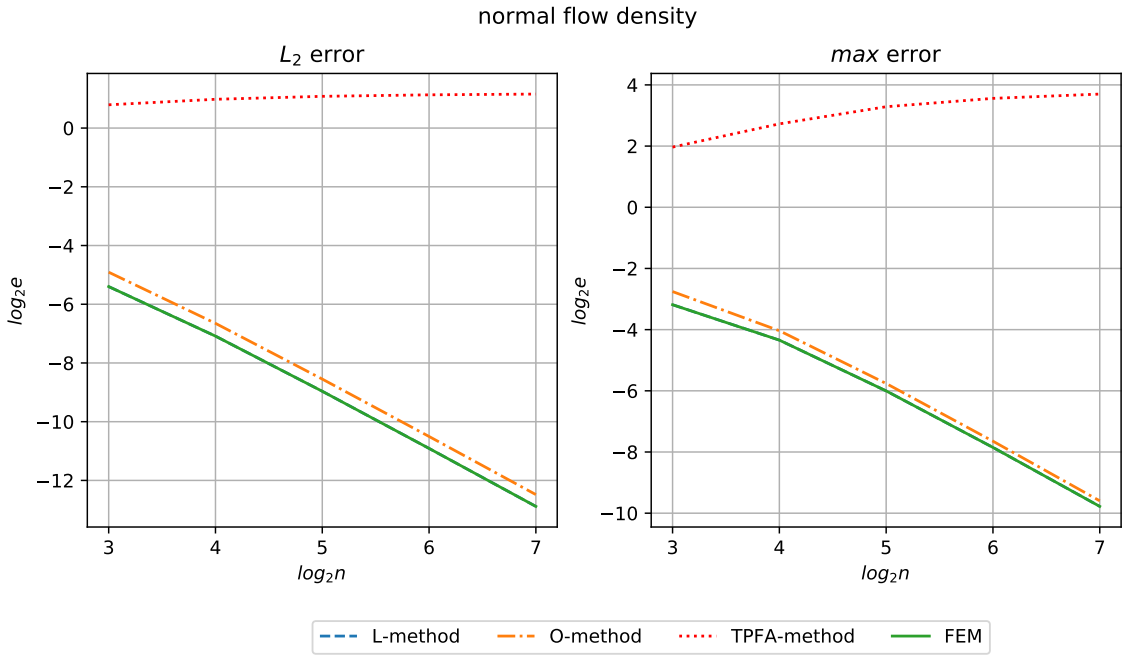


Figure 7.8: Normal flow density error on refinements of the mesh 7.6

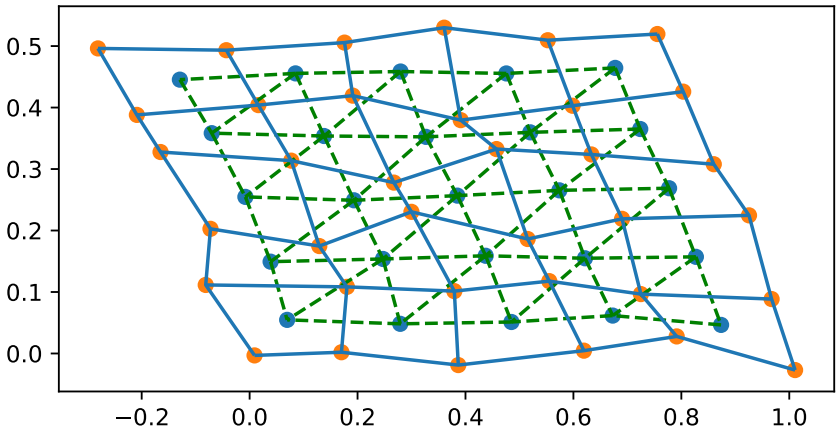


Figure 7.9: Perturbed mesh, every point in the mesh is perturbed by a random number which is $O(\frac{h}{5})$, in both x and y direction.

