Numerical Analysis for Partial Differential Equations

Andrea Bonifacio

July 4, 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 $\mathcal{L}$
- $\mathcal{L}$: 2nd 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

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

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

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

By applying Green's formula

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

So the weak formulation of the problem is

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

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

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: $\exists \, M > 0 : |a(u,v)| \leq M \|u\|_V \|v\|_V \; \forall \, u, v \in V$
- a coercive: $\exists \, \alpha > 0 : a(v,v) \geq \alpha \|v\|_V^2 \; \forall \, v \in V$

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 constant. Hence,

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

But what if some 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

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

.
Knowing that a Sobolev space

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

we can say $V_{\Gamma_D} = \left\{ v \in H^1(\Omega) : v|_{\Gamma_D} = 0 \right\}$, 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 $H^1(\Omega), a(\cdot,\cdot)$ being a continuous and coercive bilinear form 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

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

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

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

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

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

from which the claim follows. ★

Also, this can be generalized in the cases in which $a(\cdot,\cdot)$ is not symmetric. 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

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

This 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 multidimensional 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

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

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:

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

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

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

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 select a basis for the X_h^r space that is *Lagrangian*.

$$\begin{array}{llll}
\mathbb{P}^r: & \text{1D} & p(x) = \displaystyle\sum_{k=0}^{r} a_k x^k & \text{intervals} \\[2ex]
& \text{2D} & p(x_1, x_2) = \displaystyle\sum_{\substack{k,m=0 \\ k+m \leq r}}^{r} a_{km} x_1^k x_2^m & \text{triangles} \\[2ex]
& \text{3D} & p(x_1, x_2, x_3) = \displaystyle\sum_{\substack{k,m,n=0 \\ k+m+n \leq r}}^{r} a_{kmn} x_1^k x_2^m x_3^n & \text{tetrahedra}
\end{array}$$

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

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

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.

1.4 Advection Diffusion Reaction Problem

$$\begin{cases} Lu = \underbrace{-\mathrm{div}(\mu\nabla u)}_{\text{diffusion}} + \underbrace{\mathbf{b}\cdot\nabla u}_{\text{advection}} + \underbrace{\sigma u}_{\text{reaction}} = f & \text{in } \Omega \\ \text{BC} & \text{on } \partial\Omega \end{cases}$$

Lax-Milgram tells us that if $\sigma - \frac{1}{2}\mathrm{div}\mathbf{b} \geq \gamma > 0$ then $\exists!$ a solution to the problem. But what if these conditions are not satisfied? We can use Nečas theorem (1.2) with equivalent assumptions:

- Weak coercivity (Gårding inequality):

$$\exists\, \alpha, \lambda : a(v,v) \geq \alpha\|v\|_V^2 - \|v\|_{L^2(\Omega)}^2 \quad \forall\, v \in V$$

- Uniqueness condition (typically proven by maximum principle):

$$(a(u,v) = 0\ \forall\, v \in V) \Rightarrow u = 0$$

If A is spd (symmetric positive defined) then $K_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$

Proposition 1.1
If $a(\cdot,\cdot)$ is symmetric and coercive, then A is spd.

Proof. Symmetry: $A_{ij} = a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j) = A_{ji}$
$\forall\, \mathbf{v} \in \mathbb{R}^{N_h}$:

$$\begin{aligned} \mathbf{v}^T A\mathbf{v} &= \sum_{i,j} A_{ij} v_i v_j = \sum_{i,j} a(\varphi_j, \varphi_i) v_i v_j \\ &= a(\sum_j v_j\varphi_j, \sum_i v_i\varphi_i) = a(v_h, v_h) \geq \alpha\|v_h\|^2 > 0 \end{aligned}$$

if $(v_h \neq 0 \Leftrightarrow \mathbf{v} \neq \mathbf{0})$. Hence, A is positive defined. ★

Definition 1.2
If A is spd, we define the **A-norm** of $\mathbf{v}$ as

$$\begin{aligned} \|v\|_A &:= (A\mathbf{v}, \mathbf{v})^{\frac{1}{2}} \\ &= \left(\sum_{i,j} a_{ij} v_i v_j\right)^{\frac{1}{2}} \end{aligned}$$

Since A is positive defined $\Rightarrow \mathrm{Re}(\lambda_k(A)) \Rightarrow \lambda_k(A) \neq 0$. Then, by symmetry of $A \Rightarrow \lambda_k(A) \in \mathbb{R}$. Combining the two we have that A sdp $\Rightarrow \lambda_k(A) > 0 \Rightarrow \exists!$ solution of $A\mathbf{u} = \mathbf{f}$

Definition 1.3
If A is sdp, then $K_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$ is called **spectral condition number**

If $K_2(A) \gg 1 \Rightarrow A$ is ill-conditioned $\Rightarrow$ solving $A\mathbf{u} = \mathbf{f}$ is hard.
We can also prove that $\exists\, C_1, C_2 > 0 : \forall\, \lambda_h$ eigenvalue of A:

$$\alpha C_1 h^d \leq \lambda_h \leq M C_2 h^{d-2} \qquad d = 1,2,3$$

whence

$$\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leq \frac{MC_2}{\alpha C_1} h^{-2}$$

Then

$$K_2(A) = \mathcal{O}(h^{-2})$$

If we use the conjugate gradient method to solve $A\mathbf{u} = \mathbf{f}$, then:

$$\left\|\mathbf{u}^{(k)} - \mathbf{u}\right\|_A \leq 2\left(\frac{\sqrt{K_2(A)}+1}{\sqrt{K_2(A)}-1}\right)^k \left\|\mathbf{u}^{(k)} - \mathbf{u}\right\|_A$$

Same with gradient method, with $K_2(A)$ instead of $\sqrt{K_2(A)} \Rightarrow$ need for preconditioners.

1.5 Interpolant estimates

$$\|u-u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u-v_h\| \underset{\text{saturation}}{\longrightarrow} 0 \Leftrightarrow \text{convergence} \tag{1.18}$$

But how fast it saturates?
Note: $\inf_{v_h \in V_h} \|u-v_h\|_V \leq \|u-\bar{u}_h\|_V \ \forall\ \bar{u}_h$ suitable chosen in V_h and $\bar{u}_h$ is a smart guy chosen in a smart way (close enough to u).
In 1D the finite element interpolant can be defined as $\prod_h^r u(x_k) = u(x_k) \ \forall\ x_k$ node. Then $\bar{u}_h = \prod_h^r u \in V_h$.
How good is $\bar{u}_h$?

$$\prod_h^r u(x) = \sum_{j=1}^{N_h} u(x_j)\varphi_j(x)$$

which is a good approximation.

Interpolant error estimates

Then, for $m = 0, 1\ \exists\ C = C(r, m, \hat{k})$ s.t.

$$|v - \textstyle\prod_h^r v|_{H^m(\Omega)} \leq C \left(\sum_{K \in \mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{H^{r+1}(K)} \right)^{\frac{1}{2}} \tag{1.19}$$

where $h_K = \text{diam}(K)$ and $h_K \leq h\ \forall\ K$ this yields:

$$|v - \textstyle\prod_h^r v|_{H^m(\Omega)} \leq C h^{r+1-m} |v|_{H^{r+1}(K)} \quad \forall\ v \in H^{r+1}(\Omega), m = 0, 1 \tag{1.20}$$

Recall also that

$$\begin{aligned} \|u-u_h\| &= \|u-u_h\|_{H^1(\Omega)} \\ &\leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u-v_h\| \\ &\leq \frac{M}{\alpha} \|u - \textstyle\prod_h^r u\|_{H^1(\Omega)} \end{aligned}$$

Using (1.19) we obtain

$$\|u-u_h\| \leq C\frac{M}{\alpha} \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |v|^2_{H^{r+1}(K)} \right)^{\frac{1}{2}} \tag{1.21}$$

Then, by using (1.20):

$$\|u-u_h\| \leq C\frac{M}{\alpha} h^r |u|_{H^{r+1}(\Omega)} \tag{1.22}$$

Definition 1.4
Consider a bilinear form $a : V \times V \to \mathbb{R}$. The *adjoint* form a^* is defined as $a^* : V \times V \to \mathbb{R}$

$$a^*(v, w) = a(w, v) \quad \forall\ v, w \in V$$

Now let's consider the adjoint problem

$$\begin{cases} \text{find } \varphi = \varphi(g) \in V & \forall\ g \in L^2(\Omega) \\ a^*(\varphi, v) = (g, v) = \int_\Omega gv & \forall\ v \in V \end{cases} \tag{1.23}$$

Assuming that $\varphi \in H^2(\Omega) \cap V$ (elliptic regularity). Consider now, for example, $\mathcal{L} = -\Delta$. Then the solution of

$$\begin{cases} -\Delta\varphi = g & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

satisfies $\varphi \in H^2(\Omega)$. Moreover,

$$\exists\, C_1 > 0 : \|\varphi(g)\|_{H^2(\Omega)} \leq C_1 \|g\|_{L^2(\Omega)} \tag{1.24}$$

Take now $g = e_h = u - u_h$ in (1.23). Then

$$\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &= a^*(\varphi, e_h) = a(e_h, \varphi) \\
&= a(e_h, \varphi - \varphi_h) && \text{(Galerkin orthogonality)} \\
&\leq M \|e_h\|_{H^1(\Omega)} \|\varphi - \varphi_h\|_{H^1(\Omega)}
\end{aligned}$$

Take then $\varphi_h = \prod_h^1 \varphi$:

$$\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &\leq M \|e_h\|_{H^1(\Omega)} \left\| \varphi - \textstyle\prod_h^1 \varphi \right\|_{H^1(\Omega)} \\
&\leq M \|e_h\|_{H^1(\Omega)} C_2 h \, |\varphi|_{H^2(\Omega)} && \text{(for (1.20) with m=r=1)} \\
&\leq M \|e_h\|_{H^1(\Omega)} C_2 h C_1 \|e_h\|_{L^2(\Omega)} && \text{(for (1.24))}
\end{aligned}$$

Whence:

$$\begin{aligned}
\|e_h\|_{L^2(\Omega)} &\leq C_1 C_2 h h \|e_h\|_{H^1(\Omega)} \\
&\leq M C_1 C_2 h C_3 h^r \, |u|_{H^{r+1}(\Omega)} && \text{(for (1.22))}
\end{aligned}$$

So

$$\|e_h\|_{L^2(\Omega)} \leq \overline{C} h^{r+1} \, |u|_{H^{r+1}(\Omega)} \tag{1.25}$$

2 Spectral Element Method

2.1 Introduction

The problem with the Finite Element Method is that the rate of convergence is limited by the degree of the polynomials used. An alternative can be the Spectral Element Method, for which the convergence rate is limited by the regularity of the solution.

2.2 Legendre's polynomials

The Legendre polynomials $\{L_k(x) \in \mathbb{P}_k, k = 0, 1, \ldots\}$ are the eigenfunctions of the singular Sturm-Liouville problem:

$$((1-x^2)L_k'(x))' + k(k+1)L_k(x) = 0 \quad -1 < x < 1$$

So they satisfy the recurrence relation

$$\begin{aligned} &L_0(x) = 1,\ L_1(x) = x,\ \text{and for } k \geq 1 \\ &L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x) \end{aligned} \tag{2.1}$$

Given a weight function $w(x) \equiv 1$, they are mutually orthogonal with respect to it on the interval $(-1, 1)$

$$\int_{-1}^{1} L_k(x) L_m(x)\, dx = \begin{cases} \frac{2}{2k+1} & \text{if } k = m \\ 0 & \text{if } k \neq m \end{cases}$$

The expansion of $u \in L^2(-1, 1)$ in terms of L_k is

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k L_(x)$$

Given that $(f, g) = \int_{-1}^{1} fg\, dx$ we know that:

$$(u, L_m) = \sum_{k=0} \hat{u}_k (L_k, L_m) \underset{\text{orth.}}{=} \hat{u}_m \frac{2}{2m+1} \Rightarrow \hat{u}_k = \frac{2k+1}{2} \int_{-1}^{1} u L_k\, dx$$

The truncated Legendre series of u is the $L^2 -$ projection of u over $\mathbb{P}_N$ is

$$P_N u = \sum_{k=0}^{N} \hat{u}_k L_k \tag{2.2}$$

Given any $u \in H^s(-1, 1)$ with $s \in N$, the projection error $(u - P_N u)$ satisfies the estimates

$$\begin{aligned} \|u - P_N u\|_{L^2(-1,1)} &\leq C N^{-s} \|u\|_{H^s(-1,1)} && \forall\, s \geq 0 \\ \|u - P_N u\|_{L^2(-1,1)} &\leq C N^{-s} \, |u|_{H^s(-1,1)} && \forall\, s \leq N+1 \end{aligned}$$

There is also a "modified" Legendre basis for function that vanish at ± 1. This is because the Legendre basis is not suited to impose Dirichlet B.C.

$$\begin{aligned} \psi_0(x) &= \frac{1}{2}(L_0(x) - L_1(x)) = \frac{1-x}{2} \\ \psi_N(x) &= \frac{1}{2}(L_0(x) + L_1(x)) = \frac{1+x}{2} \\ \psi_{k-1}(x) &= \frac{1}{\sqrt{2(2k-1)}}(L_{k-2}(x) - L_k(x)) \\ &\text{for } k = 2, \ldots, N \ \ -1 < x < 1 \end{aligned}$$

2.3 Spectral Galerkin formulation

Given $\Omega = (-1,1), \mu, b, \sigma > 0$ const., $f : \Omega \to \mathbb{R}$. Look for $u : \Omega \to \mathbb{R}$ s.t.

$$\begin{cases} -(\mu u')' + (bu)' + \sigma u = f & \text{in } \Omega \\ u(-1) = 0 \\ u(1) = 0 \end{cases}$$

