Chapter 1

Numerical approximation techniques

Finite Element Method

The finite element method was first developed in the 1940s by Richard Courant for problems in solid mechanics. As computers became better in the 1960s the method became more mainstream [**?**]. Today there are several general purpose finite element programs being 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 $\Omega \subset \mathbb{R}^n$ be some open, bounded domain. Find u such that:

$$\begin{aligned} -\nabla \cdot \boldsymbol{K} \nabla u(x) &= f(x) \quad x \in \Omega \\ u(x) &= 0 \quad x \in \partial\Omega. \end{aligned} \tag{1.1}$$

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

Function spaces

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

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

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

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

Theorem 1.0.1 (Riesz-Fischer Theorem [**?**] 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 function spaces we develop some convenient notation for derivatives.

Definition 2 (multi index notation)**.** *Let $\overline{\alpha}$ be an ordered n-tuple. We call this a multi-index and denote the length $|\overline{\alpha}| = \sum_{i=1}^n \alpha_i$. 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
$\int_\Omega g\phi dx = (-1)^{|\overline{\alpha}|} \int_\Omega f D^{\overline{\alpha}}\phi dx \quad \forall \phi \in C^\infty$ 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*

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

We then define the sovolev spaces as

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

Theorem 1.0.2. *The Sobolev spaces $W^{k,p}(\Omega)$ are Banach spaces*

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

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

By (1.0.1) 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*

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

With the norm of H^k being 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. The following results are stated for general p based Sobolev spaces, but we will only use them for the Hilbert space H^1.

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

Theorem 1.0.3 (Trace theorem, (Evans [**?**], chapter 5))**.** *Assume U is bounded and ∂U is C^1. Then there exists a bounded, linear operator*

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

Such that

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

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

Theorem 1.0.4. *(Trace-zero functions in $W^{1,p}$,(Evans [**?**], chapter 5)) Suppose U is as in the previous theorem and $u \in W^{1,p}(U)$, then*

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

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

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

And define the norm

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

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

The variational problem

We obtain the **variational formulation** by multiplying (1.1) 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: find u such that

$$\int_\Omega (\nabla v)^T \boldsymbol{K} \nabla u \, dx = \int_\Omega v f \, dx \quad \forall v \in V. \tag{1.2}$$

A good choice of the test space V is $V = H_0^1(\Omega)$. We also choose this as the solution space. We see that if u is a solution to (1.1), it also solves (1.2). But a solution to (1.2) does not necessarily solve (1.1), that is why it is also called the *weak formulation.*
The variational problems that we will look at, that arises from PDE's, will all have the form: Find u such that

$$a(u, v) = b(v) \quad \forall v \in V, \tag{1.3}$$

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

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

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

In general, a variational formulation can we seen as finding the element in a Banach space that is mapped to an element in it's dual space by some linear map.

Boundary conditions

Let $\partial\Omega = \Gamma_D \bigcup \Gamma_N$ with $\Gamma_D \bigcap \Gamma_N = \emptyset$, then (1.1) with more complicated boundary conditions can be written:

$$\begin{aligned} -\nabla \cdot \boldsymbol{K} \nabla \hat{u}(x) &= f(x) \quad && x \in \Omega \\ \hat{u}(x) &= g_D && x \in \Gamma_D \\ \boldsymbol{K} \nabla \hat{u}(x) &= g_N && x \in \Gamma_N \end{aligned} \tag{1.4}$$

To make a variational formulation of (1.4) we first define the test space:

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

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

$$a(u+w, v) = \int_\Omega (\nabla u + \nabla w)^T \boldsymbol{K} \nabla v \, dx = \int_\Omega f v \, dx - \int_{\partial\Omega} \boldsymbol{K} \nabla (u+w) \cdot \boldsymbol{n} v \, dx. \tag{1.5}$$

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

$$a(u,v) = b(v) = \int_\Omega fv \; dx - \int_\Omega (\nabla w)^T \boldsymbol{K} \nabla v \; dx - \int_{\Gamma_N} g_N v \; dx. \tag{1.6}$$

Hence both Dirichlet and Neumann boundary conditions are incorporated into the right hand side. For homogeneous Dirichlet boundary conditions, the second term on the right hand side of (1.6) vanishes.

Existence and uniqueness

We still need to show that (1.6) has an unique solution. First we define some important properties that a variational problem should have in order to have a unique solution. Let $(V, \|\cdot\|_V)$ be a Hilbert space.

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

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

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 [**?**]

Theorem 1.0.5. *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.*
That is: an element in X can always be written as the sum of an element Y and an element in $Y^\perp$.

Theorem 1.0.6 (Riesz representation theorem)**.** *Every continuous linear functional defined on a Hilbert space X can be written $x \to \langle x, v\rangle$ for a 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 1.0.5). 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 1.0.7 (Banach fixed point theorem)**.** *Let X be a Banach space and $F : X \to X$ an operator where $\|Fx - Fy\|_X \leq \theta \|x - y\|_X$ for some $\theta \in (0, 1)$, we call this a* ***contraction****.*
Then for all $x \in X$ the sequence $[x, Fx, F^2x, ...]$ converges to a point $x^* \in X$ called the fixed point of F.

See page 177 of [**?**] for a proof.

Theorem 1.0.8 (Lax Milgram)**.** *Suppose $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is a bi linear, 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, such that*

$$a(u, v) = b(v) \tag{1.7}$$

for all $v \in V$.

