MTH651 BOOTCAMP

This bootcamp and this book of notes is mainly based on [3] for the finite element method and in [2] for the finite difference method.

Date: June 3, 2025.

Summary

1. 1D Finite difference method

1.1. **Introduction.** The finite differences method is a numerical technique for approximating solutions to differential equations, has its roots in the work of key mathematicians such as Newton and Leibniz in the 17th century, who laid the groundwork for calculus. Euler further advanced numerical methods in the 18th century, introducing techniques fundamental to finite differences. In the 19th century, George Boole's "A Treatise on the Calculus of Finite Differences" and Cauchy's contributions to differential equations formalized and expanded the method.

The 20th century saw practical applications and computational advancements, notably by Lewis Fry Richardson in numerical weather prediction and John von Neumann and Stanislaw Ulam's work on fluid dynamics and nuclear reactions during the Manhattan Project. Modern developments in numerical analysis and high-performance computing have cemented finite difference methods as essential tools in fields like computational fluid dynamics and financial modeling.

1.2. **Approximating Derivatives.** We shall start by considering the definition of derivative for a function $u \in \mathcal{C}^1(\Omega)$ given a closed interval $\Omega \subset \mathbb{R}$.

$$\frac{du}{dx}(x) = \lim_{h\to 0^+} \frac{u(x+h)-u(x)}{h} = \lim_{h\to 0^+} \frac{u(x)-u(x-h)}{h} \tag{1}$$

The definition immediately delivers an approximation intuition, this is, for some h sufficiently small but not zero one shall expect

$$\frac{du}{dx}(x) \approx \frac{u(x+h)-u(x)}{h} \approx \frac{u(x)-u(x-h)}{h} \tag{2}$$

We can formalize this idea using Taylor expansions, indeed, if $u \in \mathcal{C}^2(\Omega)$ we can see that

$$u(x+h) = u(x) + hu'(x) + \mathcal{O}(h^2) \tag{3}$$

$$u(x-h) = u(x) - hu'(x) + \mathcal{O}(h^2) \tag{4}$$

We say that this expansions have *truncation error* of order 2 (see Appendix A). Then we can obtain

$$u'(x) = \frac{u(x+h)-u(x)}{h} + \mathcal{O}(h) \tag{5}$$

$$u'(x) = \frac{u(x) - u(x-h)}{h} + \mathcal{O}(h) \tag{6}$$

So both (5) and (6) are approximations for u', we can see that the error of this approximations is proportional to h, for example, in the case of (5) we can see that

$$\left| \frac{u(x+h) - u(x)}{h} - u'(x) \right| = \mathcal{O}(h) \Leftrightarrow \left| \frac{u(x+h) - u(x)}{h} - u'(x) \right| \leq Ch \tag{7}$$

for some $C > 0$ and $0 < h \leq h_0$ sufficiently small. Moreover, we can prove that an appropriate choice for the constant C would be

$$C = \sup_{y \in [x, x+h_0]} \frac{|u''(y)|}{2}$$

Definition 1.1 (Order of approximation)**.** We say that the approximation of the derivative u' at x is of order $p \in \mathbb{N}$ if there exists $C > 0$ independent of h, such that the error between the exact derivative and its approximation is bounded by Ch^p, or, in other words, is $\mathcal{O}(h^p)$.

Let us now consider a better approximation, first consider two Taylor expansions of $u \in \mathcal{C}^3(\Omega)$ of order 3

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \mathcal{O}(h^3) \tag{8}$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) + \mathcal{O}(h^3) \tag{9}$$

if we subtract this expressions we obtain

$$u(x+h) - u(x-h) = 2hu'(x) + \mathcal{O}(h^3) \tag{10}$$

which leads to

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} + \mathcal{O}(h^2) \tag{11}$$

as we saw before, this is an approximation for the first derivative of *order 2*.

Remark. Notice that the order of the approximation that we can potentially obtain is limited by the regularity of u, indeed, if $u \in \mathcal{C}^p(\Omega)$ we only can obtain approximations of order $p-1$ for the first derivative using this method.

Now let's try to build an approximation for the second derivative of u. Let $u \in \mathcal{C}^4(\Omega)$ and consider

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \mathcal{O}(h^4) \tag{12}$$

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \mathcal{O}(h^4) \tag{13}$$

if we sum up both expressions we end up with

$$u(x+h) + u(x-h) = 2u(x) + h^2u''(x) + \mathcal{O}(h^4) \tag{14}$$

which leads to

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + \mathcal{O}(h^2) \tag{15}$$

So we have obtained an approximation for u'' of *order 2*, please notice that this only holds for the strict regularity condition $u \in \mathcal{C}^4(\Omega)$.

1.3. **A finite difference scheme.** Let $\Omega = (0,1)$ and $u : \overline{\Omega} \to \mathbb{R}$ such that it satisfy the boundary value problem

$$\begin{cases} -u''(x) + c(x)u(x) = f(x) & x \in \Omega \\ u(0) = \alpha, \quad u(1) = \beta \end{cases} \tag{16}$$

where $\alpha, \beta \in \mathbb{R}$, $c \geq 0$ and f are defined on $\overline{\Omega}$ with $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. Under this conditions we have existence of solutions.

The first step for deriving a finite difference approximation is to consider a uniform grid $\{x_j\}_{j=0}^{N+1}$ given by $x_j = jh$, here $N \in \mathbb{N}$ and $h = 1/(N+1)$ with $x_0 = 0$ and $x_{N+1} = 1$. We are interested in approximating the values of u **only** in the grid points, this is, we want $u_j \approx u(x_j)$ for all $j \in \{1..., N\}$ ($u_j = u(x_j)$ in the case of $j = 0$ and $j = N+1$). Notice that in this settings we only have N unknowns, we call the vector that stores them $\boldsymbol{u}_h \in \mathbb{R}^N$, this vector is going to be the result of the finite difference method.

Assume now that at least we have $c, f \in \mathcal{C}(\overline{\Omega})$ and recall our second order approximation for the second derivative (15), we replace it in the dirichlet inhomogeneous problem (16) to obtain

$$\begin{cases} -\frac{u_{j+1}-2u_j+u_{j-1}}{h^2} + c(x_j)u_j = f(x_j) \quad j \in \{1, ..., N\} \\ u_0 = \alpha, \;\; u_{N+1} = \beta \end{cases} \tag{17}$$

This problem can be written in matrix form as

$$\boldsymbol{A}_h \boldsymbol{u}_h = \boldsymbol{b}_h \tag{18}$$

where

$$\boldsymbol{A}_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} + \boldsymbol{C}_h, \tag{19}$$

$$\boldsymbol{b}_h = \begin{pmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \frac{\beta}{h^2} \end{pmatrix} \tag{20}$$

and

$$\boldsymbol{C}_h = \begin{pmatrix} c(x_1) & 0 & 0 & \dots & 0 \\ 0 & c(x_2) & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & c(x_{N-1}) & 0 \\ 0 & \dots & 0 & 0 & c(x_N) \end{pmatrix}. \tag{21}$$

One immediate question that arise is whether this system has a solution or not, so we are interested in the invertibility of $\boldsymbol{A}_h$. One can prove that the positive definiteness of $\boldsymbol{A}_h$ holds provided $c(x_i) \geq 0$ for all $i \in \{1, ..., N\}$, indeed, suppose that this condition holds, then if we take an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^N$ we have

$$\boldsymbol{v}^T\boldsymbol{A}_h\boldsymbol{v} = \frac{1}{h^2}\boldsymbol{v}^T\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}\boldsymbol{v} + \boldsymbol{v}^T\boldsymbol{C}_h\boldsymbol{v} \tag{22}$$

and as

$$\boldsymbol{v}^T\boldsymbol{C}_h\boldsymbol{v} = \sum_{i=1}^{N} c(x_i)v_i^2 \geq 0$$

it suffices to show that

$$\xi := \frac{1}{h^2}\boldsymbol{v}^T\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}\boldsymbol{v} \geq 0. \tag{23}$$

but if we notice

$$h^2\xi = v_1^2 + (v_2 - v_1)^2 + \cdots + (v_{N-1} - v_N)^2 + v_N^2 \tag{24}$$

then $\xi \geq 0$ and the result follows. Moreover, if $\xi = 0$ then $v_{i+1} - v_i = v_1 = v_N = 0$, and so $v_i = 0$ for all $i \in \{1, ..., N\}$.

2. 1D Finite element method

The Finite Element Method has its roots in the early structural analysis studies of Galileo and Euler, but it was formally conceptualized in 1943 when Richard Courant proposed piecewise linear approximations for torsion problems. Further, Ray W. Clough, often called the father of FEM, introduced the term and applied the method to structural engineering in the 1950s, marking its formal emergence.

The 1970s and beyond saw further theoretical advancements, with mathematicians such as Ivo Babuška, J.H. Bramble, J. Tinsley Oden, and Wolfgang Nitsche enhancing the mathematical rigor of FEM by addressing issues of convergence, error estimation, and stability. The method was extended to solve multiphysics and nonlinear problems, tackling large deformations and complex boundary conditions.

2.1. **Variational Formulation.** Let us consider the boundary value problem

$$\begin{cases} -u''(x) = f(x) & x \in \Omega \\ u(0) = u(1) = 0 \end{cases} \tag{D}$$

where $\Omega = (0,1)$ and $f \in \mathcal{C}(\overline{\Omega})$. Notice that this problem has a unique solution u, it suffices to integrate $-u'' = f$ twice, the integration constants are fixed by the boundary conditions.

This problem is widely known as it models a variety of situations in continuum mechanics:

- A. Elastic bar

  Consider an elastic bar fixed at both ends subject to a tangential load of intensity $f(x)$ as shown in figure 1. Let $\sigma(x)$ and $u(x)$ be the traction and tangential displacement at x, respectively, under the load f. If we assume small displacements and a linearly elastic material, then

$$\sigma = Eu' \qquad x \in \Omega \qquad \text{(Hooke's law)} \tag{25}$$

$$-\sigma' = f \qquad x \in \Omega \qquad \text{(Equilibrium equation)} \tag{26}$$

$$u = 0 \qquad x \in \{0,1\} \qquad \text{(Boundary conditions)} \tag{27}$$

  where E is the modulus of elasticity. If we take here $E = 1$ and eliminate σ we obtain (D).

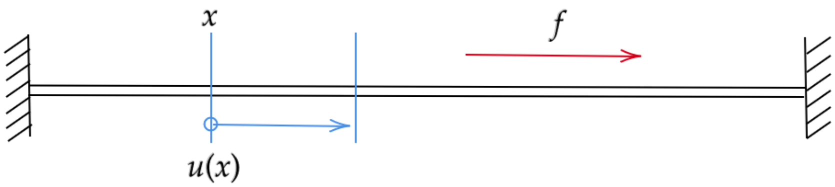

FIGURE 1. Elastic bar

- B. Elastic cord

  Consider an elastic cord with tension 1, fixed at both ends and subject to transversal load of intensity f as shown in figure 2. Assuming again small displacements, we have that the transversal displacement u satisfies (D).

