Chapter 1

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[**?**] . 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 $\Omega \subset \mathbb{R}^n$

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

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 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)**.** *Let $L^p(\Omega)$ be the set of all functions u defined on Ω for wich the norm $\|u\|_p = (\int_\Omega u^p dx)^{1/p} < \infty$ is defined, where $p \in [1, \infty]$*

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.*

Theese spaces have the property that they are complete.

Theorem 1.0.1 (Riesz-Fischer Theorem [**?**] chapter 8)**.** *Each L^p space is a Banach space.*

The L^2 space has the property that it is a **Hilbert space**, ie. a complete inner product space.

Definition 2. *Let $u, v \in L^2(\Omega)$, then $\langle u, v\rangle_{L^2(\Omega)} = \int_\Omega uvdx$ defines an inner product.*

Before we continue the study of funtion spaces we develop some convinient notation for derivatives.

Definition 3 (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 4 (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 5 (Sobolev space)**.** *Let k be a non-negative integer, define the Sobolev norm as*

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

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 $W^{k,p}$ 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 6. *We rename the L^2 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 7. *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 1.0.3 (Trace theorem, (Evans [**?**], chapter 5))**.** *Assume U is bounded and ∂U is C^1. Then there exists a bounded, linear operator*

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

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 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*

$$u \in W_0^{1,p} \Leftrightarrow Tu = 0 \text{ on } \partial U \tag{1.2}$$

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 (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:

$$\text{find } u \text{ such that}$$
$$\int_\Omega (\nabla v)^T \boldsymbol{K}\nabla u \, dx = \int_\Omega vF \, dx \tag{1.3}$$
$$\forall v \in V$$

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 (1.3). But a solution to (1.3) does not necessarily solve (1.1), 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

$$\begin{gathered}\textit{find } u \textit{ such that}\\ a(u, v) = b(v)\\ \forall v \in V\end{gathered} \tag{1.4}$$

Where $a(\cdot, \cdot)$ is a *bilinear form* on a V and $b(\cdot)$ is a **linear functional** on V, where V is a Hilbert space.

We still need to show that (1.4) has a unique solution.

Existence and uniqueness

First we define some important properties that a variational problem should have in order to have a uniqe solution. Let $(V, \|\cdot\|_V)$ be a Hilbert space.

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

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

It is clear that our variational problem (1.3) has both these properties. 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.*
In plain english: an element in X can always be written as the sum of an element Y and an element in $Y^{\perp}$

Proof. □

Definition 9. *We denote the space of all continuous linear functionals on a Banach space X by X'*

Remark 2. *It's easy to check that X' is also a Banach space*

Remark 3. *In a Hilbert space, the inner product by a fixed vector uniqely determines a member of the dual space.*

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 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 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 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 [**?**] for a proof.

Theorem 1.0.8 (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.*

$$a(u, v) = b(v) \tag{1.5}$$

For all $v \in V$

Remark 4. *If $a(\cdot, \cdot)$ also is symmetric, and defines an inner product on V giving a complete space. We can 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

$$A : V \to V$$
$$w \mapsto Aw$$

- 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 (1.5):
find u such that

$$Au = f \tag{1.6}$$

Now we need to show that (1.6) 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 (1.6) 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. *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 (1.4) has an unique solution for suitable a and b. The variational form of poisson equation (1.3) 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* $\|\cdot\|_{H_0^1}$

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

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

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

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* $f \in L^2(\Omega)$*:*

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

Hence (1.3) *is well posed and we get a solution* $u \in H_0^1(\Omega)$.

Galerkin FEM

Now we want to discretizise the variational equation (1.4). We do this by replacing the test space V by a finite dimensional subspace V_h, this is called the *Galerkin method.*

$$\begin{aligned}&\textit{find } u \in V_h \textit{ such that} \\ &a(u, v_h) = b(v_h) \\ &\textit{for all } v_h \in V_h\end{aligned} \tag{1.7}$$

Since a and b both are linear, it's easy to see that if (1.7) 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 10 (two dimensional triangulation, page 56 of Knaber [**?**])**.** *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 11 (Linear ansatz space)**.** *Let $P_1(K)$ be the space of polynomials of one degree in two variables on $K \subset \mathbb{R}^2$, then the ansatz space*

$$V_h = \left\{u_h \in C(\overline{\Omega}) : u_{h|K} \in P_1(K)\ \forall K \in \tau_h, u|_{\partial\Omega} = 0\right\}$$

Are the space of piecewise linear functions on each K

Remark 6. *If we have $V_h \in V$ we call it a* conformal finite element method*. Note that **(T3)** together with the definition of linear ansatz space ensures that we have no discontinuities and we have a* conformal finite element method.