Set $V = H_0^1(\Omega)$, then the weak form of the differential problem reads:

$$\text{find } u \in V \text{ s.t } a(u,v) = (f,v)_{L^2(\Omega)} \quad \forall\, v \in V,\ f \in L^2(\Omega)$$

where

$$a(u,v) = \int_\Omega (\mu u' - bu)v'\,dx + \int_\Omega \sigma uv\,dx$$
$$(f,v)_{L^2(\Omega)} = \int_\Omega fv\,dx$$

Now set $V_N = \mathbb{P}_N^0$

$$\text{find } u_N \in V_N : a(u_N, v_N) = (f, v_N)_{L^2(\Omega)} \tag{2.3}$$

Now expand $u_N(x) = \sum_{k=1}^{N-1} \tilde{u}_k \psi_k(x)$ and chose $v_N = \psi_i(x)$ for any $i = 1, \dots, N-1$. The discretization of the problem reads:

$$\text{find } u = [\tilde{u}]_{k=1}^{N-1} : \sum_{k=1}^{N-1} a(\varphi_k, \psi_i)\tilde{u}_k = (f, \psi_i)_{L^2(\Omega)} \quad \text{for any } i = 1, \dots, N-1$$

Given $u_N \in V_N$ the solution of the problem, then if $u \in H^{s+1}(\Omega)$ with $s \geq 0$, thanks to Ceà Lemma, holds that:

$$\|u - u_N\|_{H^1(\Omega)} \leq C(s) \left(\frac{1}{N}\right)^s \|u\|_{H^{s+1}(\Omega)}$$

So u_N converges with spectral accuracy with respect to N. But doing so, we would have two full matrices, the stiffness one and the mass one $M_{ij} = (\psi_j, \psi_i)_{L^2(-1,1)}$ are quite expensive to compute or invert.
To solve this we can use a Lagrange nodal basis instead of a modal one, by using the Legendre-Gauss-Lobatto quadrature formulas. In this case we need a Legendre polynomial $L_N(x)$.
Given an $L_N(x)$ polynomial, we can put one node at each end of the domain, so $x_0 = -1, x_N = 1$ and $x_j =$ zeros of L_N' with $j = 1, \dots, N-1$. We also need a set of weights $w_j = \frac{2}{N(N+1)} \frac{1}{[L_N(x_j)]^2}$ with $j = 0, \dots, N$.
With this set of nodes and weights it's possible to obtain the following interpolatory quadrature formula

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=0}^{N} f(x_j) * w_j$$

The degree of exactness of this method is $2N-1$, meaning that

$$\int_{-1}^{1} f(x)\,dx = \sum_{j=0}^{N} f(x_j) w_j \quad \forall\, f \in \mathbb{P}_{2N-1}$$

Some useful operation with LGL nodes

- Discrete inner product in $L^2(-1,1)$:

$$(u,v)_N = \sum_{j=0}^{N} u(x_j)v(x_j)w_j$$

  with degree of exactness $2N-1$

$$(u,v)_{L^2(\Omega)} = (u,v)_N \quad \text{only if } u, v \in \mathbb{P}_{2N-1}$$

- Discrete norm in $L^2(-1,1)$

$$\|u\|_N = (u,u)_N^{\frac{1}{2}}$$

with the following norm equivalence: $\exists\, c_1, c_2 > 0$ s.t.

$$c_1\|v_N\|_{L^2(-1,1)} \leq \|v_N\|_N \leq c_2\|v_N\|_{L^2(-1,1)} \quad \forall\, v_N \in \mathbb{P}_N$$

Given $\{\varphi_0, \dots, \varphi_N\}$ characteristic Lagrange polynomials in $\mathbb{P}_N$ w.r.t the LGL nodes. Then

$$\varphi_j = \frac{1}{n(n+1)} \frac{(1-x^2)}{(x_j - x)} \frac{L_N'(x)}{L_N(x_j)} \quad \text{for } j = 0, \dots, N$$

Also true that $\varphi_j(x_k) = \delta_{kj}$ and $\{\varphi_j\}$ are orthogonal w.r.t. the discrete inner product $(\cdot,\cdot)_N$, meaning that the mass matrix M is diagonal. Given $\{w_i\}$ the set of weights, then

$$M_{ij} = (\varphi_j, \varphi_i)_N = \delta_{ij} w_i \quad i,j = 0, \dots, N$$

2.4 Galerkin with Numerical Integration

We can now define the spectral Galerkin method with numerical integration (GNI), by setting our bilinear discrete form as $a_N(u_N, v_N) = (\mu u_N' - b u_n, v_N')_N + (\sigma u_n, v_n)_N$, and the problem as

$$\text{find } u_N^{\text{GNI}} \in V_N : a_N(u_N^{\text{GNI}}, v_N) = (f, v_N)_N \quad \forall\, v_N \in V_N$$

Then, by the same expansion w.r.t. the Lagrange basis: $u_N^{\text{GNI}}(x) = \sum_{i=0}^N u_N^{\text{GNI}}(x_i)\varphi_i(x)$ and choose $v_N(x) = \varphi_i(x)$ for any $i = 1, \dots, N-1$.
The GNI discretization of the weak problem reads:

$$\text{look for } u^{\text{GNI}} = \left[u_N^{\text{GNI}}(x_j)\right]_{j=0}^N : \begin{cases} u_N^{\text{GNI}}(x_0) = u_N^{\text{GNI}}(x_N) \\ \sum_{j=0}^N a_N(\varphi_j, \varphi_i) u_N^{\text{GNI}}(x_j) = (f, \varphi_i)_N \quad \forall\, i = 1, \dots, N-1 \end{cases}$$

Now let's have a closer look to the $\{\varphi_j\}$:

$$\varphi_j \in \mathbb{P}_N : \varphi_j(x_i) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Given the discrete inner product $(u,v)_N = \sum_{j=0}^N u(x_j)v(x_j)w_j$ we can write:

$$\begin{aligned} (\varphi_k, \varphi_m)_N &= \sum_{j=0}^N \underbrace{\varphi_k(x_j)}_{\delta_{kj}} \underbrace{\varphi_m(x_j)}_{\delta_{mj}} w_j \quad 0 \leq k, m \leq N \\ &= \sum_{k=0}^N = \begin{cases} w_m & \text{if } k = m \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

so $\{\varphi_k\}$ is orthogonal under the discrete inner product.
The GNI solution is

$$u_N(x) = \sum_{i=0}^N \alpha_i \varphi_i(x) \quad \{\alpha_i\} \text{ unknown coefficients}$$

Set now $x = x_j$ with LGL nodes:

$$u_N(x_j) = \sum_{i=0}^N \alpha_i \underbrace{\varphi_i(x_j)}_{\delta ij} = \alpha_j$$

So, given $u_n^{\text{GNI}}(x_j)$ the nodal values, we obtain the nodal expansion:

$$u_N^{\text{GNI}}(x) = \sum_{j=0}^N u_N^{\text{GNI}}(x_j)\varphi_j(x)$$

Algebraic form of Spectral GNI

Now it's about solving the following linear system

$$A^{\text{GNI}}\mathbf{u}^{\text{GNI}} = \mathbf{f}^{\text{GNI}}$$

with $A_{ij}^{\text{GNI}} = a_N(\varphi_j, \varphi_i)$ for $i = 1, \dots, N-1, j = 0, \dots, N$ and $\mathbf{f}^{\text{GNI}} = (f, \varphi_i)_N$ for $i = 1, \dots, N-1$:

$$A^{\text{GNI}} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & & & \vdots \\ \vdots & & a_N(\varphi_j, \varphi_i) & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad \mathbf{f}^{\text{GNI}} = \begin{bmatrix} 0 \\ \vdots \\ f_i^{\text{GNI}} \\ \vdots \\ 0 \end{bmatrix}$$

Given that $a(u,v) = \int_{-1}^{1} \mu u'v' - \int_{-1}^{1} buv' + \int_{-1}^{1} \sigma uv$ and $(f,v) = \int_{-1}^{1} fv$. We established that $a_n(u,v) = (\mu u', v')_N - (bu, v')_N + (\sigma u, v)_N$ and that $(f,v)_N = (f,v)_N$, so we obtain

$$A_{ij}^{\text{GNI}} = a_N(\varphi_j, \varphi_i) = \underbrace{\left(\mu\varphi_j', \varphi_i'\right)_N}_{\text{A}} - \underbrace{\left(b\varphi_j, \varphi_i'\right)_N}_{\text{B}} + \underbrace{\left(\sigma\varphi_j, \varphi_i\right)_N}_{\text{C}}$$

Assuming $\mu, b, \sigma \in \mathbb{R}$ we have that

$$C : \sigma(\varphi_j, \varphi_i)_N = \sigma\delta_{ij}w_i = \begin{cases} \sigma w_i & i = j \\ 0 & i \neq j \end{cases} \qquad \rightarrow \quad M = \sigma \underbrace{\begin{bmatrix} w_0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_N \end{bmatrix}}_{\text{diagonal weight matrix}}$$

$$B : -b(\varphi_j, \varphi_i')_N = -b\sum_{k=0}^{N} \underbrace{\varphi_j(x_k)}_{\delta_{jk}} \underbrace{\varphi_i'(x_k)}_{D_{ki}\neq 0} w_k \qquad \rightarrow \quad \text{full matrix}$$

$$A : \mu(\varphi_j', \varphi_i')_N = \mu\sum_{k=0}^{N} \underbrace{\varphi_j'(x_k)}_{D_{kj}} \underbrace{\varphi_i'(x_k)}_{D_{ki}} w_k \qquad \rightarrow \quad \text{full matrix}$$

where $D = (D_{ki}) = \varphi_k'(x_i)$ is the differentiation matrix that can be computed only once. The computation of $(f, \varphi_i)_N$ can be made this way

$$(f, \varphi_i)_N = \sum_m w_m f(x_m) \underbrace{\varphi_i(x_m)}_{\delta_{im}} = w_i f(x_i)$$

In conclusion the GNI method is still as full as the spectral one, but much easier to compute thanks to the nodal expansion.

Accuracy

We can define the Global Lagrange polynomial of degree N that interpolates u at LGL nodes as:

$$I_n u(x) = \sum_{j=0}^{N} u(x_j)\varphi_j(x)$$

And the interpolation error, for any $u \in H^{s+1}(-1,1)$ with $s \geq 0$, the interpolation error $u - I_N u$ satisfies the estimate:

$$\|u - I_N u\|_{H^k(-1,1)} \leq C(s)\left(\frac{1}{N}\right)^{s+1-k} \|u\|_{H^{s+1}(-1,1)} \quad \text{for } k = 0, 1$$

One important feature of LGL nodes is that they are not uniformly spaced (otherwise there could be problems), so that

$$I_n u(x_k) = u(x_k) \quad 0 \le k \le N$$

It's also possible to estimate the L^2 norm of the error as:

$$\|u - I_N u\|_{L^2(-1,1)} \le C(s) \left(\frac{1}{N}\right)^{s+1} \|u\|_{H^{s+1}(-1,1)} \quad s \ge 1$$

Theorem 2.1 (Quadrature error)
$\exists\, c > 0 : \forall\, f \in H^q(-1,1)$, with $q \ge 1$, $\forall\, v_N \in \mathbb{P}_N$ it holds

$$\left| \int_{-1}^{1} f v_N \, dx - (f, v_N)_N \right| \ge c \left(\frac{1}{N}\right)^q \|f\|_{H^q(-1,1)} \|v_N\|_{L^2(-1,1)}$$

Let now $u_N^{\mathrm{GNI}} \in V_N$ be the solution of

$$a_N(u_N^{\mathrm{GNI}}, v_N) = (f, v_N)_N \quad \forall\, v_N \in V_N$$

If $u \in H^{s+1}(\Omega)$ and $f \in H^s(\Omega)$ with $s \ge 0$, then:

$$\left\|u - u_N^{\mathrm{GNI}}\right\|_{H^1(\Omega)} \le C(s) \left(\frac{1}{N}\right)^s \left(\|u\|_{H^{s+1}(\Omega)} + \|f\|_{H^s(\Omega)}\right)$$

So u_N^{GNI} converges with spectral accuracy w.r.t. to N to the exact solution when the latter is smooth.