Remark 4. *If $a(\cdot, \cdot)$ also is symmetric, it defines an inner product on V giving a complete space. We can then use Riesz representation theorem 1.0.6 to show that it has an unique solution.*

proof of Lax Milgram theorem 1.0.8.
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 1.0.6 $a_w(\cdot)$ 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 (1.7):
find u such that

$$Au = f. \tag{1.8}$$

Now we need to show that (1.8) has an unique solution, and for that we need the Banach fixed point theorem. Let $\epsilon > 0$, we define the operator

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

If T has a fixed point u^*, then $u^* - \epsilon(Au^* - f) = u^* \Rightarrow Au^* = f$ and we have solved (1.8) 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 5. *The solution, u to our bi linear problem depends on the data $b(\cdot)$. To see this we use the coercivity:*

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

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

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

Now we have proved that (1.3) has an unique solution for suitable a and b. The variational form of Poisson equation (1.2) satisfies this:

Example 1 (Well posedness of variational form of Poisson equation)**.** *Let $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$. Then 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 Schwarz 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 (1.2) *is well posed and we get a solution* $u \in H_0^1(\Omega)$*.*

Galerkin FEM

Now we want to discretize the variational equation (1.3). We do this by replacing the test space V by a finite dimensional subspace V_h, this is called the *Galerkin method*. The discretization now reads: Find $u \in V_h$ such that

$$a(u, v_h) = b(v_h) \tag{1.9}$$

for all v_h in V_h. Since a is bi linear and b is linear, it is easy to see that if (1.9) holds for the basis functions of V_h, it holds for all elements in V_h. In the *finite element method*, the finite dimensional subspace are determined by the *triangulation*. In this thesis, we only consider problems in two spatial dimensions, so let $\Omega \subset \mathbb{R}^2$.

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

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

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

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

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

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

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

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

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

Are the space of piecewise linear functions on each K

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

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

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

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

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

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

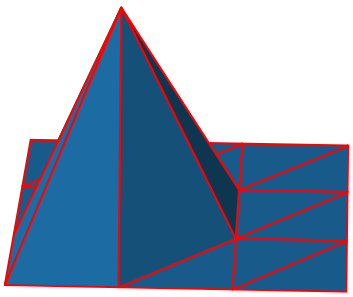


Figure 1.1: A hat function .

Now, we demonstrate how the method works in practice. We seek a solution $u_h \in V_h$. Write this in terms of the basis functions: $u_h = \sum_{i=1}^{n} \hat{u}_i \phi_i$. Now, (1.9) can be written as an equation with a solution vector with real coefficients: Find $\hat{\boldsymbol{u}}_h$ in $\mathbb{R}^n$ such that

$$\sum_{i=1}^{n} \hat{u}_i a(\phi_i, \phi_j) = b(\phi_j). \tag{1.10}$$

So we get a system of linear equations $\boldsymbol{A}\hat{\boldsymbol{u}}_h = \boldsymbol{b}$, where we have one equation for each interior node. If we solve (1.2), our variational problem, and also matrix, will be symmetric. The matrix is then often called a *stiffness matrix*. These names originated from mechanics and structural analysis, where the solution represents displacement and the force function represents load. The stiffness matrix is also sparse, which is a very important property when designing algorithms to solve it.

With the setup described in this subsection, the degrees of freedom are the same as the dimension of V_h. If we in definition 10 instead had chosen a space of quadratic polynomials on each element, we 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 are not expected to be very regular.

Implementation

In this subsection we explain the most important parts of the algorithm for discretizing elliptic PDE's with linear triangular elements. We consider the homogenous elliptic model problem (1.2) 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 [**?**]. If we have N nodes, our triangulation would be stored as a $N \times 2$ array of floats, being the coordinates of the nodes. And a $E \times 3$ array of ints being the elements, where each entry is the index of a coordinate in the coordinate matrix, E is the number of elements.

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

3. Define the basis functions on a reference element, this is also called the shape functions, see figure 1.2 and (1.11). Also compute the gradients of the shape functions.

$$\begin{aligned} N_1(x,y) &= 1 - x - y \\ N_2(x,y) &= x \\ N_3(x,y) &= y \end{aligned} \tag{1.11}$$

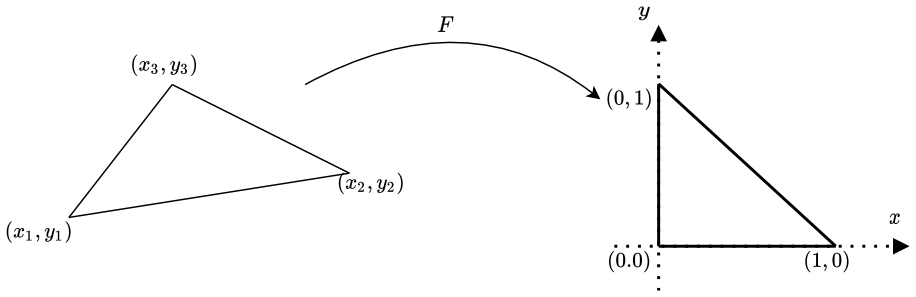


Figure 1.2: The map F from element K to the reference element $\hat{K}$.

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

$$\begin{aligned} F : K &\to \hat{K} \\ x &\mapsto Bx + d \end{aligned} \tag{1.12}$$

To achieve this we set up a system of equations inspired by figure 1.2

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ d_1 & d_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{1.13}$$

So for each element we solve (1.13) 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 $u : K \to \mathbb{R}$ we get by the chain rule:

$$\nabla^T_{\hat{x}} u(F^{-1}(\hat{x})) = \nabla^T_x u(F^{-1}(\hat{x})) \nabla^T_{\hat{x}} F^{-1}(\hat{x}) = \nabla^T_x u(F^{-1}(\hat{x})) B^{-1}. \tag{1.14}$$

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

$$\nabla_x u(F^{-1}(\hat{x})) = B^T \nabla_{\hat{x}} u(F^{-1}(\hat{x})). \tag{1.15}$$

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

$$\begin{aligned} a(\phi_i, \phi_j)|_K &= \int_K (\nabla \phi_i)^T \nabla \phi_j dx \\ &= \int_{\hat{K}} (\nabla_x \phi_i(F^{-1}(\hat{x})))^T \nabla_x \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(J(F^{-1}))| d\hat{x} \\ &= \int_{\hat{K}} (B^T \nabla_{\hat{x}} \phi_i(F^{-1}(\hat{x})))^T B^T B \nabla_{\hat{x}} \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(B^{-1})| d\hat{x} \\ &= \int_{\hat{K}} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) |\mathrm{Det}(B^{-1})| d\hat{x} \\ &= \frac{1}{2} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) \frac{1}{|\mathrm{Det}(B)|} \end{aligned} \tag{1.16}$$

