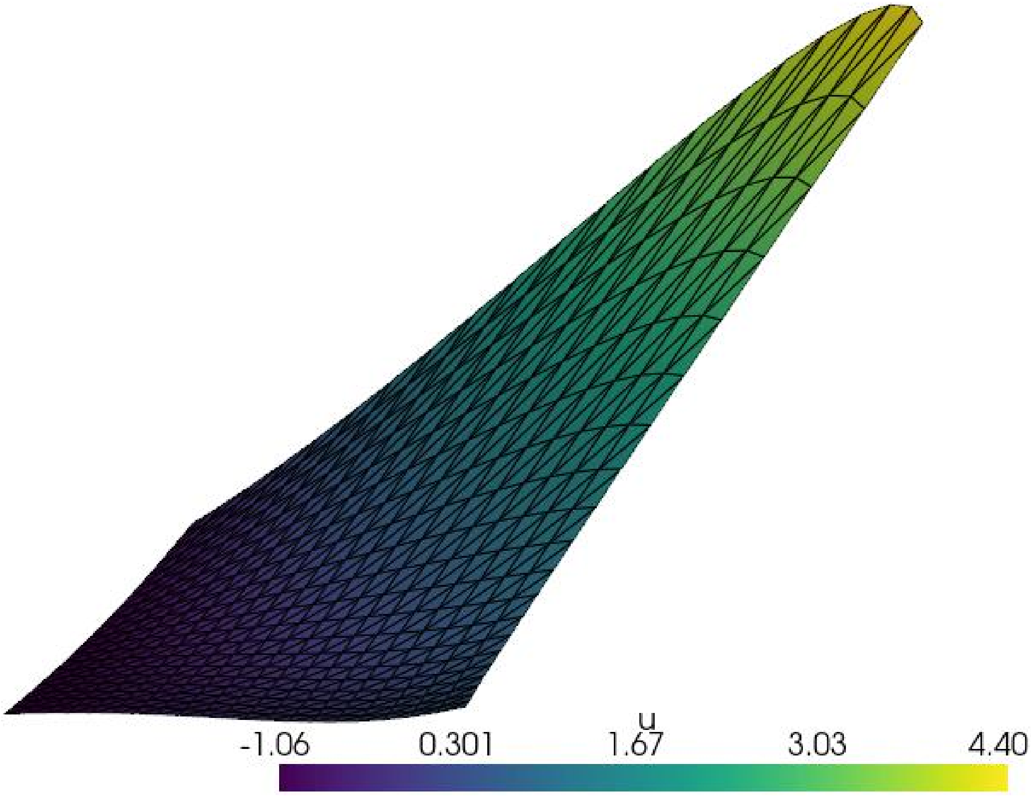
An introductory examination of the Finite Element Method



Group 5.216c
Aalborg University
Mathematics

AALBORG UNIVERSITY
STUDENT REPORT

Department of Mathematical Sciences
Mathematics
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

Title:

An introductory examination of the Finite Element Method

Project:

P6-project

Project Period:

February 2024-May 2024

Project Group:

5.216c

Participants:

Jacob Engberg
William Dam Høedt-Rasmussen
Patrick Guldberg

Supervisor:

Anton Evgrafov
Fynn Jerome Aschmoneit

Pages: 56

Finished: 23$^{\text{rd}}$ of May 2024

Abstract:

I denne projektrapport gennemgår vi den basale teori for ‘Finite Element Method’, som bruges til at approksimere løsninger til partiale differentialligninger med grænsebetingelser på kanten. For at kunne det har vi beskrevet diverse Hilbert rum, og opsætningen af disse, samt vurderinger af fejlen af approksimationen. Afslutningsvis eksaminerer vi et numerisk eksempel ved hjælp af programeringsplatformen FEniCSx.

Preface

This student report has been written by group 5.216c throughout the course of 6$^{\text{th}}$ semester during the project period at the Department of Mathematical Sciences at Aalborg University.

The authors would like to express their gratitude to Anton Evgrafov and Fynn Jerome Aschmoneit for supervision during the course of the project period.

List of Symbols

$(u,v)_0$	The scalar/inner product on $L_2(\Omega)$, meaning $\int_\Omega u(x)v(x)dx = (u,v)_{L_2}$
$\|f\|_0$	The norm on $L_2(\Omega)$, meaning $\sqrt{(f,f)_0}$
$(u,v)_m$	The scalar/inner product on $H^m(\Omega)$, meaning $\sum_{\lvert\alpha\rvert\leq m}(\partial^\alpha u, \partial^\alpha v)_0$
$\|f\|_m$	The norm on $H^m(\Omega)$, meaning $\sqrt{\sum_{\lvert\alpha\rvert\leq m}\|\partial^\alpha f\|^2_{L_2(\Omega)}}$
$\lvert f\rvert_m$	The semi-norm on $H^m(\Omega)$, meaning $\sqrt{\sum_{\lvert\alpha\rvert=m}\|\partial^\alpha f\|^2_{L_2(\Omega)}}$
V'	Dual space over V
Ω	An open subset of $\mathbb{R}^n$ with piecewise smooth boundary, which obeys the cone condition
$\partial\Omega$	The boundary of Ω
div f	$(\partial f/\partial x_1 + \partial f/\partial x_2 + \ldots + \partial f/\partial x_n)$
$\mathbf{n}$	The outwards facing normal on $\partial\Omega$
$\mathcal{P}_t$	$\{f(x,y) = \sum_{i+k\leq t} c_{ik}x^iy^k\}$, for $i,k \geq 0$; I.e. polynomials of degree t
$\mathcal{Q}_t$	$\{f(x,y) = \sum_{0\leq i,k\leq t} c_{ik}x^iy^k\}$

Contents

1 Introduction

When examining complex physical problems, partial differential equations often occur. These equations are often difficult and in some case impossible to calculate algebraically. We therefore approximate the solution numerically. One way to do this is the Finite Element Method. This is done by looking at smaller parts of the domain where we can restrict the problem, and from these restrictions we are able to approximate solutions. By combining these solutions we can obtain an approximated solution to the original problem. This process is called the Finite Element Method. Furthermore we are able to give an error estimate of the approximated solution, which is important when we want to know how accurate our solution is. In this project we will examine the basic theory regarding the Finite Element Method and use it in a numerical analysis.

2 | Examination of Partial Differential Equations

This chapter is based on [1]. The theory of Finite Element Method (FEM) stems from a wish to solve Partial Differential Equations (PDE's), which may or may not be analytically solvable. To do this we will introduce a specific kind of PDE, various types of problems, and examine the spaces containing solutions. The functions used in this chapter wil be of n variables and Ω will be an open subset of $\mathbb{R}^n$ with a piecewise smooth boundary. The first PDE we introduce is a second order PDE and as such has the form

$$c(x)u + \sum_{i=1}^{n} b_i(x)u_{x_i} - \sum_{i,k=1}^{n} a_{ik}(x)u_{x_i x_k} = f(x). \tag{2.1}$$

The coefficients in (2.1) can be organized in a vector containing every b_i in $b(x)$ and a matrix $A(x)$ containing every a_{ik}. Assuming u in (2.1) is sufficiently smooth, then $u_{x_i x_k} = u_{x_k x_i}$, and we can assume without loss of generality $A(x)$ is symmetric. We use $A(x)$ to define the type of equation we are working with and therefore to characterize the problem we are working with.

Definition 2.1. Classification of a PDE
The Equation (2.1) is classified as follows:

- Elliptic at x, if $A(x)$ is positive definite
- Hyperbolic at x, if $A(x)$ has one negative and $n-1$ postive eigenvalues
- Parabolic at x, if $A(x)$ is positive semidefinite, not positive definite, and $\operatorname{rank}([A(x), b(x)]) = n$

If for a given equation the above conditions hold for all points of the domain, the equation is called elliptic, hyperbolic or parabolic respectively.

Example 2.2.
The PDE $c(x)u + u_{x_1x_1} + u_{x_2x_2} = f(x)$ is an example of a simple Elliptic PDE, since

$$A(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b(x) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.2}$$

> Had entry a_{11} been negative, the PDE would have been Hyperbolic, since the eigenvalues would be -1 and 1. To convert (2.2) to matrices for a Parabolic PDE, we can simply change a_{11} to 0, which results in $A(x)$ being positive semidefinite and $\text{rank}([A(x), b(x)]) = 2 = n$. ◊

In this chapter we will work with several different spaces, and with different norms and seminorms. These norms are defined in [1] and in the List of Symbols. The spaces in which these norms are from should be obvious from context, but from time to time a space will be included in the subscript to specify exactly which space the norm is from. The theory of FEM stems from elliptic PDE's, so we will start by describing these problems for elliptic PDE's.

> **Definition 2.3.** Partial Differential Operator
> Let L denote the Partial Differential Operator defined as
>
> $$Lu = -\sum_{i,k=1}^{n} \partial_i(a_{ik}\partial_k u) + a_0 u, \tag{2.3}$$
>
> where a_{ik} are entries in a matrix denoted $A(x)$.

The ellipticity requirement in Definition 2.1 only makes demands on the coefficients for the second order partial derivatives, and we will ignore the other parts for now. Unless otherwise stated, L is always defined as in Definition 2.3.

2.1 Spaces of functions

The goal in this section is to develop some fundamental knowledge of the spaces in which we work. It is often the case that a classical solution to a given PDE can not be found, so we define a 'weak derivative' using Lebesgue integrals, meaning the derivative exists in some strict mathematical sense. To use these integrals, we define $L_2(\Omega)$, which are all function f over Ω such that

$$\int_{\Omega} f^2 dx < \infty.$$

We use this space to define a weak derivative.

> **Definition 2.4.** Weak Derivative
> Let α be a n-dimensional multi-index, i.e. an n-dimensional vector, with entries from

the natural numbers. A function $f \in L_2(\Omega)$ has a weak derivative $g \in L_2(\Omega)$ if

$$\int_\Omega g(x)\phi(x)dx = (-1)^{|\alpha|} \int_\Omega \partial^\alpha \phi(x) f(x)dx \qquad \forall \phi \in C_0^\infty(\Omega).$$

We then write $g = \partial^\alpha f$.

We will use derivative and weak derivative interchangeably in this text. If the derivative must be the classical, or strong, derivative, we will specify so explicitly. We use Definition 2.4 and $L_2(\Omega)$ to define the Hilbert spaces we are going to be working with.

Definition 2.5. Sobolev Space
Let $m \geq 0$ be an integer. Then $H^m(\Omega)$ is

$$H^m(\Omega) = \{f \in L_2(\Omega) \mid \partial^\alpha f \in L_2(\Omega) \quad \forall |\alpha| \leq m\}.$$

In other words, $H^m(\Omega)$ is the set of functions in $L_2(\Omega)$ which possess weak derivatives up to degree m.

Since $H^m(\Omega)$ is complete with respect to the $\| \cdot \|_m$ norm, $H^m(\Omega)$ is also a Hilbert space, hence the H. In [2] one can find a proof for the following theorem, which we use for defining a specific subspace of $H^m(\Omega)$.

Theorem 2.6.
Let $\Omega \subset \mathbb{R}^n$ be an open set with piecewise smooth boundary, and let $m \geq 0$. Then $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$.

The reason we state Theorem 2.6 is that it implies a similar relationship for a smaller subset, namely $C_0^\infty(\Omega)$, continuous differentiable functions with a compact support. So the following definition makes sense.

Definition 2.7.
We define $H_0^m(\Omega)$ to be the completion of $C_0^\infty(\Omega)$ with respect to the Sobolev norm $(\| \cdot \|_m)$.

Since $H_0^m(\Omega)$ is a restriction on the functions in $H^m(\Omega)$, we obviously know $H_0^m(\Omega) \subset H^m(\Omega)$ and, in the same sense, obviously $H^{n+1}(\Omega) \subset H^n(\Omega)$. We now move on to discussing the norms on these spaces, which is the first step in figuring out how to 'get closer' to the solution.

Theorem 2.8. Friedrichs' Inequality
Let Ω be contained in a hypercube with sidelength s. Then

$$\|v\|_0 \leq s|v|_1 \qquad \forall v \in H_0^1(\Omega).$$

Proof. We will start by assuming $v \in C_0^\infty(\Omega)$, and later use the completeness of $H_0^1(\Omega)$ to finish the proof. Let $W = \{(x_1, \dots, x_n) \mid 0 < x_i < s\}$ be the box which contains Ω, and let $v(x) = 0 \;\; \forall x \in W \setminus \Omega$. Otherwise we can simply translate Ω such that this is true. We can then write

$$v(x) = v(0, x_2, \dots, x_n) + \int_0^{x_1} \partial_1 v(t, x_2, \dots, x_n) dt.$$

Due to the definition of W, the first term (the boundary term) disappears. Applying the Cauchy-Schwartz inequality results in

$$\begin{aligned} |v(x)|^2 &= \left| \int_0^{x_1} \partial_1 v(t, x_2, \dots, x_n) dt \right|^2 \\ &= |(1, \partial_1 v)_{L^2([0,x_1])}|^2 \\ &\leq \int_0^{x_1} 1^2 dt \int_0^{x_1} |\partial_1 v(t, x_2, \dots, x_n)|^2 dt \\ &\leq s \int_0^s |\partial_1 v(t, x_2, \dots, x_n)|^2 dt. \end{aligned}$$

After expanding the integral and getting the last inequality, we know the right hand side is independent of x_1, and so by integrating we find

$$\int_0^s |v(x)|^2 dx_1 \leq s^2 \int_0^s |\partial_1 v(x)|^2 dx_1.$$

This can be done for every variable, and finding the integral over the entire domain gives us

$$\|v\|_0 = \int_W |v|^2 dx \leq s^2 \int_W |\partial_1 v|^2 \leq s^2 |v|_1^2.$$

Since $C_0^\infty(\Omega)$ is complete in $H_0^1(\Omega)$, for every $u \in H_0^1(\Omega) \setminus C_0^\infty(\Omega)$ there exists some Cauchy sequence of smooth functions $u_k \in C_0^\infty(\Omega)$ such that

$$\lim_{k\to\infty} \|u - u_k\|_1 \to 0. \tag{2.4}$$

We have proven Theorem 2.8 for every one of these u_k, and since s^2 is independent of k, we can write

$$\|u\|_0 \leq \|u_k\|_0 + \|u - u_k\|_0 \leq s^2|u_k|_1 + \|u - u_k\|_0 \leq s^2|u|_1 + s^2|u - u_k|_1 + \|u - u_k\|_0.$$

By (2.4), both of the last terms in the previous equation must tend to 0 as $k \to \infty$, and we have the proof. □

This result can be generalized to a higher order.