General ideas

The ideas proposed until now are the followings:

(WP)	V Hilbert	a bilinear form	F functional
(SG)	V_h instead of V	same a	same F
(GNI)	V_N	a_N	F_N

- For the Galerkin method one can use Ceà Lemma

$$\begin{aligned} \|u - u_N\|_{H^1(\Omega)} &\le \underbrace{\inf_{v_N \in V_N} \|u - v_N\|_{H^1(\Omega)}}_{\text{distance of } V \text{ from } V_N} \\ &\le \|u - I_n u\|_{H^1(\Omega)} \end{aligned}$$

- For the Galerkin with Numerical Integration we need something more:

$$\begin{aligned} \|u - u_N\|_{H^1(\Omega)} \le \quad & \text{“distance” of } V \text{ from } V_N \\ & + \text{“distance” of } a(\cdot,\cdot) \text{ from } a_N(\cdot,\cdot) \\ & + \text{“distance” of } F(\cdot) \text{ from } F_N(\cdot) \end{aligned}$$

2.5 Strang Lemma

Lemma 2.1 (Strang lemma)
Consider the problem

$$\text{find } u \in V : a(u,v) = F(v) \quad \forall\, v \in V \tag{2.4}$$

and its approximation

$$\text{find } u_h \in V_h : a_h(u_h, v_h) = F_h(v_h) \quad \forall\, v_h \in V_h \tag{2.5}$$

with $\{V_h\}$ being a family of subspaces of V. Suppose that $a_h(\cdot,\cdot)$ is continuous on $V_h \times V_h$ and uniformly

coercive on V_h meaning that:

$$\exists\, \alpha^* > 0 \text{ independent of } h : a_h(v_h, v_h) \geq \alpha^* \|v_h\|_V^2 \quad \forall\, v_h \in V_h$$

Also suppose that F_h is linear and bounded on V_h. Then:

- Exist a unique solution u_h to the problem.
- Such solution depends continuously on the data, i.e. we have

$$\|u_h\|_V \leq \frac{1}{\alpha^*} \sup_{v_h \in V_h \setminus \{0\}} \frac{F_h(v_h)}{\|v_h\|_V}$$

- finally, the following a priori error estimate holds

$$\begin{aligned}
\|u - u_h\|_V \quad &\leq \inf_{w_h \in V_h} \left\{ \left(1 + \frac{M}{\alpha^*}\right) \|u - w_h\|_V \right. \\
&\quad \left. + \frac{1}{\alpha^*} \sup_{v_h \in V_h \setminus \{0\}} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_V} \right\} \\
&\quad + \frac{1}{\alpha^*} \sup_{v_h \in V_h \setminus \{0\}} \frac{|F(v_h) - F_h(v_h)|}{\|v_h\|_V}
\end{aligned}$$

with M being the continuity constant of $a(\cdot, \cdot)$

Proof. The assumption of Lax-Milgram are satisfied for (2.5), so the solution exists and is unique. Moreover,

$$\|u_h\|_V \leq \frac{1}{\alpha^*} \|F_h\|_{V_h'}$$

with $\|F_h\|_{V_h'} = \sup_{v_h \in V_h \setminus \{0\}} \frac{F_h(v_h)}{\|v_h\|_V}$ being the norm of the dual space V_h'.
Now the only thing missing is the error inequality. Let w_h be any function of the subspace V_h. Setting $\sigma_h = u_h - w_h \in V_h$, we have:

$$\begin{aligned}
\alpha^* \|\sigma_h\|_V^2 &\leq a_h(\sigma_h, \sigma_h) && \text{(by coercivity of } a_h) \\
&= a_h(u_h, \sigma_h) - a_h(w_h, \sigma_h) \\
&= F_h(\sigma_h) - a_h(w_h, \sigma_h) && \text{(by (2.5))} \\
&= F_h(\sigma_h) - F(\sigma_h) + F(\sigma_h) - a_h(w_h, \sigma_h) \\
&= [F_h(\sigma_h) - F(\sigma_h)] + a(u, \sigma_h) - a_h(w_h, \sigma_h) && \text{(by (2.4))} \\
&= [F_h(\sigma_h) - F(\sigma_h)] + a(u - w_h, \sigma_h) + [a(w_h, \sigma_h) - a_h(w_h, \sigma_h)]
\end{aligned}$$

If $\sigma_h \neq 0$, we can divide everything by $\alpha^* \|\sigma_h\|_V$

$$\begin{aligned}
\|\sigma_h\|_V &\leq \frac{1}{\alpha^*} \left\{ \frac{|F_h(\sigma) - F(\sigma_h)|}{\|\sigma_h\|_V} + \frac{|a(u - w_h, \sigma_h)|}{\|\sigma_h\|_v} + \frac{|a(w_h, \sigma_h) - a_h(w_h, \sigma_h)|}{\|\sigma_h\|_V} \right\} \\
&\leq \frac{1}{\alpha^*} \left\{ M \|u - w_h\|_V + \sup_{v_h \in V_h \setminus \{0\}} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_V} + \sup_{v_h \in V_h \setminus \{0\}} \frac{|F_h(\sigma_h) - F(\sigma_h)|}{\|v_h\|_V} \right\}
\end{aligned}$$

Clearly, if $\sigma_h = 0$, the inequality still holds.

We can now estimate the error between u and u_h. Since $u - u_h = (u - w_h) - \sigma_h$ we obtain

$$\begin{aligned}
\|u - u_h\| &\leq \|u - w_h\|_V + \|\sigma_h\|_V \\
&\leq \|u - w_h\|_V + \frac{1}{\alpha^*}\Bigg\{ M\|u - w_h\|_V + \sup_{v_h \in V_h \setminus \{0\}} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_V} \\
&\quad + \sup_{v_h \in V_h \setminus \{0\}} \frac{|F_h(\sigma_h) - F(\sigma_h)|}{\|v_h\|_V} \Bigg\} \\
&= \left(1 + \frac{M}{\alpha^*}\right) \|u - w_h\|_V + \frac{1}{\alpha^*} \sup_{v_h \in V_h \setminus \{0\}} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_V} \\
&\quad + \sup_{v_h \in V_h \setminus \{0\}} \frac{|F_h(\sigma_h) - F(\sigma_h)|}{\|v_h\|_V}
\end{aligned}$$

If this inequality holds $\forall\, w_h \in V_h$, then it holds when taking the infimum. ★

Now we should try to apply Strang's lemma to GNI method in one dimension, to verify its convergence. Obviously, we will have V_N instead of V_h and everything that follows from there.
First, the error of the LGL numerical integration formula

$$E(g, v_N) = (g, v_N) - (g, v_N)_N$$

with g and v_N being a generic continuous function and a generic polynomial of $\mathbb{Q}_N$ respectively. Introducing the interpolation polynomial $I_N g$, we obtain:

$$\begin{aligned}
E(g, v_n) &= (g, v_N) - (I_N g, v_N) \\
&= (g, v_N) - (I_{N-1} g, v_N) + \underbrace{(\overbrace{I_{N-1} g}^{\in \mathbb{Q}_{N-1}}, \overbrace{v_N}^{\in \mathbb{Q}_N})}_{\in \mathbb{Q}_{2N-1}} \\
&= (g, v_N) - (I_{N-1} g, v_N) + (I_{N-1} g, v_N)_N - (I_N g, v_N)_N \\
&= (g - I_{N-1} g, v_N) + (I_{N-1} g - I_N g, v_N)_N
\end{aligned}$$

The first summand of the right-hand side can be bounded from above using Cauchy-Schwartz:

$$|(g - I_{N-1} g, v_N)| \leq \|g - I_{N-1} g\|_{L^2(-1,1)} \|v_N\|_{L^2(-1,1)}$$

For the second term, it's a bit more difficult, we need to introduce two new lemmas

Lemma 2.2
The discrete scalar product $(\cdot, \cdot)_N$ is a scalar product on $\mathbb{Q}_N$ and, as such, it satisfies the Cauchy-Schwartz inequality

$$|(\varphi, \psi)_N| \leq \|\varphi\|_N \|\psi\|_N$$

where the discrete norm is defined as

$$\|\varphi\|_N = \sqrt{(\varphi, \varphi)_N} \quad \forall\, \varphi \in \mathbb{Q}_N$$

Lemma 2.3
The "continuous" norm of $L^{"}(-1, 1)$ and the "discrete" norm $\|\cdot\|_N$ verify the inequalities

$$\|v_N\|_{L^2(-1,1)} \leq \|v_N\|_N \leq \sqrt{3} \|v_N\|_{L^2(-1,1)}$$

hence they are uniformly equivalent on $\mathbb{Q}_N$

By using these two lemmas we are able to obtain

$$\begin{aligned}
&|(I_{N-1} g - I_N g, v_N)_N| \leq \|I_{N-1} g - I_N g\|_N \|v_N\|_N \\
&\leq 3 \left[\|I_{N-1} g - g\|_{L^2(-1,1)} + \|I_N g - g\|_{L^2(-1,1)} \right] \|v_N\|_{L^2(-1,1)}
\end{aligned}$$

Putting all together we obtain the upper bound

$$|E(g, v_N)| \leq \left[4\|I_{N-1}g - g\|_{L^2(-1,1)} + 3\|I_{N-1}g - g\|_{L^2(-1,1)}\right] \|v_N\|_{L^2(-1,1)}$$

Using then the interpolation estimate

$$\|f - I_N f\|_{H^k(-1,1)} \leq C(s) \left(\frac{1}{N}\right)^{s-k} \|f\|_{H^s(-1,1)} \quad s \geq 1, k = 0, 1$$

we can bound $|E(g, v_N)|$ even more

$$|E(g, v_N)| \leq C(s) \left[\left(\frac{1}{N-1}\right)^s + \left(\frac{1}{N}\right)^s\right] \|g\|_{H^s(-1,1)} \|v_N\|_{L^2(-1,1)}$$

assuming that $g \in H^s(-1, 1)$.
Then, since for each $N \geq 2$ we have that $\frac{1}{N-1} \leq \frac{2}{N}$, the error for the LGL integration can be written as

$$|E(g, v_N)| \leq C(s) \left(\frac{1}{N}\right)^s \|g\|_{H^s(-1,1)} \|v_N\|_{L^2(-1,1)}$$

2.6 GNI as Collocation method

Let us introduce a problem

$$\begin{cases} Lu = -(\mu u')' + (bu)' + \sigma u = f \quad -1 < x < 1 \\ u(-1) = u(1) = 0 \end{cases}$$

that has the usual weak formulation

$$\text{find } u \in V = H_0^1(-1, 1) : a(u, v) = F(v), \forall\, v \in V$$

The GNI formulation follows

$$\begin{cases} \text{find } u_N \in V_N = \mathbb{P}_N^0 = \{v_N \in \mathbb{P}_N : v_N(\pm 1) = 0\} \\ a_N(u_N, v_N) = F_N(v_N) \quad \forall\, v_N \in V_N \end{cases}$$

Note that, thanks to the exactness of LGL quadrature formula

$$\begin{aligned} a_N(u_N, v_N) &\overset{\text{(def. of } I_N)}{=} (I_N \underbrace{(\mu u_N' - bu_N)}_{\in \mathbb{P}_N}, \underbrace{v_N'}_{\in \mathbb{P}_{N-1}})_N + (\sigma u_N, v_N)_N \\ &\overset{\text{(exactness)}}{=} (I_N(\mu u' - bu), v_N) + (\sigma u_N, v_N)_N \\ &\overset{\text{(int. by parts)}}{=} -(\underbrace{I_N(\mu u' - bu)'}_{\in \mathbb{P}_{N-1}}, \underbrace{v_N}_{\in \mathbb{P}_N}) + (\sigma u_N, v_N)_N \\ &\overset{\text{(exactness)}}{=} (\underbrace{-(I_N(\mu u' - bu_N))' + \sigma u_N}_{= L_N u_N}, v_N)_N \end{aligned}$$

So it's obvious that (GNI) $\iff$ $(L_N u_N, v_N)_N = F_N(v_N)\ \forall\, v_N \in V_N$, so it's a collocation method.

2.7 1D Spectral Elements

Let $p \geq 1$ integer and $\mathbb{P}_p$ the space of polynomials of degree $\leq p$. We can divide the domain $\Omega = \bigcup_{n=1}^{N_e} I_k$ with I_k disjoint elements s.t. $I_k = F_k((-1, 1))$ and

$$F_k : \xi \mapsto x = \frac{b_k - a_k}{2}\xi + \frac{b_k + a_k}{2}$$

with $N_p = p \cdot N_e + 1$ the total number of nodes in Ω. Then we use the Lagrange basis functions $\{\varphi_i\}_{i=1}^{N_p}$ w.r.t. the LGL nodes.
Now set $X_\delta = \left\{v \in \mathcal{C}^0 : v|_{I_k} \in \mathbb{P}_p, \forall\, I_k\right\}$ with $h_k = meas(I_k)$, mesh size $h = max_k h_k$ and polynomial degree p we can define $\delta = (h, p)$ and

$$v_\delta(x) = \sum_{i=1}^{N_p} v_\delta(x_i)\varphi_i(x) \quad \forall\, v_\delta \in X_\delta$$

Let now $(\hat{\xi}_j, \hat{w}_j)$ for $j = 0, \dots, p$ be the LGL nodes and respective weights in $\hat{\Omega} = (-1, 1)$. We can define the local LGL quadrature as

$$\int_{I_k} u(x)v(x)\,dx \approx (u, v)_{\delta, I_k} = \sum_{j=0}^{p} u(\xi_j)v(\xi_j)w_j$$

with $\xi_j = \frac{b_k - a_k}{2}\hat{\xi}_j + \frac{b_k + a_k}{2}$ and $w_j = \frac{b_k - a_k}{2}\hat{w}_j$. Meanwhile, we can pass this quadrature to the whole domain, obtaining the composite LGL quadrature:

$$\int_\Omega u(x)v(x)\,dx \approx (u, v)_{\delta,\Omega} = \sum_{k=1}^{N_e} (u, v)_{\delta, I_k}$$

with its relative error $\exists\, c > 0 : \forall\, f \in H^r(\Omega), r \geq 1, p \geq 1 : \forall\, v_\delta \in X_\delta$:

$$\left| \int_\Omega f v_\delta\,dx - (f, v_\delta)_{\delta,\Omega} \right| \leq c h^{\min(p,r)} \left(\frac{1}{p}\right)^r \|f\|_{H^r(\Omega)} \|v_\delta\|_{L^2(\Omega)}$$

and its interpolation error as: $\exists\, c > 0 : \forall\, \in H^{s+1}(\Omega), s \geq 1$

$$\left\| v - \prod_\delta^{LGL} v \right\|_{H^k(\Omega)} \leq C h^{\min(p+1, s+1) - k} \left(\frac{1}{p}\right)^{s+1-k} \|v\|_{H^{s+1}(\Omega)}$$

2.8 Spectral Element Method with Numerical Integration

Let's go back to the problem

$$\begin{cases} -(\mu u')' - (bu)' + \sigma u = f & \text{in } \Omega \\ u(a) = u(b) = 0 \end{cases}$$

Given $V = H_0^1(\Omega)$, the weak formulation reads

$$\text{find } u \in V : a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall\, v \in V, f \in L^2(\Omega)$$

with

$$a(u, v) = \int_\Omega (\mu u' - bu)v'\,dx + \int_\Omega \sigma u v\,dx$$
$$(f, v)_{L^2(\Omega)} = \int_\Omega f v\,dx$$

Now set $a_\delta(\varphi_j, \varphi_i) = (\mu\varphi_j' - b\varphi_j, \varphi_i)_{\delta,\Omega} + (\sigma\varphi_j, \varphi_i)_{\delta,\Omega}$ to get the SEM-GNI formulation:

$$\text{find } u_\delta^{\text{GNI}} \in V_\delta : a_\delta(u_\delta^{\text{GNI}}, v_\delta) = (f, v_\delta)_{\delta,\Omega} \quad \forall\, v_\delta \in V_\delta \tag{2.6}$$