So for each element we evaluate the last line of (1.16) 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

(1.16) we compute the integral on the reference element:

$$\begin{aligned} b(\phi_i)|_K &= \int_{\hat{K}} f(F^{-1}(\hat{x}))\phi_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})N_i(F^{-1}(\hat{x}))\frac{1}{\mathrm{Det}(B)}d\hat{x} \\ &\approx \frac{1}{\mathrm{Det}(B)}\sum_k \omega_k \hat{f}(\hat{p}_k)N_i(\hat{p}_k) \end{aligned} \tag{1.17}$$

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 convergence behaviour.

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

Remark 9. *If we have inhomogeneous Dirichlet boundary conditions this is in practice done the same way as in the homogenous case, eliminating the degrees of freedom on the boundary. For Neumann conditions one has to evaluate integrals along the boundary as in* (1.6)*, using one-dimensional elements.*

Convergence

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

Theorem 1.0.9 (Cèa's lemma)**.** *Let u solve the variational problem* (1.3) *and u_h solve the corresponding Galerkin approximation* (1.9)*, where the bi linear form a is bounded and coercive. Then we have:*

$$\|u - u_h\|_V \leq \frac{C_b}{C_c} min\left\{\|u - v_h\| : \ v_h \in V_h\right\}. \tag{1.18}$$

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 [**?**], 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 1.3. A direct proof with techniques from calculus is possible in this case.

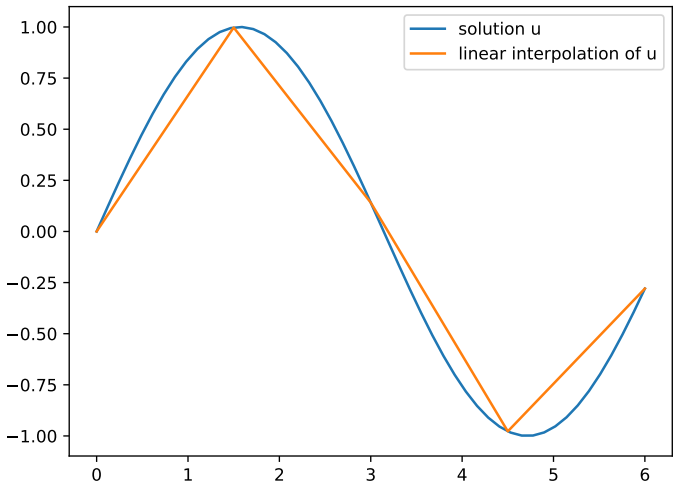


Figure 1.3: The unique linear interpolation of a function in one dimension.

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

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

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

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

Remark 10. *The global interpolator operator 11 maps from continuous functions, so we need to make sure our solution is continuous. By the Sobolev embedding theorem, (Evans [?],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 (1.2), we hope to reach an estimate on the form

$$\|u - u_h\|_1 \leq C \|u - I_h(u)\|_1 \leq C^* h^k |u|_{k+1} \tag{1.19}$$

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 [?], there is a detailed proof of (1.19).
Note that this means that our linear finite element method has a linear convergence in the $\|\cdot\|_1$ norm, if our variational problem admits a solution with sufficient regularity. We tie these observations together in a theorem:

Theorem 1.0.10 (energy norm estimate)**.** *Consider a finite element discretization as described by* (1.10) *in $\mathbb{R}^d$ for $d \leq 3$ on a family of triangulations with an uniform upper bound on the maximal angle. Suppose we have a linear ansatz space as in 10, then*

$$\|u - u_h\|_1 \leq Ch|u|_2. \tag{1.20}$$

Often we are happy with a convergence rate estimate in the $\|\cdot\|_0$ norm, which do not measure an error in the approximation of the derivative. We then expect a better convergence rate, as can be shown by the *duality trick*. We consider the dual problem of our variational problem (1.2): $a(v, u_f) = \langle f, v \rangle_0$, and assume some uniqueness and stability of the solution u_f of this.

Theorem 1.0.11 (L^2 estimate)**.** *Suppose the situation of theorem 1.0.10 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:*

$$\|u - u_h\|_0 \leq C^* h \|u - u_h\|_1 . \tag{1.21}$$

See [?] for a proof. When it comes to the assumption on the dual problem, this is satisfied for our elliptic model problem 1.1. If we put the last two theorems together we obtain quadratic convergence in the L^2 norm

Remark 11. *In this chapter we have only discussed the convergence behaviour of the solution to the Galerkin problem* (1.9)*. 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.*

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 (1.1): Find u such that:

$$\begin{aligned} -\nabla \cdot \boldsymbol{K} \nabla u(x) &= f(x) \quad x \in \Omega \\ u(x) &= 0 \quad x \in \partial\Omega. \end{aligned} \tag{1.22}$$

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:

$$\int_{\Omega_i} -\nabla \cdot \boldsymbol{K} \nabla u dx = -\int_{\partial\Omega_i} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} ds = \int_{\Omega_i} f dx \tag{1.23}$$

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

$$q_{E_{i,j}} = -\int_{E_{i,j}} \boldsymbol{K} \nabla u \cdot \hat{\boldsymbol{n}} ds \tag{1.24}$$

by a linear combination of u_i at neighbouring cell centers

$$q_{E_{i,j}} \approx \tilde{q}_{E_{i,j}} = \sum_k t_{i,j}^k u^k. \tag{1.25}$$

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

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

$$\sum_{j \in \mathcal{S}_i} \tilde{q}_{E_{i,j}} = |\Omega_i| f(x_i), \tag{1.26}$$

where $\mathcal{S}_i$ is the set of indexes of neighbouring cells. The system of equations (1.26) ensures local mass conservation. It can also be written in matrix form as:

$$\boldsymbol{A}^V \tilde{\boldsymbol{u}}_h = \boldsymbol{f}. \tag{1.27}$$

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

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

When it comes to boundary conditions, this is usually straightforward for Neumann boundaries. Especially no-flow boundary conditions, where one makes creates a strip of cells outside the boundary with zero permeability. The same discretization algorithm can the be applied everywhere in the domain. Dirichlet boundary conditions often require more special care. If one, however, knows the solution somewhere outside the domain, one could make a strip of cells outside the boundary where the potential values are known, this is known as ghost Dirichlet boundary conditions. We will focus on discretizing the interior of the domain in the following sections.