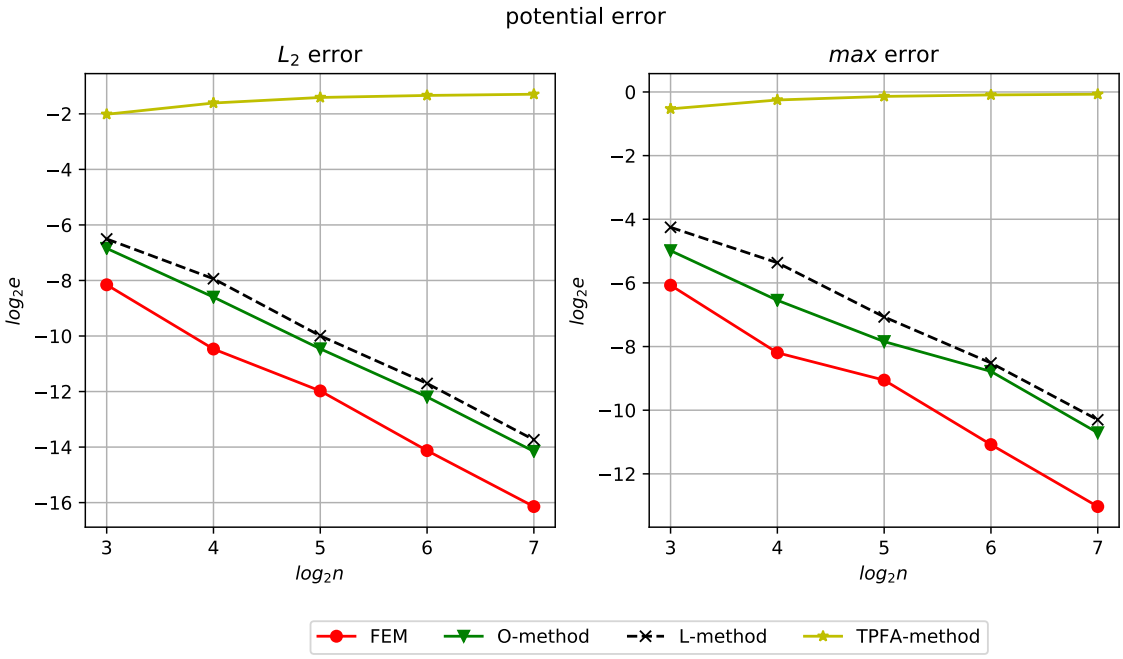


Figure 7.10: The pressure error of perturbed mesh.

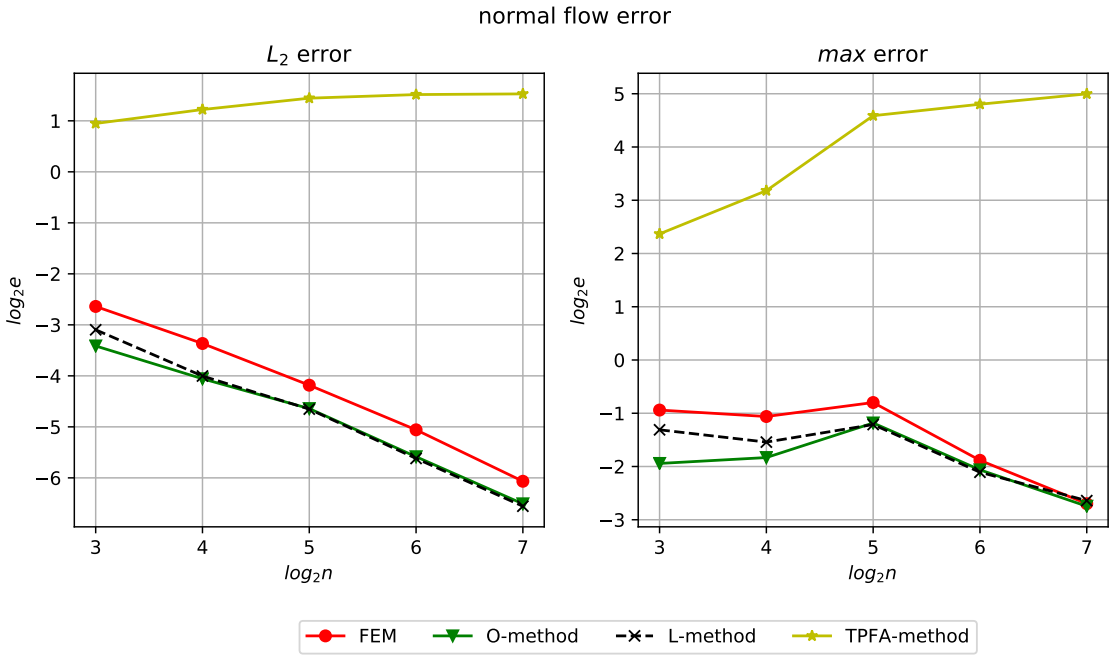


Figure 7.11: The normal flow density error of perturbed mesh

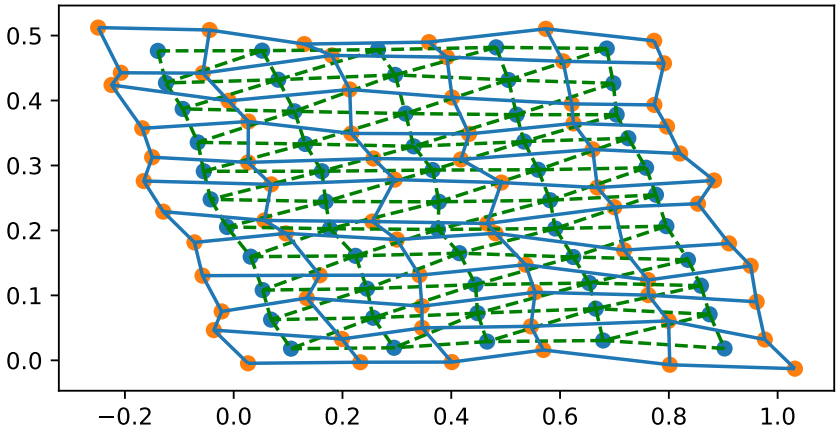


Figure 7.12: Perturbed mesh with aspect ratio 0.5, there are half as many points in the x-direction as in the y-direction.