Now expand u_δ^{GNI} w.r.t the Lagrange basis $u_\delta^{\text{GNI}}(x) = \sum_{i=1}^{N_p} u_\delta^{\text{GNI}}(x_i)\varphi_i(x)$ and choose $v_\delta(x) = \varphi_i(x)$ for any $i = 1, \dots, N_p$. We can now write the SEM-GNI discretization of the weak formulation:

$$\begin{cases} \text{find } u^{\text{GNI}} = \left[u_\delta^{\text{GNI}}(x_j)\right]_{j=1}^{N_p} \\ u_\delta^{\text{GNI}}(x_1) = u_\delta^{\text{GNI}}(x_{N_p}) = 0 \\ \sum_{j=1}^{N_p} a_\delta(\varphi_j, \varphi_i) u_\delta^{\text{GNI}}(x_j) = (f, \varphi_i)_{\delta,\Omega} \quad \forall\, i = 1, \dots, N_p \end{cases}$$

or, in algebraic form $A^{\mathrm{GNI}}u^{\mathrm{GNI}} = f^{\mathrm{GNI}}$ with $A_{ij}^{\mathrm{GNI}} = a_\delta(\varphi_j, \varphi_i)$ and $f_i^{\mathrm{GNI}} = (f, \varphi_i)_{\delta,\Omega}$.
Now, for the error analysis, we will apply the Strang lemma, so:

$$\begin{aligned}\left\|u - u_\delta^{\mathrm{GNI}}\right\|_V \leq \|u - u_\delta\|_V & \\ &+ \frac{1}{\mu^*} \sup_{v_\delta \in V_\delta \setminus \{0\}} \frac{|a(u_\delta, v_\delta) - a_\delta(u_\delta, v_\delta)|}{\|v_\delta\|_V} \\ &+ \frac{1}{\mu^*} \sup_{v_\delta \in V_\delta \setminus \{0\}} \frac{|f, v_\delta|_{L^2(\Omega)} - (f, v_\delta)_{\delta,\Omega}}{\|v_\delta\|_V}\end{aligned}$$

where μ^* is the coercivity constant of a_δ: $a_\delta(v_\delta, v_\delta) \geq \mu^* \|v_\delta\|_V^2$ and u_δ the SEM-GNI solution. Thus, for any $u \in H^{s+1}(\Omega)$ and $f \in H^r(\Omega)$

$$\left\|u - u_\delta^{\mathrm{GNI}}\right\|_{H^1(\Omega)} \leq C \left[h^{\min(p,s)} \left(\frac{1}{p}\right)^s \|u\|_{H^{s+1}(\Omega)} + h^{\min(p,r)} \left(\frac{1}{p}\right)^r \|f\|_{H^r(\Omega)} \right]$$

So u_δ^{GNI} converges with spectral accuracy w.r.t. p and algebraic accuracy w.r.t. h to the exact solution.

2.9 Convergence rate of SEM-GNI

When s, r are large ($s, r > p$):

$$\|u - u_\delta\|_{H^1\Omega} \leq C \left[h^p \left(\frac{1}{p}\right)^s \|u\|_{H^{s+1}(\Omega)} + h^p \left(\frac{1}{p}\right)^r \|f\|_{H^r(\Omega)} \right]$$

when s is small ($s \leq p$):

$$\|u - u_\delta\|_{H^1(\Omega)} \leq C \left(\frac{h}{p}\right)^s \|u\|_{H^{s+1}(\Omega)}$$

3 Discontinuous Galerkin methods

The idea behind DG methods is to seek the solution in a discrete space made of polynomials that are completely discontinuous across the elements of the mesh.

$$V_h \subsetneq V$$

3.1 1D case

Let us consider a Poisson problem

$$\begin{cases} -u'' = f & a < x < b \\ u(a) = u(b) = 0 \end{cases}$$

The aim is to use discontinuous piecewise polynomials, so that between every interval I_k from one node to another we obtain

$$\int_a^b -u''v = \int_a^b fv \Rightarrow -\sum_{k=0}^{N-1} \int_{I_k} u''v = \sum_{k=0}^{N-1} \int_{I_k} fv$$

We must now integrate by parts, but our test functions are discontinuous at the nodes, so we must acknowledge it. Let's call x_k^- and x_k^+ the left and right side of the x_k node. Then we can:

$$-\sum_{k=0}^{N-1} \int_{I_k} u''v = \sum_{k=0}^{N-1} \left[\int_{I_k} u'v' - \left(u'v|_{x_{k+1}^-} - u'v|_{x_k^+} \right) \right] \tag{3.1}$$

$$\begin{aligned} \sum_{k=0}^{N-1} (u'v|_{x_{k+1}^-} - u'v|_{x_k^+}) &= u'(x_1^-)v(x_1^-) - u'(x_0^+)v(x_0^+) \\ &+ u'(x_2^-)v(x_2^-) - u'(x_1^+)v(x_1^+) \\ &+ \dots \\ &+ u'(x_N^-)v(x_N^-) - u'(x_{N-1}^+)v(x_{N-1}^+) \\ &= \sum_{k=0}^{N} [\![u'(x_k)v(x_k)]\!] \end{aligned} \tag{3.2}$$

where we have defined the jump function

$$\begin{aligned} [\varphi(x_0)] &:= -\varphi(x_0^+) \\ [\varphi(x_k)] &:= \varphi(x_k^-) - \varphi(x_k^+) \qquad x_k: \text{ interior node} \\ [\varphi(x_N)] &:= \varphi(x_N^-) \end{aligned} \tag{3.3}$$

By using (3.1) and (3.3) we obtain

$$\sum_{k=0}^{N-1} \int_{I_k} u'v' - \sum_{k=0}^{N} [\![u'(x_k)v(x_k)]\!] = \sum_{k=0}^{N-1} \int_{I_k} fv \tag{3.4}$$

Now define the average operator

$$\begin{aligned} \{\!\{\varphi(x_0)\}\!\} &:= \varphi(x_0^+) \\ \{\!\{\varphi(x_k)\}\!\} &:= \frac{1}{2}\varphi(x_k^-) + \varphi(x_k^+) \qquad x_k: \text{ interior node} \\ \{\!\{\varphi(x_N)\}\!\} &:= \varphi(x_N^-) \end{aligned} \tag{3.5}$$

This way we obtain this formula (magic formula)

$$\sum_{k=0}^{N} [\![u'(x_k)v(x_k)]\!] = \sum_{k=0}^{N} \{\!\{u'(x_k)\}\!\} [v(x_k)] + \sum_{k=1}^{N-1} [\![u'(x_k)]\!] \{\!\{v(x_k)\}\!\} \tag{3.6}$$

If u is the exact solution and $u \in \mathcal{C}^1([a,b])$, then $[u'(x_k)] = 0$ for every interior node, and the second sum in (3.6) drops.
We end up with the formulation (by collecting (3.4) and (3.6))

$$\begin{aligned} \underbrace{\sum_{k=0}^{N-1} \int_{I_k} u'v' - \sum_{k=0}^{N} \{\!\{u'(x_k)\}\!\} [v(x_k)] - \sum_{k=1}^{N-1} [\![u'(x_k)]\!] \{\!\{v(x_k)\}\!\}}_{\mathcal{A}(u,v)} \\ = \sum_{k=0}^{N-1} \int_{I_k} fv \quad \forall\, v \in V \end{aligned} \tag{3.7}$$

where

$$V = H^1_{\text{broken}}(\Omega) := \left\{ v \in L^2(\Omega) : v|_{I_k} \in H^1 I_k \ \forall\, k = 0, \dots, N-1 \right\}$$

with the broken norm

$$\|v\|_{H^1_{\text{broken}}(\Omega)} = \left(\sum_{k=0}^{N} \|v|_{I_k}\|^2_{H^1(\Omega)} \right)^{\frac{1}{2}}$$

Let now $V_h \subset V$

$$\text{find } u_h \in V_h : \mathcal{A}(u_h, v_h) = \sum_{k=0}^{N-1} \int_{I_k} f v_h \quad \forall\, v_h \in V_h \tag{3.8}$$

Remark 3.1
V_h is not a subspace of $H^1(\Omega)$

But (3.8) is not well posed, so the (3.7) must be modified such that:

- drop 3rd term because $[u'(x_k)] = 0$
- add symmetrization term ($= 0$ if u is the exact solution)

$$-\sum_{k=0}^{N} \theta \{\!\{v'(x_k)\}\!\} [u(x_k)]$$

  with

  $\theta = 1$ SIP (Symmetric Interior Penalty)
  $\theta = -1$ NIP (Non-symmetric Interior Penalty)
  $\theta = 0$ IIP (Incomplete Interior Penalty)

- add the stabilization term ($= 0$ if u is the exact solution)

$$+\sum_{k=0}^{N} \gamma [u(x_k)] [v(x_k)]$$

We can now obtain a new bilinear form

$$\begin{aligned} \mathcal{A}^*(u_h, v_h) = \underbrace{\sum_{k=0}^{N-1} \int_{I_k} u_h' v_h'}_{(1)} \underbrace{- \sum_{k=0}^{N} \{\!\{u_h'(x_k)\}\!\} [v_h(x_k)]}_{(2)} \\ \underbrace{- \sum_{k=0}^{N} \theta \{\!\{v_h'(x_k)\}\!\} [u_h(x_k)]}_{(3)} \underbrace{+ \sum_{k=0}^{N} \gamma [u_h(x_k)] [v_h(x_k)]}_{(4)} \end{aligned} \tag{3.9}$$

Neumann BC

Impose Neumann BC through $\{\!\{u'(x_k)\}\!\}$ in (2). In this case we have $\sum_{k=1}^{N-1}$ in (2) and, consequently, we write $\sum_{k=1}^{N-1}$ in (3) for symmetry.

Non-homogeneous Dirichlet BC

Impose Dirichlet BC as follows. In (3) and (4) replace $[\![u_h(x_0)]\!]$ and $[\![u_h(x_N)]\!]$ with the following definition:

$$\begin{aligned} [\![u_h(x_0)]\!] &:= \alpha - u_h(x_0^+) && \text{if } u(a) = \alpha \\ [\![u_h(x_N)]\!] &:= u_h(x_N^-) - \beta && \text{if } u(b) = \beta \end{aligned}$$

In case $\alpha = \beta = 0$ we have homogeneous Dirichlet.
Now in (3.9) split sums as follows:

$$\begin{aligned} &\mathcal{A}^*(u_h, v_h) \\ &= \sum_{k=0}^{N-1} \int_{I_k} u_h' v_h' \\ &- \sum_{k=1}^{N-1} \{\!\{u_h'(x_k)\}\!\} [\![v_h(x_k)]\!] + u_h'(x_0^+) v_h(x_0^+) - u_h'(x_N^-) v_h(x_N^-) \\ &- \sum_{k=1}^{N-1} \theta \{\!\{v_h'(x_k)\}\!\} [\![u_h(x_k)]\!] - \left[\theta v_h'(x_0^+)(\alpha - u_h(x_0^+)) + \theta v_h'(x_N^-)(u_h(x_N^-) - \beta)\right] \\ &- \sum_{k=1}^{N-1} \gamma [\![u_h(x_k)]\!] [\![v_h(x_k)]\!] + \gamma(\alpha - u_h(x_0^+))(-v_h(x_0^+)) + \gamma(u_h(x_N^-) - \beta) v_h(x_N^-) \end{aligned} \tag{3.10}$$

Now move terms, including α and β to the right-hand side of the formulation.
On the left-hand side it remains

$$\begin{aligned} &\tilde{\mathcal{A}}(u_h, v_h) \\ &= \sum_{k=0}^{N-1} \int_{I_k} u_h' v_h' \\ &- \sum_{k=1}^{N-1} \{\!\{u_h'(x_k)\}\!\} [\![v_h(x_k)]\!] + u_h'(x_0^+) v_h(x_0^+) - u_h'(x_N^-) v_h(x_N^-) \\ &- \sum_{k=1}^{N-1} \theta \{\!\{v_h'(x_k)\}\!\} [\![u_h(x_k)]\!] - \left[\theta v_h'(x_0^+) u_h(x_0^+) + \theta v_h'(x_N^-) u_h(x_N^-)\right] \\ &- \sum_{k=1}^{N-1} \gamma [\![u_h(x_k)]\!] [\![v_h(x_k)]\!] + \gamma u_h(x_0^+) v_h(x_0^+) + \gamma u_h(x_N^-) v_h(x_N^-) \end{aligned} \tag{3.11}$$

On the right-hand side instead

$$\mathcal{F}(v_h) = \sum_{k=0}^{N-1} \int_{I_k} f v_h + \theta(\alpha v_h'(x_0^+) - \beta v_h'(x_N^-)) + \gamma(\alpha v_h(x_0^+) + \beta v_h(x_N^-)) \tag{3.12}$$

Remark 3.2
Note that for $\theta = 1$, $\tilde{\mathcal{A}}(u_h, v_h) = \tilde{\mathcal{A}}(v_h, u_h)$, so it's symmetric.