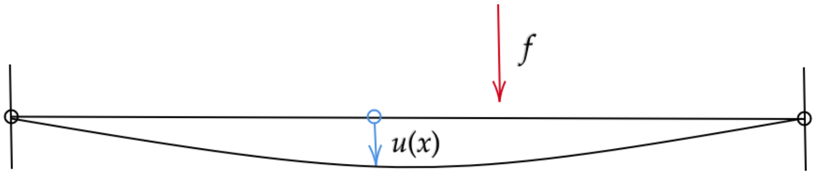

FIGURE 2. Elastic cord

- C. Heat conduction

  Let u be the temperature and q the heat flow in a heat conducting bar with diffusion constant k, subject to a distributed heat source of intensity f. Assuming the temperature to be zero at the end points, we have in the stationary case

$$-q = ku' \qquad x \in \Omega \qquad \text{(Fourier's law)} \tag{28}$$

$$q' = f \qquad x \in \Omega \qquad \text{(Conservation of energy)} \tag{29}$$

$$u = 0 \qquad x \in \{0,1\} \qquad \text{(Boundary conditions)} \tag{30}$$

We shall now show that the solution u of the boundary value problem (D) also is the solution of a minimization problem (M) and a variational problem (V). To formulate the problems (M) and (V) we introduce the notation

$$(v, w) = \int_0^1 v(x)w(x)\, dx \tag{31}$$

for real-valued piecewise continuous bounded functions. We also introduce the linear space

$$V_0 \;=\; \{v \;:\; v \in \mathcal{C}_0([0,1]),\; v' \text{ is piecewise continuous and bounded on } [0,1]\} \tag{32}$$

and the linear functional $F : V_0 \to \mathbb{R}$ given by

$$F(v) = \frac{1}{2}(v', v') - (f, v). \tag{33}$$

The problems (M) and (V) are the following

$$\text{Find } u \in V_0 \text{ such that } F(u) \leq F(v) \qquad \forall v \in V_0 \tag{M}$$

$$\text{Find } u \in V_0 \text{ such that } (u', v') = (f, v) \qquad \forall v \in V_0 \tag{V}$$

Let us notice that in the context of the problems A and B above, the quantity $F(v)$ represents the *total potential energy* associated with the displacement $v \in V_0$. The term $\frac{1}{2}(v', v')$ represents the internal elastic energy and (f, v) the load potential. Thus, the minimization problem (M) corresponds to the fundamental *Principle of minimum potential energy* in mechanics. Further the variational problem (V) corresponds to the *Principle of virtual work*.

Let us now first show that the solution u of (D) also is a solution of (V). To see this we multiply the equation $-u'' = f$ by an arbitrary function $v \in V_0$, a so-called *test function* v, and integrate over the interval $(0, 1)$ which gives

$$-(u'', v) = (f, v) \tag{34}$$

We now integrate by parts to obtain

$$-(u'', v) = -u'(1)v(1) + u'(0)v(0) + (u', v') \tag{35}$$

and as $v(0) = v(1) = 0$ we obtain

$$(u', v') = (f, v) \qquad \forall v \in V \tag{36}$$

which shows that u is a solution of (V).

Next, we show that the problems (V) and (M) have the same solutions. Suppose then first that u is a solution to (V), let $v \in V_0$ and set $w = v - u$ so that $v = u + w$ and $w \in V_0$. We have

$$\begin{aligned} F(v) &= F(u + w) \\ &= \frac{1}{2}(u' + w', u' + w') - (f, u + w) \\ &= \frac{1}{2}(u', u') - (f, u) + (u', w') - (f, w) + \frac{1}{2}(w', w') \\ &\geq F(u) \end{aligned}$$

since by (36), $(u', w') - (f, w) = 0$ and $(w', w') \geq 0$, which shows that u is a solution of (M). On the other hand, if u is a solution of (M) the we have that for any $v \in V_0$ and $\epsilon > 0$.

$$F(u) \leq F(u + \epsilon v), \tag{37}$$

since $u + \epsilon v \in V_0$. Thus, the differentiable function

$$g(\epsilon) := F(u + \epsilon v) = \frac{1}{2}(u', u') + \epsilon(u', v') + \frac{\epsilon^2}{2}(v', v') - (f, u) - \epsilon(f, v) \tag{38}$$

has a minimum at $\epsilon = 0$ and hence $g'(0) = 0$. But

$$g'(0) = (u', v') - (f, v), \tag{39}$$

and we see that u is a solution of (V).

Let us also show that a solution to (V) is uniquely determined. Suppose by contradiction that u_1 and u_2 are solutions of (V), i.e., $u_1, u_2 \in V_0$ and

$$\begin{aligned} (u_1', v') &= (f, v) && \forall v \in V_0 \\ (u_2', v') &= (f, v) && \forall v \in V_0 \end{aligned}$$

Subtracting these equations and picking $v = u_1 - u_2 \in V_0$, we get

$$\int_0^1 (u_1' - u_2')^2 \, dx = 0 \tag{40}$$

this shows that

$$u_1'(x) - u_2'(x) = (u_1 - u_2)'(x) = 0 \qquad \forall x \in [0,1] \tag{41}$$

It follows that $(u_1 - u_2)(x)$ is a constant function on $[0,1]$ which together with the boundary condition $u_1(0) = u_2(0) = 0$ gives $u_1(x) = u_2(x)$ for all $x \in [0,1]$, and the uniqueness follows.

To sum up, we have shown that if u is the solution to (D), then u is the solution to the equivalent problems (M) and (V) which we write symbolically as

$$(D) \Rightarrow (V) \Leftrightarrow (M)$$

Let us finally also indicate how to see that if u is the solution of (V) then u also satisfies (D). Thus, we assume that $u \in V_0$ satisfies

$$\int_0^1 u'v' \, dx = \int_0^1 fv \, dx = 0 \qquad \forall v \in V_0 \tag{42}$$

If, in addition, we also assume that u'' exists and is continuous, then we can integrate the first term by parts, and using the fact that $v(0) = v(1) = 0$ one gets

$$-\int_0^1 (u'' + f)v \, dx = 0 \qquad \forall v \in V_0 \tag{43}$$

But with the assumption that $(u'' + f)$ is continuous, by lemma 2.1, this relation can only hold if

$$(u'' + f)(x) = 0 \qquad x \in (0,1) \tag{44}$$

it follows that u is the solution of (D).

Lemma 2.1 (Othogonality to V_0)**.** *Let $w \in \mathcal{C}([0,1])$ and*

$$\int_0^1 wv \; dx = 0 \qquad \forall v \in V_0$$

then $w(x) = 0$ for $x \in [0,1]$.

Proof. We will use the fact that the space V_0 is dense in $\mathcal{C}_0([0,1])$ or the space of continuous functions that vanish at $\{0,1\}$ (see Appendix B). This means that for every $\psi \in \mathcal{C}_0([0,1])$ and every $\epsilon > 0$ there exists $v \in V_0$ such that $\|v - \psi\|_{L^\infty} < \epsilon$. Furthermore, for all $\psi \in \mathcal{C}_0([0,1])$ we have that

$$\begin{aligned}\left|\int_0^1 \psi w \; dx - \int_0^1 vw \; dx\right| &= \left|\int_0^1 (\psi - v)w \; dx\right| \\ &\leq \int_0^1 |(\psi - v)w| \; dx \\ &\leq \|\psi - v\|_{L^\infty}\|w\|_{L^1} && \text{(Holder's inequality)} \\ &< \epsilon\|w\|_{L^1}\end{aligned}$$

but recall that $\int_0^1 wv \; dx = 0$ for all $v \in V_0$, thus

$$\left|\int_0^1 \psi w \; dx\right| < \epsilon\|w\|_{L^1}$$

which gives us that $\int_0^1 \psi w \; dx = 0$ for all $\psi \in \mathcal{C}_0([0,1])$. Then the result follows by the *Fundamental lemma of the calculus of variations.* ■

Lemma 2.2 (Fundamental lemma of the calculus of variations)**.** *Let $w \in \mathcal{C}([0,1])$ and*

$$\int_0^1 wv \; dx = 0 \qquad \forall v \in \mathcal{C}_0([0,1])$$

then $w(x) = 0$ for $x \in [0,1]$.

Proof. By contradiction fix some $x_0 \in [0,1]$ such that $w(x_0) \neq 0$ and without loss of generality assume that $w(x_0) > 0$, as $w \in \mathcal{C}([0,1])$ there exists some sub interval $[a,b]$ such that $w([a,b]) > 0$. Now let

$$\alpha(x) := \begin{cases} (x-a)(b-x) & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases} \tag{45}$$

and notice that $\alpha([a,b]) > 0$. It follows that

$$\begin{aligned} \int_0^1 w\alpha \, dx &= \int_0^a w\alpha \, dx + \int_a^b w\alpha \, dx + \int_b^1 w\alpha \, dx \\ &= \int_a^b w\alpha \, dx \\ &= \int_a^b w(x-a)(b-x) \, dx \\ &> 0 \end{aligned}$$

This is a contradiction since $\alpha \in \mathcal{C}_0([0,1])$. ■

Thus we have seen that if u is the solution of (V) and in addition satisfies a regularity assumption (u'' is continuous), then u is the solution of (D). It is now possible to show that if u is the solution of (V), then u in fact satisfies the desired regularity assumption and thus we have $(V) \Rightarrow (D)$ which shows that the three problems (D), (V) and (M) are equivalent.

