Numerical Analysis for Partial Differential Equations

Andrea Bonifacio

March 21, 2023

1 Boundary Value Problems

1.1 Weak Formulation

Let's consider a problem

$$\begin{cases} \mathcal{L}u = f & \text{in } \Omega \\ +\text{B.C.} & \text{on } \partial\Omega \end{cases} \tag{1.1}$$

- Ω: open bounded domain in $\mathbb{R}^d$, with $d = 2, 3$
- $\partial\Omega$: boundary of Ω
- f: given
- B.C. accordingly to $\mathcal{L}$
- $\mathcal{L}$: 2^{nd} order operator, like:

  (1) $\mathcal{L}u = -\text{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u$ (non-conservative form)

  (2) $\mathcal{L}u = -\text{div}(\mu\nabla u) + \text{div}(\mathbf{b}u) + \sigma u$ (conservative form)

  - $\mu \in L^\infty(\Omega), \quad \mu(\mathbf{x}) \geq \mu_0 > 0$ uniformly bounded from below
  - $\mathbf{b} \in (L^\infty(\Omega))^d$ transport term
  - $\sigma \in L^2(\Omega)$ reaction term
  - $f \in L^2(\Omega)$ can be less regular

General elliptic problems

Consider

$$\begin{cases} -\text{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u = f & \text{in } \Omega \quad g \in L^2(\Gamma_N) \\ u = 0 & \text{on } \Gamma_D \quad \partial\Omega = \Gamma_D \cup \Gamma_N \\ \mu\nabla u\cdot\mathbf{n} = g & \text{on } \Gamma_N \quad {\Gamma_D}^{\text{o}} \cap {\Gamma_N}^{\text{o}} = \varnothing \end{cases} \tag{1.2}$$

Suppose that $f \in L^2(\Omega)$ and $\mu, \sigma \in L^\infty(\Omega)$. Also suppose that $\exists\ \mu_0 > 0$ s.t. $\mu(\mathbf{x}) \geq \mu_0$, and $\sigma(\mathbf{x}) \geq 0$ a.e. on Ω. Then, given a test function v, we multiply the equation by v, and integrate on the domain Ω

$$\int_\Omega \left[-\text{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u\right] v \ = \int_\Omega fv$$

By applying Green's formula

$$\underbrace{\int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega \mathbf{b}\cdot\nabla u v + \int_\Omega \sigma uv}_{=:a(u,v)} = \int_\Omega fv + \underbrace{\int_{\Gamma_D} \mu\nabla u\cdot\mathbf{n} v}_{=0 \text{ if } v|_{\Gamma_D}=0} + \int_{\Gamma_N} \underbrace{\mu\nabla u\cdot\mathbf{n}}_{=g} v$$

So the weak formulation of the problem is

$$\begin{cases} \text{Find } u \in V & V = \{v \in H^1(\Omega), v|_{\Gamma_D} = 0\} =: H^1_{\Gamma_D}(\Omega) \\ a(u,v) = \langle F, v\rangle & \forall\ v \in V \end{cases} \tag{1.3}$$

where $a : V \times V \to \mathbb{R}$ is a bilinear form and $F : V \to \mathbb{R}$ is a linear form s.t. $\langle F, v\rangle \equiv F(v) = \int_\Omega fv + \int_{\Gamma_N} gv$.

Theorem 1.1 (Lax-Milgram)
Assume that

- V Hilbert space with $\|\cdot\|$ and inner product $(\cdot,\cdot)$
- $F \in V^* : |F(v)| \leq \|F\|_{V^*}\|v\| \ \forall\, v \in V$
- a continuous: $\exists\, M > 0 : |a(u,v)| \leq M\|u\|\,\|v\| \ \forall\, u, v \in V$
- a coercive: $\exists\, \alpha > 0 : a(v,v) \geq \alpha\|v\|^2 \ \forall\, v \in V$

Then, there exists a unique solution u of 1.3

Moreover

$$\alpha\|u\|^2 \leq a(u,u) = F(u) \leq \|F\|_{V^*}\|u\|$$

where α is the coercivity costant. Hence

$$\|u\| \leq \frac{\|F\|_{V^*}}{\alpha} \rightarrow \text{stability/continuous dependence on data}$$