Non-homogeneus Dirichlet conditions

$$\text{find } u_h \in V_h : \tilde{\mathcal{A}}(u_h, v_h) = \mathcal{F}(v_h) \quad \forall\, v_h \in V_h \tag{3.13}$$

with $\mathcal{F}$ depending on f, α and β.
Note that in (3.11), if we define $[\![u_h(x_0)]\!]$ and $[\![u_h(x_N)]\!]$ as $[\![v_h(x_0)]\!]$ and $[\![v_h(x_N)]\!]$

$$\begin{aligned}
-\sum_{k=1}^{N-1} \{\!\{u_h'(x_k)\}\!\}\, [\![v_h(x_k)]\!] + u_h'(x_0^+)v_h(x_0^+) - u_h'(x_N^-)v_h(x_N^-) \\
= -\sum_{k=0}^{N} \{\!\{u_h'(x_k)\}\!\}\, [\![v_h(x_h)]\!] \\
-\sum_{k=1}^{N-1} \theta\, \{\!\{v_h'(x_k)\}\!\}\, [\![u_h(x_k)]\!] + (\theta u_h(x_0^+)v_h'(x_0^+) - \theta u_h(x_N^-)v_h'(x_N^-)) \\
= -\sum_{k=0}^{N} \theta\, \{\!\{v_h'(x_k)\}\!\}\, [\![u_h(x_k)]\!] \\
+\sum_{k=1}^{N-1} \gamma\, [\![u_h(x_k)]\!]\, [\![v_h(x_k)]\!] + \gamma u_h(x_0^+)v_h(x_0^+) + \gamma u_h(x_N^-)v_h(x_N^-) \\
+\sum_{k=0}^{N} \gamma\, [\![u_h(x_k)]\!]\, [\![v_h(x_k)]\!]
\end{aligned}$$

3.2 Multidimensional case

We can take our Poisson problem in multidimension

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{3.14}$$

with the triangulation $\mathcal{T}_h$, but this time we cannot assume that the conformity constraint is present. So we need to take a test function v (element-wise smooth), and integrate over an element $\mathcal{K} \in \mathcal{T}_h$

$$\int_{\mathcal{K}} -\Delta u v = \int_{\mathcal{K}} f v$$

As usual, integrate by parts, and sum over all the elements $\mathcal{K} \in \mathcal{T}_h$

$$\sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\mathcal{K}} \nabla u \cdot \nabla v - \sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\partial\mathcal{K}} \nabla u \cdot \mathbf{n}_{\mathcal{K}} v = \int_{\Omega} f v$$

since for any $F \in \mathcal{F}_h'$ which is the set of interior faces shared by two elements $\mathcal{K}^+$ and $\mathcal{K}^-$

$$\{\!\{v\}\!\} = \frac{(v^+ + v^-)}{2} \qquad [\![v]\!] = v^+\mathbf{n}^+ + v^-\mathbf{n}^-$$
$$\{\!\{\boldsymbol{\tau}\}\!\} = \frac{(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-)}{2} \qquad [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}^+\mathbf{n}^+ + \boldsymbol{\tau}^-\mathbf{n}^-$$

while, for the set of boundary faces $F \in \mathcal{F}_h^B$

$$\{\!\{v\}\!\} = v \qquad [\![v]\!] = v\mathbf{n}$$
$$\{\!\{\boldsymbol{\tau}\}\!\} = \boldsymbol{\tau} \qquad [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau} \cdot \mathbf{n}$$

in this way we can obtain the following formula $\forall\, \boldsymbol{\tau}$ vector function:

$$\sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\partial\mathcal{K}} \boldsymbol{\tau} \cdot \mathbf{n}_{\mathcal{K}} v = \sum_{F\in\mathcal{F}_h} \int_F \{\!\{\boldsymbol{\tau}\}\!\} \cdot [\![v]\!] + \sum_{F\in\mathcal{F}_h'} \int_F [\![\boldsymbol{\tau}]\!]\, \{\!\{v\}\!\} \tag{3.15}$$

and thanks to that we obtain

$$-\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\partial\mathcal{K}}\boldsymbol{\tau}\cdot\mathbf{n}_{\mathcal{K}}v=-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\sum_{F\in\mathcal{F}'_h}\int_F[\nabla u]\{v\}$$

then

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\cdot\nabla v-\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\partial\mathcal{K}}\nabla u\cdot\mathbf{n}_{\mathcal{K}}v=\int_\Omega fv,$$

so it becomes

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\sum_{F\in\mathcal{F}'_h}\int_F[\nabla u]\{v\}=\int_\Omega fv$$

but, if we assume $u\in H^2(\Omega)$, then $[\nabla u]=0\ \forall\ F\in\mathcal{F}'_h$. This regularity assumption is fulfilled if $f\in L^2$ and the domain is a convex polygon, thanks to the property of elliptic regularity.

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\cancel{\sum_{F\in\mathcal{F}'_h}\int_F[\nabla u]\{v\}}=\int_\Omega fv$$

Now we can assume that $[u]=0\ \forall\ F\in\mathcal{F}_h$ (since $u\in H^2(\Omega)\cap H^1_0(\Omega)$) to add a symmetry term

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\cdot\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\sum_{F\in\mathcal{F}'_h}\int_F\{\nabla_h v\}[u]=\int_\Omega fv$$

where ∇_h is the element wise gradient (v is only piecewise smooth). We also add a stabilization term that controls the jumps

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\cdot\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\sum_{F\in\mathcal{F}_h}\int_F[u]\cdot\{\nabla_h v\}+\sum_{F\in\mathcal{F}_h}\int_F\gamma\,[u]\cdot[v]=\int_\Omega fv$$

where γ is a stabilization function.
Now we can define the DG discrete space

$$V_h^p=\left\{v_h\in L^2(\Omega):v_h|_{\mathcal{K}}\in\mathcal{P}^{p_{\mathcal{K}}}(\mathcal{K})\ \forall\ \mathcal{K}\in\mathcal{T}_h\right\}\not\subseteq H^1_0(\Omega)$$

Discretize $u\rightsquigarrow u_h, v\rightsquigarrow v_h$ and obtain the following weak formulation

$$\text{find } u_h\in V_h^p \text{ s.t. } \mathcal{A}(u_h,v_h)=\int_\Omega fv\quad\forall\ v_h\in V_h^p$$

where

$$\mathcal{A}(u,v)=\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\cdot\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\sum_{F\in\mathcal{F}_h}\int_F[u]\cdot\{\nabla_h v\}$$
$$+\sum_{F\in\mathcal{F}_h}\int_F\gamma\,[u]\cdot[v]$$

Interior Penalty DG methods

$$\text{find } u_h\in V_h^p \text{ s.t. } \mathcal{A}(u_h,v_h)=\int_\Omega fv\quad\forall\ v_h\in V_h^p$$

Note that $\mathcal{A}$ depends on the triangulation, and it differs from the original weak formulation in the infinite dimension problem.

$$\mathcal{A}(u,v)=\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}\nabla u\cdot\nabla v-\sum_{F\in\mathcal{F}_h}\int_F\{\nabla u\}\cdot[v]-\theta\sum_{F\in\mathcal{F}_h}\int_F[u]\cdot\{\nabla_h v\}$$
$$+\sum_{F\in\mathcal{F}_h}\int_F\gamma\,[u]\cdot[v]$$

where

- $\theta = 1$ Symmetric Interior Penalty (SIP)
- $\theta = -1$ Non-symmetric Interior Penalty (NIP)
- $\theta = 0$ Incomplete Interior Penalty (IIP)

Dirichlet BC

The above formulation holds when applying homogeneous Dirichlet BC, but in the case of non-homogeneous BC, such as

$$u = g_D \quad \text{on } \partial\Omega$$

the right-hand side must be modified as

$$\int_\Omega fv - \theta \sum_{F \in \mathcal{F}_h^B} \int_F g_D \nabla_h v \cdot \mathbf{n} + \sum_{F \in \mathcal{F}_h^B} \int_F \gamma g_D v$$

Neumann BC

In the case of Neumann BC, like

$$\nabla u \cdot \mathbf{n} = g_N \quad \text{on } \partial\Omega$$

the bilinear form has to be modified such as

$$\mathcal{A}(u,v) = \sum_{\mathcal{K} \in \mathcal{T}_h} \int_{\mathcal{K}} \nabla u \cdot \nabla v - \sum_{F \in \mathcal{F}_h'} \int_F \{\nabla u\} \cdot [v] - \theta \sum_{F \in \mathcal{F}_h'} \int_F [u] \cdot \{\nabla_h v\} + \sum_{F \in \mathcal{F}_h'} \int_F \gamma [u] \cdot [v]$$

and the right-hand side

$$\int_\Omega fv - \sum_{F \in \mathcal{F}_h^B} \int_F g_N v$$

The stabilization function γ

$$\sum_{F \in \mathcal{F}_h} \int_F \gamma [u] \cdot [v] \quad \gamma = \alpha \frac{p^2}{h}$$

where

$$p = \begin{cases} \max\{p_{\mathcal{K}^+}, p_{\mathcal{K}^-}\} & \text{if } F \in \mathcal{F}_h' \\ p_{\mathcal{K}} & \text{if } F \in \mathcal{F}_h^B \end{cases}$$

and

$$h = \begin{cases} \min\{h_{\mathcal{K}^+}, h_{\mathcal{K}^-}\} & \text{if } F \in \mathcal{F}_h' \\ h_{\mathcal{K}} & \text{if } F \in \mathcal{F}_h^B \end{cases}$$

Since we can make some assumptions

$$h_F \approx h_{\mathcal{K}^+} \approx h_{\mathcal{K}^-},\ p_{\mathcal{K}^+} \approx p_{\mathcal{K}^-} \Rightarrow \gamma = \mathcal{O}\left(\frac{p^2}{h}\right)$$

3.3 Theoretical reminders

For an integer $s \geq 1$ define the broken Sobolev space

$$H^s(\mathcal{T}_h) = \left\{ v \in L^2(\Omega) : v|_{\mathcal{K}} \in H^s(\mathcal{K}) \ \forall \ \mathcal{K} \in \mathcal{T}_h \right\}$$
$$\|v\|^2_{H^s(\mathcal{K})} = \sum_{\mathcal{K} \in \mathcal{T}_h} \|v\|^2_{H^s(\mathcal{K})}$$

Define also

$$\|v\|^2_{L^2(\mathcal{F}_h)} = \sum_{F \in \mathcal{F}_h} \|v\|^2_{L^2(F)}$$

We define then the following norms

$$\|v\|^2_{DG} = \|\nabla_h v\|^2_{L^2(\Omega)} + \left\| \gamma^{\frac{1}{2}} [v] \right\|^2_{L^2(\mathcal{F}_h)} \quad \forall \ v \in H^2(\mathcal{T}_h)$$
$$|||v|||_{DG} = \|v\|^2_{DG} + \left\| \gamma^{\frac{1}{2}} \{\nabla_h v\} \right\|^2_{L^2(\mathcal{F}_h)} \quad \forall \ v \in H^2(\mathcal{T}_h)$$

where $\nabla_h v$ is the element wise gradient:

$$(\nabla_h v)|_{\mathcal{K}} = \nabla(v|_{\mathcal{K}}) \quad \forall \ \mathcal{K} \in \mathcal{T}_h$$

Notice that $V_h^p \subset H^2(\mathcal{T}_h)$. It can be shown that

$$\|v\|_{DG} \underset{(trivial)}{\leq} |||v|||_{DG} \not\lesssim \|v\|_{DG} \quad \forall \ v \in H^2(\mathcal{T}_h)$$
$$\|v_h\|_{DG} \underset{(trivial)}{\leq} |||v_h|||_{DG} \underset{(on \ slides)}{\not\lesssim} \|v_h\|_{DG} \quad \forall \ v_h \in V_h^p$$

Some key properties:

- Continuity on $H^2(\mathcal{T}_h) \times V_h^p$:

  $$|\mathcal{A}(v, w_h)| \lesssim |||v|||_{DG} \, \|w_h\|_{DG} \quad \forall \ v \in H^2(\mathcal{T}_h), \ \forall \ w_h \in V_h^p$$

  Also remind that $|\mathcal{A}(v, w_h)| \not\lesssim \|v\|_{DG} \|w_h\|_{DG}$

- Coercivity on $V_h^p \times V_h^p$:

  $$\mathcal{A}(v_h, v_h) \gtrsim \|v_h\|_{DG} \quad \forall \ v_h \in V_h^p$$

  For SIP and IIP, the penalty parameter α should be large enough.

- Strong-consistency (Galerkin orthogonality):

  $$\mathcal{A}(u, v_h) = \int_\Omega f v_h \ \ \forall \ v_h \in V_h^p \Rightarrow \mathcal{A}(u - u_h, v_h) = 0 \quad \forall \ v_h \in V_h^p$$

- Approximation. Let $\prod_h^p u \in V_h^p$ be a suitable approximation of u, then

  $$\left|\left|\left| u - \textstyle\prod_h^p u \right|\right|\right|_{DG} \lesssim \frac{h^{\min(p,s)}}{p^{s-\frac{1}{2}}} \|u\|_{H^{s+1}(\mathcal{T}_h)}$$

  If $p \geq s$

  $$\left|\left|\left| u - \textstyle\prod_h^p u \right|\right|\right|_{DG} \lesssim \left(\frac{h}{p}\right)^s p^{\frac{1}{2}} \|u\|_{H^{s+1}(\mathcal{T}_h)}$$

3.4 Error estimates

Recall the abstract error estimate $\|u - u_h\|_{DG} \lesssim \left|\left|\left| u - \prod_h^p u \right|\right|\right|_{DG}$.
If u is sufficiently regular then

$$\|u - u_h\|_{DG} \lesssim \frac{h^{\min(p,s)}}{p^{s-\frac{1}{2}}} \|u\|_{H^{s+1}(\mathcal{T}_h)}$$

Then, by using a duality argument, one can obtain an estimate for the L^2 norm.
Assuming that Ω is such that the following elliptic regularity result holds: for any $g \in L^2(\Omega)$, the solution z of the problem

$$\begin{cases} -\Delta z = g & \text{in } \Omega \\ z = 0 & \text{on } \partial\Omega \end{cases}$$

satisfies $z \in H^2(\Omega)$ and

$$\|z\|_{H^2(\Omega)} \lesssim \|g\|_{L^2(\Omega)}$$

If the exact solution $u \in H^s(\Omega), s \geq 2$ and, if u_h is obtained with the SIP method, it holds

$$\|u - u_h\|_{L^2(\Omega)} \lesssim \frac{h^{\min(p,s)+1}}{p^{s+\frac{1}{2}}} \|u\|_{H^{s+1}(\Omega)}$$

while for NIP and IIP holds

$$\|u - u_h\|_{L^2(\Omega)} \lesssim \frac{h^{\min(p,s)}}{p^{s-\frac{1}{2}}} \|u\|_{H^{s+1}(\Omega)}$$