> **Corollary 2.9.** Higher order Friedrichs' Inequality
> Let Ω be contained in a hypercube with sidelength s. Then
>
> $$\|v\|_k \leq C|v|_{k+1} \qquad \forall v \in H_0^{k+1}(\Omega),$$
>
> where $C \in \mathbb{R}$.

Proof. We will prove this using induction, where the base case is Theorem 2.8. By the definition of these norms, the following equation is true.

$$\|u\|_k \leq C_k|u|_k \quad \forall u \in H_0^k(\Omega).$$

As in Theorem 2.8, we assume $u \in C_0^\infty(\Omega)$, and use completeness later. As $u \in C_0^\infty(\Omega) \implies u \in H^{k+1}(\Omega) \subset H^k(\Omega)$. Since $u \in C^\infty \implies \partial^k u \in C^\infty$, and thus $\partial^\alpha u \in H_0^1(\Omega)$. Then by the base case, we have

$$\|\partial^\alpha u\|_{L_2(\Omega)}^2 \leq C^2|\partial^\alpha u|_1^2 \qquad \forall \alpha : |\alpha| = k.$$

Summing over all α such that $|\alpha| = k$ and letting C_α be the C corresponding to each inequality, we get

$$\begin{aligned}\sum_{|\alpha|=k} \|\partial^\alpha u\|_{L_2(\Omega)}^2 &\leq \sum_{|\alpha|=k} C_\alpha^2|\partial^\alpha u|_1^2 \\ &\leq \tilde{C}^2 \sum_{|\tilde{\alpha}|=k+1} |\partial^{\tilde{\alpha}} u|_0^2 \\ &= \tilde{C}^2|u|_{k+1},\end{aligned} \tag{2.5}$$

where $\tilde{C} = \max\{C_\alpha : |\alpha| = k\}$. We thus get

$$\|u\|_{k+1}^2 \leq \tilde{C}^2|u|_{k+1}^2, \tag{2.6}$$

proving the corollary. □

Furthermore, Friedrichs' Inequality can be used to show $|\cdot|_m$ and $\|\cdot\|_m$ are equivalent on a bounded Ω. This can be shown thusly:

$$\begin{aligned}\|u\|_{k+1}^2 &= \|u\|_k^2 + |u|_{k+1}^2 \\ &\leq C^2|u|_{k+1}^2 + |u|_{k+1}^2 \\ &= (C^2+1)|u|_{k+1}^2 \\ &\leq \tilde{C}^2|u|_{k+1}^2,\end{aligned}$$

where we used Corollary 2.9 in the second line. Theorem 2.8 and Corollary 2.9 both deal with functions which are zero on the boundary. This is restrictive, and we would like to be able to examine functions which are not zero on the boundary. To remedy this, we introduce the Trace Theorem, which requires cones and the cone condition, which we now define.

Definition 2.10. Cone
Let $\mathbf{a}_i \in \mathbb{R}^m \setminus \vec{0}$ be the columns in some matrix $A \in \mathbb{R}^{m\times n}$, and let $x_i \geq 0$ be the entries for some vector $\mathbf{x} \in \mathbb{R}^n$. Then

$$V(A, \mathbf{x}) = \left\{ \sum_{i=1}^{n} x_i \mathbf{a}_i \right\},$$

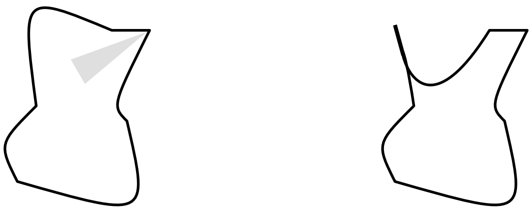
is a cone.

In $\mathbb{R}^2$ this reduces to a simple triangle, as seen in Figure 2.1a

Definition 2.11. Cone Condition
Let E be a Euclidean space and let G be an open subset of E. Then G fulfills the cone condition if

$$\forall g \in G \; \exists x \in \mathbb{R}^n \; \exists A \in \mathbb{R}^{m\times n} \;\; : V(A, \mathbf{x}) \neq \varnothing \wedge g + V(A, \mathbf{x}) \subset G.$$

This means: at all points in G, we can create a non-empty cone, such that the cone is contained in G.



(a) Space which fulfill the Cone Condition **(b)** Space which does not fulfill the Cone Condition

Figure 2.1: The Cone Condition visualized

In Figure 2.1, two spaces are visualized. The space in Figure 2.1a fulfills the Cone Condition as seen by the grey cone, which can be translated, rotated and scaled to any point, and still be contained in the space. The space in Figure 2.1b does not fulfill the Cone Condition, as in the upper left section of the space a cone cannot be drawn which is contained in the space.

With Definition 2.11, we can proceed to the Trace Theorem.

> **Theorem 2.12.** Trace Theorem
> Let Ω be bounded, and suppose Ω has a piecewise smooth boundary. In addition, suppose Ω satisfies the cone condition. Then there exists a bounded linear mapping
>
> $$\gamma : H^1(\Omega) \to L_2(\partial\Omega),$$
>
> where $\|\gamma(v)\|_{0,\partial\Omega} \leq c\|v\|_{1,\Omega}$, and $\gamma(v) = v|_{\partial\Omega}$, for all $v \in C^1(\Omega)$.

Proof. The presented proof will only be done in two dimensions, but can be generalised to higher dimensions. Suppose the boundary is piecewise smooth, and for the finitely many points where the boundary is not smooth the cone condition is satisfied. We can split the boundary into a finite number of smooth pieces $\partial\Omega_1, \partial\Omega_2, \dots, \partial\Omega_m$, where each $\partial\Omega_i$ after rotation of the coordinate system gives us;

1. For some function $\phi = \phi_i \in C^1[y_1, y_2]$, we have

$$\Gamma_i = \{(x, y) \in \mathbb{R}^2 \quad | \quad x = \phi(y), y_1 \leq y \leq y_2\}.$$

2. The domain $\Omega_i = \{(x, y) \subset \mathbb{R}^2 \quad | \quad \phi(y) < x < \phi(y) + r, y_1 < y < y_2\}$ is contained in Ω, where $r > 0$ is sufficiently small.

Firstly a function $v \in C^1(\bar{\Omega})$ and a point $(x, y) \in \partial\Omega$ can be written as

$$v(\phi(y), y) = v(\phi(y) + t, y) - \int_0^t \partial_1 v(\phi(y) + s, y)\, ds \tag{2.7}$$

where $0 \leq t \leq r$. By integrating (2.7) and switching the integration order we obtain

$$\begin{aligned}
\int_0^r v(\phi(y), y) dt &= \int_0^r \left(v(\phi(y) + t, y) - \int_0^t \partial_1 v(\phi(y) + s, y)\, ds \right) dt \\
r v(\phi(y), y) dt &= \int_0^r v(\phi(y) + t, y) dt - \int_0^r \int_0^t \partial_1 v(\phi(y) + s, y)\, ds dt \\
r v(\phi(y), y) dt &= \int_0^r v(\phi(y) + t, y) dt - \int_0^r \int_s^r \partial_1 v(\phi(y) + s, y)\, dt ds \\
r v(\phi(y), y) dt &= \int_0^r v(\phi(y) + t, y) dt - \int_0^r \partial_1 v(\phi(y) + t, y)(r - t) dt.
\end{aligned}$$

By squaring the former equation, using a form of Young's inequality $(a+b)^2 \leq 2a^2 + 2b^2$, and applying the Cauchy-Schwarz inequality to the squares of the integrals results in

$$r^2 v^2(\phi(y), y) \leq 2 \int_0^r 1dt \int_0^r v^2(\phi(y)+t, y)\, dt + 2 \int_0^r t^2 \int_0^r |\partial_1 v(\phi(y)+t, y)|^2 dt. \tag{2.8}$$

By subsituting $\int 1dt = r$ and $\int t^2 dt = \frac{r^3}{3}$, dividing by r^2 and integrating over y, we obtain

$$\begin{aligned} \int_{y_1}^{y_2} v^2(\phi(y), y) dy &\leq 2r^{-1} \int_{\Omega_i} v^2 dxdy + \frac{2r}{3} \int_{\Omega_i} |\partial_1 v|^2 dxdy \\ &\leq 2r^{-1} \int_{\Omega_i} v^2 dxdy + r \int_{\Omega_i} |\partial_1 v|^2 dxdy. \end{aligned}$$

On $\partial\Omega$ the arc length differential is given by $ds = \sqrt{1+\phi'(y)^2} dy$. Thus, we have

$$\int_{\partial\Omega} v^2 ds \leq c_i \left[2r^{-1} \|v\|_0^2 + r|v|_1^2 \right], \tag{2.9}$$

where $c_i = \max\{\sqrt{1+\phi'^2} \,|\, y_1 \leq y \leq y_2\}$. Setting $c = (r + 2r^{-1}) \sum_{i=1}^m c_i$, results in

$$\|v\|_{0,\partial\Omega} \leq c \|v\|_{1,\Omega}. \tag{2.10}$$

Thus, the restriction $\gamma : H^1(\Omega) \cap C^1(\bar{\Omega}) \to L_2(\partial\Omega)$ is a bounded linear mapping on a dense subset of $H^1(\Omega)$. Because of the completness of $L_2(\partial\Omega)$, it can be extended to all of $H^1(\Omega)$ without enlarging the bound. □

Theorem 2.12 guarantees that functions which are not zero on the boundary are at least L_2 functions on the boundary, and as such are bounded.

2.1.1 Attributes of Solutions

In this section we examine certain qualities of solutions, and the first step is the ability to characterize solutions to certain problems. To do that, we introduce the following theorem.

Theorem 2.13. Characterization Theorem
Let V be a linear space, such that $a : V \times V \to \mathbb{R}$ is a strictly positive symmetric bilinear form and $\ell : V \to \mathbb{R}$ is a linear functional. Then the quantity

$$J(v) = \frac{1}{2} a(v, v) - \ell(v)$$

attains its minimum over V at u if and only if

$$a(u, v) = \ell(v) \quad \text{for all } v \in V. \tag{2.11}$$

There is at most one solution to (2.11).

Proof. For $u, v \in V$ and $t \in \mathbb{R}$ we have

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - \ell(u+tv) \\ &= \frac{1}{2}\big(a(u,u) + a(tv,tv) + 2a(u,tv)\big) - \big(\ell(u) + \ell(tv)\big) \\ &= J(u) + t\big(a(u,v) - \ell(v)\big) + \frac{1}{2}t^2 a(v,v). \qquad (2.12) \end{aligned}$$

If $u \in V$ satisfies (2.11) and $t = 1$, then from (2.12) we have

$$\begin{aligned} J(u+v) &= J(u) + \frac{1}{2}a(v,v) && \text{for all } v \in V \\ &> J(u) && \text{for } v \neq 0, \qquad (2.13) \end{aligned}$$

since a is positive. Thus u is a unique minimal point. To prove the opposite way, we assume that J has a minimum at u. Then for every $v \in V$, the function $f(t) = J(u+tv)$ must fulfill the condition

$$\frac{d}{dt}f(0) = 0,$$

since J has a minimum at u, and as such any $t > 0$ must increase the value of J. Equation (2.12) is an equivalent statement to the usual definition of differentiability, which gives us that the derivative is $a(u,v) - \ell(v)$, and (2.11) then follows. □

To use Theorem 2.13, we only make assumptions on the linearity of the space, however we do not guarantee the existence of solutions, only their characterization. It is possible to set up a variational problem such that J does not attain its minimum. To engage with the existence of solutions, we make more assumptions on the space in which we work. We start by specifying the bilinear form in Theorem 2.13.

Definition 2.14.
Let H be a Hilbert space. A bilinear form $a : H \times H \to \mathbb{R}$ is called continuous if there exists some $C > 0$ such that

$$|a(u,v)| \leq C\|u\|\,\|v\| \quad \forall u, v \in H. \qquad (2.14)$$

If a bilinear form a is symmetric and continuous, and there exists som $\alpha > 0$ such that

$$a(v,v) \geq \alpha\|v\|^2 \quad \forall v \in H,$$

a is called elliptic.

From this point forward a will be referring to the bilinear form given by

$$a(u,v) = \int_\Omega \left(\sum_{i,k} a_{ik}\partial_i u \partial_k v + a_0 uv\right) dx, \qquad (2.15)$$

unless otherwise stated.

2.2 Homogeneous Dirichlet Conditions

Since we are trying to solve the PDE over some domain Ω, we write

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.16}$$

The assumption that $u = 0$ is often far too restrictive, but we will soften this later. We now use Theorem 2.13 to show a link between classical solutions and solutions of appropriate variational problems.

Theorem 2.15. Minimal Property
Let an elliptic PDE be given, and let a_{ik} be the entries in the positive definite matrix A for the PDE. Every classical solution of the boundary-value problem given by

$$\begin{aligned} -\sum_{i,k} \partial_i(a_{ik}\partial_k u) + a_0 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.17}$$

is a solution to the variational problem given by

$$J(v) = \int_\Omega \left[\frac{1}{2} \sum_{i,k} a_{ik}\partial_i v \partial_k v + \frac{1}{2} a_0 v^2 - fv \right] dx \longrightarrow \min,$$

among all functions in $C^2(\Omega) \cap C^0(\bar{\Omega})$ with zero boundary values.

Proof. We start off by applying Green's formula,

$$\int_\Omega v \partial_i w dx = -\int_\Omega w \partial_i v dx + \int_{\partial\Omega} vw\mathbf{n}_i ds. \tag{2.18}$$

Here we assume v and w to be $C^1(\Omega)$ functions, and $\mathbf{n}_i$ is the i'th component of the outward-pointing normal $\mathbf{n}$. Now if we insert $w = a_{ik}\partial_k u$ in (2.18), we get

$$\int_\Omega v \partial_i (a_{ik}\partial_k u) dx = -\int_\Omega a_{ik}\partial_i v \partial_k u dx, \tag{2.19}$$

given that $v = 0$ on $\partial\Omega$. Now let

$$a(u, v) = \int_\Omega \left[\sum_{i,k} a_{ik}\partial_i u \partial_k v + a_0 uv \right] dx \tag{2.20}$$

and

$$\ell(v) = \int_\Omega fv dx. \tag{2.21}$$

Then by summing (2.19) over i and k we get that for every $v \in C^1(\Omega) \cap C(\bar{\Omega})$ with $v = 0$ on $\partial\Omega$ we have

$$\begin{aligned} a(u,v) - \ell(v) &= \int_\Omega v \left[-\sum_{i,k} \partial_i (a_{ik} \partial_k u) + a_0 u - f \right] dx \\ &= \int_\Omega v[Lu - f]dx \\ &= 0, \end{aligned} \tag{2.22}$$

given that $Lu = f$, where L is as defined in (2.3). This property holds if u is a classical solution. The minimal property is then implied by Theorem 2.13 □

The same kind of proof can be used to show that a solution u to (2.22) which fulfills $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ is also a classical solution. When $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ the variational problem and homogeneous boundary value problem are thusly equivalent. We can therefore use the variational problem to find solutions, making the the following theorem useful.