Two point flux approximation

The simplest way of constructing $t_{i,j}^k$ is also the most popular in the industry. As the name suggests, we only use the function value at two points, x_0 and x_1, to compute the numerical flux $\tilde{q}_{E_{0,1}}$.

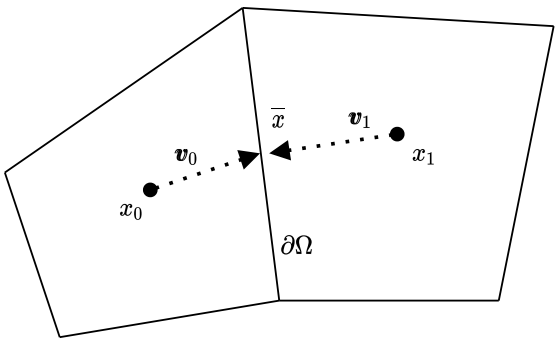


Figure 1.4: The two point flux approximation (TPFA) setup.

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

$$\tilde{q}_{E_{0,1},0} = -\boldsymbol{n}_0^T \boldsymbol{K}_0 \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}(u(\overline{x}) - u(x_0))ds \tag{1.28}$$

or as

$$\tilde{q}_{E_{0,1},1} = -\boldsymbol{n_1}^T \boldsymbol{K}_1 \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}(u(x_1) - u(\overline{x}))ds \tag{1.29}$$

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

$$\tilde{q}_{E_{0,1},0} = \tilde{q}_{E_{0,1},1} = t^0 u(x_0) + t^1 u(x_1) \tag{1.30}$$

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

$$T_0(u(\overline{x}) - u(x_0)) = T_1(u(x_1) - u(\overline{x})) \Rightarrow u(\overline{x}) = \frac{T_0 u(x_0) + T_1 u(x_1)}{T_0 + T_1}. \tag{1.31}$$

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

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} =& -T_0(u(\overline{x}) - u(x_0)) \\
=& -T_0\left(\frac{T_0 u(x_0) + T_1 u(x_1)}{T_0 + T_1} - u(x_0)\right) \\
=& -T_0\left(\frac{T_0 u(x_0) + T_1 u(x_1) - u(x_0)T_0 - u(x_0)T_1}{T_0 + T_1}\right) \\
=& -T_0\left(\frac{T_1 u(x_1) - u(x_0)T_1}{T_0 + T_1}\right) \\
=& \frac{u(x_0) - u(x_1)}{\frac{1}{T_1} + \frac{1}{T_0}}.
\end{aligned} \tag{1.32}$$

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

$$\begin{aligned}
\tilde{q}_{E_{0,1},0} &= t^0 u(x_0) + t^1 u(x_1) \\
\frac{u(x_0) - u(x_1)}{\frac{1}{T_1} + \frac{1}{T_0}} &= t^0 u(x_0) + t^1 u(x_1) \\
\Rightarrow t^0 &= \frac{1}{\frac{1}{T_1} + \frac{1}{T_0}}.
\end{aligned} \tag{1.33}$$

Hence, the transmissibility is the *harmonic mean* of the local transmissivities. One way of looking at this discretization, is that we assume the potential to be a linear function of one variable in the v_i direction between the cell center and the edge in figure 1.4. So for each edge, we have two linear functions on each side, which

gives us four degrees of freedom. Two of them are used to respect the cell center potential values, the other two are used on pressure and flux continuity across the edge. With these assumptions, expressions (1.28) and (1.29) are exact. And we only have to solve for the transmissibility coefficients.

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

$$\boldsymbol{n}_2 \cdot \boldsymbol{K}\boldsymbol{n}_1 = 0 \tag{1.34}$$

for a uniform parallelogram mesh with the normal vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$. We then call the grid **K-orthogonal**. In the setting of figure 1.4, our grid would not be K-orthogonal as the control volumes are not a parallelograms. All meshes with orthogonal control volumes are K-orthogonal if the permeability is isotropic.

reference a figure showing the failed convergence of TPFA

O-method

The O-method is a multi-point flux approximation method, these types of methods were developed to make control volume methods converge for grids that are not K-orthogonal. It is described in detail in [**?**], we only give a brief introduction. Consider the control volumes in 1.5.

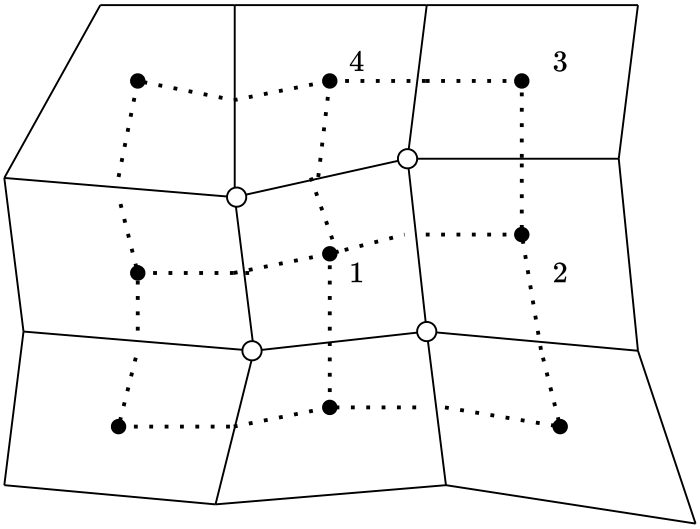


Figure 1.5: The solid lines are the control volumes, the dashed lines are the dual mesh connecting the cell centers, going through the midpoints of each edge. The solid circles are cell centers, the white circles are grid points .