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:

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

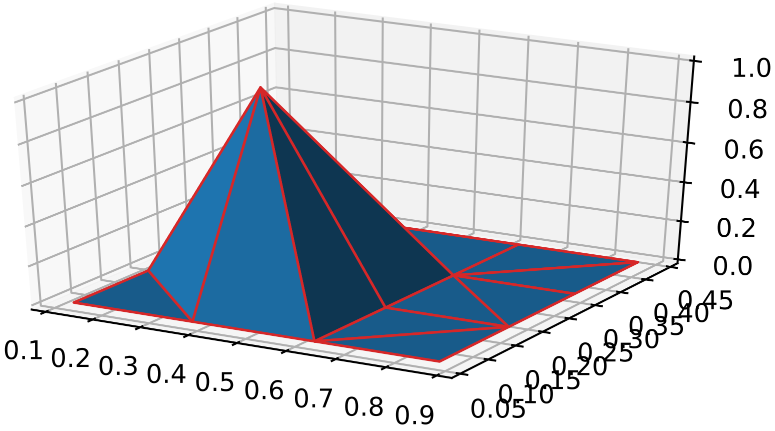

Figure 1.1: A hat function .

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 (1.7) becomes

$$\begin{aligned} &\textit{find } \boldsymbol{u}^* \in \mathbb{R}^n \textit{ such that} \\ &\sum_{i=1}^{n} u_i^* a(\phi_i, \phi_j) = b(\phi_j) \end{aligned} \tag{1.8}$$

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 (1.3) 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 11 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 (1.8). 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 (1.3) 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 [?]. 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 1.2 and (1.9). Also compute the gradients of the shape functions.

$$\begin{aligned} N_1(x, y) &= 1 - x - y \\ N_2(x, y) &= x \\ N_3(x, y) &= y \end{aligned} \tag{1.9}$$

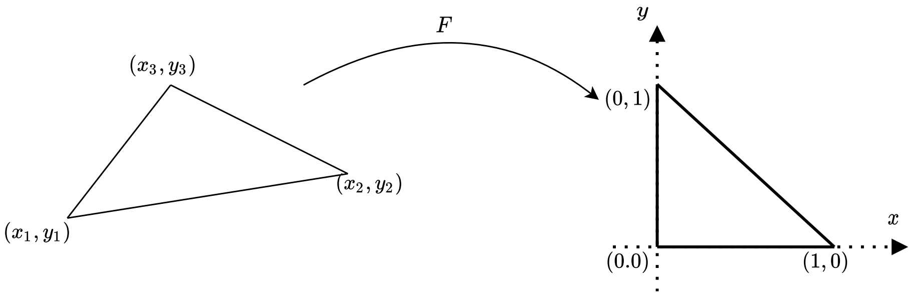

Figure 1.2: The map F from element K to the reference element $\hat{K}$.

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

$$\begin{aligned} F : K &\to \hat{K} \\ x &\mapsto Bx + d \end{aligned} \tag{1.10}$$

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

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ d_1 & d_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{1.11}$$

So for each element we solve (1.11) for B and d, that means computing an inverse of a three by three matrix and a matrix product. Note that this only needs to be done once 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 $u : K \to \mathbb{R}$

$$\nabla_{\hat{x}}^T u(F^{-1}(\hat{x})) = \nabla_x^T u(F^{-1}(\hat{x})) \nabla_{\hat{x}}^T F^{-1}(\hat{x}) = \nabla_x^T u(F^{-1}(\hat{x})) B^{-1} \tag{1.12}$$

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

$$\nabla_x u(F^{-1}(\hat{x})) = B^T \nabla_{\hat{x}} u(F^{-1}(\hat{x})) \tag{1.13}$$

Now we can compute the product of the gradients of the bassis functions on an element.