> **Theorem 2.16.** Lax-Milgram Theorem
> Let V be a closed, convex, non-empty set in a Hilbert space H, and let $a : H \times H \to \mathbb{R}$ be an elliptic bilinear form. Then for every $\ell \in H'$, the variational problem given by
>
> $$J(v) = \frac{1}{2}a(v,v) - \ell(v) \longrightarrow \min$$
>
> has a unique solution in V.

Proof. J is bounded from below, since

$$\begin{aligned} J(v) &\geq \frac{1}{2}\alpha\|v\|^2 - \|\ell\|\,\|v\| \\ &= \frac{1}{2\alpha}(\alpha\|v\| - \|\ell\|)^2 - \frac{\|\ell\|^2}{2\alpha} \\ &\geq -\frac{\|\ell\|^2}{2\alpha}. \end{aligned}$$

We then let $c_1 = \inf\{J(v); v \in V\}$ and $\{v_n\}_{n=1}^{\infty}$ be a minimizing sequence. Then we get

$$\begin{aligned}\alpha \|v_n - v_m\|^2 &\leq a(v_n - v_m, v_n - v_m) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m) \\ &= 4J(v_n) + 4J(v_m) - 8J(\frac{v_m + v_n}{2}) \\ &\leq 4J(v_n) + 4J(v_m) - 8c_1.\end{aligned}$$

By the convexity assumption on V, we get that $\frac{1}{2}(v_n + v_m) \in V$, and from that we get the last inequality. We now get that $\{v_n\}_{n=1}^{\infty}$ is a Cauchy sequence in H and $u = \lim_{n\to\infty} v_n$ exists given the fact that $J(v_n), J(v_m) \to c_1$ implies $\|v_n - v_m\| \to 0$ for $n, m \to \infty$.

Since J is continuous, we see $J(u) = \lim_{n\to\infty} J(v_n) = \inf_{v\in V} J(v)$. Moreover, we also have that $u \in V$, because we assumed V to be closed. Lastly, we need to show that the solution is unique. To do that, we assume both u_1 and u_2 to be solutions of the variational problem. Then we set up a minimizing sequence $u_1, u_2, u_1, u_2, \ldots$, which we know to be a Cauchy sequence from earlier. This implies $u_1 = u_2$. □

As all Hilbert spaces are vector spaces, they will be convex by definition. In addition we will limit ourselves to complete Hilbert spaces. As such, the subsets of Hilbert spaces we are going to be working will be as well, and we can use the following corollary.

> **Corollary 2.17.**
> Let H be a complete Hilbert space, and let V be a closed, linear subspace of H. The variational problem from Theorem 2.16 has a unique solution in V given by
>
> $$a(u, v) = \ell(v) \quad \forall v \in V.$$

Proof. From the assumptions we know V is closed and convex. Since $0 \in V$, V is also non-empty. Then we can use Theorem 2.13 to find the form in the corollary. □

Theorem 2.15, Theorem 2.16 and Corollary 2.17 together shows that solving a Dirichlet problem is the same as finding the minimum of the corresponding variational problem. We can use this in our approach both theoretically and computationally.

When working with PDE's the equation shown in Corollary 2.17 is called the variational formulation of a boundary condition problem. To find the solution u, we split a Hilbert space of functions, H, into finite dimensional subspaces, and use these to approximate u. We test this approximatation using test functions, typically from the same subspace, using the equality in Corollary 2.17, which must be true for all test functions. This widens the space in which we look for solutions. For example, by using the variational formulation, we can define a weak solution which no longer requires $u \in C^2(\Omega)$.

Definition 2.18. Weak Solution
A function $u \in H_0^1(\Omega)$ is called a weak solution of the Dirichlet problem with homogeneous boundary conditions, (2.16), if

$$a(u,v) = (f,v)_0 \quad \forall v \in H_0^1(\Omega).$$

Here $a(u,v)$ is the bilinear from from (2.20), and $(f,v)_0$ is a linear bounded functional provided $f \in L^2$, and it can therefore be used as $\ell(v)$ in (2.21).

A weak solution only requires $u \in H_0^1(\Omega)$, and is clearly less strict. Thus there is a higher chance of a solution existing for any given problem, and is therefore desirable. So we now look at the existence of weak solutions.

Theorem 2.19. Existence theorem
Let L be the second order uniformly elliptic partial differential operator from (2.3). Then the homogeneous Dirichlet problem, (2.16), always has a weak solution in $H_0^1(\Omega)$. It is a minimum of the variational problem

$$J(v) = \frac{1}{2}a(v,v) - (f,v)_0 \to \min \tag{2.23}$$

over $H_0^1(\Omega)$.

Proof. Let $c = \sup\{\left|a_{ik}(x)\right| \mid x \in \Omega, 1 \leq i,k \leq n\}$. Then owing to the Cauchy-Schwarz inequality we get

$$\begin{aligned}\left|\sum_{i,k}\int a_{ik}\partial_i u \partial_k v\right| dx &\leq c\sum_{i,k}\int |\partial_i u \partial_k v|\, dx \\ &\leq c\sum_{i,k}\left[\int (\partial_i u)^2 dx \int (\partial_k v)^2 dx\right]^{1/2} \\ &\leq C|u|_1|v|_1\end{aligned} \tag{2.24}$$

where $C = cn^2$. We also assume $C \geq \sup\{|a_0(x)| \mid x \in \Omega\}$ giving us

$$\left|\int a_0 uv dx\right| \leq C\int |uv| dx \leq C\|u\|_0\|v\|_0. \tag{2.25}$$

Then by (2.24) and (2.25) we get continuity of a. Furthermore the uniform ellipticity implies the pointwise estimate

$$\sum_{i,k} a_{ik}\partial_i v \partial_k v \geq \alpha \sum_i (\partial_i v)^2,$$

for $C^1(\Omega)$ functions. Following the same density arguments as in the proof for Theorem 2.8, $C^1(\Omega)$ is dense in $H^1(\Omega)$, therefore all $u \in H^1(\Omega)$ are represented in $C^1(\Omega)$, and thus the previous statement can be generalized to $H^1(\Omega)$. Integrating both sides and using $\mathrm{a}_0 \geq 0$ leads to

$$a(v,v) \geq \alpha \sum_i \int_\Omega (\partial_i v)^2 dx = \alpha |v|_1^2, \quad \forall v \in H^1(\Omega). \tag{2.26}$$

We know from Friedrichs' Inequality $|\cdot|_1$ and $\|\cdot\|_1$ are equivalent norms on H_0^1, resulting in a being an H^1-elliptic bilinear form on $H_0^1(\Omega)$. From Theorem 2.16 there exists a unique weak solution which is also a solution of the variational problem. □

2.3 Inhomogeneous Dirichlet Conditions

The theory in the previous section have been regarding homogeneous Dirichlet problems; however, if we soften the requirement for u on $\partial\Omega$, and write

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.27}$$

we get an inhomogeneous Dirichlet boundary condition. However, inhomogeneous Dirichlet problems can be turned into homogeneous Dirichlet problems using Theorem 2.19. In the inhomogeneous Dirichlet problem, Equation (2.27), let $\tilde{u} \in C^2(\Omega) \cap C^0(\bar{\Omega}) \cap H^1(\Omega)$ be a function such that

$$\tilde{u} = g \quad \text{on } \partial\Omega,$$

and $w = u - \tilde{u}$ and $\tilde{f} = f - L\tilde{u}$. We now have the homogeneous Dirichlet problem

$$\begin{aligned} Lw &= \tilde{f} \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.28}$$

By Theorem 2.19, a weak solution $w \in H_0^1(\Omega)$ exists. As $(L\tilde{u}, v)_0 = a(\tilde{u}, v)$, we get

$$\begin{aligned} a(w,v) &= (\tilde{f}, v)_0 \qquad\qquad \forall v \in H_0^1(\Omega) \\ a(u-\tilde{u}, v) &= (f - L\tilde{u}, v)_0 \\ a(u,v) - a(\tilde{u}, v) &= (f,v)_0 - (L\tilde{u}, v)_0 \\ a(u,v) &= (f,v)_0. \end{aligned}$$

We stil have the condition $u - \tilde{u} = w \in H_0^1(\Omega)$, and the weak formulation of the inhomogeneous Dirichlet boundary condition can then be written as

$$\begin{aligned} a(u,v) = (f,v)_0 \quad \forall v \in H_0^1(\Omega) \\ u - \tilde{u} \in H_0^1(\Omega). \end{aligned}$$

By Theorem 2.6, $C^2(\Omega) \cap H^1(\Omega)$ is dense in $H^1(\Omega)$, and $C^0(\bar{\Omega}) \cap H_0^1(\Omega)$ is dense in $H_0^1(\Omega) \subset H^1(\Omega)$. Following the same density arguments as have been done several times before, it suffices to look at $\tilde{u} \in H^1(\Omega)$, however it is not always possible to find an admissible $\tilde{u}$.

As seen in the two previous sections, a Dirichlet condition on the boundary directly prescribe the solution. We cannot always guarantee the behaviour of the function on the boundary and thus the Dirichlet condition can be too restrictive. A less restrictive requirement deals with the derivatives on the boundary, which the next section will deal with.

2.4 Neumann Boundary Conditions

We remind the reader of the bilinear form we are working with,

$$a(u,v) = \int_\Omega \left(\sum_{i,k} a_{ik} \partial_i u \partial_k v + a_0 uv \right) dx. \tag{2.29}$$

We now require $a_0(x)$ in the Elliptic PDE, see 2.1, to be bounded from below by a positive number (and so also in a). Inequality (2.26) will also be true for smaller α, so if necessary we can reduce it, and assume

$$a_0(x) \geq \alpha > 0 \quad \forall x \in \Omega.$$

This gives us the lower bound on a as

$$a(v,v) \geq \alpha |v|_1^2 + \alpha \|v\|_0^2 = \alpha \|v\|_1^2 \quad \forall v \in H^1(\Omega).$$

We can therefore assume from this point on that a is H^1 elliptic.

Definition 2.20. Neumann Boundary condition
Let $\frac{\partial u}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla u$ be the normal derivative. Then the Neumann boundary condition is given by

$$\frac{\partial u}{\partial \mathbf{n}} = g \quad \text{on } \partial\Omega. \tag{2.30}$$

With this definition, we can start looking at the existence of solutions to problems which are defined using this type of condition.

Theorem 2.21.
Let Ω be bounded, have a piecewise smooth boundary, and satisfy the cone condition.

Let $f \in L_2(\Omega)$ and $g \in L_2(\partial\Omega)$. There exists a unique $u \in H^1(\Omega)$ that solves the variational problem

$$J(v) = \frac{1}{2}a(v,v) - (f,v)_{0,\Omega} - (g,v)_{0,\partial\Omega} \to \min .$$

Also, $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ if and only if a classical solution of

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ \sum_{i,k} \mathbf{n}_i a_{ik} \partial_k u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.31}$$

exists, in which case these 2 solutions are the same. Here $\mathbf{n}$ is outward pointing normal on $\partial\Omega$, defined almost everywhere.

Proof. Due to the Trace Theorem, 2.12, the functional $(g,v)_{0,\partial\Omega}$ is bounded, and thus $(f,v)_{0,\Omega} - (g,v)_{0,\partial\Omega}$ is a bounded linear functional, and since a is H^1 elliptic, Theorem 2.16 gives us the uniqueness and existence of u. Corollary 2.17 gives us

$$a(u,v) = (f,v)_{0,\Omega} + (g,v)_{0,\partial\Omega} \quad \forall v \in H^1(\Omega). \tag{2.32}$$

We now move on to the "if and only if" part of the theorem. Start by assuming $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$. If we look in the interior, and as such at $v \in H_0^1(\Omega)$, we get $\gamma(v) = 0$ and (2.32) simply becomes a a Dirichlet problem where u is used to define the boundary condition, see (2.16). On the interior we then have

$$Lu = f \quad \text{in } \Omega, \tag{2.33}$$

which is the first condition in (2.31). We now examine the boundary, and we assume $v \in H^1(\Omega)$. In the proof of Theorem 2.15, we used Green's formula,

$$\int_\Omega v \partial_i w dx = -\int_\Omega w \partial_i v dx + \int_{\partial\Omega} vw\mathbf{n}_i ds. \tag{2.34}$$

In that proof we could assume the integral on the boundary to be zero, which we cannot do here. However, with the same arguments regarding summing and substituting, we get

$$a(u,v) - (f,v)_{0,\Omega} - (g,v)_{0,\partial\Omega} = \int_\Omega v[Lu - f]dx + \int_{\partial\Omega} \left[\sum_{i,k} \mathbf{n}_i a_{ik} \partial_k u - g \right] vds. \tag{2.35}$$

Using (2.32) and (2.33), the last integral in (2.35) must be 0.

Suppose that for some i and k the function $v_0 = \mathbf{n}_i a_{ik} \partial_k u - g$ does not vanish on $\partial\Omega$. Then $\int_{\partial\Omega} v_0^2 ds > 0$, and by the density of $C^1(\bar{\Omega})$ in $C^0(\bar{\Omega})$, there exists a $v \in C^1(\bar{\Omega})$ such that

$\int_{\partial\Omega} v_0 \cdot v ds > 0$, which cannot happen, due that what we found using (2.35). We therefore have the second condition in (2.31), and u is a classical solution of this boundary problem.

On the other hand, assuming u is a classical solution and satisfy (2.31), we see from (2.35) that u also satisfy (2.32), and is therefore a solution to the variational problem, and by Theorem 2.16 unique. □

Not all Neumann problems fulfill the assumption that $a_0(x) > 0$. For example in the Neumann Poisson problem,

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{2.36}$$

one can find $a_0(x) = 0$. We remedy that by constructing a subspace V. Due to the fact that (2.36) only contains derivatives with respect to u, additive constants independent of u are unimportant. Thus if u satisfies (2.36), then $u + c$ also satisfies (2.36) for any constant c. This means that the solution to (2.36) is only unique up to a constant, which extends to other equations with Neumann boundary conditions.