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

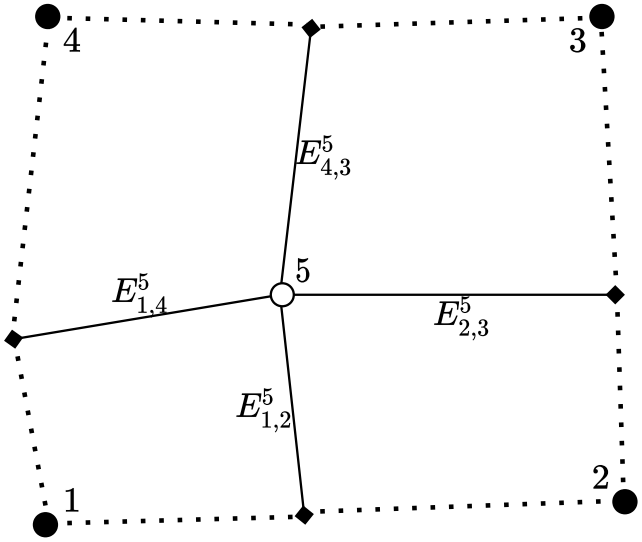


Figure 1.6: The four subcells in the interaction region corresponding to cells $1, 2, 3, 4$ and grid point 5. Here, $\mathcal{R}_5 = \{1, 2, 3, 4\}$.

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

expression

$$\tilde{q}^n_{E_{i,j}} = \sum_{k\in\mathcal{R}_n} t^{k,n}_{i,j} u^k \approx \int_{E_{i,j}} \hat{\boldsymbol{n}}_j^T \boldsymbol{K}\nabla u \; ds \quad i,j \in \mathcal{R}_n \tag{1.35}$$

for the flux through each half edge $E^n_{i,j}$ in the interaction region corresponding to grid point n (figure 1.6). Where $\mathcal{R}_n$ is the index set of the four cells neighbouring grid point n.

We assume for now that the potential is linear in each of the four sub cells in the interaction region, figure 1.6. This gives $4 \cdot 3 = 12$ degrees of freedom. The linear potential must of course equal the cell center values of the potential in the cell centres, this removes four degrees of freedom. We also require flux continuity on the four half edges in the interaction region, this removes an additional four degrees of freedom. The last four degrees of freedom are spent on potential continuity of the midpoints of the edges.

By these assumptions on flux and potential continuity, the linear potential in each sub cell is well defined given values at the cell center. We can now use this to compute the four by four matrix of transmissibility coefficients for each of the four half edges. In the situation of figure 1.6 and equation (1.35) it would look like

$$\boldsymbol{T}^5 = \begin{bmatrix} t^{1,5}_{1,2} & t^{2,5}_{1,2} & t^{3,5}_{1,2} & t^{4,5}_{1,2} \\ t^{1,5}_{2,3} & t^{2,5}_{2,3} & t^{3,5}_{2,3} & t^{4,5}_{2,3} \\ t^{1,5}_{4,3} & t^{2,5}_{4,3} & t^{3,5}_{4,3} & t^{4,5}_{4,3} \\ t^{1,5}_{1,4} & t^{2,5}_{1,4} & t^{3,5}_{1,4} & t^{4,5}_{1,4} \end{bmatrix}. \tag{1.36}$$

Computing (1.36) involves inverting a four by four matrix with coefficients depending on the mesh and permeability, see [**?**] for details. Finally, we assemble the system of equations (1.26) with the transmissibility coefficients. Note that we write the flux over the jth edge of cell i, $\tilde{q}_{i,j}$ as the flux over the two half edges:

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

$$\sum_{j=1}^{4} (\sum_{k=1}^{4} t^{k,1}_{i,j} u^k + \sum_{k=1}^{4} t^{k,2}_{i,j} u^k) = |\Omega_i| f(x_i).$$

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

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

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

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

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

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

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

L-method

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

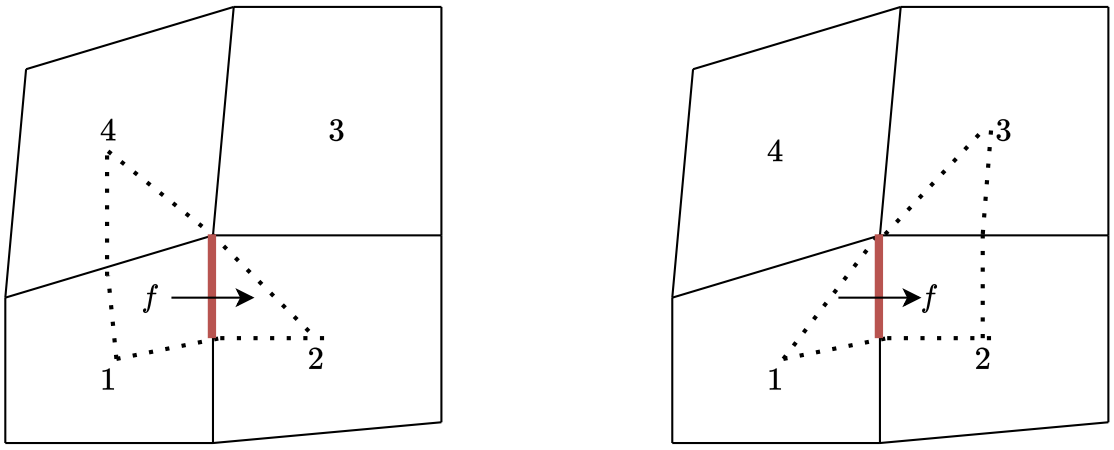
As the O-method, the L-method is also a multipoint flux approximation method. It was introduced in [**?**], where the authors demonstrate improved monotonicity properties with numerical experiments. This method is similar to the O-method,

in that it goes through the half edges and uses information from the same interaction regions. But instead of using four points for the flux across each half edge, we use three, with two half edges between them.