$$\begin{aligned} a(\phi_i, \phi_j)|_K &= \int_K (\nabla \phi_i)^T \nabla \phi_j dx \\ &= \int_{\hat{K}} (\nabla_x \phi_i(F^{-1}(\hat{x})))^T \nabla_x \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(J(F^{-1}))| d\hat{x} \\ &= \int_{\hat{K}} (B^T \nabla_{\hat{x}} \phi_i(F^{-1}(\hat{x})))^T B^T \nabla_{\hat{x}} \phi_j(F^{-1}(\hat{x})) |\mathrm{Det}(B^{-1})| d\hat{x} \\ &= \int_{\hat{K}} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) |\mathrm{Det}(B^{-1})| d\hat{x} \\ &= \frac{1}{2} (\nabla_{\hat{x}} N_i(\hat{x}))^T B^T B \nabla_{\hat{x}} N_j(\hat{x}) \frac{1}{|\mathrm{Det}(B)|} \end{aligned} \tag{1.14}$$

So for each element we evaluate the last line of (1.14) 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

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

$$\begin{aligned} b(\phi_i)|_K &= \int_{\hat{K}} f(F^{-1}(\hat{x}))\phi_i(F^{-1}(\hat{x}))\frac{1}{\text{Det}(B)}d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})N_i(F^{-1}(\hat{x}))\frac{1}{\text{Det}(B)}d\hat{x} \\ &\approx \frac{1}{\text{Det}(B)}\sum_k \omega_k \hat{f}(\hat{p}_k)N_i(\hat{p}_k) \end{aligned} \tag{1.15}$$

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 inhomogenous dirichlet boundary conditions our solution space is no longer* $H_0^1(\Omega)$

explain inhomogenous boundary conditions

Convergence

In this section we review the most important concepts in studying the convergence, 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. *Let u solve the variational problem* (1.4) *and u_h solve the corresponding galerkin approximation* (1.7)*, where the bi-linear form a is bounded and coercive. Then we have*

$$\|u - u_h\|_V \le \frac{C_b}{C_c} min\{\|u - v_h\| : \; v_h \in V_h\} \tag{1.16}$$

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

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

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 $a(\cdot,\cdot)$:

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

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.

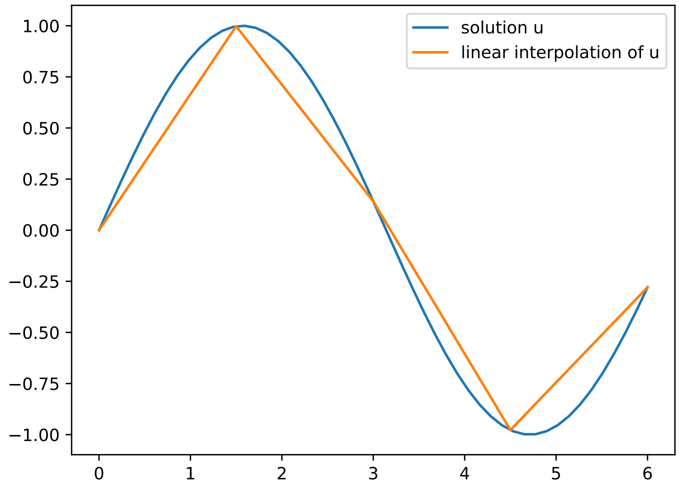

Figure 1.3: The unique linear interpolation of a function in one dimension.

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

Definition 12 (Global interpolation operator)**.**

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

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

Remark 8. *The global interpolator operator 12 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 (1.3), we hope to reach an estimate on the form

$$\|u - u_h\|_1 \leq C \|u - I_h(u)\|_1 \leq C^* h^k |u|_{k+1} \tag{1.17}$$

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 (1.17).
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 1.0.10 (energy norm estimate)**.** *Consider a finite element discretization as described by* (1.8) *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 11, then*

$$\|u - u_h\|_1 \leq Ch|u|_2 \tag{1.18}$$

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 (1.3): $a(v, u) = \langle f, v\rangle_0$, and assume some uniqueness and stability of the solution $u = u_f$ 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*

$$\|u - u_h\|_0 \leq C^* h \|u - u_h\|_1 \tag{1.19}$$

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 obtian quadratic convergence in the L^2 norm

Stability

A stability property for the solution of the Galerkin problem (1.7) follows from remark (5):

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