It turns out that we can formulate the weak form of this problem by restricting to the subspace $V = \{v \in H^1(\Omega) : \int_\Omega v dx = 0\}$. Define $a(u, v) = \int_\Omega \nabla u \cdot \nabla v dx$. Examining a version of the Friedrich inequality, Theorem 2.8, which can be found on page 46 of [1], a can be found to be V-elliptic. We now want to assert that every classical solution to the variational problem in Theorem 2.21 satisfies (2.36). To do this, we let $w = \nabla u$. Thus (2.36) can be written as

$$\begin{aligned} -\text{div } w &= f \quad \text{in } \Omega, \\ \mathbf{n}^T w &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{2.37}$$

From the Gauss Integral Theorem, we have that

$$\int_\Omega \text{div } w dx = \int_{\partial\Omega} \mathbf{n}^T w ds.$$

Then from (2.37) we get

$$\int_\Omega -f dx = \int_{\partial\Omega} g ds. \tag{2.38}$$

The integral used in V is a bounded linear functional, and thus V is a closed and linear subspace of a Hilbert space, V is complete with respect to the norm, and is therefore also a Hilbert space. We can then use Corollary 2.17 and get $u \in V$, where

$$a(u, v) = (f, v)_{0,\Omega} + (g, v)_{0,\partial\Omega} \quad \forall v \in V. \tag{2.39}$$

Equation (2.36) has now been shown to be true for every classical solution of the variational problem for $u \in V$. However, for some $f \in H^1(\Omega)\backslash V$, we can use the following construction for the constant c:

$$\forall f \in H^1(\Omega): \; c = \frac{\int_\Omega f(x)dx}{\int_\Omega 1dx}.$$

Define $v(x) = f(x) - c$, and then

$$\int_\Omega v(x)dx = \int_\Omega f(x)dx - c\int_\Omega 1dx = \int_\Omega f(x)dx - \frac{\int_\Omega f(x)dx}{\int_\Omega 1dx}\int_\Omega 1dx = 0,$$

so $v \in V$. For every $u \in H^1(\Omega) \setminus V$ we can find a fitting constant such that $u + c \in V$. Every classical solution of the variational problem corresponding to a Neumann Boundary condition therefore satisfies (2.36), and (2.36) becomes the weak formulation.

The end result is we look for the solution $u \in V$, and use test functions $v \in H^1(\Omega)$.

> **Example 2.22.** Mixed conditions
> Consider the following problem: In a room with two open windows, and a vacuum cleaner placed in the middle of the room aimed at a fixed point. The wind flowing through the windows can be described using a Neumann boundary condition, while the flow around the vacuum cleaner would result in a Dirichlet boundary condition. $\Diamond$

We have now examined properties of elliptic PDE's, some of the boundary problems originating from elliptic PDE's, and the existence of solutions of these problems. From this we can now move on to how to find these solutions using FEM.

3 | Finite Element Method

In this chapter we examine ways to numerically approximate the solutions described in the previous chapter. The idea is essentially to partition the domain Ω into a finite amount of subsets, preferably with certain qualities, which we will discuss later.

Using these subsets, we can define a subspace of $H^m(\Omega)$ or $H_0^m(\Omega)$ with a finite dimension, called S_h. This could be a space of piecewise polynomials. The variational problem can then be solved over this space, since an infinite number of dimensions can make computation difficult, or impossible.

There will be 2 parts; in one we will examine how the minimizers of J behaves in S_h versus the full space, and in the other we will discuss partitioning Ω, using a mesh. An *element* or *cell* is a subset of Ω, meaning a geometric object, and a *finite element* is a triple consisting of; a subset of Ω, a space of functions, and a set of linearly independent functionals on these functions. When the context makes the meaning clear, we might deviate from this convention. We will start by limiting our discussion to a polygonial $\Omega \subset \mathbb{R}^2$, which can be partitioned into triangles or quadrilaterals.

3.1 Partitioning

Partitioning some domain simply means splitting it into a number of subsdomains. There are different ways to do that and different desirable qualities. The important one will be covered later.

By partitioning Ω, we also create a new space of functions over Ω. By splicing the function over each cell, we can create all kinds of discontinuous functions ; this, however, does not help us reach our goal, as many of these functions are not necessarily differentiable. For each cell, we create certain requirements on the edges for the functions over this cell. The functions which obey these requirements make up S_h. The idea of S_h is to solve the same variational problem as in the global case, but in a smaller space; so the problem becomes

$$J(v) = \frac{1}{2}a(v,v) - \ell(v) \to \min_{S_h}.$$

The only difference is in the space we look for a solution, namely S_h. From earlier theory, we know that the solution is given by an u_h which solves

$$a(u_h, v) = \ell(v) \quad \forall v \in S_h.$$

From this point on, for a solution u to some variational problem, u_h will be the solution in S_h. Since we assumed S_h to be finite dimensional, we can let $\{\psi_1, \psi_2, \dots, \psi_N\}$ be a basis for S_h. Since ℓ is a linear functional and a is bilinear, u_h can be found using the following weak formulation:

$$a(u_h, \psi_i) = \ell(\psi_i) \quad i = 1, 2, \dots, N.$$

As $u_h \in S_h$, u_h can be found to be a linear combination of the basis-functions, with coefficients z_i, $i = 1, 2, \dots, N$. As a is bilinear, we get the system of equations

$$\sum_{k=1}^{N} z_k a(\psi_k, \psi_i) = \ell(\psi_i) \quad i = 1, 2, \dots, N.$$

Letting $a(\psi_i, \psi_j)$ be the j, i-entries in a matrix A, z_i and $\ell(\psi_i)$ be the entries the vectors z and b respectively, this can be written in matrix form

$$Az = b. \tag{3.1}$$

When using FEM, Equation (3.1) is what we use to find the solution; the other big part of FEM is constructing A, which is dependent on the partitioning of Ω, as this partitioning decides what S_h and the basisvectors are.

It can be shown that the matrix A is symmetric and positive definite, when a is an H^m-elliptic bilinear form, which we do now. The symmetry of A follows directly from the symmetry of a, as a is H^m-elliptic. To show that A is positive definite, we need to show that $x^T Ax > 0$ for all $x \neq 0$. Thus

$$\begin{aligned} x^T Ax &= \sum_{i,k} x_i A_{ik} x_k \\ &= a\left(\sum_k x_k \psi_k, \sum_i x_i \psi_i\right) \\ &= a(u_h, u_h) \geq \alpha \|u_h\|^2 > 0. \end{aligned}$$

From now on, we will assume for simplicity that $V \subset H^m(\Omega)$ and a is V-elliptic.

Definition 3.1. Admissible Partition
A partition of Ω into triangles or quadrilaterals, $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$, is called admissible if:

1. $\bar{\Omega} = \cup_{i=1}^{M} T_i$
2. If $T_i \cap T_j$ is exactly one point, it is a common vertex of T_i and T_j
3. If, for $i \neq j$, $T_i \cap T_j$ is more than one point, it is a common edge of T_i and T_j.

We further describe these partitions. If every element in $\mathcal{T}$ has points which have at most a diameter of $2h$, we write $\mathcal{T}_h$. We also call $\mathcal{T}$ a *family of partitions.*

Intuitively, an admissible partition of Ω is nice; we can partition Ω into triangles (or quadrilaterals) without losing parts of Ω, and no elements contain random points scattered around Ω without being connected to the element itself.

Next we examine a way to dominate the error of using a solution in S_h instead of V.

> **Lemma 3.2.** Cea's Lemma
> Let the bilinear form a be V-elliptic with $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. In addition, let u and u_h be solutions to the variational problem in V and $S_h \subset V$, respectively. Then
>
> $$\|u - u_h\|_m \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_m. \tag{3.2}$$

Proof. The way we define u and u_h yields

$$\begin{aligned} a(u, v) &= \ell(v) \quad \forall v \in V, \\ a(u_h, v) &= \ell(v) \quad \forall v \in S_h. \end{aligned} \tag{3.3}$$

We assumed $S_h \subset V$, and we therefore by subtraction get

$$a(u, v) - a(u_h, v) = 0 \quad \forall v \in S_h. \tag{3.4}$$

which by linearity of a yields

$$a(u - u_h, v) = 0 \quad \forall v \in S_h. \tag{3.5}$$

Now let $v_h \in S_h$ and $v = v_h - u_h$. Then combining this with (3.5), we get

$$a(u - u_h, v_h - u_h) = 0 \quad \forall v_h \in S_h, \tag{3.6}$$

and

$$\begin{aligned} \alpha\|u - u_h\|_m^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &\leq C\|u - u_h\|_m \|u - v_h\|_m. \end{aligned}$$

Rearranging the inequality yields

$$\|u - u_h\|_m \leq \frac{C}{\alpha}\|u - v_h\|_m, \tag{3.7}$$

and therefore

$$\|u - u_h\|_m \leq \frac{C}{\alpha} \inf_{v_h \in s_h} \|u - v_h\|_m. \tag{3.8}$$

□

Lemma 3.2 gives us an understanding of possible errors; the next theorem shows a connection between our Hilbert spaces and continuous functions.

Theorem 3.3.
Let $k \geq 1$ and suppose Ω is bounded. Then a piecewise infinitely differentiable function $v : \bar{\Omega} \to \mathbb{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\bar{\Omega})$.

Proof. For simplicity we restrict ourselves to domains in $\mathbb{R}^2$. Start by assuming $k = 1$. Assuming $v \in C^0(\bar{\Omega})$ we wish to prove $v \in H^1(\Omega)$. Let $\mathcal{T} = \{T_j\}_{j=1}^M$ be a partition of Ω. Define the piecewise functions $w_i : \Omega \to \mathbb{R}$ for $i = 1, 2$ as $w_i(x) = \partial_i v(x)$ for $x \in \Omega$, where the function can take either of the two limiting values on the edges of T_j.
Let $\phi \in C_0^\infty(\Omega)$ and we get

$$\begin{aligned} \int_\Omega \phi w_i dxdy &= \sum_j \int_{T_j} \phi \partial_i v dxdy \\ &= \sum_j \left(-\int_{T_j} \partial_i \phi v dxdy + \int_{\partial T_j} \phi v \mathbf{n}_i ds \right). \end{aligned} \tag{3.9}$$

The first equality follows from linearity of integrals and qualities of a partition, and the second equality is Greens identity. Since v is continuous, the integrals over the interior edges cancel out, as the normal vectors point in opposite directions. The function ϕ has compact support, meaning it is zero on the boundary and outside Ω. The integrals on the boundary therefore vanish and we are left with

$$\int_\Omega \phi w_i dxdy = -\int_\Omega \partial_i \phi v dxdy. \tag{3.10}$$

This, by Definition 2.4, implies that w_i is the first order weak derivative of v. As $v \in C^0(\bar{\Omega})$ we get $v \in L_2(\Omega)$. These facts together results in $v \in H^1(\Omega)$ by definition.

We now examine the other implication. Assuming $v \in H^1(\Omega)$ and piecewise differentiable, we wish to show $v \in C^0(\bar{\Omega})$. First we examine v in the neighborhood of an edge of an element, and rotate the edge so it lies on the y-axis. On the edge we define the interval $[y_1 - \delta, y_2 + \delta]$ for y_1 and y_2 on the edge, and $y_1 < y_2$, and $\delta > 0$. Define the auxiliary function

$$\psi(x) = \int_{y_1}^{y_2} v(x, y) dy. \tag{3.11}$$

The auxiliary function has the following properties

$$\psi' = \int_{y_1}^{y_2} \partial_1 v dy, \quad \psi(x_2) - \psi(x_1) = \int_{x_1}^{x_2} \psi' dx. \tag{3.12}$$

Suppose $v \in C^{\infty}(\Omega)$. From the Cauchy-Scwarz inequality we get

$$|\psi(x_2) - \psi(x_1)|^2 = \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} \partial_1 v dxdy \right|^2 \tag{3.13}$$

$$\leq \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} 1 dxdy \right|^2 \cdot |v|_{1,\Omega}^2 \tag{3.14}$$

$$\leq |x_2 - x_1|^2 \cdot |y_2 - y_1|^2 \cdot |v|_{1,\Omega}^2, \tag{3.15}$$

which is Lipschitz continuity. Reusing the same density argument, since $C^{\infty}(\Omega)$ is dense in $H^1(\Omega)$, the previous inequality also holds for $v \in H^1(\Omega)$. The function ψ is thus continuous for the entire domain, and at $x = 0$. Since y_1 and y_2 was arbitrarily chosen, it must be true for the entire edge, so v is continuous on Ω. The boundary of Ω will be included in the edges of the elements, and therefore we get $v \in C^0(\bar{\Omega})$.

For $k > 1$, assume the theorem holds for $k - 1$. We can then, by assumption, differentiate $k - 1$ times, where the bi-implication holds after each differentiation. After differentiating $k - 1$ times, we arrive at the case we have just proven. The theorem is thus true for $k > 1$.

□

In Theorem 3.3 Ω could be the entire domain; however it could also just be a cell. This gives us some local requirements. If we want a solution which is locally C^2 (not on a cell edge), every function in the finite elements must be H^3. We now move from discussing Ω, to what a finite element is, and how to construct them.

3.2 Finite Elements

We now start describing the finite elements. These are the building blocks used to construct both S_h, partition Ω, and solve the variational problem. We start by defining them precisely.

> **Definition 3.4.** Finite Element
> A finite element is a triple (T, Π, Σ) which has the following properties:
>
> 1. $T \subset \mathbb{R}^d$ is a polyhedron
> 2. $\Pi \subset C(T)$ with finite dimension s
> 3. Σ is a set of s linearly independent functionals on Π. Every $p \in \Pi$ is uniquely defined by the values of the s functionals in Σ.

The vernacular can be extended as follows:

- The parts of ∂T which lies in different hyperplanes are called faces

- A set of functions in Π which form a basis are called shape functions
- s is the number of local degrees of freedom or local dimension

If taken head on, whenever we partition Ω, we would have to examine each finite element individually, which would make this method completely untenable. What we do instead is examine a single finite element, which can represent all the other finite elements. In the following text we will use the set of polynomials of two variables with degree t, which are defined as

$$\mathcal{P}_t = \{f(x,y) = \sum_{\substack{i+k\leq t\\ 0\leq i,k}} c_{ik}x^i y^k\}.$$

If a finite element (T, Π_i, Σ_i) fulfills $\mathcal{P}_t \subset \Pi_i$, we call it a finite element with complete polynomials (of degree t, if it is not clear).