But what if some of the assumptions of Lax-Milgram (in particular coercivity) are not satisfied? We need a slightly more general problem to formulate Nečas theorem:

$$\begin{cases} \text{find } u \in V \\ a(u,w) = \langle F, w\rangle \quad \forall w \in W \end{cases} \tag{1.4}$$

They belong to different spaces: W for the test function, V the solutions

Theorem 1.2 (Nečas)
Assume that $F \in W^*$. Consider the following conditions:

- a continuous: $\exists\, M > 0 : |a(u,w)| \leq M\|u\|_V\|w\|_W \ \forall\, u \in V, w \in W$
- $\inf - \sup$ condition: $\exists\, \alpha > 0 : \forall\, v \in V \quad \sup_{w \in W\setminus\{0\}} \frac{a(v,m)}{\|w\|_W} \geq \alpha\|v\|_V$
- $\forall\, w \in W, w \neq 0, \exists\, v \in V : a(v,w) \neq 0$

These conditions are necessary and sufficient for the existence and uniqueness of a solution of 1.4, for any $F \in W^*$. Moreover

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{W^*}$$

When $W = V$ Lax-Milgram provides necessary and sufficient conditions for existence and uniqueness of solutions.
Going back to

$$\begin{cases} \mathcal{L}u = f & \text{in } \Omega \\ +\text{B.C.} & \text{on } \partial\Omega \end{cases}$$

What could be our choice of V? Given that

$$u \in V : a(u,v) = F(v) \quad \forall\, v \in V$$

and

$$a(u,v) = \int_\Omega \mu \underbrace{\nabla u \nabla v}_{\nabla u, \nabla v \in L^2} + \int_\Omega b \underbrace{\nabla u v}_{\in L^1} + \int_\Omega \sigma \underbrace{uv}_{\in L^1}$$

We want to choose v in order to have all of these integrable

$$\Rightarrow V = \left\{ v \in L^2(\Omega), \nabla u \in \left[L^2(\Omega)\right]^d, v|_{\Gamma_D} = 0 \right\} = V_{\Gamma_D}$$

.
Knowing that a Sobolev space

$$H^1 = \left\{ v \in L^2(\Omega), \nabla u \in \left[L^2(\Omega)\right]^d \right\}$$

we can say $V_{\Gamma_D} = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}$, and if $\Gamma_D = \partial\Omega$, then $V_{\Gamma_D} = H_0^1$

1.2 Approximation

Recall for a moment the weak formulation of a generic elliptic problem

$$\begin{cases} \text{Find } u \in V \\ a(u, v) = \langle F, v \rangle \quad \forall\, v \in V \end{cases} \tag{1.5}$$

with V being an appropriate Hilbert space, subset of $H^1()$, $a(\cdot, \cdot)$ being a continuous and coercive bilinear forrm from $V \times V \to \mathbb{R}$, $F(\cdot)$ being a continuous linear functional from $V \to \mathbb{R}$.
Let $V_h \subset V$ be a family of spaces that depends on a parameter $h > 0$, such that $\dim V_h = N_h < \infty$. We can rewrite the weak formulation

$$\begin{cases} \text{Find } u_h \in V_h \\ a(u_h, v_h) = \langle F, v_h \rangle \quad \forall\, v_h \in V_h \end{cases} \tag{1.6}$$

and is called a **Galerkin problem**. Denoting with $\{\varphi_j, j = 1, 2, \ldots, N_h\}$ a basis of V_h, it is sufficient that the (1.6) is verified for each function of the basis. Also we need that

$$a(u_h, \varphi_i) = F(\varphi_i) \quad i = 1, 2, \ldots, N_h$$

Since $u_h \in V_h$

$$u_h(\mathbf{x}) = \sum_{j=1}^{N_h} u_j \varphi_j(\mathbf{x})$$

where u_j are unknown coefficients. Then

$$\sum_{j=1}^{N_h} u_j a(\varphi_j, \varphi_i) = F(\varphi_i)$$

We denote by A the matrix made by $a_{ij} = a(\varphi_j, \varphi_i)$ and $\mathbf{f}$ the vector of $F(\varphi_i) = f_i$ components. If we denote the vector $\mathbf{u}$ made by the unknown coefficients u_h.

$$A\mathbf{u} = \mathbf{f} \tag{1.7}$$

Theorem 1.3
The stiffness matrix A associated to the Galerkin discretization of an elliptic problem, whose bilinear form is coercive is positive definite.