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

We have two choices of flux stencil for each half edge, see figure 1.7. We compute the transmissibility coefficients for both, then we choose the one "best" aligned with the flow: Let t_1^i be the ith transmissibility coefficient of T_1, then

$$\begin{aligned} &\text{if } |t_1^1| < |t_2^2| \\ &\text{choose } T_1 \text{ else} \\ &\text{choose } T_2. \end{aligned} \tag{1.37}$$



(a) T_1 (b) T_2

Figure 1.7: The two choices of which cell centers to use for computing the flux over the half edge in red. We call the triangles spanned T_1 and T_2 for L-triangles.

A cheap intuition behind (1.37) is that if $|t_1^1| < |t_2^2|$, it is more likely that $sgn(t_1^1) = sgn(t_1^4)$ and if not, $sgn(t_2^2) = sgn(t_2^3)$ is more likely. This is due to the fact that $\sum t^i = 0$. Choosing L-triangle as in (1.37) increases the chances that we get the same sign of t^i on the same side of the half edge, thus increasing the chance that we get a monotone discretization. See [**?**] for a more detailed geometric intuition of choosing L-triangle in the case of homogenous permeability.

To compute transmissibility coefficients in a given L-triangle, we use the assumptions on flux and potential continuity, to construct a linear system. The

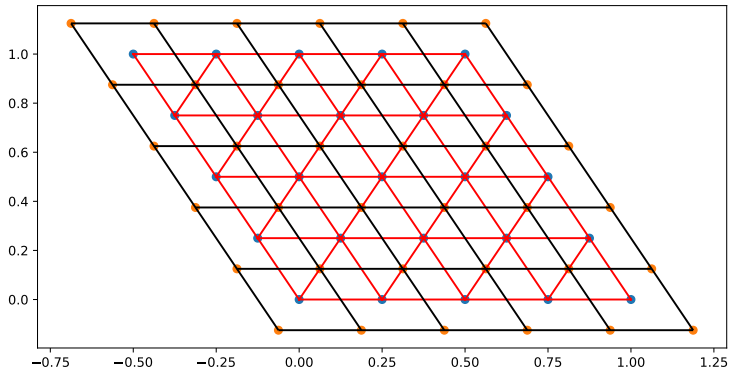
coefficients depends on mesh and permeability in the three cells. In general, we need to solve two linear systems for each half edge, as there are always two choices. In the O-method it is enough to solve a linear system per four half edges.

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

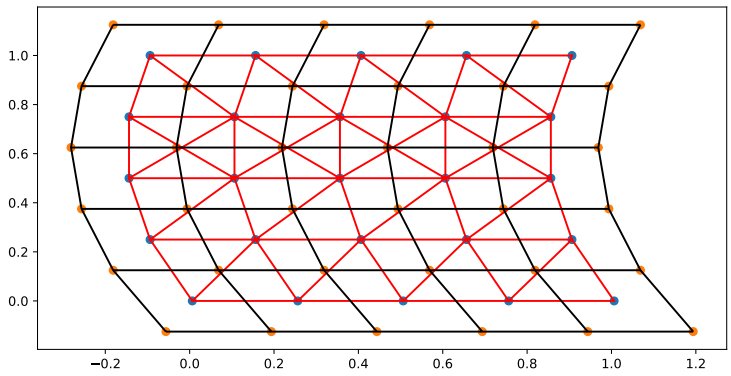
$$\begin{aligned} \sum_{j=1}^{4}(\tilde{q}_{i,j}^{1} + \tilde{q}_{i,j}^{2}) &= |\Omega_i| f(x_i) \\ \sum_{j=1}^{4}(\sum_{k=1}^{3} t_{i,j}^{k,1} u^k + \sum_{k=1}^{3} t_{i,j}^{k,2} u^k) &= |\Omega_i| f(x_i). \end{aligned} \tag{1.38}$$

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

In figure 1.8 we see the criterion in practice for a homogenous medium: In figure 1.8a all L-triangles are used by two half-edges, and they are chosen in the same way throughout the domain. In figure 1.8b there are some triangles that overlap, this is due to the fact that some L-triangles are used by only one half edge.



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



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

Figure 1.8: Examples of L-triangles(in red) in a domain with homogenous permability tensor.

The observation in figure 1.8a can be stated as a theorem:

Theorem 1.0.12 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[**?**])**.** *For homogeneous media and uniform parallelogram grids, the MPFA L-method has a seven-point cell stencil for the discretization of each interior cell, ie. the discretization of each cell is a seven point stencil including the center cell and the six closest potential cells, as shown in 1.8a.*

In case of parallelogram grid with heterogeneous permeability, it may also happen that one gets overlapping L-triangles. This is the case even if the permeability only changes as a scalar in the domain. In figure 1.9 the L-triangles are shown for a random, scalar permeability. Let $K_{m,n}$ be the permeability of the mth cell in y

direction and nth cell in x direction. Then the random permeability used in figure 1.9 given by

$$K_{n,m} = (e^{\hat{x}} - 1)^2 \tag{1.39}$$

where $\hat{x}$ is a random sample drawn from a uniform distribution over $[0, 1)$. We see that two of the L-triangles overlap. This is due to some combination of permeability at four neighbouring cells. Also note that the permeability is not so low that it causes numerical rounding errors, as $\min_{m,n} K_{m,n} = 0.0017$ in figure 1.9.