> **Definition 3.5.** Reference Finite Element
> Let $(T_{\text{ref}}, \Pi_{\text{ref}}, \Sigma_{\text{ref}})$ be a finite element with $T_{\text{ref}} \in \mathcal{T}$ for som admissible partition of Ω, and F an affine transformation. Assume that for $T_i \in \mathcal{T}$ the following is true for the corresponding finite element (T_i, Π_i, Σ_i):
>
> - $F(T_{\text{ref}}) = T_i$
> - $\{f \circ F \,|\, f \in \Pi_i\} = \Pi_{\text{ref}}$
> - $\{s(f \circ F) \,|\, f \in \Pi_i, s \in \Sigma_{\text{ref}}\} = \Sigma_i$
>
> If the previous equalities are true for all $T_i \in \mathcal{T}$ we call $(T_{\text{ref}}, \Pi_{\text{ref}}, \Sigma_{\text{ref}})$ the finite reference element.

The previous definition might be a bit dense, and so we will expand a bit further here. The first equality in the definition requires that the different cells are "similar enough" - meaning affine transformations of each other. The second equality makes sure that the functions in the i'th finite element corresponds to using the same functions on the transformed reference cell - and that these are the same functions used in the reference finite element. The final equality ensures that the functionals in the reference cell that uniquely identify the functions in the reference finite elements, remain in the transformed cell.

> **Remark 3.6.**
> Let $f \in \mathcal{P}_t$ and V an affine linear transformation. Expressing $f \circ V$ in the appropiate coordinates, we find $f \circ V \in \mathcal{P}_t$. Shifting and scaling vairables every which way still results in polynomials. The set $\mathcal{P}_t$ is therefore invariant under affine linear transformations.

Remark 3.6 shows that using polynomials as shape functions in a finite reference element is possible.

3.2.1 Continuity of Finite Elements

When designing finite elements, it can be easy to be seduced by the locality of each finite element, and forget to keep the original problem in mind. In this section we examine how different requirements on the solution can affect how the finite elements are constructed using different examples. Say, to start, we only require the solution to be continuous, everywhere on Ω. To do so, when the solution transitions from element to element, the solution must 'match up', coming from each element. One way to include this in the construction can be seen in the following example.

> **Example 3.7.**
> For some $\Omega \subset \mathbb{R}^2$, let $\mathcal{T}$ be an admissible partition into triangles. Let $t > 0$, and for each finite element, let $\Pi = \mathcal{P}_t$, resulting in finite elements with complete polynomials. In each T_i place $s = (t+1)(t+2)/2$ points such there are $t+1$ points on each edge of T_i, with shared edges also sharing points. An illustration of this can be seen in Figure 3.1, for $t = 5$. The points inside the triangle would not have to be as regular as they have been depicted here, nor would the triangle hvave to be.
>
> By choosing values for each point, every polynomial on each T_i is uniquely determined (the proof of uniqueness can be found in [1], page 64). By restricting the polynomials to an edge, they become polynomials of one variable, and are uniquely determined by the values at the $t + 1$ points on that edge.
>
> Since the values and the points at each edge is the same, polynomials from neighbouring elements reduce to the same polynomial, and we get global continuity. $\Diamond$

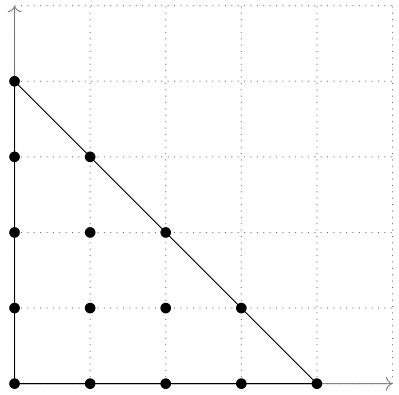


Figure 3.1: A possible triangle element

Example 3.7 demonstrates, that global continuity is somewhat easy to obtain, and constructing finite elements using polynomials of any degree grants this property. Letting

$t = 1$ results in finite elements where $\dim(\Pi_i) = 3$, and finding the solution becomes computationally easy.

If, however, we search for solutions which are $C^1(\Omega)$, difficulty increases. Along each edge the functions themselves must match up, but the first derivatives must now also match. We express this using the normal derivative, and showcase this in the following example.

> **Example 3.8.**
> As in Example 3.7, for some Ω, let $\mathcal{T}$ be an admissible partition into triangles, and let $\Pi_i = \mathcal{P}_5$ for each finite element. Remember that $\dim(\mathcal{P}_5) = 21$. Let the values of the derivatives at each vertex be given up to the 2nd order, as well as the value of the normal derivative at the midpoint of each side of each element.
>
> To ensure a solution in $C^1(\Omega)$, we now check the solution three ways; inside each element (1), on the vertices (2), and on the edges (3).
>
> **1:** Inside each element, the solution is polynomial of degree 5, and therefore also $C^1(T_i)$.
>
> **2:** At each vertex, the function and the derivatives in the direction of the edge is is known. On the edge a polynomial of degree 5 have 6 degrees of freedom, so the derivatives at the vertices guarantee global continuity.
>
> **3:** Along the edge, the polynomials reduce to 1 variable, and the normal derivative is a polynomial of degree 4. Since it follows the derivatives up to the first order at each end of the edge, and the value is given at the midpoint, the normal derivative is uniquely determined as well, ensuring the continuity of the normal derivative.
>
> Thus the normal derivative is continuous everywhere, resulting in a solution in $C^1(\Omega)$. This element is well known, and called Argyris element. ◊

As we can see from Example 3.8, by requiring a solution be differentiable, construction of finite elements and computation becomes harder.

Many different types of finite elements exists, and can differ wildly in how they are constructed. In Example 3.7 and 3.8 the finite elements are constructed using a nodal system-specifying all elements of Π uniquely through values at a number of points. However something like Hsieh–Clough–Tocher element is constructed by subdividing triangles into further triangles and using cubic polynomials.

These previous examples show how to construct finite elements 'well'-as the solution we find obeys the requirement, namely either being $C^0(\Omega)$ or $C^1(\Omega)$. However elements can also be constructed badly, as we will show here. In rectangular elements we use a polynomial family with tensor products:

$$\mathcal{Q}_t = \{f(x,y) = \sum_{0 \leq i,k \leq t} c_{ik} x^i y^k\}.$$

In the following example we will show how constructing finite elements can go wrong.

> **Example 3.9.**
> For some Ω, let $\mathcal{T}$ be an admissible partition into rectangles, and for every T_i let $\Pi_i = \mathcal{Q}_1$. Every function in Π_i then has the form
>
> $$f(x, y) = a + bx + cy + dxy.$$
>
> By specifying the values at every vertex of the element, every function in Π_i is uniquely determined. As neighbouring elements share two vertices and values, and every $f \in \Pi_i$ restricted to an edge is a linear function, we are guaranteed global continuity. However, as seen in Figure 3.2, if the element is rotated at 45°, the term dxy disappears and global continuity is longer guaranteed.
>
> This can of course be remedied by either choosing elements differently, or an appropiate transformation of variables. $\diamondsuit$

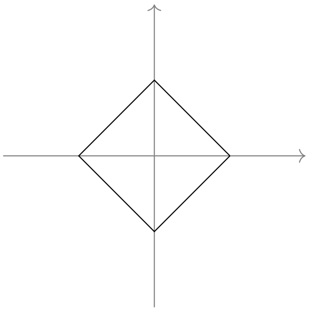


Figure 3.2: A quadrilateral element at 45° from the axis

As can be seen in Examples 3.7, 3.8, and 3.9, there are many different ways to construct finite elements, and doing so is not an exercise detached from Ω or the requirements on the solution. In this chapter we have looked at what FEM is practically, and how we construct meshes using different partitions, and how those partitions are created using finite elements. We now move on to examining how the error of using a FEM solution behaves.

4 | Aproximations and errors

We have, so far, used the norms $|\cdot|$ and $\|\cdot\|$ defined over Ω. This becomes too restrictive, as in approximating the solution we examine a reference element. The norms will therefore change depending on the constructed mesh. We do so using the following definition.

Definition 4.1.
Let $\mathcal{T}_h = \{T_1, T_2, \dots, T_M\}$ be a partition of Ω, and let $m \geq 1$. Then define

$$\|v\|_{m,h} = \sqrt{\sum_{T_j \in \mathcal{T}_h} \|v\|^2_{m,T_j}}$$

Note that if $v \in H^m(\Omega)$ then $\|v\|_{m,h} = \|v\|_{m,\Omega}$. We can now move on to defining the interpolation operator as follows.

Definition 4.2.
Let a finite element (T, Π, Σ) be given, with $v \in \Pi$. Let z_i be the s points in T that, through all $p \in \Sigma$, uniquely determines the functions in Π. Define $I : H^t \to \mathcal{P}_{t-1}$ as the function which maps $v \in H^t$ to the polynomial which solves

$$p(z_i) = v(z_i), \quad 1 \leq i \leq s.$$

Lemma 4.3.
Let Ω be a domain with a Lipschitz continuous boundary which satisfies a cone condition. Furthermore, let $t \geq 2$, and suppose $z_1, z_2, \dots z_s$ are the $s = t(t+1)/2$ prescribed points in $\bar{\Omega}$ such that the interpolation operator $I : H^t \to \mathcal{P}_{t-1}$ is well defined. Then there exists a constant c such that

$$\|u - Iu\|_t \leq c|u|_t \quad \forall u \in H^t(\Omega). \tag{4.1}$$

Proof. We endow $H^t(\Omega)$ with the norm

$$|||v||| = |v|_t + \sum_{i=1}^{s} |v(z_i)|.$$

It can then be shown that the norms $|||\cdot|||$ and $\|\cdot\|_t$ are equivalent on $H^t(\Omega)$. Thus

$$\begin{aligned}\|u-Iu\|_t &\le c|||u-Iu||| \\ &= c(|u-Iu|_t+\sum_{i=1}^{s}|(u-Iu)(z_i)|) \\ &= c|u-Iu|_t \\ &= c|u|_t.\end{aligned}$$

This holds due to the facts that Iu is a polynomial of degree $t-1$, which means that $D^\alpha Iu=0$ for all $|\alpha|=t$, and $(Iu)(z_i)=u(z_i)$, by definition of I.

We now move on to proving one direction of the equivalence of the norms. It can be shown that the imbedding $H^t\hookrightarrow H^2\hookrightarrow C^0$ is continuous (see the Sobolev Inequality in [1], page 49). Thus

$$|v(z_i)|\le c\|v\|_t \quad \text{for } i=1,2,\dots,s.$$

and

$$|||v|||\le(1+cs)\|v\|_t.$$

For the opposite direction, suppose

$$\|v\|_t\le c|||v||| \quad \text{for all } v\in H^t(\Omega)$$

fails for every positive number c. Then there exists some sequence $\{v_k\}_{k=1}^{\infty}$ in $H^t(\Omega)$ such that

$$\|v_k\|_t=1, \quad |||v_k|||\le\frac{1}{k}, \quad k=1,2,\dots$$

Now it can be shown that a subsequence of $\{v_k\}_{k=1}^{\infty}$ converges in $H^{t-1}(\Omega)$ (see [1] page 32). We can therefore assume that $\{v_k\}_{k=1}^{\infty}$ itself converges, thus $\{v_k\}_{k=1}^{\infty}$ is a Cauchy sequence in $H^{t-1}(\Omega)$. This together with the fact that $|v_k|_t\to 0$ and

$$\begin{aligned}\|v_k-v_\ell\|_t^2 &= \|v_k-v_\ell\|_{t-1}^2+|v_k-v_\ell|_t^2 \\ &\le \|v_k-v_\ell\|_{t-1}^2+(|v_k|_t+|v_\ell|_t)^2,\end{aligned}$$

we get that $\{v_k\}_{k=1}^{\infty}$ is a Cauchy sequence in $H^t(\Omega)$, and thus converges in $H^t(\Omega)$. Thus by the continuity, we get

$$\|v^*\|_t=1 \quad \text{and } |||v^*|||=0. \tag{4.2}$$

If $|v^*|_t=0$ then v^* is a polynomial of degree $t-1$, and from $v^*(z_i)=0$ for $i=1,2,\dots,s$, we get $v^*=0$, which is a contradiction as $\|0\|=0\neq 1$, and proves the equivalence of the norms. □

Using Lemma 4.3, we can now define and prove the Bramble-Hilbert Lemma.

Lemma 4.4. Bramble-Hilbert Lemma
Let Ω be a domain with a Lipschitz continuous boundary. Suppose $t \geq 2$ and that L is a bounded linear mapping from $H^t(\Omega)$ into a normed linear space Y. If $\mathcal{P}_{t-1} \subset \ker L$, then there exists a constant c such that

$$\|Lv\| \leq c|v|_t \quad \forall v \in H^t(\Omega). \tag{4.3}$$

Proof. Let $I : H^t(\Omega) \to \mathcal{P}_{t-1}$ be an interpolation operator of the same type as in Lemma 4.3. Then by the same lemma and the fact that $Iv \in \ker L$, we get

$$\|Lv\| = \|L(v - Iv)\| \leq \|L\| \cdot \|v - Iv\|_t \leq c\|L\| \cdot |v|_t, \tag{4.4}$$

where the constant c is the same as in Equation 4.1. □

It is sometimes easier to look at a transformed domain, rather than the original, which leads to the introduction of the Transformation Formula.

Theorem 4.5. Transformation Formula
Let Ω and $\tilde{\Omega}$ be affine equivalents; meaning there exists some square, nonsingular matrix B, such that

$$\begin{aligned} &F : \tilde{\Omega} \to \Omega, \\ &F(\tilde{x}) = x_0 + B\tilde{x}, \end{aligned}$$

which is affine and bijective. For $v \in H^m(\Omega)$ define $\tilde{v}(\tilde{x}) = v(F(\tilde{x}))$. Then there exists some c such that

$$|\tilde{v}|_{m,\tilde{\Omega}} \leq c\frac{\|B\|^m}{\sqrt{|\det(B)|}}|v|_{m,\Omega}.$$

Here $\|B\|$ is the operator norm, and in our case can be thought of as

$$\|B\| = \sup\{\|By\| : \|y\| = 1, y \in \mathbb{R}^n\},$$

as $\tilde{\Omega} \subset \mathbb{R}^n$.