2.2. **The Lagrange finite element space.** We shall now construct a finite dimensional subspace V_h of the space V_0 defined in the previous section. To this end let

$$0 = x_0 < x_1 < ... < x_M < x_{M+1} = 1$$

be a partition of the interval $(0,1)$ into sub-intervals $I_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$ for $j = 1, ..., M+1$ and set $h = \max_j h_j$. The quantity h is then a measure of how fine the partition is. We now define the *Lagrange finite element space*

Definition 2.1 (Lagrange finite element space)**.** We define the space $V_h \subset V$ as

$$V_h = \{v \mid v \in \mathcal{C}_0([0,1]) \text{ and } v \in \mathcal{P}^1(I_j), \ \forall j \in \{1, ..., M+1\}\} \tag{46}$$

In figure 3 we can find an example of a function in V_h. As parameters to describe a function $v \in V_h$ we may choose values $\eta_j = v(x_j)$ at the node points x_j for

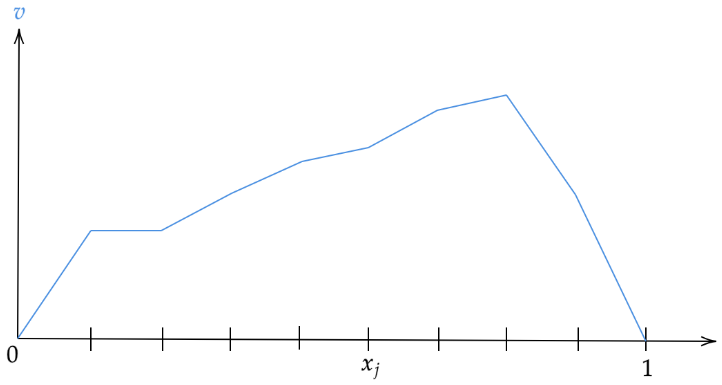

FIGURE 3. Example of a function $v \in V_h$

$j = 0, ..., M+1$. Let us introduce the *basis functions* $\varphi_j \in V_h$ for $j = 1, ..., M$ defined by

$$\varphi_j(x_i) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{47}$$

i.e. φ_j is the continuous piecewise linear function that takes the value 1 at node point x_j and the value 0 at other node points, we can see an example of this in figure 4

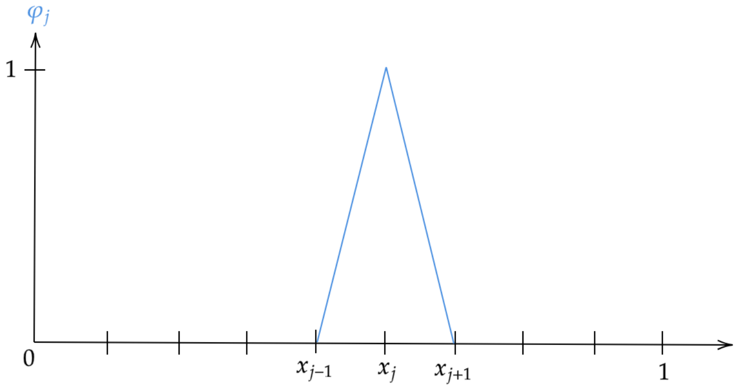

FIGURE 4. The basis function φ_j

A function $v \in V_h$ then has the representation

$$v(x) = \sum_{i=1}^{M} \eta_i \varphi_i(x) \qquad x \in [0,1] \tag{48}$$

where $\eta_i = v(x_i)$, i.e., each $v \in V_h$ can be written in a unique way as a *linear combination* of the basis functions φ_i. In particular it follows that V_h is a *linear space of dimension* M with *basis* $\{\varphi_i\}_{i=1}^{M}$.

The finite element method for the boundary value problem (D) can now be formulated as follows:

$$\text{Find } u_h \in V_h \text{ such that } F(u_h) \leq F(v) \qquad \forall v \in V_h \tag{M_h}$$

In the same way as in section 2.1 for the problems (M) as (V), we see that M_h is equivalent to the finite-dimensional variational problem (V_h):

$$\text{Find } u_h \in V_h \text{ such that } (u_h', v') = (f, v) \qquad \forall v \in V_h \tag{V_h}$$

Thus the finite element method for (D) can be formulated as (V_h) or equivalently (M_h). The problem (V_h) is usually referred to as *Galerkin's method* and (M_h) as *Ritz' method.*

We observe that if u_h satisfies (V_h) then in particular

$$(u_h', \varphi_j') = (f, \varphi_j) \qquad \forall j \in \{1, ..., M\} \tag{49}$$

and since

$$u_h = \sum_{i=1}^{M} \xi_i \varphi_i(x), \qquad \xi_i = u_h(x_i)$$

we can write

$$\sum_{i=1}^{M} \xi_i (\varphi_i', \varphi_j') = (f, \varphi_j) \qquad \forall j \in \{1, ..., M\} \tag{50}$$

which is a linear system of equations with M equations and M unknowns $\xi_1, ..., \xi_M$. In matrix form the linear system (50) can be written as

$$A\boldsymbol{\xi} = \boldsymbol{b} \tag{51}$$

where $A = (a_{ij})$ is the $M \times M$ matrix with elements $a_{ij} = (\varphi_i', \varphi_j')$ and where $\boldsymbol{\xi} = (\xi_1, ..., \xi_M)$ and $\boldsymbol{b} = (b_1, ..., b_M)$ with $b_i = (f, \varphi_i)$ are M-dimensional vectors:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{M1} & \dots & a_{MM} \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_M \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix} \tag{52}$$

The matrix A is called the *stiffness matrix* and $\boldsymbol{b}$ the *load vector*, with terminology from early applications of FEM in structural mechanics.

The elements $a_{ij} = (\varphi_i', \varphi_j')$ in the stiffness matrix A can easily be computed: First, observe that $(\varphi_i', \varphi_j') = 0$ if $|i - j| > 1$ since in this case for all $x \in [0, 1]$ either $\varphi_i(x)$ or $\varphi_j(x)$ is equal to zero. Thus, the matrix A is tri-diagonal.

Notice that for $j \in \{1, ..., M\}$ we have that

$$(\varphi_j', \varphi_j') = \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2}\, dx + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2}\, dx = \frac{1}{h_j} + \frac{1}{h_{j+1}} \tag{53}$$

and for $j \in \{2, ..., M\}$ we have that

$$(\varphi_j', \varphi_{j-1}') = (\varphi_{j-1}', \varphi_j') = -\int_{x_{j-1}}^{x_j} \frac{1}{h_j^2}\, dx = -\frac{1}{h_j} \tag{54}$$

Note also that the matrix A is *symmetric* and *positive definite* since $(\varphi_j', \varphi_{j-1}') = (\varphi_{j-1}', \varphi_j')$ and with $v(x) = \sum_{j=1}^M \eta_j \varphi_j(x)$ it follows

$$\boldsymbol{\eta}^T A \boldsymbol{\eta} = \sum_{i=1}^M \sum_{j=1}^M \eta_i (\varphi_i', \varphi_j') \eta_j = \left(\sum_{i=1}^M \eta_i \varphi_i', \sum_{j=1}^M \eta_j \varphi_j' \right) = (v', v') = \|v'\|_{L^2}^2 \geq 0$$

with equality only if $v' \equiv 0$. Then the linear system (51) has a unique solution.

We also notice that A is *sparse*, i.e., only a few elements of A are different from zero (A is tridiagonal). This very important property depends, as we have seen, on the fact that the support of a basis function $\varphi_j \in V_h$ only intersects the support of its neighbors $\varphi_{j-1}, \varphi_{j+1} \in V_h$. The fact that the basis functions may be chosen in this way is an important distinctive feature of the finite element method.

In the special case of a uniform partition with $h = h_j = \frac{1}{M+1}$ the system (51) takes the form

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \boldsymbol{\xi} = \boldsymbol{b} \tag{55}$$

Notice that, after division by h at both sides, one obtain the finite difference linear system (18) for the problem (D) (which is a slightly variant of the problem (16)), in this case one can see that the elements of the right hand side vector $b_j/h = \frac{1}{h}(f, \varphi_j)$ are mean values of f over the intervals (x_{j-1}, x_{j+1}).

2.3. **An error estimate.** We shall now study the error $u - u_h$ where u is the solution of (D) and u_h is the solution of the finite element problem (V_h). For this so, Lemma 2.3 provides a key property.

Lemma 2.3 (Galerkin's Orthogonality)**.** *If u is the solution of* (D) *and u_h is the solution of (V_h) then*

$$((u - u_h)', v') = 0 \qquad \forall v \in V_h \tag{56}$$

Proof. If u is solution of (D) then it is solution of (V), i.e.

$$(u', v') = (f, v) \qquad \forall v \in V_0$$

So in particular, as $V_0 \subset V_h$ it follows that

$$(u', v') = (f, v) \qquad \forall v \in V_h \tag{57}$$

and furthermore for u_h one have

$$(u_h', v') = (f, v) \qquad \forall v \in V_h \tag{58}$$

The result follows from subtracting (58) from (57). ■

With this, we can now prove that in some sense $u_h \in V_h$ is the best possible approximation to the exact solution u.

Theorem 2.1 (A version of Cea's Lemma)**.** *For any $v \in V_h$ we have*

$$\|(u-u_h)'\|_{L^2} \leq \|(u-v)'\|_{L^2}$$

Proof. Let $v \in V_h$ be arbitrary and set $w = u_h - v$, then $w \in V_h$ and furthermore

$$\begin{aligned}\|(u-u_h)'\|_{L^2}^2 &= ((u-u_h)',(u-u_h)') + ((u-u_h)',w') \quad \text{(Galerkin's Orthogonality)}\\ &= ((u-u_h)',(u-u_h+w)')\\ &= ((u-u_h)',(u-v)')\\ &\leq \|(u-u_h)'\|_{L^2}\,\|(u-v)'\|_{L^2} \quad \text{(Cauchy's inequality)}\end{aligned}$$

The result follows from dividing both sides by $\|(u-u_h)'\|_{L^2}$. ■

We will now be interested in estimating $\|(u-\tilde{u}_h)'\|_{L^2}$ where $\tilde{u}_h \in V_h$ is a suitable chosen function. We shall choose $\tilde{u}_h \in V_h$ to be *interpolant* of u, i.e., $\tilde{u}_h$ interpolates u at the nodes x_j, i.e.

$$\tilde{u}_h(x_j) = u(x_j) \qquad \forall j \in \{0, ..., M+1\} \tag{59}$$

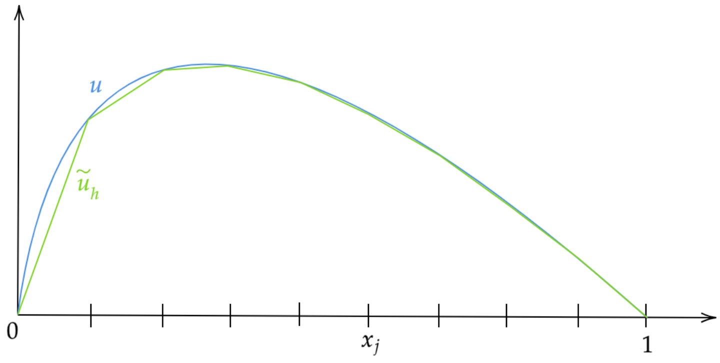

FIGURE 5. The interpolant $\tilde{u}_h$

To provide an estimate for the error $u - \tilde{u}_h$ we shall first consider the following bound

Lemma 2.4. *Let w sufficiently regular and $I = (a, b)$ or $I = [a, b]$, then*

$$|w(x) - w(y)| \leq |x - y|^{1/2} \|w'\|_{L^2(I)}$$

for all $x, y \in I$.