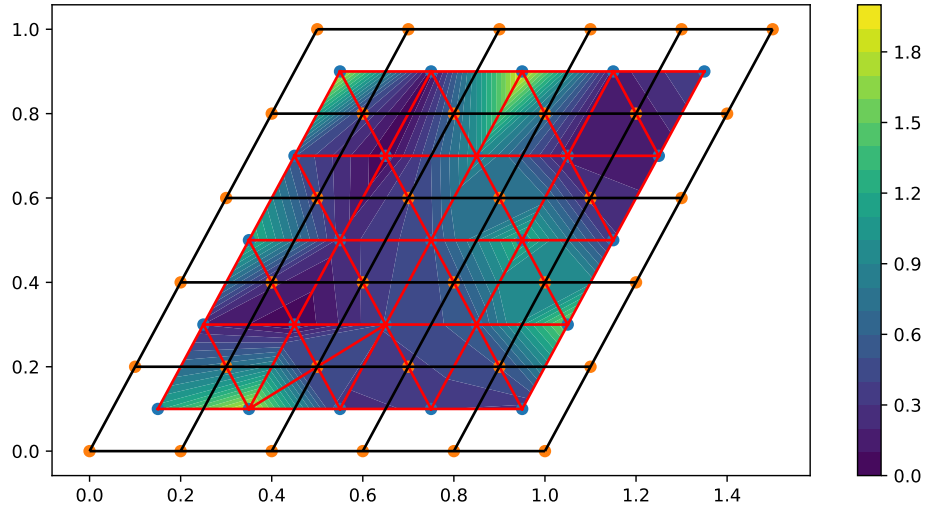


Figure 1.9: L-triangles on a random permeability.

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

Lemma 1.0.13 (Cao, Y., Helmig, R. and Wohlmuth, B.I. (2009),[**?**])**.** *Assume that the permeability* $\boldsymbol{K}$ *is homogenous on* Ω*, then the flux through each half edge* e*, computed by the L-method, can be written as*

$$\tilde{q}_e = -\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_e \tag{1.40}$$

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

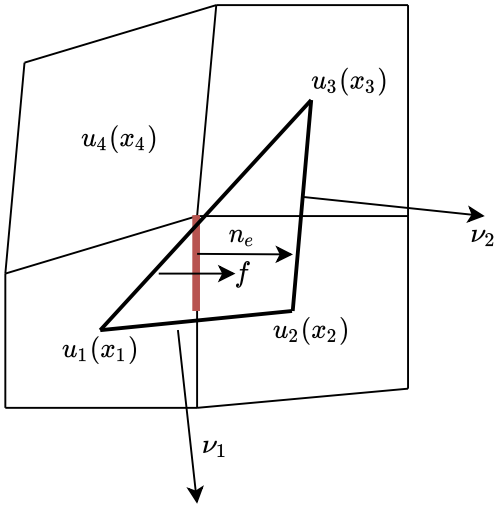


Figure 1.10: Simplified L-triangle, the original L-triangle i shown in figure 1.7b. The vector ν_1 is perpendicular to the edge between x_1 and x_2, with the same length as the edge it is perpendicular to. Same for ν_2.

Moreover, the gradient ∇u, is given by:

$$\nabla u = -\frac{1}{2F}[(u_1 - u_2)\nu_2 + (u_3 - u_2)\nu_1]. \tag{1.41}$$

Where F is the the area of the simplified L-triangle with corners x_1, x_2 and x_4, see figure 1.10. An expression like (1.41) *can be obtained for the other choice of L-triangle as well.*

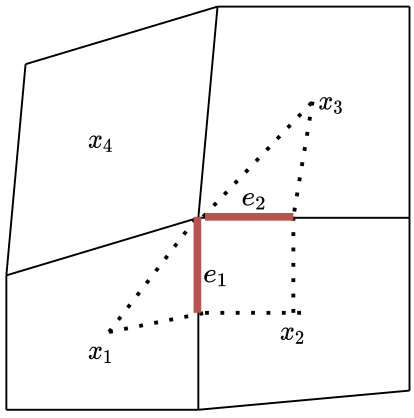


Figure 1.11: Original L-triangle with notations in proof.

Proof. It is enough to check that the jump $[\nabla u]$ is zero on e_1 and e_2 on the original L-triangle in figure 1.11. Let $\boldsymbol{t}_{e_1}$ and $\boldsymbol{n}_{e_1}$ be the tangent and normal vector to e_1.

Since we require potential continuity on each half edge, we get:

$$[\nabla u \cdot \boldsymbol{t}_{e_1}] = 0. \tag{1.42}$$

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

$$[\boldsymbol{K}\nabla u \cdot \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}^T \boldsymbol{n}_{e_1}] = [\nabla u \cdot \boldsymbol{K}\boldsymbol{n}_{e_1}] = 0. \tag{1.43}$$

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

Remark 12. *The above lemma suggests that we can obtain the transmissibility coefficients without solving a system of equations for each half edge. This simplifies implementation, but it's only possible for homogenous media.*

Remark 13. *In the case of heterogeneous media, we do not easily get meaningful, physical interpretations of the transmissibility coefficients. Like the harmonic mean, as we did for TPFA.*

To conclude; the L-method is the most sophisticated method. It has the best monotonicity properties, it is consistent for non K orthogonal grids, but it requires more work for assembling the matrix. It is also more complicated to implement, especially for more dimensions.

Time Discretization

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

$$\begin{aligned} \partial_t u - \nabla \cdot \boldsymbol{K}\nabla u &= f \quad && x \in \Omega \quad && t \in (0,T] \\ u &= 0 && x \in \partial\Gamma_D && t \in (0,T] \\ \boldsymbol{K}\nabla u &= g_N && x \in \partial\Gamma_N && t \in (0,T] \\ u &= u_0 && x \in \Omega && t = 0 \end{aligned} \tag{1.44}$$

The well-posedness of (1.44) is discussed in chapter seven of [**?**], it requires a more detailed discussion of Sobolev spaces and Bochner spaces, ie. spaces containing functions from the real numbers to some Sobolev space.