Proof. Writing the mth derivative as a multilinear form, we get

$$(D^m\tilde{v}(\tilde{x}))(\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m) = (D^m v(x))(B\tilde{y}_1, B\tilde{y}_2, \ldots, B\tilde{y}_m).$$

We can extract B, and get $\|D^m\tilde{v}\|_{\mathbb{R}^{nm}} \leq \|B\|^m\|D^m v\|_{\mathbb{R}^{nm}}$. As

$$\partial_{i_1}\partial_{i_2}\ldots\partial_{i_m}v = D^m v(e_{i_1}, e_{i_2}, \ldots, e_{i_m})$$

is a nm dimensional vector, we get the following;

$$\begin{aligned}\sum_{|\alpha|=m} |\partial^\alpha \tilde{v}|^2 &\leq n^m \max_{|\alpha|=m} |\partial^\alpha \tilde{v}|^2 \\ &\leq n^m \|D^m \tilde{v}\|^2 \\ &\leq n^m \|B\|^{2m} \|D^m v\|^2 \\ &\leq n^{2m} \|B\|^{2m} \sum_{|\alpha|=m} |\partial^\alpha v|^2.\end{aligned}$$

We now wish to integrate; however, under affine linear transformation in integrals, we must include the transformation constant, which results in

$$\int_{\tilde{\Omega}} \sum_{|\alpha|} |\partial^\alpha \tilde{v}|^2 d\tilde{x} \leq n^{2m} \|B\|^{2m} \int_{\Omega} \sum_{|\alpha|=m} |\partial^\alpha v|^2 |\det B^{-1}| dx.$$

Which, after taking the square root, gives the theorem. □

When using a transformation from one shape-regular grid to another, we do not get extra terms which can be seen by looking at the geometric interpretation. Let $F : T_1 \to T_2$, defined as $\hat{x} \mapsto B\hat{x} + x_0$ be a bijective affine mapping. We define ρ_i and r_i to be the radius of the largest inscribed circle and the radius of the smallest circle containing T_i respectively. Given $x \in \mathbb{R}^2$, with $||x|| \leq 2\rho_1$, find y_1, z_1 such that $x = y_1 - z_1$, where y_1 and z_1 are points in T_1. Since $F(y_1), F(z_1) \in T_2$, we have that $||Bx|| \leq 2r_2$, thus

$$||B|| \leq \frac{r_2}{\rho_1}. \tag{4.5}$$

Theorem 4.6.
Let $t \geq 2$, and suppose $\mathcal{T}_h$ is a shape-regular triangulation of Ω. Then there exists a constant c such that

$$\|u - I_h u\|_{m,h} \leq c h^{t-m} |u|_{t,H^t(\Omega)}, \quad \forall u \in H^t(\Omega), \quad 0 \leq m \leq t, \tag{4.6}$$

where $I_h u$ denotes the interpolation by a piecewise polynomial of degreee $t-1$.

Proof. For this proof, we use the r_i as the radius of the smallest circle containing $T_i \in \mathcal{T}_h$, and ρ_i as the radius of the largest inscribed circle in T_i. We start by proving the following inequality;

$$||u - I_h u||_{m,T_j} \leq c h^{t-m} |u|_{t,T_j}, \quad \forall u \in H^t(T_j) \tag{4.7}$$

for every triangle T_j of a shape-regular triangulation $\mathcal{T}_h$. By choosing a reference triangle with $\hat{r} = 2^{-1/2}$, $\hat{\rho} = (2 + \sqrt{2})^{-1}$ and letting $F : \hat{T} \to T$ with $T = T_j \in \mathcal{T}_h$, we can apply

Lemma 4.3 on the reference triangle and using the transformations formula, Theorem 4.5, in both directions, we get

$$\begin{aligned} ||u - I_h u||_{m,T_j} &\leq c||B^{-1}||^m|\det B|^{1/2}|\hat{u} - I_h\hat{u}|_{m,T_{\text{ref}}} \\ &\leq c||B^{-1}||^m|\det B|^{1/2}c|\hat{u}|_{t,T_{\text{ref}}} \\ &\leq c||B^{-1}||^m|\det B|^{1/2} \cdot c||B||^t \cdot |\det B|^{-1/2}|u|_{t,T} \\ &\leq c\big(||B|| \cdot ||B^{-1}||\big)^m ||B||^{t-m}|u|_{t,T}. \end{aligned} \tag{4.8}$$

Furthermore, by shape regularity, $r_i/\rho_i \leq \kappa$ for all T_i in $\mathcal{T}$, and $||B||\cdot||B^{-1}|| \leq \left(2+\sqrt{2}\right)\kappa$. From Equation (4.5) we get $||B|| \leq h/\hat{\rho} \leq 4h$. With all this we have

$$||u - I_h u||_{\ell,T_j} \leq ch^{t-\ell}|u|_{t,T} \tag{4.9}$$

Lastly by squaring and summing the terms from ℓ to m we get the desired result. □

Theorem 4.7. Inverse Estimates
Let S_h be an affine family of finite elements, consisting of piecewise polynomials of degreee k associated with uniform partitions. Then there exists a constant c such that for all $0 \leq m \leq t$,

$$\|v_h\|_{t,h} \leq ch^{m-t}\|v_h\|_{m,h} \quad \forall v \in \Pi_{\text{ref}} \tag{4.10}$$

Proof. By the Transformation Formula 4.5 we can reduce the proof to only consider the reference element. Since all norms on finite dimensional vector spaces are equivalent it suffices to show

$$|v|_{t,T_{\text{ref}}} \leq c|v|_{m,T_{\text{ref}}} \quad \forall v \in \Pi_{\text{ref}} \tag{4.11}$$

for some constant c. To obtain the inequality we consider the vector space $\Pi_{\text{ref}} \oplus \mathcal{P}_{m-1}$. Let $Iv \in \mathcal{P}_{m-1}$ be a polynomial that interpolates v at fixed points. $|Iv|_t = 0$, since $t > m-1$ and by Bramble-Hilbert, 4.4, we get

$$\begin{aligned} |v|_t = |v - Iv|_t &\leq ||v - Iv||_t \\ &\leq c||v - Iv||_m \\ &\leq c'|v|_m \end{aligned} \tag{4.12}$$

where (4.12) follows from the Bramble-Hilbert lemma with $L = 1 - I$. To extend to elements of size h, the same procedure as 4.6 should be applied. This results in the factor ch^{m-t} in the estimate. Lastly by summing the square of the expressions over all triangles or quadrilaterials leads to the desired result. □

If a domain is regular enough it results in more requirements for the solution, which we will introduce here. To introduce the Regularity Theorem, we must first define when a domain can be classified as regular

Definition 4.8.
Let $m \geq 1, H_0^m(\Omega) \subset V \subset H^m(\Omega)$, and suppose $a(\cdot,\cdot)$ is a V-elliptic bilinear form. Then the variational problem

$$a(u,v) = (f,v)_0 \quad \text{for all } v \in V \tag{4.13}$$

is called H^s-regular provided that there exists a constant C such that for every $f \in H^{s-2m}(\Omega)$, there is a solution $u \in H^s(\Omega)$ satisfying

$$\|u\|_s \leq c\|f\|_{s-2m}. \tag{4.14}$$

We now introduce the theorem which uses regularity, without proof. Further information can be found in [1].

Theorem 4.9. Regularity Theorem
Let $a(\cdot,\cdot)$ be a H_0^1-elliptic bilinear form with sufficiently smooth coefficient functions. Then the following holds;

1. if Ω is convex the Dirichlet problem is H^2-regular,
2. if Ω has a C^s boundary with $s \geq 2$ the Dirichlet problem is H^s-regular.

From this point on we will assume Ω to be convex and polygonal to ensure that a triangulation of Ω is possible. In extension to this we can now look at the error of the finite element approximation. To do this, let $\mathcal{T}_h$ be a triangulation of Ω, and

$$\begin{aligned}\mathcal{M}_0^k =& \{v \in L_2(\Omega) \,|\, v|_T \in \mathcal{P}_k \text{ for every } T \in \mathcal{T}_h\} \cap C^0(\Omega)\\ =& \{v \in L_2(\Omega) \,|\, v|_T \in \mathcal{P}_k \text{ for every } T \in \mathcal{T}_h\} \cap H^1(\Omega).\end{aligned}$$

Theorem 4.10.
Suppose $\mathcal{T}_h$ is a family of shape-regular triangulation of Ω. Then the finite element approximation $u_h \in S_h = \mathcal{M}_0^k$, $k \geq 1$ satisfies

$$\begin{aligned}\|u - u_h\|_1 &\leq ch\|u\|_2\\ &\leq ch\|f\|_0.\end{aligned} \tag{4.15}$$

Proof. Since Ω is convex the problem is H^2-regular and that $\|u\|_2 \leq c_1\|f\|_0$. From Theorem 4.6 some $v_h \in S_h$ exists, which satisfies

$$\|u - v_h\|_{1,\Omega} = \|u - v_h\|_{1,h} \leq c_2 h\|u\|_{2,\Omega}. \tag{4.16}$$

Combining the prior statements with Céa's Lemma 3.2, and define $c = (1 + c_1)c_2C/\alpha$ to obtain Equation 4.15. □

Previously we introduced different bounds to specific problems and we now expand on these to obtain a stronger error bounds for the finite element approximation. We do this by introducing the Aubin–Nitsche Theorem.

> **Theorem 4.11.** Aubin–Nitsche Theorem
> Let H be a Hilbert space with norm $|\cdot|$ and a scalar product $(\cdot,\cdot)$. Let $V \subset H$ be a Hilbert space for another norm $\|\cdot\|$, let the imbedding $V \hookrightarrow H$ be continuous, and $\forall g \in H$, let $\varphi_g \in V$ denote the unique weak solution to
>
> $$a(w, \varphi_g) = (g, w) \quad \forall w \in V. \tag{4.17}$$
>
> Here $a(\cdot,\cdot)$ is a bilinear continuous form. Then the finite element solution $u_h \in S_h \subset V$ obeys
>
> $$|u - u_h| \leq C\|u - u_h\| \sup_{g \in H} \left\{ \frac{1}{|g|} \inf_{v \in S_h} \|\varphi_g - v\| \right\},$$
>
> where sup is over all $g \in H$ such that $|g| \neq 0$.

Proof. Assume $|g| \neq 0$ for some $g \in H$. Then, by Cauchy–Schwarz, we can write

$$(g, w) \leq |g| \cdot |w| \implies \frac{(g, w)}{|g|} \leq |w|.$$

By using $g = w$ we get

$$\frac{(w, w)}{|w|} = |w| \leq |w|.$$

By taking supremum, we get

$$\sup_{g \in H} \frac{(g, w)}{|g|} = |w|. \tag{4.18}$$

The solution u and finite element solution u_h is given by

$$\begin{aligned} a(u, v) &= f(v) \quad \forall v \in V, \\ a(u_h, v) &= f(v) \quad \forall v \in S_h. \end{aligned}$$

From this we get $a(u - u_h, v) = 0$, $\forall v \in S_h$, which we use with the assumption in Equation (4.17), and the continuity of $a(\cdot,\cdot)$ to get

$$\begin{aligned}(g, u-u_h) &= a(u-u_h, \varphi_g)\\ &= a(u-u_h, \varphi_g) - 0\\ &= a(u-u_h, \varphi_g) - a(u-u_h, v)\\ &= a(u-u_h, \varphi_g - g)\\ &\leq C\|u-u_h\| \cdot \|\varphi_g - v\|.\end{aligned}$$

Using $v \in S_h$ which gives the smallest norm, we get

$$(g, u-u_h) \leq C\|u-u_h\| \cdot \inf_{v \in S_h} \|\varphi_g - v\|.$$

Using Equation 4.18 on $u - u_h$, we then get

$$\begin{aligned}|u-u_h| &= \sup_{g \in H} \frac{(g, u-u_h)}{|g|}\\ &\leq C\|u-u_h\| \sup_{g \in H} \left\{ \frac{1}{|g|} \inf_{v \in S_h} \|\varphi_g - v\| \right\}.\end{aligned}$$

□

We use the previous theorem to prove an inequality more relevant to our situation.

Corollary 4.12.
Assume $\mathcal{T}_h$ is a family of shape regular triangulations of Ω. If $u \in H^1(\Omega)$ is the solution of the variational problem, then

$$\|u-u_h\|_0 \leq cCh\|u-u_h\|_1.$$

If $f \in L_2(\Omega)$ and $u \in H^2(\Omega)$, then

$$\|u-u_h\|_0 \leq cC^2h^2\|f\|_0.$$

Proof. We want to use Theorem 4.11, which we can by the following observations. Set

$$H = H^0(\Omega) = L_2(\Omega) \quad \text{and} \quad V = H_0^1(\Omega).$$

Then $V \subset H$, and the norms

$$|\cdot| = \|\cdot\|_0 \quad \text{and} \quad \|\cdot\| = \|\cdot\|_1,$$

by $\|\cdot\|_0 \leq \|\cdot\|_1$ gives us continuity of the imbedding. We then get

$$\|u-u_h\|_0 \leq C\|u-u_h\|_1 \sup_{g \in H} \left\{ \frac{1}{|g|} \inf_{v \in S_h} \|\varphi_g - v\|_1 \right\},$$

By shape regularity, we can use Theorem 4.10 and evaluate

$$\sup_{g \in H} \left\{ \frac{1}{|g|} \inf_{v \in S_h} \|\varphi_g - v\|_1 \right\} \leq ch,$$

and Theorem 4.11 implies the result. □

If the variational problem is sufficiently regular, Corollary 4.12 gives us a connection between f in our original problem, and how big the error of our FEM approximation can be. However, the norms used to calculate this error does not describe the error at singular points; there still might be points where the error grows uncontrollably large, if we just use Corollary 4.12. To remedy that we introduce the next theorem.