Proof. Fix $x, y \in I$, then by the fundamental theorem of calculus we have that

$$w(x) = w(y) + \int_y^x w'(s)\, ds \tag{60}$$

so it follows

$$\begin{aligned} |w(x) - w(y)| &= \left| \int_y^x w'(s)\, ds \right| \\ &\leq \int_y^x |w'(s)|\ ds \\ &= \int_a^b \mathbb{1}_{(x,y)}(s)\, |w'(s)|\ ds \\ &\leq |x - y|^{1/2} \left(\int_a^b |w'(s)|^2\ ds \right) \end{aligned}$$

Notice that the last inequality is result of applying the Cauchy-Schwarz inequality for $|w'(s)|$ and $\mathbb{1}_{(x,y)}(s)$ (the indicator function of the interval (x, y)). ■

Theorem 2.2 (V_h Interpolation error)**.** *Let $u \in V_0$, then*

$$\|(u - \tilde{u}_h)'\|_{L^2} \leq h \|u''\|_{L^2} \tag{61}$$

$$\|u - \tilde{u}_h\|_{L^2} \leq h^2 \|u''\|_{L^2} \tag{62}$$

with $\tilde{u}_h \in V_h$ the interpolant of u.

Proof. Fix some sub-interval $I_j = (x_{j-1}, x_j)$, then by Lemma 2.4 for every $x, \xi \in I_j$ one have

$$|(u - \tilde{u}_h)'(x) - (u - \tilde{u}_h)'(\xi)| \leq |x - \xi|^{1/2} \|(u - \tilde{u}_h)''\|_{L^2(I_j)} \tag{63}$$

Furthermore, notice that $(u - \tilde{u}_h)(x_{j-1}) = (u - \tilde{u}_h)(x_j) = 0$, as $\tilde{u}_h$ exactly approximates u in the partition nodes, and so, by the Rolle's Theorem[1] there exists some $\xi \in I_j$ such that $(u - \tilde{u}_h)'(\xi) = 0$. Moreover, $(u - \tilde{u}_h)'' = u'' - \tilde{u}_h'' = u''$, because u_h is linear in every I_j. Then it follows that

$$|(u - \tilde{u}_h)'(x)| \leq |x - \xi|^{1/2} \|u''\|_{L^2(I_j)} \tag{64}$$

We calculate now the L^2 norm of the functions at both sides to obtain

$$\begin{aligned}
\|(u - \tilde{u}_h)'\|_{L^2(I_j)} &\leq \left(\int_{I_j} |x - \xi| \|u''\|^2_{L^2(I_j)} \, dx \right)^{1/2} \\
&\leq \sup_{x \in I_j} |x - \xi|^{1/2} \left(\int_{I_j} \|u''\|^2_{L^2(I_j)} \, dx \right)^{1/2} \\
&\leq h^{1/2} \|u''\|_{L^2(I_j)} \left(\int_{I_j} 1 \, dx \right)^{1/2} \\
&= h \|u''\|_{L^2(I_j)}
\end{aligned}$$

This proves the first estimate. For the second one consider again Lemma 2.4, this time apply it to $u - \tilde{u}_h$ and $x \in \bar{I}_j = [x_{j-1}, x_j]$

$$|(u - \tilde{u}_h)(x) - (u - \tilde{u}_h)(x_{j-1})| \leq |x - x_{j-1}|^{1/2} \|(u - \tilde{u}_h)'\|_{L^2(\bar{I}_j)} \tag{65}$$

as $(u - \tilde{u}_h)(x_{j-1}) = 0$ then it follows that

$$|(u - \tilde{u}_h)(x)| \leq |x - x_{j-1}|^{1/2} \|(u - \tilde{u}_h)'\|_{L^2(\bar{I}_j)} \tag{66}$$

and we can calculate again the L^2 norm over I_j at both sides, in the same spirit than the first estimate, thus

$$\|u - \tilde{u}_h\|_{L^2(I_j)} \leq h \|(u - \tilde{u}_h)'\|_{L^2(\bar{I}_j)} \tag{67}$$

finally, apply the first estimate to the right hand side

[1] Mean-Value Theorem

$$\|u - \tilde{u}_h\|_{L^2(I_j)} \leq h^2 \|u''\|_{L^2(\bar{I}_j)} \tag{68}$$

Summing over all sub-intervals I_j concludes the proof. ■

Notice that as $\tilde{u}_h \in V_h$, Theorem 2.1 and Theorem 2.2 together provide an immediate estimate for the finite element approximation

$$\|(u - u_h)'\|_{L^2} \leq \|(u - \tilde{u}_h)'\|_{L^2} \leq h\|u''\|_{L^2} \tag{69}$$

Consider now the Lemma 2.4 for $u - u_h$ over $[0, 1]$ to obtain

$$\begin{aligned} |(u - u_h)(x) - (u - u_h)(0)| &\leq |x|^{1/2} \|(u - u_h)'\|_{L^2} \\ &\leq \|(u - u_h)'\|_{L^2} \end{aligned}$$

furthermore, we have that $u(0) = u_h(0) = 0$, and so

$$|(u - u_h)(x)| \leq \|(u - u_h)'\|_{L^2} \qquad \forall x \in [0, 1]$$

and our second error estimate for the finite element solution u_h follows from equation (69)

$$|(u - u_h)(x)| \leq h\|u''\|_{L^2} \qquad \forall x \in [0, 1] \tag{70}$$

we observe that this latter estimate is less sharp than the estimate for the interpolant of Theorem 2.2 where we have a factor of h^2, with a more precise analysis it is possible to show that in fact also the finite element method provides a factor h^2 for the error $u - u_h$ (see Appendix C). On the other hand, this is our first formal proof of convergence of the finite element solution

$$|(u - u_h)(x)| \xrightarrow[h \to 0^+]{} 0 \qquad \forall x \in [0, 1] \tag{71}$$

or, in other words

$$\|(u - u_h)\|_{L^\infty} \xrightarrow[h \to 0^+]{} 0 \tag{72}$$

so we have uniform convergence.

2.4. **Implementation and computational aspects.** We can split the implementation of the finite element method in two main parts:

(1) Assemble the linear system $A\boldsymbol{\xi} = \boldsymbol{b}$ introduced in (51).

(2) Solve the linear system $A\boldsymbol{\xi} = \boldsymbol{b}$.

We will focus in the first part, as the solution of ill-conditioned linear systems is a wide field and goes beyond the scope of the bootcamp.

In order to assemble the linear system $A\boldsymbol{\xi} = \boldsymbol{b}$ we first need to consider the definitions for A and $\boldsymbol{b}$

$$A_{ij} = (\varphi_i, \varphi_j) \qquad \boldsymbol{b}_i = (f, \varphi_i) \tag{73}$$

for all $i, j \in \{1, ..., N\}$. It is convenient to assemble the *load vector* $\boldsymbol{b}$ by iterating over the intervals I_k, as one just need to numerically perform the integration with an appropriate quadrature rule for each involved function (there are only two involved basis functions for each interval, as shown in figure 6), for this so, we define a local load vector[2]

$$b^k = \begin{pmatrix} b_1^k \\ b_2^k \end{pmatrix} = \begin{pmatrix} (\varphi_{k-1}, f)_{I_k} \\ (\varphi_k, f)_{I_k} \end{pmatrix} \qquad \forall k \in \{1, ..., M\} \tag{74}$$

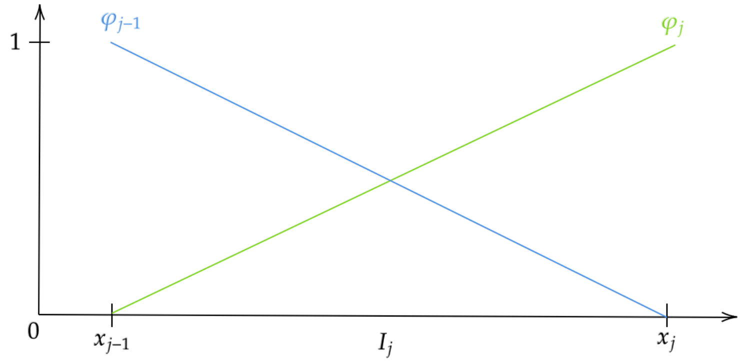

FIGURE 6. Basis functions that match in element I_j

[2] Denote $(a, b)_{I_k} = \int_{I_k} ab \, dx$.

With a slightly difference for the first and last elements, in which we only supports one basis function. One can notice that in this settings the component $\boldsymbol{b}_j$ of the global vector is given by

$$\boldsymbol{b}_j = b_2^j + b_1^{j+1} \qquad \forall j \in \{1, ..., M\} \tag{75}$$

and further $\boldsymbol{b}^1 = (0, (\varphi_1, f)_{I_1})^T$ and $\boldsymbol{b}^{M+1} = ((\varphi_M, f)_{I_{M+1}}, 0)^T$.

In the same fashion we can assemble the matrix A, recall that this matrix is highly *sparse*, and so for computational efficiency it makes sense to only compute the nonzero integrals. For this purpose we build first a local stiffness matrix for each element

$$A^k = \begin{pmatrix} A_{11}^k & A_{12}^k \\ A_{21}^k & A_{22}^k \end{pmatrix} = \begin{pmatrix} (\varphi'_{k-1}, \varphi'_{k-1})_{I_k} & (\varphi'_{k-1}, \varphi'_k)_{I_k} \\ (\varphi'_k, \varphi'_{k-1})_{I_k} & (\varphi'_k, \varphi'_k)_{I_k} \end{pmatrix} = \frac{1}{h_k} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{76}$$

here k runs from 1 to M, analogously to the local load vector case we have the boundary local stiffness matrix auxiliary defined as

$$A^1 = \begin{pmatrix} 0 & 0 \\ 0 & (\varphi'_1, \varphi'_1)_{I_1} \end{pmatrix} = \frac{1}{h_1} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{77}$$

and

$$A^{M+1} = \begin{pmatrix} (\varphi'_M, \varphi'_M)_{I_{M+1}} & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{h_{M+1}} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{78}$$

Then an arbitrary global *stiffness matrix* is given by

$$A_{i,i+1} = A_{12}^{i+1} \qquad A_{i,i-1} = A_{21}^{i} \qquad A_{i,i} = A_{22}^{i} + A_{11}^{i+1} \tag{79}$$

Or analogously

$$A_{i+1,i} = A_{21}^{i+1} \qquad A_{i-1,i} = A_{12}^{i} \qquad A_{i,i} = A_{22}^{i} + A_{11}^{i+1} \tag{80}$$

by symmetry. In this settings we iterate over the elements and store the coefficients efficiently, for example in a COO format matrix [3].

[3] See associated code.

3. 2D Finite element method

3.1. **Poisson equation and vector calculus.** We will now consider the following boundary value problem for the Poisson equation:

$$\begin{cases} -\Delta u(x) = f(x) & x \in \Omega \\ u \equiv 0 & x \in \Gamma \end{cases} \tag{81}$$

where Ω is a bounded open domain in the plane $\mathbb{R}^2 = \{x = (x_1, x_2) \ : \ x_i \in \mathbb{R}\}$ with boundary Γ, f is a given function and as usual

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \tag{82}$$

Let us now before continuing recall a certain *Green's formula* which will be of fundamental importance in what follows. Let us start from the *divergence theorem* (in two dimensions)

Theorem 3.1 (Divergence theorem)**.** *Let $\boldsymbol{A} = (A_1, A_2)$ be a vector-valued function defined on Ω. then*

$$\int_\Omega div\ \boldsymbol{A}\, dx = \int_\Gamma \boldsymbol{A} \cdot \boldsymbol{n}\, dS$$

where $\boldsymbol{n} = (n_1, n_2)$ is the outward unit normal to Γ.