We expect low regularity in time, so there is not much 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 through the domain instantaneously. A careful analysis of time discretizations of parabolic equations is done in ([?], 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 $\tau = \frac{T}{N}$ be the time-step. Then we state the semi-discrete version of (1.44) by exchanging the time derivative by a difference quotient $(\partial_t u)^n = \frac{u^n - u^{n-1}}{\tau}$. Note that this difference quotient is implicit because u^n is not explicitly given by terms of the previous time-step. We end up with: Given u^{n-1} and f^n, find u^n such that

$$\begin{aligned} u^n - \tau \nabla \cdot \boldsymbol{K} \nabla u^n &= \tau f^n + u^{n-1} && x \in \Omega \\ u^n &= 0 && x \in \partial\Gamma_D \\ \boldsymbol{K}\nabla u &= g_N && x \in \partial\Gamma_N \\ u^0 &= u_0 && x \in \Omega. \end{aligned} \tag{1.45}$$

Now we have an elliptic problem (1.45) for each time-step. This has almost the same structure as the elliptic model problem (1.1) 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 (1.45) is achieved as before by multiplying by test functions in $H_0^1(\Omega)$: Given u^{n-1} in V, f^n in the dual of V, find u^n in V such that

$$\langle u^n, v\rangle_0 + \tau \langle \boldsymbol{K}\nabla u^n, \nabla v\rangle_0 = \tau \langle f^n, v\rangle_0 + \left\langle u^{n-1}, v\right\rangle_0 \tag{1.46}$$

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

$$\begin{gathered} \text{find } \hat{\boldsymbol{u}}^n \in \mathbb{R}^d \text{ such that} \\ \boldsymbol{B}\hat{\boldsymbol{u}}^n + \tau \boldsymbol{A}\hat{\boldsymbol{u}}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}\hat{\boldsymbol{u}}^{n-1} \end{gathered} \tag{1.47}$$

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(1.44) in conservation form on each control volume

$$\partial_t \int_{\Omega_i} u \, dx - \int_{\partial\Omega_i} \boldsymbol{K}\nabla u \cdot \hat{\boldsymbol{n}} \, dx = \int_{\Omega_i} f \, dx, \tag{1.48}$$

and discretize the first term with backward Euler. Or one could make sure the semi-discrete heat equation (1.44) holds for each control volume and use the divergence theorem. Both ways, we end up with

$$\int_{\Omega_i} u^n \, dx - \tau \int_{\partial\Omega_i} \boldsymbol{K}\nabla u^n \cdot \hat{\boldsymbol{n}} \, dx = \tau \int_{\Omega_i} f^n \, dx + \int_{\Omega_i} u^{n-1} \, dx. \tag{1.49}$$

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

$$(\boldsymbol{B}^V + \tau \boldsymbol{A}^V)\boldsymbol{u}^n = \tau \boldsymbol{f}^n + \boldsymbol{B}^V \boldsymbol{u}^{n-1} \tag{1.50}$$

The matrix $\boldsymbol{A}^V$ is as in chapter 3, with the fluxes through the edges of cell i described by the jth row of $\boldsymbol{A}^V$. The matrix $\boldsymbol{B}^V$ is diagonal with the entry i being the volumes of the volume of cell i.

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. As we will see later, this is the challenging part.

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

$$\text{find } x \in U \text{ such that } \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \text{ where } f : U \subset \mathbb{R}^n \to \mathbb{R}^n \tag{1.51}$$

The solution in (1.51) is called a *root*, it is almost always found using an iterative method.

A common iterative scheme to solve (1.51) is the *Newton method*, let $D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ be the Jacobian of $\boldsymbol{f}(\boldsymbol{x}_{j-1})$.

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} - D\boldsymbol{f}(\boldsymbol{x}_{j-1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{1.52}$$

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

Theorem 1.0.14. *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*

$$|x_{j+1} - \overline{x}| \leq \frac{1}{\delta}|x_j - \overline{x}|^2 < |x_j - \overline{x}| \tag{1.53}$$

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

$$0 = f(\overline{x}) = f(x_j - e_j) = f(x_j) - f'(x_j)e_j + \frac{f''(\psi)e_j^2}{2} \tag{1.54}$$

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

$$\begin{aligned} e_{j+1} = x_{j+1} - \overline{x} &= x_n - \frac{f(x_j)}{f'(x_j)} - \overline{x} \\ &= e_j - \frac{f(x_j)}{f'(x_j)} \\ &= \frac{e_j f'(x_j) - f(x_j)}{f'(x_j)} \end{aligned} \tag{1.55}$$

By the Taylor expansion around x_j, (1.54), we get

$$e_{j+1} = \frac{e_j^2 f''(\psi)}{2f'(x_j)} \tag{1.56}$$

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

$$|e_1| \leq \frac{2}{2\delta}|e_0|^2 < |e_0| \tag{1.57}$$

By the same reasoning we get convergence

$$|e_{j+1}| < |e_j| \tag{1.58}$$

And the quadratic convergence

$$|e_{j+1}| \leq \frac{1}{\delta}|e_j|^2 \tag{1.59}$$

□

For a similar result in more dimensions see (Knabner [**?**], 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

$$D\boldsymbol{f}(\boldsymbol{x}_{j-1})\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{1.60}$$

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

$$L\boldsymbol{\delta}_j = -\boldsymbol{f}(\boldsymbol{x}_{j-1}) \tag{1.61}$$

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 1.0.15. *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

$$e_{j+1} = x_j - \frac{f(x_j)}{L} - \overline{x} = e_j - \frac{f(x_j)}{L} \tag{1.62}$$

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

$$0 = f(\overline{x}) = f(x_j - e_j) = f(e_j) - f'(\psi)e_j \Rightarrow e_j = \frac{f(x_j)}{f'(\psi)} \tag{1.63}$$

Using this and the assumption on L we get the estimate

$$|e_{j+1}| = |e_j(1 - \frac{f'(\psi)f(x_j)}{f(x_j)L})| \leq |e_j||1 - \frac{f'(\psi)}{L}| < |e_j| \tag{1.64}$$

□