> **Theorem 4.13.**
> Let the L_∞-norm be
>
> $$\|v\|_{\infty,\Omega} = \operatorname*{ess\,sup}_{x \in \Omega} |v(x)|.$$
>
> Then for a H^2-regular variational problem with a solution u, the following holds;
>
> $$\|u - u_h\|_{\infty,\Omega} \leq ch|u|_2. \tag{4.19}$$

Proof. For a function $v \in H^2(T_{\text{ref}})$, we let Iv be its interpolant in the polynomial space Π_{ref}. We have that $H^2 \subset C^0$ and therefore

$$\|v - Iv\|_{\infty,T_{\text{ref}}} \leq c|v|_{2,T_{\text{ref}}} \tag{4.20}$$

by Lemma 4.4. Now let u be the solution to the variational problem, and $I_h u$ be the interpolant in S_h. Then we choose an element T in the triangulation, which we without loss of generality assume to be uniform. Let $\hat{u}$ be the affine transformation of $u|_T$ to the reference triangle. Then by Equation 4.20 and the transformation formula, 4.5;

$$\begin{aligned} \|u - I_h u\|_{\infty,T} &= \|\hat{u} - I\hat{u}\|_{\infty,T_{\text{ref}}} \\ &\leq c|\hat{u}|_{2,T_{\text{ref}}} \\ &\leq ch|u|_{2,T} \\ &\leq ch|u|_{2,\Omega}. \end{aligned} \tag{4.21}$$

From taking the maximum over all triangles, we get

$$\|u - I_h u\|_{\infty,\Omega} \leq ch|u|_{2,\Omega}. \tag{4.22}$$

Then by the transformation formula and Theorem 4.7, we can obtain the inverse estimate as

$$\|v_h\|_{\infty,\Omega} \leq ch^{-1}\|v_h\|_{0,\Omega} \quad \text{for all } v_h \in S_h. \tag{4.23}$$

Then using Theorem 4.6 and Corollary 4.12 for $u_h - I_h u \in S_h$ we get

$$\begin{aligned}\|u_h - I_h u\|_{0,\Omega} &= \|(u - I_h u) - (u - u_h)\|_{0,\Omega} \\ &\leq \|u - I_h u\|_{0,\Omega} + \|u - u_h\|_{0,\Omega} \\ &\leq ch^2 |u|_{2,\Omega}.\end{aligned}$$

We then use the inverse estimate from earlier. Thus

$$\begin{aligned}\|u - u_h\|_{\infty,\Omega} &\leq \|u - I_h u\|_{\infty,\Omega} + \|u_h - I_h u\|_{\infty,\Omega} && (4.24) \\ &\leq \|u - I_h u\|_{\infty,\Omega} + ch^{-1}\|u_h - I_h u\|_{0,\Omega} && (4.25) \\ &\leq ch|u|_{2,\Omega} + ch^{-1}ch^2|u|_{2,\Omega} && (4.26) \\ &\leq \tilde{c}h|u|_{2,\Omega}, && (4.27)\end{aligned}$$

which is the theorem. □

Theorem 4.13 gives us, as the observant reader might have guessed, a maximum size of the error of our FEM approximation. If the variational problem is sufficiently regular (meaning H^2) the error should become smaller as h becomes smaller. We have in this chapter focused on the possible error of our FEM approximation, and how big it can be. When the variational problem is regular, we have found both an upper maximum in general for the error, but we have also shown other error estimates, if the problem is not that easy to work with. We now proceed to the next chapter, and examine an implimentation of FEM.

5 | Application and numerical examination

We will in this chapter apply the theory developed throughout the project to a specific example, which will be a modified version of some of the examples from [3]. Here we will consider a problem with Dirichlet boundary conditions. We wish to study convergence rate of the approximate solution, both in context of the size of cells and in degree used in the finite elements. The domain we will be working with is $\Omega = \{(x, y) \in \mathbb{R}^2 \mid -1 \leq x, y \leq 1\}$, which is the square in $\mathbb{R}^2$ with sidelength 2, and center in origo. We then define the solution

$$k(x, y) = \mathrm{e}^{x+y} \cos(x) \sin(y) + x \quad \text{on } \Omega. \tag{5.1}$$

We work backwards from the solution to be able to compare the numerical approximation with the real function. Using this k, we define the Dirichlet boundary problem to be

$$\begin{aligned} Lu &= -\mathrm{div}\ \nabla k && \text{on } \Omega \\ u &= k && \text{on } \partial\Omega, \end{aligned} \tag{5.2}$$

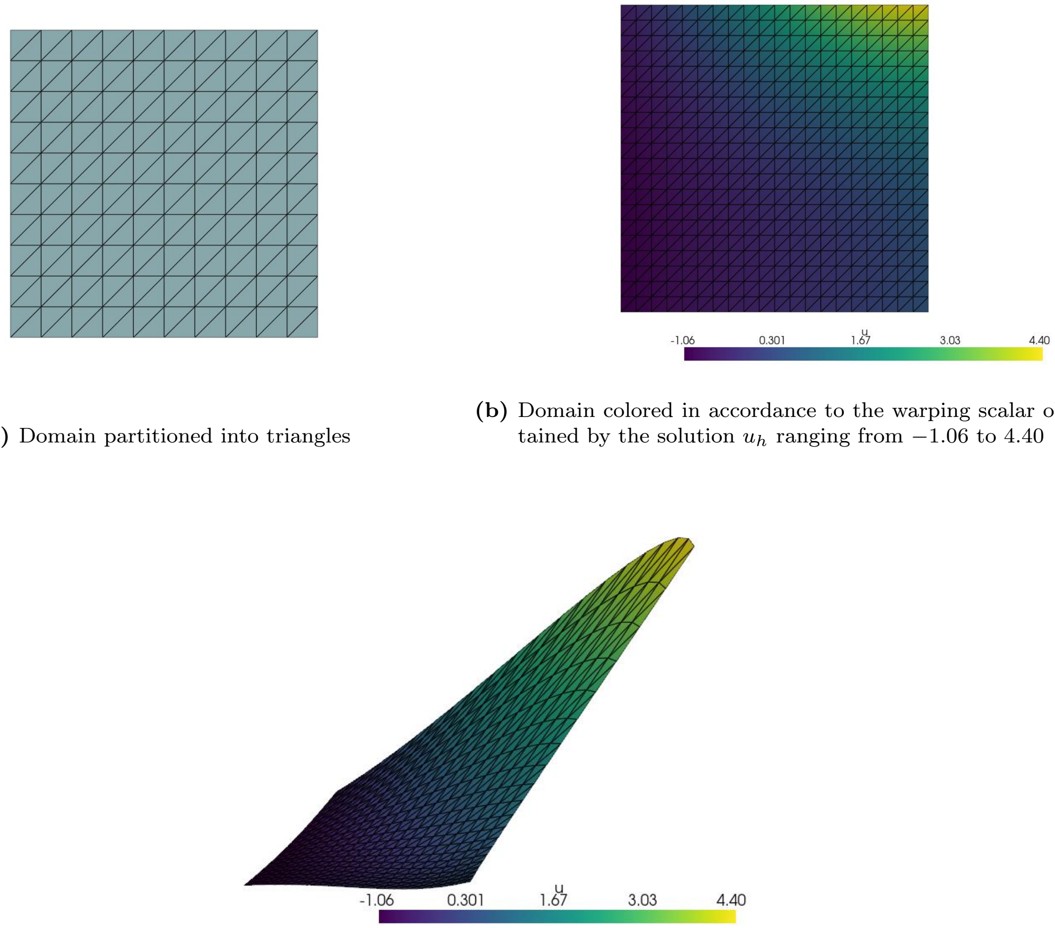
Note in this example that, for simplicity, we have chosen the solution k to be the same on the boundary as in the domain. We will use the finite element method to approximate a solution for the problem. In this context we use the framework FEniCSx, which is a computational tool for solving PDE's. FEniCSx consists of several different parts, which can be seen in [4], [5], and [6].

All references in the following section will be refering to the code listed in Appendix A.

There are three big parts, contained in section A.1. The other 2 code files, found in section A.2, are for the plots found in this chapter. The first part in A.1 is a solver, which solves a FEM-problem, using a certain number of cells, and a certain degree of polynomials. The second is two functions which calculates the error in L_2 and H^1, which comes directly from [3]. The last part is two functions, which loop over various numbers of cells and polynomial degrees, and outputs a table.

We start on line 11 by defining our solution u as the function from Equation (5.1), after which we define two different interpolations of u using the packages "numpy" and "ufl". We do this for convience, as numpy are better for numerical apporiximations, and ufl is easier to work with in FEniCSx. Next, we need to define the mesh and the function space we will be working in. We do this on line 20-23, where N defines the number of cells in each directions of our mesh, resulting in N^2 elements in total. Then 'degree' defines the degree of polynomials we use in each cell. Next step in the process is to define the space of trial functions and the space of test functions, which we do on line 26-27. In a different situation, these two spaces could have been different.

We then define the boundary condition, which is simply the function k, as mentioned earlier. With the boundary condition established we can move on to defining the bilinear form a and the linear form L. With all this in place, we can now solve the problem using the PETSc interface from FEniCSx.



(a) Domain partitioned into triangles

(b) Domain colored in accordance to the warping scalar obtained by the solution u_h ranging from -1.06 to 4.40

(c) The same domain warped with respect to the scalar value, ranging from -1.06 to 4.40

Figure 5.1: Plots of the domain

As discussed earlier the first step in approximating the solution to a PDE is to establish a fitting domain. In our example we have chosen a square ranging from $(-1, -1)$ to $(1, 1)$. We have chosen to partition into triangles, as can be seen in Figure 5.1a. These triangles are of equal size making the domain regular. After the domain is partitioned and we have estimated the solution, we can plot the solution over the domain to obtain Figure 5.1b where each cell is coloured in accordance to the value in the given point. We can then warp the domain to give a clear look on what exactly our solution looks like in 3D, which is shown in Figure 5.1c.

Examining the error can show us the convergence rate, which tells us how fast the error decreases as we increase the number of elements in our mesh and degree of polynomials. We will show this imperically by computing the error for different mesh sizes and degrees and comparing them. In Table 5.2, the error for five different mesh sizes are shown. For each different mesh, the solution have been approximated using polynomials of degree one through ten. As can be seen in Table 5.2, every time the degree of polynomials used in approximations are increased by 1, the error changes with a factor 10^{-1} or factor 10^{-2}, while a similar trend can be seen every time the cell size is halved.

However, when the error is very small, an increase in degree or decrease in cell size results in an increase in error; this can for example be seen at $h = 0.0625$ from degree 9 to 10, and at $h = 0.0156$ from degree 5 to 6. Also, for degree 6, the error also increases as the cell size is halved from 0.0312 to 0.0156. This could be explained both by computational accuracy for numbers of this size, meaning rounding error and truncation, and the ratio between the radii of the inner and outer circles might increase, which could also be a factor here. There can be some question as to whether using polynomials of degree 10 and 64^2 cells over a 2 by 2 square is computationally practical in other settings. This is where an evaluation of physical properties, like observational accuracy, desired precision and other factors would come into play. There also exists algorithms for refining a coarse mesh. It involves finding hanging nodes, triangle points which dissects other triangle edges, and splitting the triangles which are dissected. However, making cells distorted can cause the error estimates to diverge, as the ratio between the radii of the inner and outer circles might

Degree	Size of h				
	0.25	0.125	0.0625	0.0312	0.0156
1	0.159	0.0418	0.0106	0.00266	0.000666
2	0.0114	0.00145	0.000181	2.26e-05	2.83e-06
3	0.00103	6.26e-05	3.79e-06	2.31e-07	1.43e-08
4	6.4e-05	2.02e-06	6.3e-08	1.97e-09	6.14e-11
5	2.24e-06	3.51e-08	5.48e-10	8.55e-12	5.63e-13
6	9.71e-08	8.23e-10	6.57e-12	8.43e-13	3.39e-12
7	6.09e-09	2.45e-11	4.88e-13	1.84e-12	7.35e-12
8	2.31e-10	4.62e-13	2.83e-13	1.01e-12	3.76e-12
9	5.1e-12	1.83e-13	7.46e-13	3.04e-12	1.22e-11
10	2.01e-13	5.18e-13	2.02e-12	7.97e-12	3.17e-11

Figure 5.2: Table of convergence rate of errors using the L_2-norm

increase. To get a better overview of the error, we have made a graph of the error for different degrees of the polynomial and different sizes of h. This makes it more managable to see how the the different degrees and mesh sizes affect the error. We start by discussing the L_2 error. We see on Figure 5.4 that in general the error decreases as the degree of the polynomial increases and as the size of h decreases, so does the norm, although this is not the case for all mesh sizes. We see that the error decreases as the mesh size decreases up to polynomials of degree 8, which can be seen more accurately in Table 5.2, but this changes for polynomials of degree 8, 9, and 10. This should not be the case, but could be due to the size of the error, which is so small the rounding errors could become more significant. The H^1 error shown on Figure 5.5 generally has the same tendency as the L_2 error, but as mentioned earlier the error is generally higher. However, the order of polynomial, where the tendency changes is different for the H^1 error, as it changes at degree 9 instead of 8, see Table 5.3. Furthermore, after the change in tendency, the error does not increase as drastically as for the L_2 error.

Degree	Size of h				
	0.25	0.125	0.0625	0.0312	0.0156
1	1.2	0.61	0.307	0.153	0.0768
2	0.162	0.0428	0.0109	0.00272	0.000682
3	0.0201	0.00255	0.000316	3.93e-05	4.89e-06
4	0.0015	9.42e-05	5.88e-06	3.67e-07	2.29e-08
5	6.32e-05	1.99e-06	6.23e-08	1.94e-09	6.07e-11
6	3.2e-06	5.38e-08	8.56e-10	1.36e-11	9.13e-12
7	2.27e-07	1.83e-09	1.44e-11	3.29e-12	1.26e-11
8	9.66e-09	3.8e-11	1.93e-12	7.28e-12	2.87e-11
9	2.37e-10	9.49e-13	2.38e-12	7.94e-12	2.96e-11
10	7.5e-12	6.09e-13	1.35e-12	3.34e-12	9.48e-12

Figure 5.3: Table of convergence rate of errors using the H^1-norm

In general the numerical experimentation agrees with the theory, though for high degree of polynomials we get conflicting information. We are unsure as to the exact reason, as changing the way the errors are calculated, both using the H^1-norm and L_2 norm, and increasing the degree used to calculate errors in the practical function did not change this outcome.

In the same sense, decreasing the degree used in calculating error did not result in greater error, and the same trends for polynomials of degree 4 through 10. For degree 1 through 3, a smaller error was seen for small h, suggesting the way the errors have been calculated to have been misused.