Proof. Recall that a matrix $B \in \mathbb{R}^{n \times n}$ is said to be positive definite if

$$\mathbf{v}^T B \mathbf{v} \geq 0 \quad \forall\, \mathbf{v} \in \mathbb{R}^n$$

and

$$\mathbf{v}^T B \mathbf{v} = 0 \Leftrightarrow \mathbf{v} = \mathbf{0}$$

The correspondence

$$\mathbf{v} = (v_i) \in \mathbb{R}^{N_h} \longrightarrow v_h(x) = \sum_{j=1}^{N_h} v_j \varphi_j \in V_h$$

defines a bijection between V_h and $\mathbb{R}^{N_h}$. Given a generic vector $\mathbf{v} = (v_i)$ of $\mathbb{R}^{N_h}$, thanks to the bilinearity and coercivity of a we obtain

$$\begin{aligned}
\mathbf{v}^T A \mathbf{v} &= \sum_{j=1}^{N_h} \sum_{i=1}^{N_h} v_i a i_j v_j \\
&= \sum_{j=1}^{N_h} \sum_{i=1}^{N_h} v_i a(\varphi_j, \varphi_i) v_j \\
&= \sum_{j=1}^{N_h} \sum_{i=1}^{N_h} a(v_j \varphi_j, v_i \varphi_i) \\
&= a\left(\sum_{j=1}^{N_h} v_j \varphi_j \sum_{i=1}^{N_h} v_i \varphi_i \right) \\
&= a(v_h, v_h) \geq \alpha \| v_h \|_V^2 \geq 0
\end{aligned}$$

Moreover, if $\mathbf{v}^T A \mathbf{v} = 0$, then $\|v_h\|_V^2 = 0$. ★

Existence and uniqueness

Corollary 1.1
The solution of the Galerkin problem (1.6) exists and is unique.

To prove this we can prove that the solution to (1.7) exists and is unique. The matrix A is invertible as the unique solution of $A\mathbf{u} = \mathbf{0}$ is the null solution, meaning that A is definite positive.

Stability

Corollary 1.2
The Galerkin method is stable, uniformly with respect to h, by virtue of the following upper bound for the solution

$$\|u_h\|_V \leq \frac{1}{\alpha} \|F\|_{V^*}$$

The stability of the method guarantees that the norm $\|u_h\|_V$ of the discrete solution remains bounded for $h \to 0$. Equivalently it guarantees that $\|u_h - w_h\|_V \leq \frac{1}{\alpha} \|F - G\|_{V^*}$ with u_h and w_h being numerical solution corresponding to different data F and G.

Convergence

Lemma 1.1 (Galerkin orthogonality)

The solution u_h of the Galerkin method satisfies

$$a(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h \tag{1.8}$$

Proof. Since $V_h \subset V$, the exact solution u satisfies the weak problem (1.5) for each element $v = v_h \in V_h$, hence we have

$$a(u, v_h) = F(v_h) \forall\, v_h \in V_h \tag{1.9}$$

By subtracting side by side (1.6) from (1.9), we obtain

$$a(u, v_h) - a(u_h, v_h) = 0 \forall\, v_h \in V_h$$

from which the claim follows. ★

Also this can be generalized in the cases in which $a(\cdot, \cdot)$ is not symmetrcic. Consider the value taken by the bilinear form when both its arguments are $u - u_h$. If v_h is an arbitrary element of V_h we obtain

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h)$$

The last term is null by (1.8). Moreover

$$|a(u - u_h, u - v_h)| \leq M \|u - u_h\|_V \|u - v_h\|_V$$

having exploited the continuity of the bilinear form. Also by the coercivity

$$a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|_V^2$$

hence

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \|u - v_h\|_V \quad \forall\, v_h \in V_h$$