Recall that

$$\text{div } A = \frac{\partial A_1}{\partial x_1} + \frac{\partial A_2}{\partial x_2} \tag{83}$$

If we apply the *divergence theorem* to $A = (vw, 0)$ and $A = (0, vw)$, we find that

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx + \int_\Omega v \frac{\partial w}{\partial x_i}\, dx = \int_\Gamma v w n_i\, dS, \qquad i = 1, 2 \tag{84}$$

Denoting by ∇v the *gradient* of v, i.e., $\nabla v = \left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}\right)$, we get from (84) the following Green's formula:

$$\begin{aligned}\int_\Omega \nabla v \cdot \nabla w \; dx &\equiv \int_\Omega \left[\frac{\partial v}{\partial x_1}\frac{\partial w}{\partial x_1} + \frac{\partial v}{\partial x_2}\frac{\partial w}{\partial x_2}\right] \; dx \\ &= \int_\Gamma \left[v\frac{\partial w}{\partial x_1}n_1 + v\frac{\partial w}{\partial x_2}n_2\right] \; dS - \int_\Omega v \left[\frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2}\right] \; dx \\ &= \int_\Gamma v\frac{\partial w}{\partial \boldsymbol{n}} \; dS - \int_\Omega v\Delta w \; dx\end{aligned}$$

thus

$$\int_\Omega \nabla v \cdot \nabla w \; dx = \int_\Gamma v\frac{\partial w}{\partial \boldsymbol{n}} - \int_\Omega v\Delta w \; dx \tag{85}$$

here

$$\frac{\partial w}{\partial \boldsymbol{n}} = \frac{\partial w}{\partial x_1}n_1 + \frac{\partial w}{\partial x_2}n_2$$

is the *normal derivative*, i.e. the derivative in the outward normal direction to the boundary Γ.

3.2. **Variational Formulation.** We shall now give a variational formulation of problem (81). We shall first show that if u satisfies (81), then u is the solution of the following variational problem:

$$\text{Find } u \in V \text{ such that } a(u,v) = (f,v) \qquad \forall v \in V \tag{86}$$

here

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \; dx \tag{87}$$

$$(f,v) = \int_\Omega fv \; dx \tag{88}$$

and

$$V = \left\{v \;:\; v \in \mathcal{C}_0(\Omega) \text{ and } \frac{\partial v}{\partial x_1}, \; \frac{\partial v}{\partial x_2} \text{ are piecewise continuous on } \Omega\right\} \tag{89}$$

In exactly the same way as in the 1D case, we see that $u \in V$ satisfies (86) if and only if u is the solution of the following minimization problem

$$\text{Find } u \in V \text{ such that } F(u) \leq F(v) \qquad \forall v \in V \tag{90}$$

where

$$F(v) = \frac{1}{2}a(v,v) - (f,v) \tag{91}$$

is the total potential energy.

To see that (86) follows from (81) we multiply the equation over Ω with an arbitrary test function $v \in V$ and integrate over Ω. According to Green's formula (84) we then have

$$(f,v) = -\int_\Omega \Delta uv \; dx = -\int_\Gamma \frac{\partial u}{\partial n} v \; dS + \int_\Omega \nabla u \cdot \nabla v \; dx = a(u,v) \tag{92}$$

where the boundary integral vanishes since $v \in \mathcal{C}_0(\Omega)$ i.e. $v = 0$ on Γ. On the other hand if $u \in V$ satisfies (86) and if u is sufficiently regular, then we see as in the 1D case that u also satisfies (81).

3.3. **Finite dimensional approximation.** Let us now construct a finite-dimensional subspace V_h of V. for simplicity we shall assume that Γ is a polygonal curve, in which case we say that Ω is a polygonal domain (if Γ is curve we may first approximate Γ with a polygonal curve). Let us now make a *triangulation* of Ω, by subdividing Ω into a set $\mathcal{T}_h = \{K_1, K_2, ..., K_m\}$ of non-overlapping triangles K_i:

$$\Omega = \bigcup_{K \in \mathcal{T}_h} K \tag{93}$$

such that no vertex of one triangle lies on the edge of another triangle (see figure 7). We introduce the mesh parameter

$$h = \max_{K \in \mathcal{T}_h} \text{diam } (K) \tag{94}$$

where diam (K) i.e. the diameter of K is the longest side of K. We now define V_h as follows

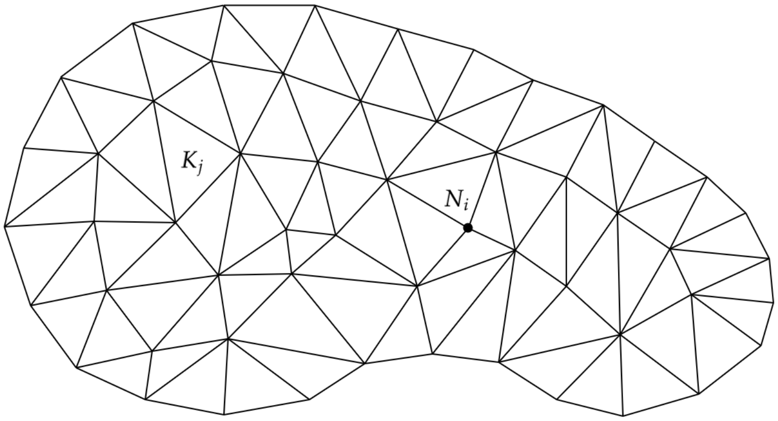

FIGURE 7. A finite element triangulation

$$V_h = \{v \,:\, v \in \mathcal{C}_0(\Omega), \;\; v|_K \text{ is linear for } K \in \mathcal{T}_h\} \tag{95}$$

here $v|_K$ denotes the restriction of v to K, i.e. the function defined of K agreeing with v on K. The space V_h consistes of all continuous functions that are linear on each triangle K and vanish on Γ. We notice that $V_h \subset V$. As parameters to describe a function $v \in V_h$ we choose the values $v(N_i)$ of v at the *nodes* N_i, $i = 1, ..., M$, of $\mathcal{T}_h$ (see figure 7) but exclude the nodes of the boundary since $v = 0$ on Γ. The corresponding basis functions $\varphi_j \in V_h$, $j = 1, ..., M$ are then defined by (see fig 9)

$$\varphi_j(N_i) = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{96}$$

We see that the *support* of φ_j (the set of points x for which $\varphi_j(x) \neq 0$) consists of the triangles with the common node N_j (the gray area in figure 9). A function $v \in V_h$ now has the representation

$$v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x), \quad \eta_j = v(N_j), \qquad \forall x \in \Omega \cup \Gamma \tag{97}$$

We can now formulate the following finite element method for (81) starting from the variational formulation (86):

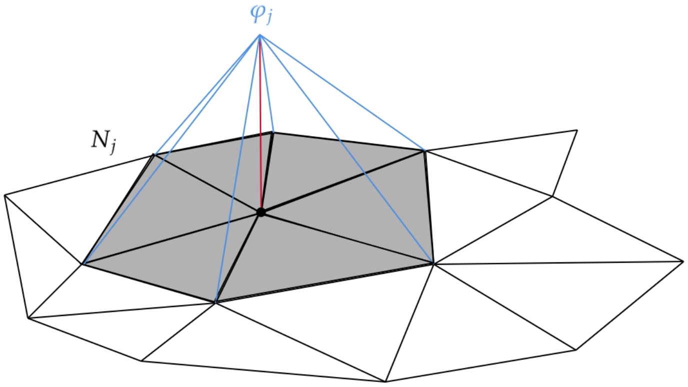

FIGURE 8. The basis function φ_j

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v) = (f, v) \qquad \forall v \in V_h \tag{98}$$

Exactly as in 1D we see that (98) is equivalent to the linear system of equations

$$A\boldsymbol{\xi} = \boldsymbol{b} \tag{99}$$

where $A = a_{ij}$, the stiffness matrix, is an $M \times M$ matrix with elements $a_{ij} = a(\varphi_i, \varphi_j)$ and $\boldsymbol{\xi}$, $\boldsymbol{b}$ are M-vectors with elements $\xi_i = u_h(N_i)$, $b_i = (f, \varphi_i)$.

Clearly A is symmetric and as in the 1D case we see that A is positive definite and thus in particular non-singular so that (99) admits a unique solution $\boldsymbol{\xi}$. Moreover. A is again sparse; we have that $a_{ij} = 0$ unless N_i and N_j are nodes of the same triangle.

3.4. **Error convergence.** In the same way that in the 1D case we realize that $u_h \in V_h$ is the best approximation of the exact solution u in the sense that

$$\|\nabla u - \nabla u_h\|_{L^2} \leq \|\nabla u - \nabla v\|_{L^2} \qquad \forall v \in V_h \tag{100}$$

notice that

$$\|\nabla v\|_{L^2} = a(v, v)^{1/2} \tag{101}$$

In particular we have

$$\|\nabla u - \nabla u_h\|_{L^2} \leq \|\nabla u - \nabla \tilde{u}_h\|_{L^2} \tag{102}$$

where $\tilde{u}_h$ is the interpolant of u, i.e. $\tilde{u}_h \in V_h$ and

$$\tilde{u}_h(N_i) = u(N_i) \qquad \forall i = 1, ..., M \tag{103}$$

We can further prove that, if the triangles $K \in \mathcal{T}_h$ are not allowed to become too thin, then

$$\|\nabla u - \nabla \tilde{u}_h\|_{L^2} \leq Ch \tag{104}$$

Here $C > 0$ is independent of h, and in a similar fashion that section 2 one can show that

$$\|u - u_h\|_{L^2} \leq Ch^2 \tag{105}$$

where, again, $C > 0$ is independent of h, see Appendix C. In particular these estimates show that if the exact solution u is sufficiently regular, then

$$\|u - u_h\|_{L^2} \xrightarrow[h \to 0]{} 0 \qquad \|\nabla u - \nabla u_h\|_{L^2} \xrightarrow[h \to 0]{} 0 \tag{106}$$

Example 3.1. Let Ω be the unit square and let $\mathcal{T}_h$ be the uniform triangulation of Ω according to figure 9 with the indicated enumeration of the nodes of $\mathcal{T}_h$. In this case the linear system (99) reads as follows:

$$\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & \dots & \vdots \\ 0 & -1 & 4 & 0 & 0 & -1 & \dots & \vdots \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & -1 \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & -1 & 0 \\ \vdots & \ddots & \ddots & -1 & 0 & -1 & 4 & -1 \\ 0 & \dots & \dots & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \vdots \\ \vdots \\ \xi_{M-1} \\ \xi_M \end{pmatrix} = h^2 \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ \vdots \\ \vdots \\ b_{M-1} \\ b_M \end{pmatrix} \tag{107}$$

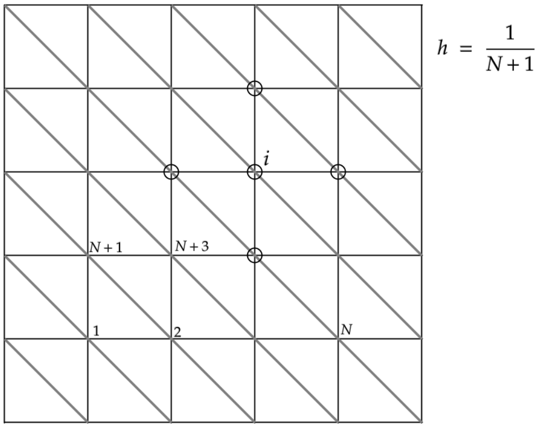

FIGURE 9. Uniform mesh over unit square.