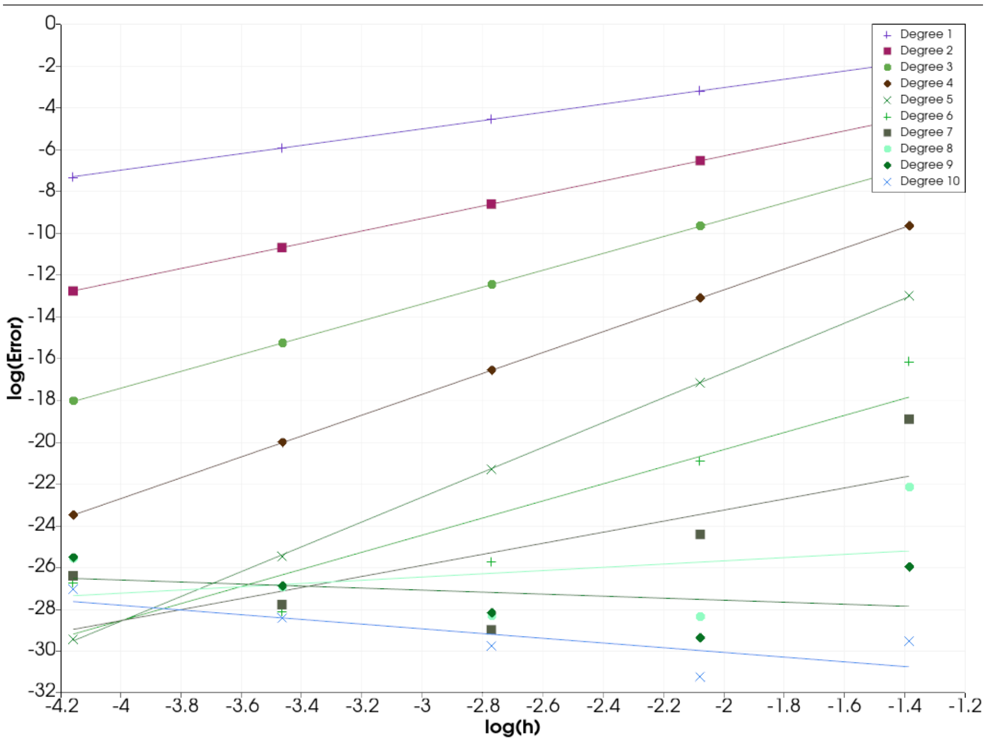


Figure 5.4: Plot of the L_2 error for the different degrees of the polynomial and different sizes of h.

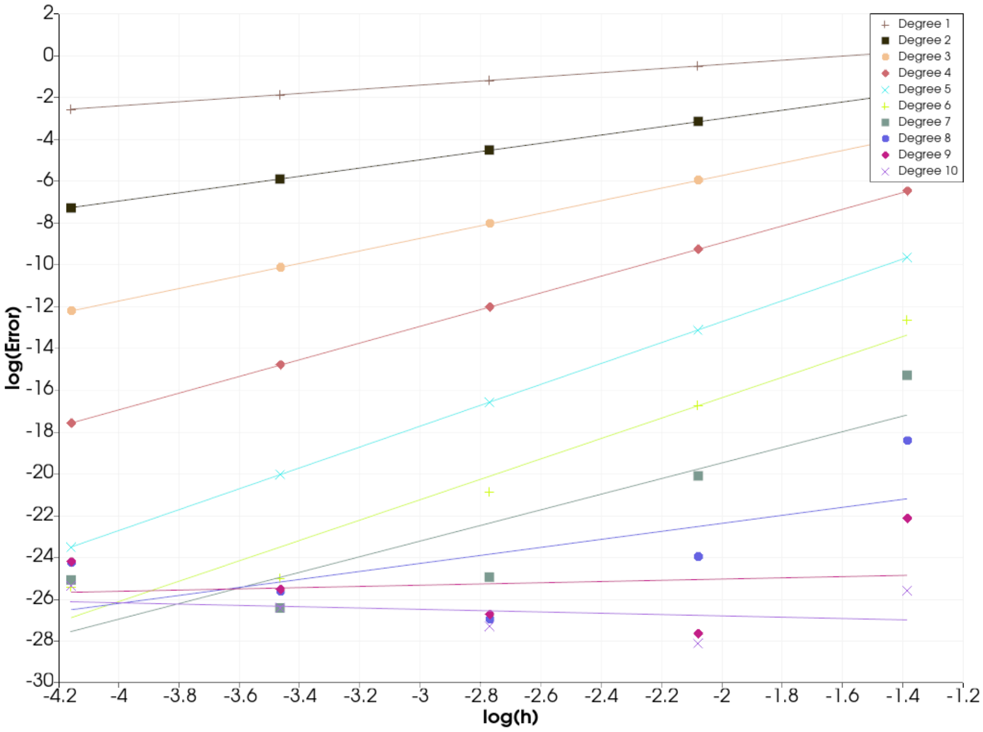


Figure 5.5: Plot of the H^1 error for the different degrees of the polynomial and different sizes of h.

6 | Conclusion

We have examined how the space containing the solution to a PDE with boundary conditions can be constructed, and how a solution can be approximated using piecewise polynomials, and by partitioning the domain into smaller pieces. We have focused mostly on partitioning using triangles, though rectangles could also have been chosen. After partitioning the domain we are able to use the Finite Element Method to approximate numerical solutions, as we've done in the application section of this project, to varying degrees of success, with some behaviour in the convergence rate which were not explained.

7 | Bibliography

[1] Braess D. Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics. 3rd ed. Cambridge University Press; 2007.

[2] Brezis H. Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer New York, NY; 2010.

[3] The FEniCSx tutorial [Interactive Jyputer-notebook]. Jørgen S. Dokken; 2024 [cited 13-May-2024]. Available at: `https://jsdokken.com/dolfinx-tutorial/index.html`.

[4] Baratta IA, Dean JP, Dokken JS, Habera M, Hale JS, Richardson CN, et al.. DOLFINx: the next generation FEniCS problem solving environment; 2023. preprint.

[5] Scroggs MW, Dokken JS, Richardson CN, Wells GN. Construction of arbitrary order finite element degree-of-freedom maps on polygonal and polyhedral cell meshes. ACM Transactions on Mathematical Software. 2022;48(2):18:1–18:23.

[6] Scroggs MW, Baratta IA, Richardson CN, Wells GN. Basix: a runtime finite element basis evaluation library. Journal of Open Source Software. 2022;7(73):3982.

A Code

A.1 FEM solver

```
import ufl
import dolfinx
import numpy as np
from dolfinx import mesh, fem, plot, default_scalar_type
from mpi4py import MPI
import pyvista as pv
import dolfinx.fem.petsc
from tabulate import tabulate


def u_ex(mod):
    return lambda x: mod.exp(x[0] + x[1]) * mod.cos(x[0]) * mod.sin(x[1]) + x[0]

u_ufl = u_ex(ufl)
u_numpy = u_ex(np)


def solver(N=10, degree=2):
    ## Define domain and functionspace over it
    domain = mesh.create_rectangle(
        comm=MPI.COMM_WORLD, points=((-1, -1), (1, 1)), n=(N, N)
    )
    V = fem.functionspace(domain, ("Lagrange", degree))

    ## Define trial and test functions
    u = ufl.TrialFunction(V)
    v = ufl.TestFunction(V)

    ## Source term of the poisson equation
    x = ufl.SpatialCoordinate(domain)
    f = -ufl.div(ufl.grad(u_ufl(x)))

    ## Applying boundary conditions
    uD = fem.Function(V)
    uD.interpolate(lambda x: np.exp(x[0] + x[1]) * np.cos(x[0]) * np.sin(x[1]) +
        x[0])
```

```python
    tdim = domain.topology.dim
    fdim = tdim - 1
    domain.topology.create_connectivity(fdim, tdim)
    boundary_facets = mesh.exterior_facet_indices(domain.topology)
    boundary_dofs = fem.locate_dofs_topological(V, fdim, boundary_facets)
    bc = fem.dirichletbc(uD, boundary_dofs)

    # bilinear form
    a = ufl.dot(ufl.grad(u), ufl.grad(v)) * ufl.dx
    L = f * v * ufl.dx

    # set PETSc solver options
    sol_opts = {"ksp_type": "preonly", "pc_type": "lu"}
    # formulate the problem
    problem = dolfinx.fem.petsc.LinearProblem(a, L, bcs=[bc],
        petsc_options=sol_opts)
    return problem.solve(), u_ufl(x), domain, V, tdim


def error_L2(uh, u_ex, degree_raise=3):
    degree = uh.function_space.ufl_element().degree()
    family = uh.function_space.ufl_element().family()
    mesh = uh.function_space.mesh
    W = fem.FunctionSpace(mesh, (family, degree + degree_raise))
    u_W = fem.Function(W)
    u_W.interpolate(uh)
    u_ex_W = fem.Function(W)
    if isinstance(u_ex, ufl.core.expr.Expr):
        u_expr = fem.Expression(u_ex, W.element.interpolation_points)
        u_ex_W.interpolate(u_expr)
    else:
        u_ex_W.interpolate(u_ex)
    e_W = fem.Function(W)
    e_W.x.array[:] = u_W.x.array - u_ex_W.x.array
    error = fem.form(ufl.inner(e_W, e_W) * ufl.dx)
    error_local = fem.assemble_scalar(error)
    error_global = mesh.comm.allreduce(error_local, op=MPI.SUM)
    return np.sqrt(error_global)


def error_H10(uh, u_ex, degree_raise=3):
    degree = uh.function_space.ufl_element().degree()
    family = uh.function_space.ufl_element().family()
    mesh = uh.function_space.mesh
    W = fem.FunctionSpace(mesh, (family, degree + degree_raise))
    u_W = fem.Function(W)
    u_W.interpolate(uh)
    u_ex_W = fem.Function(W)
    if isinstance(u_ex, ufl.core.expr.Expr):
        u_expr = fem.Expression(u_ex, W.element.interpolation_points)
        u_ex_W.interpolate(u_expr)
    else:
        u_ex_W.interpolate(u_ex)
    e_W = fem.Function(W)
    e_W.x.array[:] = u_W.x.array - u_ex_W.x.array
    error_H10 = fem.form(
        ufl.dot(ufl.grad(e_W), ufl.grad(e_W)) * ufl.dx + ufl.dot(e_W, e_W) *
            ufl.dx
    )
    E_H10 = np.sqrt(mesh.comm.allreduce(fem.assemble_scalar(error_H10), MPI.SUM))
    return E_H10


def convergence_rate(error, Ns=[4, 8, 16, 32, 64], deg=1):
    result = []
    Es = np.zeros(len(Ns), dtype=default_scalar_type)
    hs = np.zeros(len(Ns), dtype=np.float64)
    for i, N in enumerate(Ns):
        uh, u_ex, domain, V, tdim = solver(N, deg)
        comm = uh.function_space.mesh.comm
        Es[i] = f"{error(uh, u_numpy, degree_raise=0):.2e}"
        hs[i] = f"{(1. / Ns[i]):.2e}"
    return [list(hs), list(Es)]


def tabulate_convergence_rate(error, start=1, end=11):
    for i in range(start, end):
        if i == start:
            data = convergence_rate(error=error, deg=i)
            data[0].insert(0, " ")
            data[1].insert(0, f"{i}")
        else:
            data.append(convergence_rate(error=error, deg=i)[1])
            data[i].insert(0, f"{i}")
    print(tabulate(data, tablefmt="latex_raw"))


# tabulate_convergence_rate(error_H10)
```

A.2 Code to plots

```

import ufl
import dolfinx
import numpy as np
from dolfinx import mesh, fem, plot, default_scalar_type
from mpi4py import MPI
import pyvista as pv
import dolfinx.fem.petsc
import FEM

uh, u_ex, domain, V, tdim = FEM.solver()


pv.off_screen = True
topology, cell_types, geometry = dolfinx.plot.vtk_mesh(domain, tdim)
grid = pv.UnstructuredGrid(topology, cell_types, geometry)

plotter = pv.Plotter()
plotter.add_mesh(grid,show_edges=True)
plotter.view_xy()
plotter.show(cpos="xy", screenshot=’./screenshot_1.jpeg’)
#plotter.save_graphic(’mesh.svg’)

u_topology, u_cell_types, u_geometry = plot.vtk_mesh(V)
u_grid = pv.UnstructuredGrid(u_topology, u_cell_types, u_geometry)
u_grid.point_data["u"] = uh.x.array.real
u_grid.set_active_scalars("u")
u_plotter = pv.Plotter()
u_plotter.add_mesh(u_grid,show_edges=True)
u_plotter.view_xy()
u_plotter.show(cpos="xy", screenshot=’./screenshot_2.jpeg’)
#u_plotter.save_graphic(’contour.svg’)

warped = u_grid.warp_by_scalar()
plotter2 = pv.Plotter()
plotter2.add_mesh(warped, show_edges=True, show_scalar_bar=True)
plotter2.show(cpos="xy", screenshot=’./screenshot_3.jpeg’)
```

```
#import ufl
#import dolfinx
#from dolfinx import mesh, fem, plot, default_scalar_type
#from mpi4py import MPI
```

```
import pyvista as pv
#import dolfinx.fem.petsc
import FEM
import numpy as np
import random
from matplotlib import pyplot as plt

def tabulate_convergence_rate(start=1, end=11):
    chart = pv.Chart2D()
    # for error in [FEM.error_H10, FEM.error_L2 ]:
    for i in range(start, end):
        h_and_error = FEM.convergence_rate(error=FEM.error_H10, deg=i)
        # chart.scatter(np.log(h_and_error[0]), np.log(h_and_error[1]),
            style="o", color=[random.randrange(1, 255), random.randrange(1, 255),
            random.randrange(1, 255)], label=f"Degree {i}")
        #find line of best fit
        x = np.log(h_and_error[0])
        y = np.log(h_and_error[1])
        a, b = np.polyfit(x, y, 1)

        color = [random.randrange(1, 255), random.randrange(1, 255),
            random.randrange(1, 255)]

        styles = [ "x","+","s","o","d"]

        chart.line([np.log(0.0156), np.log(0.25)], [a*np.log(0.0156)+b,
            a*np.log(0.25)+b], color=color)
        chart.scatter(x, y, style=styles[i % len(styles)], color=color,
            label=f"Degree {i}")

        # Increase tick label size
        chart.x_axis.tick_label_size = 20
        chart.y_axis.tick_label_size = 20

        # Increase axis label size
        chart.x_axis.label_size = 20
        chart.y_axis.label_size = 20

        # Set marker size
        chart.marker_size = 20

        chart.x_label = "log(h)"
        chart.y_label = "log(Error)"

    #chart.save_graphic("h1.pdf")
    #chart.show(screenshot='./H-convergence.jpeg')
    chart.show()

tabulate_convergence_rate()
```