Such inequality holds for all functions $v_h \in V_h$ and therefore we find

$$\underbrace{\|u - u_h\|_V}_{\text{Galerkin error}} \leq \frac{M}{\alpha} \underbrace{\inf_{w_h \in V_h} \|u - w_h\|_V}_{\text{Best Approximation Error}} \tag{1.10}$$

In order for the method to converge, it is sufficient that, for $h \to 0$ the space V_h tends to saturate the entire space V.

$$\lim_{h \to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0 \quad \forall\, v \in V \tag{1.11}$$

In that case the Galerkin method is convergent and it can be written that

$$\lim_{h \to 0} \|u - u_h\|_V = 0 \Leftrightarrow \text{convergence}$$

Thie space V_h must be chosen carefully to satisfy the saturation property (1.11).

1.3 Finite Element Method

Partitions

1D Let us suppose that Ω is an interval (a, b). How to create an approximation of the space $H^1(a, b)$ that depend on a parameter h. Consider a partition $\mathcal{T}_h$ in $N + 1$ subintervals $K_j = x_{j-1}, x_j$, having width $h_j = x_j - x_{j-1}$ with

$$a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b \tag{1.12}$$

and set $h = \max_j h_j$.

2D Now we can extend the FEM for multi-dimensional problems. For simplicity we will consider $\Omega \subset \mathbb{R}^2$ with polygonal shapes $\mathcal{T}_h$. In this case the partition is called a triangulation. We can define the discretized domain

$$\Omega_h = \text{int}\left(\bigcup_{K\in\mathcal{T}_h} K\right)$$

in a way that the internal part of the union of the triangles $\mathcal{T}_h$. Having set $\text{diam}(K) = \max_{x,y\in K}|x-y| = h_k$. Also, given ρ_K the measure of the diameter of the circle inscribed in the triangle K, must be satisfied the condition that, for a suitable $\delta > 0$

$$\frac{h_k}{\rho_k} \leq \delta \quad \forall\, K \in \mathcal{T}_h \tag{1.13}$$

The condition (1.13) excludes very deformed triangles.

Definition 1.1 (Seminorms)
A seminorm is defined as

$$|f|_k = |f|^k_H(\Omega) = \sqrt{\sum_{|\alpha|=k}\int_\Omega (D^\alpha f)^2\, d\Omega}$$

In particular

$$\begin{aligned}
\text{1D:}\quad & |u|_{H^1(a,b)} = \left(\|u_x\|^2_{L^2(a,b)}\right)^{\frac{1}{2}} = \|u_x\|^2_{L^2(a,b)} \\
& |u|_{H^2(a,b)} = \|u_{xx}\|^2_{L^2(a,b)} \\
\text{2D:}\quad & |u|_{H^1(a,b)} = \left(\|u_x\|^2_{L^2(a,b)} + \|u_y\|^2_{L^2(a,b)}\right)^{\frac{1}{2}} \\
& |u|_{H^1(a,b)} = \left(\|u_{xx}\|^2_{L^2(a,b)} + \|u_{xy}\|^2_{L^2(a,b)} + \|u_{yx}\|^2_{L^2(a,b)} + \|u_{yy}\|^2_{L^2(a,b)}\right)^{\frac{1}{2}}
\end{aligned}$$

Always true that $|u|_{H^q} \leq \|u\|_{H^q}$

The problem is always:

$$\begin{aligned}
&\text{find } u_h \in V_h : a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h \\
&\qquad\qquad \downarrow \\
&\qquad\qquad V_h = \{v_h \in X^r_h : v_h|_{\Gamma_D} = 0\} \quad r \geq 1
\end{aligned} \tag{1.14}$$

Since the functions of $H^1(a,b)$ are continuous on $[a,b]$, it is possible to create the family of spaces

$$X^r_h = \left\{v_h \in C^0\left(\overline{\Omega}\right) : v_h|_{K_j} \in \mathbb{P}_r \;\; \forall\, K_j \in \mathcal{T}_h\right\}, \quad r = 1, 2, \ldots \tag{1.15}$$

having denoted by $\mathbb{P}_r$ the space of polynomials with degree lower or equal to r in the variable x. All these spaces are subspaces of $H^1(a,b)$ as they are constituted by differentiable functions except for at most a finite number of points (the vertices of the partition). It is convenient to sekect a basis for the X^r_h space that is *Lagrangian*.