Note that here the left-hand side of equation i is a linear combination of the values of u_h at the 5 nodes indicated in figure 9 with coefficients given in figure 10. We can recognize this as the linear system obtained by applying the so-called 5-point difference method for (81) with the components if the right-hand side being weighted averages of f around the nodes N_i, the second order approximation induced by the stencil presented in figure 10 reads

$$\frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \approx \frac{u(x_1 + h, x_2) - 2u(x_1, x_2) + u(x_1 - h, x_2)}{h^2} \tag{108}$$

for x_1 and analogously for the x_2 variable.

The elements $a_{ij} = a(\varphi_i, \varphi_j)$ in the stiffness matrix A are usually in practice computed by summing the contributions from the different triangles:

$$a(\varphi_i, \varphi_j) = \sum_{K \in \mathcal{T}_h} a_K(\varphi_i, \varphi_j) \tag{109}$$

where

$$a_K(\varphi_i, \varphi_j) = \int_K \nabla\varphi_i \cdot \nabla\varphi_j \; dx \tag{110}$$

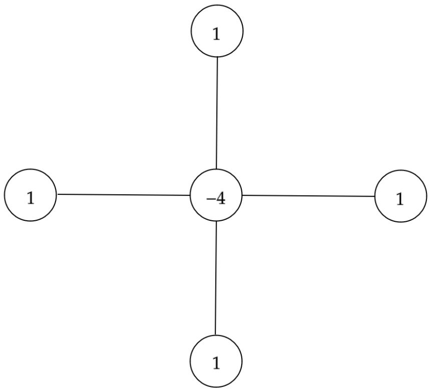

Figure 10. 2D Finite difference stencil.

Notice that this is analogous to the efficient implementation of the *stiffness matrix* assemble we saw in the 1D case. In the same spirit we notice now that $a_K(\varphi_i, \varphi_j) = 0$ unless both nodes N_i and N_j are vertices of K. Let N_i, N_j and N_k be the vertices of the triangle K we then call the 3×3 symmetric matrix

$$A^K = \begin{pmatrix} a_K(\varphi_i, \varphi_i) & a_K(\varphi_i, \varphi_j) & a_K(\varphi_i, \varphi_k) \\ & a_K(\varphi_j, \varphi_j) & a_K(\varphi_j, \varphi_k) \\ & & a_K(\varphi_k, \varphi_k) \end{pmatrix} \tag{111}$$

the *local stiffness matrix* for the element K, as we similarly did in the 1D case.

The *global stiffness matrix* A may thus be computed by first computing the element stiffness matrices for each $K \in \mathcal{T}_h$ and then summing the contributions from each triangle according to (109). The assemble of the right hand side $\boldsymbol{b}$ follows in the same fashion than explained in the 1D case[4].

To compute the elements in the stiffness matrix (111) we clearly work with the restrictions of the basis functions φ_i, φ_j and φ_k to the triangle K. Denoting these restrictions by ψ_i, ψ_j and ψ_k, we have that each ψ is a linear function on K that takes the value one at one vertex and vanishes at the other two vertices of K. We call ψ_i, ψ_j and ψ_k the *basis functions on the triangle* K (see figure 11). If w is a linear function on K, then w has the representation

[4] We formally define this process of computing A and $\boldsymbol{b}$ by summation as the *assembly* of A and $\boldsymbol{b}$.

$$w(x) = w(N_i)\psi_i(x) + w(N_k)\psi_k(x) + w(N_k)\psi_k(x) \qquad x \in K \tag{112}$$

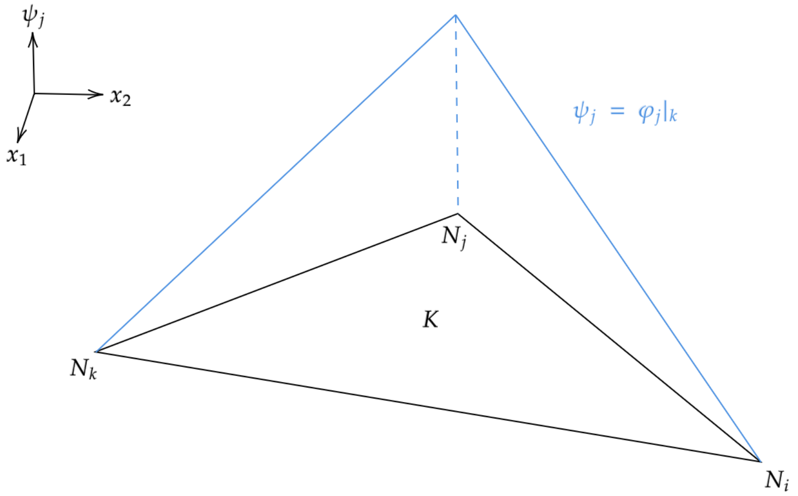

FIGURE 11. The basis function ψ_j asociated with K.

4. A geometric interpretation of FEM

We shall now give an interpretation of the finite element method in geometric terms in the function space $H_0^1(\Omega)$.

Definition 4.1 (Orthogonality)**.** Two elements v and w in a linear space with *inner product* $(\cdot,\cdot)$ are said to be *orthogonal* if $(v,w)=0$.

Let us for simplicity consider the following variant of our previous problem (81)

$$\begin{cases} -\Delta u + u = f & x \in \Omega \\ u = 0 & x \in \Gamma \end{cases} \tag{113}$$

The corresponding variational problem reads:

$$\text{Find } u \in H_0^1(\Omega) \text{ s.t. } (u,v)_{H^1} = (f,v) \qquad \forall v \in H_0^1(\Omega) \tag{114}$$

where (see Appendix D):

$$(u,v)_{H^1} = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx \tag{115}$$

Let V_h be a finite-dimensional subspace of $H_0^1(\Omega)$, e.g. the space of piecewise linear functions of section 3.3 and consider the following finite element method for (113):

$$\text{Find } u_h \in V_h \text{ s.t. } (u_h,v)_{H^1} = (f,v) \qquad \forall v \in V_h \tag{116}$$

Since $V_h \subset H_0^1(\Omega)$ we may choose $v \in V_h$ in (114) and on substraction from (116) we obtain

$$(u - u_h, v)_{H^1} = 0 \qquad \forall v \in V_h \tag{117}$$

i.e. the error $u - u_h$ is orthogonal to V_h with respect to $(\cdot,\cdot)_{H^1}$. We may also express this fact as follows: The finite element solution u_h is the *projection* with respect to $(\cdot,\cdot)_{H^1}$ of the exact solution u *on* V_h, i.e., u_h is the element in V_h closest to u with respect to the $H^1(\Omega)$ norm $\|\cdot\|_{H^1}$ (see Appendix D), or in other words

$$\|u - u_h\|_{H^1} \le \|u - v\|_{H^1} \qquad \forall v \in V_h \tag{118}$$

This situations is symbolically illustrated in figure 12 where $H_0^1(\Omega)$ is represented by the whole plane while the straight line through the origin represents V_h

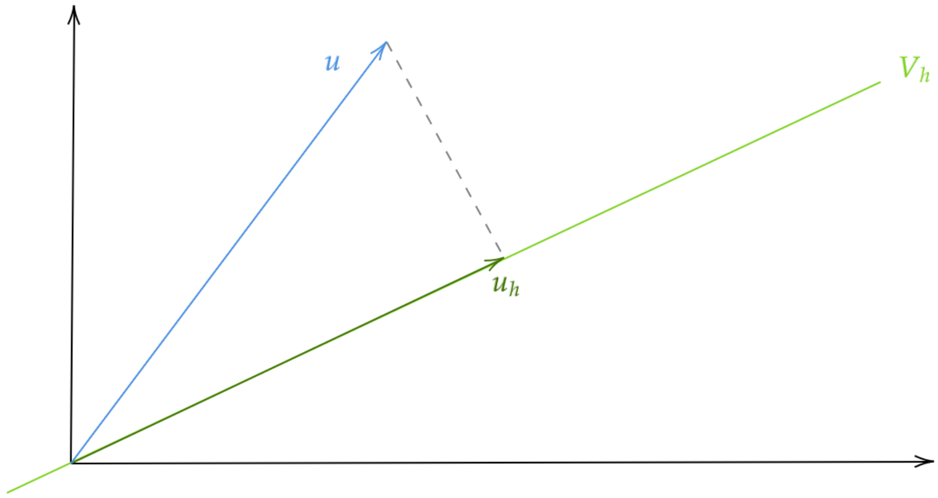

FIGURE 12. FEM projection

According to (118), u_h is the best approximation of the exact solution u, in the sense that for no other function $v \in V_h$, the error $u - v$ is smaller when measured in the $H^1(\Omega)$-norm. we have seen that u_h can be found by solving a linear system of equations with right hand side depending on the given function f. Thus, we can compute a best approximation u_h of u, without knowing u itself, knowing only that $-\Delta u + u = f$ in Ω and $u = 0$ on Γ.

5. Miscellaneous

5.1. **Neumann problem.**

We shall now consider a problem with another type of boundary condition, namely the following *Neumann problem* (NBVP):

$$\begin{cases} -\Delta u + u = f & x \in \Omega \\ \frac{\partial u}{\partial n} = g & x \in \Gamma \end{cases} \tag{NBVP}$$

where again Ω is a bounded domain with boundary Γ and $\frac{\partial}{\partial n}$ denotes the outward normal derivative to Γ. The boundary condition is a *Neumann condition* while the boundary condition $u = u_0$ on Γ considered previously is said to be a *Dirichlet condition*. In mechanics or physics the Neumann condition in (NBVP) corresponds to a given force or flow on Γ.