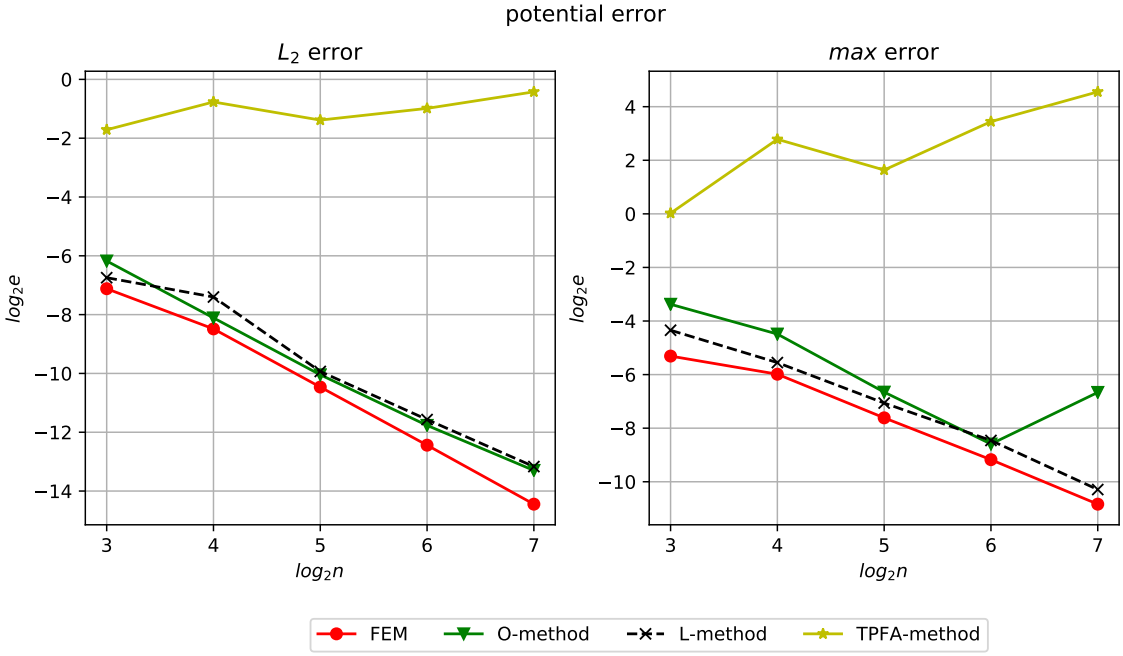


Figure 7.13: The pressure error of perturbed mesh with aspect ratio 0.1.

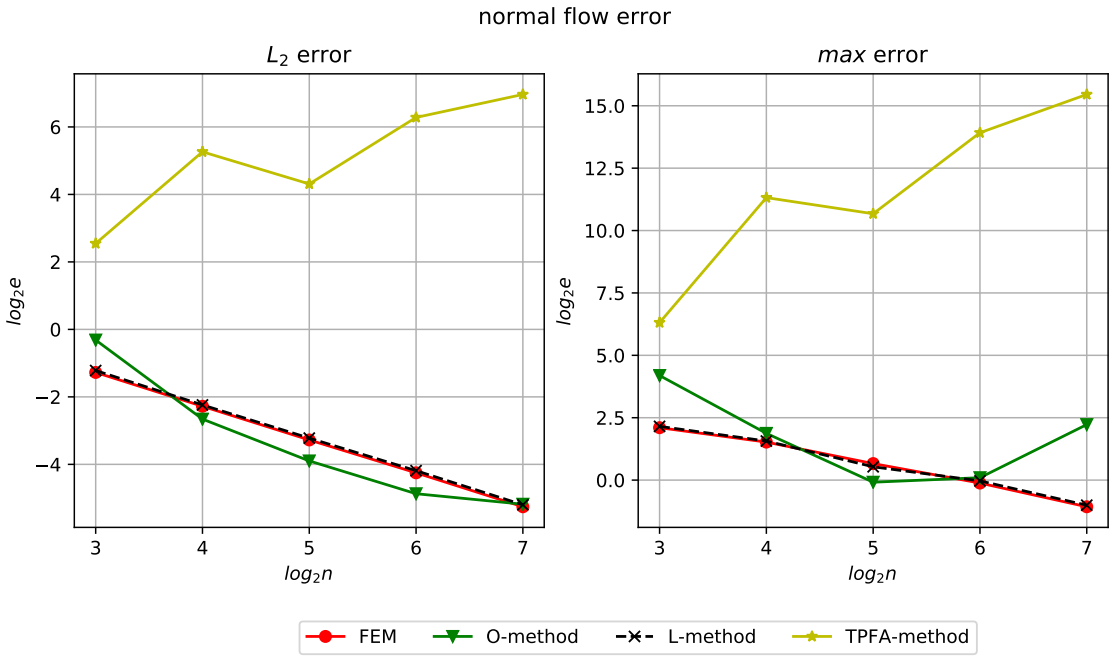


Figure 7.14: The normal flow density error of perturbed mesh with aspect ratio 0.1.

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] 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.

[8] 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.

[9] 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.

[10] 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.