4 Advection-Diffusion-Reaction equations

4.1 Formulation of the problem

Considering the problem $\mathcal{L}u = f$ in Ω, $u = 0$ on $\partial\Omega$ where

$$\begin{aligned} \mathcal{L}u &= -\text{div}\,(\mu\nabla u + \mathbf{b}u) + \sigma u && \text{(conservative form)} \\ \mathcal{L}u &= -\text{div}\,(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u && \text{(non-conservative form)} \end{aligned}$$

with the same assumptions as (1.1).
The weak formulation is written as

$$\text{find } u \in V = H_0^1(\Omega) : a(u,v) = F(v) \;\forall\, v \in V \tag{4.1}$$

with

$$F(v) = \int_\Omega fv$$

and

$$a(u,v) = \begin{cases} \displaystyle\int_\Omega (\mu\nabla u + \mathbf{b}u)\cdot\nabla v + \int_\Omega \sigma uv & \text{conservative form} \\ \displaystyle\int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega \mathbf{b}\cdot\nabla uv + \int_\Omega \sigma uv & \text{non-conservative form} \end{cases}$$

Let's verify the uniqueness of the solution:

Coercivity

Sufficient conditions for coercivity:

$$\begin{aligned} \sigma - \frac{1}{2}\text{div}\mathbf{b} &\geq 0 \text{ in } \Omega && \text{non-conservative case} \\ \sigma + \frac{1}{2}\text{div}\mathbf{b} &\geq 0 \text{ in } \Omega && \text{conservative case} \end{aligned}$$

In both cases: $a(u,v) \geq \mu_0\|\nabla u\|^2 \rightarrow$ coercivity constant $\alpha \simeq \mu_0$

Continuity

In both cases, continuity constant: $M \simeq \|\mu\|_{L^\infty} + \|\mathbf{b}\|_{L^\infty} + \|\sigma\|_{L^2}$
Given that the hypotheses of Lax-Milgram holds, the solution exists and is unique. We can now bring in the Galerkin formulation

$$\text{find } u_h \in V_h : a(u_h, v_h) = (f, v_h) \quad \forall\, v_h \in V_h$$

and move to the error estimate

$$\|u - u_h\| \underset{\text{(Ceà)}}{\leq} \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\| \underset{\substack{\text{(interpolation} \\ \text{error estimate)}}}{\leq} C\frac{M}{\alpha} h^r |u|_{H^{r+1}(\Omega)}$$

If it is a convection dominated flow (or reaction dominated), then $\frac{M}{\alpha} \gg 1$, then we need to find a trade-off between $\frac{M}{\alpha}$ and h^r. Also, it is numerically prohibitive.
The Péclet number tells us if the flow is dominated by advection or diffusion if its greater or smaller than 1. We can define it as

$$\mathbb{P}e = h\frac{M}{\alpha}$$

Should be less than 1 for stability issues.

4.2 Stabilization methods

The idea now is to stabilize the Galerkin method.

- 1D case: Upwind method $\iff$ Artificial diffusion
- 2D case: Streamline diffusion:

$$+c(h)\int_\Omega \frac{1}{\|\mathbf{b}\|}\left(\mathbf{b}\cdot\nabla u_h\right)\left(\mathbf{b}\cdot\nabla v_h\right)$$

  Artificial diffusion:

$$+c(h)\int_\Omega \nabla u_h\cdot\nabla v_h$$

Now the solution is stabilized, but is not fully consistent. So the solution is to find a way to obtain a fully consistent solution

$$\text{find } u_h\in V_h : a(u_h,v_h)+\mathscr{L}_h(u_h,f;v_h)=F(v_h)\quad\forall\, v_h\in V_h$$

with $\mathscr{L}_h$ suitably chosen such that

$$\mathscr{L}(u_h,f;v_h)=0\quad\forall\, v_h\in V_h,$$

so we obtain a strongly consistent approximation of the original problem.
One possibility could be to use an operator proportional to the residual:

$$\mathscr{L}_h(u_h,f;v_h)=\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}(\mathcal{L}u-f)\tau_{\mathcal{K}}\varphi(v_h)\quad\forall\, v_h\in V_h$$

with $\tau_{\mathcal{K}}$ as a scaling factor. Typically, is chosen, given $h_{\mathcal{K}}=diam(\mathcal{K})$:

$$\tau_{\mathcal{K}}(\mathbf{x})=\delta\frac{h_{\mathcal{K}}}{|\mathbf{b}(\mathbf{x})|}\quad\forall\,\mathbf{x}\in\mathcal{K},\mathcal{K}\in\mathcal{T}_h$$

while, for $\varphi(v_h)$ there are many possibilities. Two of them are

- $\varphi(v_h)=\mathcal{L}v_h\to$ GLS - Galerkin Least Squares method
- $\varphi(v_h)=\mathcal{L}_{ss}v_h\to$ SUPG - Streamline Upwind Petrov-Galerkin method

Brief notation remark: $\mathcal{L}=\mathcal{L}_s+\mathcal{L}_{ss}$ (symmetric + skew-symmetric part) Which we define as

$$\begin{aligned}{}_{V'}\langle\mathcal{L}_s u,v\rangle_V &= {}_V\langle v,\mathcal{L}_s u\rangle_{V'}\quad\forall\, u,v\in V\\ {}_{V'}\langle\mathcal{L}_{ss} u,v\rangle_V &= -{}_V\langle v,\mathcal{L}_{ss} u\rangle_{V'}\quad\forall\, u,v\in V\end{aligned}$$

For matrices it is $A=A_S+A_{SS}$

$$A_S=\frac{1}{2}(A+A^T)\quad A_{SS}=\frac{1}{2}(A-A^T)$$

Let us see an example in the non-conservative form

$$\begin{aligned}\mathcal{L}^1&=-\mu\Delta u+\mathbf{b}\cdot\nabla u+\sigma u\\ &=\underbrace{\left[-\mu\Delta u+\left(\sigma-\frac{1}{2}\mathrm{div}\mathbf{b}\right)u\right]}_{\mathcal{L}^1_s u}+\underbrace{\left[\frac{1}{2}\left(\mathrm{div}(\mathbf{b}u)+\mathbf{b}\cdot\nabla u\right)\right]}_{\mathcal{L}^1_{ss} u}\end{aligned}$$

Indeed we can see

$$\begin{aligned}{}_{V'}\langle\mathcal{L}^1_s,v\rangle_V &= \int_\Omega \mu\nabla u\cdot\nabla v+\left(\sigma-\frac{1}{2}\mathrm{div}\mathbf{b}\right)uv\\ &=\int_\Omega\left[-\mu\Delta v+\left(\sigma-\frac{1}{2}\mathrm{div}\mathbf{b}\right)v\right]u={}_V\langle v,\mathcal{L}^1_s\rangle_{V'}\end{aligned}$$

$$
\begin{aligned}
{}_{V'}\langle \mathcal{L}^1_{ss}, v\rangle_V &= \frac{1}{2}\int_\Omega (\operatorname{div}(\mathbf{b}u)v + (\mathbf{b}\cdot\nabla u)v) \\
&= \frac{1}{2}\int_\Omega (-(\mathbf{b}u)\nabla v + (\mathbf{b}v)\cdot\nabla u) \\
&= \frac{1}{2}\int_\Omega (-(\mathbf{b}\cdot\nabla v)u - \operatorname{div}(\mathbf{b}v)u) = -{}_V\langle u, \mathcal{L}^1_{ss}\rangle_{V'}
\end{aligned}
$$

Remark 4.1
If $\operatorname{div}\mathbf{b} = 0$, which happens if $\mathbf{b}$ is constant, then the conservative and non-conservative forms coincide.

4.3 GLS method (conservative form)

$$
\text{find } u_h \in V_h : a(u_h, v_h) + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_\Omega \mathcal{L}u_h \tau_\mathcal{K}\mathcal{L}v_h = \int_\Omega f v_h + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_\Omega f\tau_\mathcal{K}\mathcal{L}v_h \qquad \forall\, v_h \in V_h
$$

Theorem 4.1
Consider the conservative case. Suppose that

$$
\exists \gamma_0, \gamma_1 > 0 : 0 < \gamma_0 \le \gamma(\mathbf{x}) \le \gamma_1
$$

then, for a suitable constant C, independent of h, we have:

$$
\|u_h\|^2_{GLS} \le C\|f\|^2_{L^2(\Omega)}
$$

where $\|\cdot\|_{GLS}$ will be defined later

Proof. Take $u_h = v_h$. We have

$$
\begin{aligned}
a_h(u_h, u_h) &= \int_\Omega \mu|\nabla u_h|^2 + \underbrace{\int_\Omega \operatorname{div}(\mathbf{b}\,u_h)u_h}_{\substack{=-\int_\Omega \mathbf{b}\cdot(u_h\nabla u_h) \\ =-\frac{1}{2}\int_\Omega \mathbf{b}\cdot\nabla(u_h^2) \\ =\frac{1}{2}\int_\Omega \operatorname{div}\mathbf{b}\,u_h^2}} + \int_\Omega \sigma u_h^2 + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_\mathcal{K} \tau_\mathcal{K}(\mathcal{L}u_h)^2 \\
&= \int_\Omega \mu|\nabla u_h|^2 + \int_\Omega \underbrace{\left(\sigma + \frac{1}{2}\operatorname{div}\mathbf{b}\right)}_{=:\gamma(\mathbf{x})} u_h^2 + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_\mathcal{K} \tau_\mathcal{K}(\mathcal{L}u_h)^2 \\
&=: \|u_h\|^2_{GLS}
\end{aligned}
$$

On the other hand

$$
|F_h(u_h)| \le \left|\int_\Omega f u_h\right| + \left|\textstyle\sum_{\mathcal{K}\in\mathcal{T}_h}\displaystyle\int_\mathcal{K} f\tau_\mathcal{K}\mathcal{L}u_h\right|
$$

where

$$
\begin{aligned}
\left|\int_\Omega f u_h\right| = \left|\int_\Omega \frac{1}{\sqrt{\gamma}} f\sqrt{\gamma}u_h\right| &\underset{\text{Cauchy-Schwartz}}{\le} \left\|\frac{1}{\sqrt{\gamma}}f\right\|_{L^2(\Omega)} \|\sqrt{\gamma}u_h\|_{L^2(\Omega)} \\
&\underset{\text{Young}}{\le} \left\|\frac{1}{\sqrt{\gamma}}f\right\|^2_{L^2(\Omega)} + \frac{1}{4}\|\sqrt{\gamma}u_h\|^2_{L^2(\Omega)}
\end{aligned}
$$

and where

$$\left|\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}} f\tau_{\mathcal{K}}\mathcal{L}u_h\right| = \left|\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}} \sqrt{\tau_{\mathcal{K}}}f\sqrt{\tau_{\mathcal{K}}}\mathcal{L}u_h\right|$$
$$\underset{\text{Cauchy-Schwartz}}{\leq} \sum_{\mathcal{K}\in\mathcal{T}_h} \|\sqrt{\tau_{\mathcal{K}}}f\|_{L^2(\mathcal{K})}\|\sqrt{\tau_{\mathcal{K}}}\mathcal{L}u_h\|^2_{L^2(\mathcal{K})}$$
$$\underset{\text{Young}}{\leq} \sum_{\mathcal{K}\in\mathcal{T}_h} \|\sqrt{\tau_{\mathcal{K}}}f\|^2_{L^2(\mathcal{K})} + \frac{1}{4}\|\sqrt{\tau_{\mathcal{K}}}\mathcal{L}u_h\|^2_{L^2(\mathcal{K})}$$

So, $a_h(u_h,u_h) = F_h(u_h)$ implies:

$$\begin{aligned}
\|u_h\|^2_{GLS} &= \int_\Omega \mu|\nabla u_h|^2 + \int_\Omega \gamma u_h^2 + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}} \tau_{\mathcal{K}}(\mathcal{L}u_h)^2 \\
&\leq \left[\left\|\frac{1}{\sqrt{\gamma}}f\right\|_{L^2(\Omega)} + \sum_{\mathcal{K}\in\mathcal{T}_h}\|\sqrt{\tau_{\mathcal{K}}}f\|^2_{L^2(\Omega)}\right] \\
&\quad + \frac{1}{4}\left[\int_\Omega \gamma u_h^2 + \sum_{\mathcal{K}\in\mathcal{T}_h}\tau_{\mathcal{K}}(\mathcal{L}u_h)^2\right] \\
&\leq \underbrace{\left(\frac{1}{\gamma_0} + \max_{\mathcal{K}\in\mathcal{T}_h}\tau_{\mathcal{K}}\right)}_{=C(\text{if } \tau_{\mathcal{K}} \text{ uniformly bounded w.r.t. } h)}\|f\|^2_{L^2(\Omega)} + \frac{1}{4}\|u_h\|^2_{GLS}
\end{aligned}$$

In the end

$$\|u_h\|^2_{GLS} \leq \frac{4}{3}C\|f\|^2_{L^2(\Omega)}$$

★

As we already said, a smart choice for $\tau_{\mathcal{K}}$ is $\delta\frac{h_{\mathcal{K}}}{|\mathbf{b}(\mathbf{x})|}$. But another possibility may be

$$\tau_{\mathcal{K}}(\mathbf{x}) = \frac{h_{\mathcal{K}}}{2|\mathbf{b}(\mathbf{x})|}\xi(\mathbb{P}e_{\mathcal{K}})$$

with $\xi(\theta) = \coth(\theta) - \frac{1}{\theta}$. And $\mathbb{P}e_{\mathcal{K}}(\mathbf{x}) = \frac{|\mathbf{b}(\mathbf{x})|}{2\mu(\mathbf{x})}h_{\mathcal{K}}$ is the local Péclet number. Moreover, if $\theta \to 0$, then $\xi(\theta) = \frac{\theta}{3} + o(\theta)$, therefore when $\mathbb{P}e_{\mathcal{K}}(\mathbf{x}) \ll 1$, we have $\tau_{\mathcal{K}}(\mathbf{x}) \to 0$ and no stabilization is needed.