We can give the problem (NBVP) the following variational formulation (NVF)

$$\text{Find } u \in H^1(\Omega) \text{ s.t. } a(u,v) = (f,v) + \langle g, v\rangle \qquad \forall v \in H^1(\Omega) \tag{NVF}$$

where

$$a(u,v) = (u,v)_{H^1}, \qquad \langle g, v\rangle = \int_\Gamma gv \, dS \tag{119}$$

This is equivalent to the following minimization formulation (NMIN)

$$\text{Find } u \in H^1(\Omega) \text{ s.t. } F(u) \leq F(v) \qquad \forall v \in H^1(\Omega) \tag{NMIN}$$

where

$$F(v) = \frac{1}{2}a(v,v) - (f,v) - \langle g, v\rangle \tag{120}$$

To see that (NVF) follows from (NBVP) we multiply the first equation of (NBVP) with a test function $v \in H^1(\Omega)$ and integrate over Ω. According to Green's formula (85), we then get, since $\frac{\partial u}{\partial n} = g$ on Γ,

$$\begin{aligned}(f,v) &= \int_\Omega [-\Delta u + u] v \, dx \\ &= -\int_\Gamma \frac{\partial u}{\partial n} v \, dS + \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx \\ &= -\langle g, v\rangle + (u,v)_{H^1} = a(u,v) - \langle g, v\rangle\end{aligned}$$

which leads to (NVF).

Let us now also motivate why a solution $u \in H^1(\Omega)$ of the variational problem (NVF) also should satisfy (NBVP). Using Green's formula again we find from (NVF) that if u is sufficiently regular, then

$$(f,v) + \langle g, v\rangle = a(u,v) = \int_\Gamma \frac{\partial u}{\partial n} v \, dS + \int_\Omega [-\Delta u + u] v \, dx \tag{121}$$

so that, rearranging terms

$$\int_\Omega [-\Delta u + u - f] v \, dx + \int_\Gamma \left[\frac{\partial u}{\partial n} - g\right] v \, dS = 0 \qquad \forall v \in H^1(\Omega) \tag{122}$$

Now, we are in a similar situation than Lemma 2.2 but with different spaces and inner products. Depending on the regularity assumptions on u the proof may change, for example, assume that $-\Delta u + u - f \in L^2(\Omega)$ and $\frac{\partial u}{\partial n} - g \in L^2(\Omega)$, then as (122) holds for all $v \in H^1(\Omega) \subset L^2(\Omega)$, by density arguments i.e.[5]

$$\overline{H^1(\Omega)}^{\|\cdot\|_{L^2}} = L^2(\Omega) \tag{123}$$

in particular holds $v = -\Delta u + u - f$ in Ω and $v = \frac{\partial u}{\partial n} - g$ on Γ, thus

$$\int_\Omega [-\Delta u + u - f]^2 \, dx + \int_\Gamma \left[\frac{\partial u}{\partial n} - g\right]^2 \, dS = 0 \tag{124}$$

or in other words

$$\| -\Delta u + u - f\|^2_{L^2(\Omega)} + \left\|\frac{\partial u}{\partial n} - g\right\|^2_{L^2(\Gamma)} = 0 \tag{125}$$

[5] To prove this notice that $\mathcal{C}_0^\infty(\Omega) \subset H_0^1(\Omega) \subset H^1(\Omega) \subset L^2(\Omega)$

So $-\Delta u + u = f$ in Ω and $\frac{\partial u}{\partial n} = g$ on Γ in the L^2 sense.

5.2. **Natural and essential boundary conditions.** We note that the Neumann condition in (NBVP) does not appear explicitly in the variational formulation (NVF); the solution u of (NVF) is only required to belong to $H^1(\Omega)$ and is not explicitly required to satisfy (NBVP). This boundary condition is instead implicitly contained the term $\langle g, v\rangle$ in (NVF).

Such a boundary condition, that does not have to be explicitly imposed in the space of the variational formulation, is said to be a *natural boundary condition*.

This is in contrast to a so-called *essential boundary condition*, like the Dirichlet condition $u = 0$ on Γ in (113), that has to be explicitly imposed in the space of the variational formulation of the form (114).

5.3. **Finite Elements for the Neumann problem.** Let us now formulate a finite element method for the Neumann problem (NVF). Let $\mathcal{T}_h$ be a triangulation of Ω as in section 3.3 and define

$$V_h = \{v \in \mathcal{C}(\Omega) \mid v|_K \in \mathcal{P}^1(K) \;\; \forall K \in \mathcal{T}_h\} \tag{126}$$

As parameters to describe the functions in V_h we of course choose the values at the nodes, now including also the nodes on the boundary Γ. Note that the functions in V_h are not required to satisfy any boundary condition and that $V_h \subset H^1(\Omega)$. By starting from (NVF) we now have the following finite element method for (NBVP):

$$\text{Find } u_h \in V_h \text{ s.t. } a(u_h, v) = (f, v) + \langle g, v\rangle \qquad v \in V_h \tag{NFEM}$$

Analogously to the Dirichlet problem from section 3.3, we can see that this problem has a unique solution u_h that can be determined by solving a symmetric, positive definite linear system of equations. We also have the following error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq \|u - v\|_{H^1(\Omega)} \qquad \forall v \in V_h \tag{127}$$

and hence as above

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch \tag{128}$$

if u is regular enough. The function u_h will satisfy the Neumann condition of (NBVP) approximatly, i.e. $\frac{\partial u_h}{\partial n}$ will be an approximation to g on Γ.

Remark. When formulating a difference method for (NBVP) one meets severe difficulties due to the boundary condition, unless Ω has a very simple shape such as a rectangle. On the other hand, in the finite element formulation, the same boundary condition does not cause any complication.

5.4. **Non-homogeneous Dirichlet boundary conditions.** Consider the following boundary value problem (NHBVP):

$$\begin{cases} -\Delta u + u = f & x \in \Omega \\ u = u_0 & x \in \Gamma \end{cases} \tag{NHBVP}$$

where $u_0 \neq 0$. If we proceed as before, one should obtain a variational formulation like

$$\text{Find } u \in H^1_\Gamma(\Omega) \text{ s.t. } (u, v)_{H^1} = (f, v) \qquad \forall v \in H^1_0(\Omega) \tag{129}$$

where

$$H^1_\Gamma(\Omega) = \{v \in H^1(\Omega) \mid v|_\Gamma = u_0\} \tag{130}$$

but notice that $H^1_\Gamma(\Omega)$ is not a Hilbert space, moreover, it is not even a linear space, as if we take $v, w \in H^1_\Gamma(\Omega)$ then $(v + w)|_\Gamma = 2u_0$, and so $v + w \notin H^1_\Gamma(\Omega)$. To solve this issue we shall write our solution u as $u = w + U_0$, where $w \in H^1_0(\Omega)$ and U_0 is a sufficiently regular function that vanishes in Ω but $U_0|_\Gamma = u_0$, then we have

$$\text{Find } w \in H^1_0(\Omega) \text{ s.t. } (w, v)_{H^1} = (f, v) - (U_0, v)_{H^1} \qquad \forall v \in H^1_0(\Omega) \tag{131}$$

Which is an equivalent problem to the homogeneous boundary condition case. We call this condition an *essential boundary condition.*

5.5. **Robin boundary conditions.** Consider the following boundary value problem (RBVP):

$$\begin{cases} -\Delta u + u = f & x \in \Omega \\ \frac{\partial u}{\partial n} + \alpha u = g & x \in \Gamma \end{cases} \tag{RBVP}$$

where $\alpha > 0$. Notice that the boundary condition of this problem is similar to both the non-homogeneous Dirichlet problem and the Neumann problem, indeed, it is a linear combination of both of them, and it is called *Robin boundary condition.*

To obtain a variational formulation we multiply the first equation of (RBVP) with a test function $v \in H^1(\Omega)$ and integrate over Ω. According to Green's formula (85), we then get, since $u + \frac{\partial u}{\partial n} = g$ on Γ,

$$\begin{aligned}(f, v) &= \int_\Omega [-\Delta u + u] v \, dx \\ &= -\int_\Gamma \frac{\partial u}{\partial n} v \, dS + \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx \\ &= -\langle g - \alpha u, v\rangle + (u, v)_{H^1} \\ &= -\langle g, v\rangle + \alpha\langle u, v\rangle + (u, v)_{H^1}\end{aligned}$$

Then the variational formulation reads

$$\text{Find } u \in H^1(\Omega) \text{ s.t. } (u, v)_{H^1} + \alpha\langle u, v\rangle = (f, v) + \langle g, v\rangle \qquad \forall v \in H^1(\Omega) \tag{132}$$

And we can further rewrite this as

$$\text{Find } u \in H^1(\Omega) \text{ s.t. } (u, v)_{H^1_\alpha} = (f, v) + \langle g, v\rangle \qquad \forall v \in H^1(\Omega) \tag{133}$$

where

$$(u, v)_{H^1_\alpha} := (u, v)_{H^1} + \alpha \int_{\partial\Omega} uv \, dS \tag{134}$$

This new inner product is still bilinear, symmetric, and positive definite, as long as $\alpha \geq 0$. Furthermore, it can be proven[6] that the norm induced by this new inner product is equivalent to the usual H^1-norm, it is defined by

$$\| \cdot \|_{H^1_\alpha} = (\| \cdot \|^2_{H^1} + \alpha \| \cdot \|^2_{\partial\Omega})^{1/2} \tag{135}$$

where

$$\|u\|^2_{\partial\Omega} := \int_{\partial\Omega} u^2 \, dS \tag{136}$$

[6] See Poincare inequality.

6. Appendix

6.1. Appendix A: $\mathcal{O}$ and $\mathcal{o}$ notation.

For a given function g we define

$$\mathcal{O}(g) = \{f \mid \exists c, x_0 > 0 \text{ s.t. } 0 \leq f(x) \leq cg(x) \;\; \forall x \geq x_0\} \tag{137}$$

and

$$\mathcal{o}(g) = \{f \mid \forall c > 0 \;\; \exists x_0 > 0 \text{ s.t. } 0 \leq f(x) < cg(x) \;\; \forall x \geq x_0\} \tag{138}$$

Intuitively $f \in \mathcal{o}(g)$ means

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

6.2. Appendix B: V_0 is dense in $\mathcal{C}_0([0,1])$.

Proof. Let $f \in \mathcal{C}_0([0,1])$ and $\epsilon > 0$ be arbitrary, we want to show that there exists $v \in V_0$ such that $\|v - f\|_{L^\infty} < \epsilon$. For this so, first notice that from the continuity of f and the compactness of a bounded closed set in $\mathbb{R}$ it follows that f is uniformly continuous over $[0,1]$, this is, for all $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(y)| < \epsilon/2$ provided $|x - y| < \delta$.

Define now a finite subdivision of $[0,1]$ as $\Omega_n := \{a_0 = 0, a_1, a_2, \ldots, a_n = 1\}$, and let it be such that $|a_{k+1} - a_k| < \delta$ for all $k \in \{0, ..., n-1\}$. Then, notice that there exists $p \in V_0$ such that $p(a_k) = f(a_k)$, to see this, it suffices to consider a piecewise linear function that connects $(a_k, f(a_k))$ to $(a_{k+1}, f(a_{k+1}))$ with a straight line segment for each k.