$\mathbb{P}^r$: 1D $$p(x) = \sum_{k=0}^{r} a_k x^k$$ intervals

2D $$p(x_1, x_2) = \sum_{\substack{k,m=0 \\ k+m \leq r}}^{r} a_{km} x_1^k x_2^m$$ triangles

3D $$p(x_1, x_2, x_3) = \sum_{\substack{k,m,n=0 \\ k+m+n \leq r}}^{r} a_{kmn} x_1^k x_2^m x_3^n$$ tetrahedra

The space X_h^1

The space is constituted by the functions of the partition (1.12). Since only a straight line can pass through different points, the degrees of freedom (DOF, the number of values we need to assign to the basis to define the functions) of the functions will be equal to the number $N+2$ of vertices of the partition. It follows naturally that $\{\varphi_i\}, i = 0, 1, \ldots, N, N+1$. In this case the basis functions are characterized by the following properties

$$\varphi_i \in X_h^1 \text{ s.t } \varphi_i(x_j) = \delta_{ij}, \quad i, j = 0, 1, \ldots, N, N+1$$

where δ_{ij} is the Kronecker delta. So we have our basis function that have value 1 in the node x_j and 0 elsewhere.
The formula for the basis function is then given by

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{for } x_{i-1} \leq x \leq x_i \\ \frac{x - x_{i+1}}{x_{i+1} - x_i} & \text{for } x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{1.16}$$

The space X_h^2

In this case polynomials are of degree 2, so the points necessary to evaluate them are 3. The chosen points for every element of the partition $\mathcal{T}_h$. The nodes from the interval goes from $a = x_0$ to $b = x_{2N+2}$, so that midpoints are the nodes with odd indices. As the previous case the basis is Lagrangian

$$\varphi_i \in X_h^2 \text{ s.t } \varphi_i(x_j) = \delta_{ij}, \quad i, j = 0, 1, \ldots, 2N+2$$

The space V_h

This space is generated by

$$V_h = \{v_h \in X_h^r : v_h(a) = v_h(b) = 0\}$$

Having defined a basis $\{\varphi_j(\mathbf{x})\}_{j=1}^{N_h}$ for the space V_h, each v_h can be expanded as a linear combination of elements of the basis, suitably weighted by coefficients $\{v_j\}_{j=1}^{N_h}$

$$v_h(\mathbf{x}) = \sum_{j=1}^{N_h} v_j \varphi_j(\mathbf{x})$$

A basis is called Lagrangian if it satisfies the following properties

$$\varphi_i(\mathbf{x}_j) = \delta_{ij} \quad \forall\ 1 \leq i, j \leq N_h$$

and then the following property holds:

$$v_h(\mathbf{x}_j) = v_j \quad \forall\ 1 \leq i, j \leq N_h$$

The solution of the Finite Element Method, u_h can be written as

$$u_h(\mathbf{x}) = \sum_{j=1}^{N_h} u_j \varphi_j(\mathbf{x}) \tag{1.17}$$

In (1.14) take $v_h = \varphi_j\ \forall\ j = 1, \ldots, N_h$ such that $a(u_h, \varphi_i) = F(\varphi_i)\ \forall\ i = 1, \ldots, N_h$. Then use (1.17) to obtain

$$\begin{aligned}
& a\left(\sum_{j=1}^{N_h} u_j \varphi_j(\mathbf{x}), \varphi_i\right) = \underbrace{F(\varphi_i)}_{F_i} \\
& \Rightarrow \sum_{j=1}^{N_h} \underbrace{a(\varphi_j, \varphi_i)}_{\substack{a_{ij} \text{ elements} \\ \text{of } A}} u_j(\mathbf{x}) = F_i \quad i = 1, \ldots, N_h \\
& \Rightarrow A\,\mathbf{u} = \mathbf{F}
\end{aligned}$$

Which is a linear system of dimension $N_h \times N_h$ with $\mathbf{F}$ the right hand side (RHS), A the stiffness matrix and $\mathbf{u}$ a vector of unknown nodal values of the solution u_h.