If we now take an arbitrary $x \in [0,1]$, by construction there exists exactly one index $\hat{k} \in \{0, ..., n-1\}$ such that $a_{\hat{k}} \leq x \leq a_{\hat{k}+1}$, then we have

$$\begin{aligned} |p(x) - p(a_{\hat{k}})| &\leq |p(a_{\hat{k}+1}) - p(a_{\hat{k}})| \\ &= |f(a_{\hat{k}+1}) - f(a_{\hat{k}})| \\ &< \frac{\epsilon}{2} \end{aligned}$$

as $|a_{\hat{k}+1} - a_{\hat{k}}| < \delta$. And since $|x - a_{\hat{k}}| \leq |a_{\hat{k}+1} - a_{\hat{k}}| < \delta$ it also follows that $|f(x) - f(a_{\hat{k}})| < \frac{\epsilon}{2}$. Thus

$$\begin{aligned}|p(x) - f(x)| &\leq |p(x) - p(a_{\hat{k}})| + |p(a_{\hat{k}}) - f(x)| \\ &= |p(x) - p(a_{\hat{k}})| + |f(x) - f(a_{\hat{k}})| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2}\end{aligned}$$

We can conclude using the arbitrariness of $x \in [0,1]$ by taking supremum at both sides of the inequality to obtain

$$\sup_{x\in[0,1]} |p(x) - f(x)| < \epsilon$$

■

6.3. **Appendix C: Proof of optimal error estimate for the 1D/2D FEM solution.**

See Theorem 5.4.1. in [1].

■

6.4. **Appendix D: The Hilbert spaces $L^2(\Omega)$, $H^1(\Omega)$, $H^1_0(\Omega)$.**

When giving variational formulations of boundary value problems for PDE, it is from the mathematical point of view natural and very useful to work with function spaces V that are slightly larger (i.e. contain somewhat more functions) with piecewise continuous derivatives used in sections 2 and 3. It is also useful to endow the spaces V with an *inner product* such that V is a *Hilbert space.*

Before introducing Hilbert spaces let us recall a few simple concepts from linear algebra: If V is a linear space, then we say that L is a *linear form* on V if $L : V \to \mathbb{R}$ and L is *linear*, i.e., for all $v, w \in V$ and $\beta, \theta \in \mathbb{R}$:

$$L(\beta v + \theta w) = \beta L(v) + \theta L(w) \tag{139}$$

Furthermore, we say that $a(\cdot, \cdot)$ is a *bilinear form* on $V \times V$ if $a : V \times V \to \mathbb{R}$ and a is linear in each argument, i.e. for all $u, v, w \in V$ and $\beta, \theta \in \mathbb{R}$ we have:

$$a(u, \beta v + \theta w) = \beta a(u, v) + \theta a(u, w) \tag{140}$$
$$a(\beta u + \theta v, w) = \beta a(u, w) + \theta a(v, w) \tag{141}$$
$$\tag{142}$$

The bilinear form $a(\cdot, \cdot)$ on $V \times V$ is said to be *symmetric* if

$$a(v, w) = a(w, v) \qquad \forall v, w \in V \tag{143}$$

A symmetric bilinear form $a(\cdot, \cdot)$ on $V \times V$ is said to be an *inner product* on V if

$$a(v, v) > 0 \qquad \forall v \in V,\ v \neq 0 \tag{144}$$

The *norm* $\| \cdot \|_a$ associated with an inner product $a(\cdot, \cdot)$ is defined by

$$\|v\|_a := \sqrt{a(v, v)} \tag{145}$$

Further, if $\langle \cdot, \cdot \rangle$ is an inner product with corresponding norm $\| \cdot \|$, then we have *Cauchy's inequality*

$$|\langle v, w \rangle| \leq \|v\| \|w\| \qquad \forall v, w \in V \tag{146}$$

We further recall that if V is a linear space equipped with an inner product and its induced norm $\| \cdot \|$, then V is said to be a *Hilbert space* provided that it is *complete* (see definitions (6.1) and (6.2)).

Definition 6.1 (Complete Space)**.** A normed space $(V, \| \cdot \|)$ is said to be complete if every *Cauchy sequence* with respect to $\| \cdot \|$ is convergent.

Definition 6.2 (Cauchy sequence)**.** A sequence $\{v_n\}_{n=1}^{\infty}$ of elements in a normed space $(V, \| \cdot \|)$ is said to be a Cauchy sequence if for all $\epsilon > 0$ there exist $N \in \mathbb{N}$ such that $\|v_n - v_m\| < \epsilon$ for all $m, n \in \mathbb{N}$.

Further, we say that $\{v_n\}_{n=1}^{\infty}$ converges to $v \in V$ if $\|v_n - v\| \xrightarrow[n \to \infty]{} 0$.

We now introduce some Hilbert spaces that are natural to use for variational formulations of the boundary value problems we will consider. Let us start with the one-dimensional case. If $I = (a, b)$ is an interval, we define the space if "square integrable functions" on I:

$$L^2(I) = \left\{ v : I \to \mathbb{R} \;\middle|\; \int_I v^2 \, dx < \infty \right\} \tag{147}$$

The space $L^2(I)$ is a Hilbert space with the inner product

$$(v, w) = \int_I vw \, dx \tag{148}$$

and the corresponding norm

$$\|v\|_{L^2} = \sqrt{(v, v)} = \left(\int_I v^2 \, dx \right)^{1/2} \tag{149}$$

By Cauchy inequality $|(v, w)| \leq \|v\|_{L^2} \|w\|_{L^2}$ we see that (v, w) is well-defined, i.e., the integral (v, w) exists, if $v, w \in L^2(I)$.

Remark. To really appreciate the definition of $L^2(I)$ and realize that this space is complete requires some familiarity with the Lebesgue integral. In this notes, however, it is sufficient to get an idea of $L^2(I)$ by using the usual Riemann integral; from this point of view we may think of a "typical" function $v \in L^2(I)$ as a piecewise continuous function, possibly unbounded, such that $\|v\|_{L^2} < \infty$.

Example 6.1. We have that the function $v : I \to \mathbb{R}$ s.t. $v(x) = x^{-\beta}$ belongs to $L^2(I)$ if $\beta < \frac{1}{2}$. Hint: Apply the p-test of convergence for improper integrals. ■

We also introduce the space $H^1(I) = \{v \mid v, v' \in L^2(I)\}$, and we equip this space with the scalar product

$$(v, w)_{H^1} = \int_I vw \, dx + \int_I v'w' \, dx \qquad \forall v, w \in H^1(I) \tag{150}$$

and the corresponding norm

$$\|v\|_{H^1} = \int_I v^2 \, dx + \int_I (v')^2 \, dx \qquad \forall v \in H^1(I) \tag{151}$$

The space $H^1(I)$ thus consist of the functions v defined on I which together with their first derivatives are in $L^2(I)$.

In the case of boundary value problems of the form $-u'' = f$ on $I = (a, b)$ with $u(a) = u(b) = 0$, we shall use the space

$$H_0^1(I) = \{v \in H^1(I) \mid v(a) = v(b) = 0\} \tag{152}$$

with the same inner product and norm as for $H^1(I)$. Now our introductory BVP

$$\begin{cases} -u'' = f & x \in I = (0, 1) \\ u = 0 & x \in \{0, 1\} \end{cases} \tag{153}$$

con now be given the following variational formulation

$$\text{Find } u \in H_0^1(I) \text{ s.t. } (u', v') = (f, v) \qquad \forall v \in H_0^1(I) \tag{154}$$

If we compare (154) with the formulation (V) in section 2, we note that the space $H_0^1(I)$ is larger than the space V. The space $H_0^1(I)$ is specially tailored for a variational formulation of (153) and is in fact the largest space for which a variational formulation of the form (154) is meaningful.

From a mathematical point of view the “right” choice of function space is essential since this may make it easier to prove the existence of a solution to the continuous problem.

From the finite element point of view ther formulation (154) as opposed to (V) is of interest mainly because the basic error estimate for the finite element method is an estimate in the norm indicated by (154) i.e. the H^1 norm.

Now let Ω be a bounded domain $\mathbb{R}^d$ with $d = 2, 3$ and define

$$L^2(\Omega) = \left\{ v : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega v^2 \, dx < \infty \right\} \tag{155}$$

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) \;\middle|\; \frac{\partial v}{\partial x_i} \in L^2(\Omega), \;\; \forall i = 1, ..., d \right\} \tag{156}$$

and introduce the corresponding scalar products and norms

$$(v,w) = \int_\Omega vw\ dx, \qquad \|v\|_{L^2} = \sqrt{\int_\Omega v^2\ dx} \tag{157}$$

$$(v,w)_{H^1} = \int_\Omega vw\ dx + \int_\Omega \nabla v \cdot \nabla w\ dx, \qquad \|v\|_{H^1} = \sqrt{\int_\Omega v^2 + |\nabla v|^2\ dx} \tag{158}$$

We also define

$$H_0^1(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma\} \tag{159}$$

where Γ is the boundary of Ω and we equip $H_0^1(\Omega)$ with the same scalar product and norm as $H^1(\Omega)$.

The BVP

$$\begin{cases} -\Delta u = f & x \in \Omega \\ u = 0 & x \in \Gamma \end{cases} \tag{BVP}$$

can now be given the following variational formulation:

$$\text{Find } u \in H_0^1(\Omega) \text{ s.t. } a(u,v) = (f,v) \qquad \forall v \in H_0^1(\Omega) \tag{VF}$$

or equivalently

$$\text{Find } u \in H_0^1(\Omega) \text{ s.t. } F(u) \leq F(v) \qquad \forall v \in H_0^1(\Omega) \tag{MIN}$$

where

$$F(v) = \frac{1}{2}a(v,v) - f(f,v) \tag{160}$$

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v\ dx \tag{161}$$

Remark. The formulation (VF) is said to be a *weak formulation* of (BVP) and the solution of (VF) is said to be a *weak solution* of (BVP). If u is a weak solution of (BVP) then it is not immediately clear that u is also a clasical solution of (BVP), since this requires u to be sufficiently regular so Δu is defined in a classical sense.

The advantage mathematically of the weak formulation (VF) is that it is easy to prove the existence of a solution to (VF), whereas is relatively difficult to prove the existence of a classical solution to (BVP).

To prove the existence of a classical solution of (BVP) one usually starts with the weak solution of (BVP) and shows, often with considerable effort, that in fact this solution is sufficiently regular to be also a classical solution. For more complicated, e.g. non-linear problems, it may be extremely difficult or practically impossible to prove the existence of classical solutions whereas existence of weak solutions may still be within reach. ■

References

[1] Institutt for matematiske fag. Error analysis. https://wiki.math.ntnu.no/_media/tma4220/2018h/ch5.pdf, 2018.

[2] Pascal Frey. The finite difference method. https://www.ljll.fr/frey/cours/UdC/ma691/ma691_ch6.pdf, May 2017.

[3] C. Johnson. Numerical solution of partial differential equations by the finite element method. Studentlitteratur, 1994.