4.4 Convergence of GLS

To state the convergence of GLS we need the inverse inequality, defined as

$$\sum_{\mathcal{K}\in\mathcal{T}_h} h_{\mathcal{K}}^2\int_{\mathcal{K}}|\Delta v_h|^2\,d\mathcal{K} \leq C_0\|\nabla v_h\|^2_{L^2(\Omega)} \quad \forall\, v_h \in X_h^r \tag{4.2}$$

Theorem 4.2 (Convergence of GLS)
Assume that the space V_h satisfies the following local approximation property: for each $v \in V \cap H^{r+1}(\Omega)$, there exists a function $\hat{v}_h \in V_h$ s.t.

$$\|v - v_h\|_{L^2(\mathcal{K})} + h_{\mathcal{K}}\|v - \hat{v}_h\|_{H^1(\mathcal{K})} + h_{\mathcal{K}}^2|v - \hat{v}_h|_{H^2(\mathcal{K})} \leq Ch_{\mathcal{K}}^{r+1}|v|_{H^{r+1}} \tag{4.3}$$

for each $\mathcal{K} \in \mathcal{T}_h$. Moreover, we suppose that for each $\mathcal{K} \in \mathcal{T}_h$ the local Péclet number of K satisfies

$$\mathbb{P}e_{\mathcal{K}}(\mathbf{x}) = \frac{|\mathbf{b}(\mathbf{x})|h_{\mathcal{K}}}{2\mu} > 1 \quad \forall\, \mathbf{x} \in \mathcal{K} \tag{4.4}$$

that is, we are in the pre-asymptotic regime. Finally, we suppose that the inverse inequality holds and that the stabilization parameters satisfies the relation $0 < \delta \leq 2C_0^{-1}$.
Then, as long as $u \in H^{r+1}(\Omega)$, the following super-optimal estimate holds:

$$\|u - u_h\|_{GLS} \leq Ch^{r+\frac{1}{2}}|u|_{H^{r+1}(\Omega)} \tag{4.5}$$

Proof. First, rewrite the error as

$$e_h = u_h - u = \sigma_h - \eta \tag{4.6}$$

with $\sigma_h = u_h - \hat{u}_h, \eta = u - \hat{u}_h$, where $\hat{u}_h$ is a function that depends on u and that satisfies property (4.3). If, for instance, $V_h = X_h^r \cap H_0^1(\Omega)$, we can choose $\hat{u}_h = \prod_h^r u$ that is the finite element interpolant of u.
We start by estimating the norm $\|\sigma_h\|_{GLS}$. By exploiting the strong consistency of the GLS scheme we obtain

$$\|\sigma_h\|_{GLS}^2 = a_h(\sigma_h, \sigma_h) = a_h(u_h - u + \eta, \sigma_h) = a_h(\eta, \sigma_h)$$

Now, thanks to the homogeneous Dirichlet boundary conditions it follows that, by adding and subtracting $\sum_{\mathcal{K}\in\mathcal{T}_h}(\eta, \mathcal{L}\sigma_h)_{\mathcal{K}}$, suitable computations lead to:

$$\begin{aligned}
a_h(\eta, \sigma_h) &= \mu\sigma_\Omega \nabla\eta \cdot \nabla\sigma_h \, d\Omega - \int_\Omega \eta \mathbf{b} \cdot \nabla\sigma_h \, d\Omega + \int_\Omega \sigma\eta\sigma_h \, d\Omega \\
&\quad + \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\mathcal{L}\eta, \frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\sigma_h \right)_{L^2(\mathcal{K})} \\
&= \underbrace{\mu\left(\nabla\eta, \nabla\sigma_h\right)_{L^2(\Omega)}}_{(I)} - \underbrace{\sum_{\mathcal{K}\in\mathcal{T}_h} (\eta, \mathcal{L}\sigma_h)_{L^2(\Omega)}}_{(II)} + \underbrace{2\left(\gamma\eta, \sigma_h\right)_{L^2(\mathcal{K})}}_{(III)} \\
&\quad + \underbrace{\sum_{\mathcal{K}\in\mathcal{T}_h} (\eta, -\mu\Delta\sigma_h)_{L^2(\mathcal{K})}}_{(IV)} + \underbrace{\sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\mathcal{L}\eta, \frac{h_\mathcal{K}}{|\mathbf{b}|\mathcal{L}\sigma_h} \right)_{L^2(\mathcal{K})}}_{(V)}
\end{aligned}$$

Now, we bound each of these terms. By using Cauchy-Schwartz and Young's inequalities we obtain

$$\begin{aligned}
|(I)| &= \left|\mu\left(\nabla\eta, \nabla\sigma_h\right)_{L^2(\mathcal{K})}\right| \leq \frac{\mu}{4}\|\nabla\sigma_h\|_{L^2(\Omega)}^2 + \mu\|\nabla\eta\|_{L^2(\Omega)}^2 \\
|(II)| &= \left|\sum_e it\left(\eta, \mathcal{L}\sigma_h\right)_{L^2(\mathcal{K})}\right| \\
&= \left|\sum_{\mathcal{K}\in\mathcal{T}_h} \left(\sqrt{\frac{|\mathbf{b}|}{\delta h_\mathcal{K}}}\eta, \sqrt{\frac{\delta h_\mathcal{K}}{|\mathbf{b}|}}\mathcal{L}\sigma_h \right)_{L^2(\Omega)}\right| \\
&\leq \frac{1}{4}\sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_k}{|\mathbf{b}|}\mathcal{L}\sigma_h, \mathcal{L}\sigma_h \right) \\
|(III)| &= 2\left|\left(\gamma\eta, \sigma_h\right)_{L^2(\Omega)}\right| = 2\left|\left(\sqrt{\gamma\eta, \sqrt{\gamma}\sigma_h}\right)_{L^2(\Omega)}\right| \\
&\leq \frac{1}{2}\|\sqrt{\gamma}\sigma_h\|_{L^2(\Omega)}^2 + 2\|\sqrt{\gamma}\eta\|_{L^2(\Omega)}^2
\end{aligned}$$

Then, thanks to CS and Young, but also hypotheses (4.4) and (4.2), we obtain

$$
\begin{aligned}
|(IV)| &= \left| \sum_{\mathcal{K}\in\mathcal{T}_h} (\eta, -\mu\Delta\sigma_h)_{L^2(\mathcal{K})} \right| \\
&\leq \frac{1}{4} \sum_{\mathcal{K}\in\mathcal{T}_h} \delta\mu^2 \left(\frac{h_\mathcal{K}}{|\mathbf{b}|}\Delta\sigma_h, \Delta\sigma_h \right)_{L^2(\mathcal{K})} + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta h_\mathcal{K}}\eta, \eta \right)_{L^2(\mathcal{K})} \\
&\leq \frac{1}{8}\delta\mu \sum_{\mathcal{K}\in\mathcal{T}_h} h_\mathcal{K}^2 (\nabla\sigma_h, \nabla\sigma_h)_{L^2(\mathcal{K})} + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta h_\mathcal{K}} \right)_{L^2(\mathcal{K})} \\
&\leq \frac{\sigma C_0 \mu}{8} \|\nabla\sigma_h\|^2_{L^2(\Omega)} + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta h_\mathcal{K}}\eta, \eta \right)_{L^2(\mathcal{K})}
\end{aligned}
$$

The last one can be bounded once again thanks to CS and Young inequalities as follows

$$
\begin{aligned}
|(V)| &= \left| \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\mathcal{L}\eta, \frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\sigma_h \right)_{L^2(\mathcal{K})} \right| \\
&\leq \frac{1}{4} \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\sigma_h, \mathcal{L}\sigma_h \right)_{L^2(\mathcal{K})} + \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\eta, \mathcal{L}\eta \right)_{L^2(\mathcal{K})}
\end{aligned}
$$

So we can rewrite everything bounded as

$$
\begin{aligned}
\|\sigma_h\|^2_{GLS} = a_h(\eta, \sigma_h) &\leq \frac{1}{4}\|\sigma_h\|^2_{GLS} \\
&+ \frac{1}{4} \left(\|\sqrt{\gamma}\sigma_h\|^2_{L^2(\Omega)} + \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\sigma_h, \mathcal{L}\sigma_h \right)_{L^2(\mathcal{K})} \right) + \frac{\delta C_0 \mu}{8}\|\nabla\sigma_h\|^2_{L^2(\Omega)} \\
&+ \underbrace{\mu\|\nabla\eta\|^2_{L^2(\Omega)} + 2\sum_{\mathcal{K}\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta h_\mathcal{K}}\eta, \eta \right)_{L^2(\mathcal{K})} + 2\|\sqrt{\gamma}\eta\|^2_{L^2(\Omega)} + \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_\mathcal{K}}{|\mathbf{b}|}\mathcal{L}\eta, \mathcal{L}\eta \right)_{L^2(\mathcal{K})}}_{\mathcal{E}(\eta)} \\
&\leq \frac{1}{2}\|\sigma_h\|^2_{GLS} + \mathcal{E}(\eta)
\end{aligned}
$$

Having exploited the assumption that $\delta \leq 2C_0^{-1}$. We can state then

$$
\|\sigma_h\|^2_{GLS} \leq 2\mathcal{E}(\eta)
$$

It's time to estimate $\mathcal{E}(\eta)$, by bounding each of its summands separately. To do this, we will use the local approximation property (4.3) and the local Péclet (4.4). Moreover, we observe that the constant C, introduced in the remainder, depends neither on h nor on $\mathbb{P}e_\mathcal{K}$, but can depend on other quantities such as the constant γ_1, the reaction constant σ or the norm $\|\mathbf{b}\|_{L^\infty(\Omega)}$, the stabilization parameter δ.
Then we have

$$
\begin{aligned}
\mu\|\nabla\eta\|^2_{L^2(\Omega)} &\leq C\mu h^{2r}|u|^2_{H^{r+1}(\Omega)} \\
&\leq C\frac{\|\mathbf{b}\|_{L^\infty(\Omega)h}}{2} h^{2r}|u|^2_{H^{r+1}(\Omega)} \leq Ch^{2r+1}|u|^2_{H^{r+1}(\Omega)} \\
2\sum_{\mathcal{K}\in\mathcal{T}_h} \left(\frac{|\mathbf{b}|}{\delta h_\mathcal{K}}\eta, \eta \right)_{L^2(\mathcal{K})} &\leq C\frac{\|\mathbf{b}\|_{L^\infty(\Omega)h}}{2} \sum_{\mathcal{K}\in\mathcal{T}_h} \frac{1}{h_\mathcal{K}} h_\mathcal{K}^{2r+1}|u|^2_{H^{r+1}(\Omega)} \\
&\leq ch^{2r+1}|u|^2_{H^{r+1}(\Omega)} \\
2\|\sqrt{\gamma}\eta\|^2_{L^2(\Omega)} &\leq 2\gamma_1\|\eta\|^2_{L^2(\Omega)} \leq Ch^{2r+1}|u|^2_{H^{r+1}(\Omega)}
\end{aligned}
$$

For the fourth term we have

$$\begin{aligned}
\sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\frac{h_{\mathcal{K}}}{|\mathbf{b}|}\mathcal{L}\eta, \mathcal{L}\eta\right)_{L^2(\mathcal{K})} &= \sum_{\mathcal{K}\in\mathcal{T}_h} \left\| \sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\mathcal{L}\eta \right\|^2_{L^2(\mathcal{K})} \\
&= \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left\| -\mu\sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\Delta\eta + \sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\operatorname{div}(\mathbf{b}\eta) + \sigma\sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\eta \right\|^2_{L^2(\mathcal{K})} \qquad (4.7) \\
&\le C \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left(\left\| \mu\sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\Delta\eta \right\|^2_{L^2(\mathcal{K})} + \left\| \sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\operatorname{div}(\mathbf{b}\eta) \right\|^2_{L^2(\mathcal{K})} \right. \\
&\quad \left. + \left\| \sigma\sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\eta \right\|^2_{L^2(\mathcal{K})} \right)
\end{aligned}$$

Now it is easy to prove that the second and third term of the summands can be bounded using a term or the form $Ch^{2r+1}|u|^2_{H^{r+1}(\Omega)}$, for a suitable choice of the constant C. For the first term we have

$$\begin{aligned}
\sum_{\mathcal{K}\in\mathcal{T}_h} \delta \left\| \mu\sqrt{\frac{h_{\mathcal{K}}}{|\mathbf{b}|}}\Delta\eta \right\|^2_{L^2(\mathcal{K})} &\le \sum_{\mathcal{K}\in\mathcal{T}_h} \delta \frac{h^2_{\mathcal{K}}\mu}{2} \|\Delta\eta\|^2_{L^2(\mathcal{K})} \\
&\le C\delta \|\mathbf{b}\|_{L^\infty(\Omega)} \sum_{\mathcal{K}\in\mathcal{T}_h} h^3_{\mathcal{K}} \|\Delta\|^2_{L^2(\mathcal{K})} \le |u|^2_{H^{r+1}(\Omega)}
\end{aligned}$$

having used again (4.3) and (4.4). Now we can conclude that

$$\mathcal{E}(\eta) \le Ch^{2r+1}|u|^2_{H^{r+1}(\Omega)}$$

that is

$$\|\sigma_h\|_{GLS} \le Ch^{r+\frac{1}{2}}|u|_{H^{r+1}(\Omega)} \qquad (4.8)$$

Reverting to (4.6), to obtain the desired estimate for the norm $\|u_h - u\|_{GLS}$ we need to estimate $\|\eta\|_{GLS}$. But thanks to (4.7) we obtain

$$\|\eta\|_{GLS} \le Ch^{r+\frac{1}{2}}|u|_{H^{r+1}(\Omega)}$$

Combining this with (4.8) we obtain (4.5). ★

5 Parabolic equations

5.1 Introduction

Now we consider parabolic equations of the form

$$\frac{\partial u}{\partial t} + \mathcal{L}u = f \quad \mathbf{x} \in \Omega, t > 0 \tag{5.1}$$

where:

- Ω is a domain of $\mathbb{R}^d$ with $d = 1, 2, 3$
- $f = f(\mathbf{x}, t)$ is a given function
- $\mathcal{L} = \mathcal{L}(\mathbf{x})$ is a generic elliptic operator acting on $u = u(\mathbf{x}, t)$

When solved for a bounded time interval, for example $0 < t < T$, the region $Q_T = \Omega \times (0, T)$ is called cylinder in the space $\mathbb{R}^d \times \mathbb{R}^+$. In (5.1) must be assigned an initial condition

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \tag{5.2}$$

also we'll need some BC, like

$$\begin{aligned} u(\mathbf{x}, t) &= \varphi(\mathbf{x}, t) \quad \mathbf{x} \in \Gamma_D \text{ and } t > 0 \\ \frac{\partial u(\mathbf{x}, t)}{\partial n} &= \psi(\mathbf{x}, t) \quad \mathbf{x} \in \Gamma_N \text{ and } t > 0 \end{aligned} \tag{5.3}$$

where u_0, φ and ψ are given function and $\{\Gamma_D, \Gamma_N\}$ provides a boundary partition that is $\Gamma_D \cup \Gamma_N = \partial\Omega, \Gamma_D^o \cap \Gamma_N^o = \varnothing$. For obvious reasons Γ_D is the Dirichlet boundary, while Γ_N is the Neumann one. In the one dimensional case the problem becomes

$$\begin{aligned} &\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = f && 0 < x < d, t > 0 \\ &u(x, 0) = u_0(x) && 0 < x < d \\ &u(0, t) = u(d, t) = 0 && t > 0 \end{aligned} \tag{5.4}$$

which describes the evolution of the temperature $u(x, t)$ at point x and time t of a metal bar of length d occupying the interval $[0, d]$, whose thermal conductivity is ν and whose endpoints are kept at a constant temperature of zero degrees. The function u_0 describes the temperature in the initial state, while f represents the heat generated per unit of length by the bar. This is called the heat equation.

5.2 Weak formulation and approximation

We proceed as usual by multiplying for each $t > 0$ the differential equation by a test function $v = v(\mathbf{x})$ and integrating. We set $V = H^1_{\Gamma_D}(\Omega)$ and, for each $t > 0$, we seek $u(t) \in V$ s.t.

$$\int_\Omega \frac{\partial u(t)}{\partial t} v \, d\Omega + a(u(t), v) = \int_\Omega f(t) v \, d\Omega \quad \forall \, v \in V \tag{5.5}$$

where

- $u(0) = u_0$
- $a(\cdot, \cdot)$ is the bilinear form associated to the operator $\mathcal{L}$
- we have supposed for simplicity $\varphi = 0$ and $\psi = 0$

Definition 5.1

A bilinear form $a(\cdot,\cdot)$ is said weakly coercive if

$$\exists\, \lambda \geq 0, \exists\, \alpha > 0 : a(v,v) + \lambda \|v\|^2_{L^2(\Omega)} \geq \alpha \|v\|^2_V \quad \forall\, v \in V$$

Rationale for weak coercivity

$$\frac{\partial u}{\partial t} + \mathcal{L}u = f$$

now we perform a change of variable ($u = e^{\lambda t}$)

$$\frac{\partial w}{\partial t} + \mathcal{L}w + \lambda w = e^{-\lambda t} f$$

so that the bilinear form

$$\tilde{a}(w,v) := a(w,v) + \lambda (w,v)_{L^2(\Omega)} \Rightarrow \tilde{a}(w,w) := a(w,w) + \lambda \|w\|^2_{L^2(\Omega)}$$

Theorem 5.1
Suppose that the bilinear form $a(\cdot,\cdot)$ is continuous and weakly coercive. Moreover, we require $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$. Then, (5.5) admits a unique solution $u \in \mathcal{C}^0(\mathbb{R}^+; L^2(\Omega))$. Also, $u \in L^2(\mathbb{R}^+; V)$ and $\frac{\partial u}{\partial t} \in L^2(\mathbb{R}^+; V')$.

5.3 Algebraic formulation

Now we can use Galerkin to approximate, for each $t > 0$, we need to find $u_h(t) \in V_h$ s.t.

$$\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + a(u_h(t), v_h) = \int_\Omega f(t) v_h \, d\Omega \quad \forall\, v_h \in V_h \tag{5.6}$$

with $u_h(0) = u_{0h}$, where $V_h \subset V$ is a suitable space of finite dimension and u_{0h} is a convenient approximation of u_0 in the space V_h.
Now we need to discretize the temporal variable, because, as of now, we obtained a semi-discretization of the problem.
We introduce a basis $\{\varphi_j\}$ for V_h, and we observe that it suffices that (5.6) is verified for the basis function in order to be satisfied by all the functions in the subspace.
Moreover, since for each $t > 0$, the solution to the Galerkin problem belongs to the subspace as well, we will have

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

where the coefficients $\{u_j(t)\}$ represent the unknown of the problem.
Denoting by $\dot{u}_j(t)$ the temporal derivatives of $u_j(t)$, (5.6) becomes

$$\int_\Omega \sum_{j=1}^{N_h} \dot{u}_j(t) \varphi_j \varphi_i \, d\Omega + a\left(\sum_{j=1}^{N_h} u_j(t) \varphi_j, \varphi_i \right) = \int_\Omega f(t) \varphi_i \, d\Omega, \quad i = 1, 2, \ldots, N_h$$

that is

$$\sum_{j=1}^{N_h} \dot{u}_j(t) \underbrace{\int_\Omega \varphi_j \varphi_i \, d\Omega}_{m_{ij}} + \sum_{j=1}^{N_h} u_j(t) \underbrace{a(\varphi_j, \varphi_i)}_{a_{ij}} = \underbrace{\int_\Omega f(t) \varphi_i \, d\Omega}_{f_i} \quad i = 1, 2, \ldots, N_h \tag{5.7}$$

If we define the vector of unknowns $\mathbf{u} = (u_1(t), u_2(t), \ldots, u_{N_h}(t))^T$, the mass matrix $M = [m_{ij}]$, the stiffness matrix $A = [a_{ij}]$ and the right-hand side vector $\mathbf{f} = (f_1(t), f_2(t), \ldots, f_{N_h}(t))^T$, the system (5.7) can be rewritten as

$$M\dot{\mathbf{u}}(t) + A\mathbf{u}(t) = \mathbf{f}(t)$$

5.4 Time discretization

For the numerical solution of this ODE system we will use the $\theta-$method, which discretizes the temporal difference quotient and replaces the other terms with a linear combination of the value at time t^k and of the value at time t^{k+1}, depending on the real parameter $0 \leq \theta \leq 1$.

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + A[\theta\mathbf{u}^{k+1} + (1-\theta)\mathbf{u}^k] = \theta\mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k \tag{5.8}$$

The real positive parameter $\Delta t = t^{k+1} - t^k$, $k = 0, 1, \ldots$ denotes the discretization step.
Some particular cases of (5.7)

- $\theta = 0$ The forward Euler method

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + A\mathbf{u}^k = \mathbf{f}^k$$

- $\theta = 1$ The backward Euler method

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + A\mathbf{u}^{k+1} = \mathbf{f}^{k+1}$$

- $\theta = \frac{1}{2}$ The Crank-Nicolson (or trapezoidal) method

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + \frac{1}{2}A\left(\mathbf{u}^{k+1}+\mathbf{u}^k\right) = \frac{1}{2}\left(\mathbf{f}^{k+1}+\mathbf{f}^k\right)$$

  which is of the second order in Δt.

Let us consider the two extremal cases $\theta = 0$ and $\theta = 1$. In the first case the system to solve is only the mass matrix $\frac{M}{\Delta t}$, while in the other case $\frac{M}{\Delta t} + A$. M is invertible, being positively defined.
In the case $\theta = 0$ the scheme is not unconditionally stable, and in the case where V_h is a subspace of finite elements, there is the following stability condition

$$\exists\, c > 0 : \Delta t \leq ch^2 \quad \forall\, h > 0$$

so Δt cannot be chosen irrespective of h.
In this case, if we make M diagonal, we actually decouple the system. This operation is called lumping of the mass matrix.
When $\theta > 0$, the system will have the form $K\mathbf{u}^{k+1} = \mathbf{g}$, where $\mathbf{g}$ is the source term and $K = \frac{M}{\Delta t}$. The latter is invariant in time (the operator $\mathcal{L}$ being independent of time), so, if the spatial mesh doesn't change, it can be factorized only once at the beginning of the process.
Then, since M is symmetric, if A is symmetric, also K will be symmetric too. Hence, we can use, for example, the Cholesky factorization $K = HH^T$, with H being lower triangular. At each timestep, then, will be solved two triangular systems

$$\begin{aligned} H\mathbf{y} &= \mathbf{g} \\ H^T\mathbf{u}^{k+1} &= \mathbf{y} \end{aligned}$$

5.5 A priori estimate

Let us know consider (5.5). Since the corresponding equation must hold for each $v \in V$, it will be legitimate to set $v = u(t)$ with t being given. Yielding

$$\int_\Omega \frac{\partial u(t)}{\partial t}u(t) = \frac{1}{2}\frac{\partial}{\partial t}\int_\Omega |u(t)|^2\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|^2_{L^2(\Omega)} \tag{5.9}$$

Considering the individual terms, the first one is

$$\int_\Omega \frac{\partial u(t)}{\partial t}u(t)\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\int_\Omega |u(t)|^2\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|^2_{L^2(\Omega)} \tag{5.10}$$

Assuming that the bilinear form is coercive, with α coercivity constant, we obtain

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

while, thanks to the CS inequality we find:

$$(f(t), u(t)) \leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \tag{5.11}$$

So (5.9) becomes

$$\frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|_{L^2(\Omega)}^2 + \alpha \|u(t)\|_V^2 \leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)}$$

Now let's define two important inequalities

Definition 5.2 (Young's inequality)
$\forall\, a, b \in \mathbb{R}$

$$ab \leq \varepsilon \alpha^2 + \frac{1}{4\varepsilon} b^2 \quad \forall\, \varepsilon > 0 \tag{5.12}$$

Definition 5.3 (Poincaré inequality)
If Γ_D is a set of positive measure, then:

$$\exists\, C_\Omega > 0 : \|v\|_{L^2(\Omega)} \leq C_\Omega \|\nabla v\|_{L^2(\Omega)} \quad \forall\, v \in H^1_{\Gamma_D(\Omega)} \tag{5.13}$$

By using these two we obtain

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|_{L^2(\Omega)}^2 + \alpha \|\nabla u\|_{L^2(\Omega)}^2 &\leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \\ &\leq \frac{C_\Omega^2}{2\alpha}\|f(t)\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|\nabla u(t)\|_{L^2(\Omega)}^2 \end{aligned} \tag{5.14}$$

Then, by integrating in time, we obtain, for all $t > 0$

$$\|u(t)\|_{L^2(\Omega)}^2 + \alpha \int_0^t \|\nabla u(s)\|_{L^2(\Omega)}^2 \, ds \leq \|u_0\|_{L^2(\Omega)}^2 + \frac{C_\Omega^2}{\alpha}\int_0^t \|f(s)\|_{L^2(\Omega)}^2 \, ds \tag{5.15}$$

This is an a priori energy estimate. Note that

$$\frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|_{L^2(\Omega)}^2 = \|u(t)\|_{L^2(\Omega)} \frac{\partial}{\partial t}\|u(t)\|_{L^2(\Omega)}$$

then, from (5.9), using (5.10), (5.11) and (5.13)

$$\begin{aligned} \|u(t)\|_{L^2(\Omega)} \frac{\partial}{\partial t}\|u(t)\|_{L^2(\Omega)} + \frac{\alpha}{C_\Omega}\|u(t)\|_{L^2(\Omega)} \|\nabla u(t)\|_{L^2(\Omega)} \\ \leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)}, \quad t > 0 \end{aligned} \tag{5.16}$$

If $\|u(t)\|_{L^2(\Omega)} \neq 0$, we can divide by $\|u(t)\|_{L^2(\Omega)}$ and integrate in time to obtain another estimate.

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} \, ds \quad t > 0 \tag{5.17}$$

Let us now use the first inequality in (5.14) and integrate in time to yield

$$\begin{aligned} \|u(t)\|_{L^2(\Omega)}^2 + 2\alpha \int_0^t \|\nabla u(s)\|_{L^2(\Omega)}^2 \, ds &\underset{(5.14)}{\leq} \|u(t)\|_{L^2(\Omega)}^2 + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u(s)\|_{L^2(\Omega)} \, ds \\ &\underset{(5.17)}{\leq} \|u_0\|_{L^2(\Omega)}^2 + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \cdot \Bigg(\|u_0\|_{L^2(\Omega)} \\ &\qquad + \int_0^s \|f(\tau)\|_{L^2(\Omega)} \, d\tau \Bigg) \, ds \\ &= \|u_0\|_{L^2(\Omega)}^2 \int_0^t \|f(s)\|_{L^2(\Omega)} \|u_0\|_{L^2(\Omega)} \, ds \\ &\qquad + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} \, d\tau ds \end{aligned} \tag{5.18}$$

Now, noticing that

$$\|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)}\, d\tau = \frac{\partial}{\partial s}\left(\int_0^s \|f\|_{L^2(\Omega)}\, d\tau\right)^s$$

we can rewrite the right-hand side of (5.18) as:

$$\begin{aligned}\|u_0\|^2_{L^2(\Omega)} \int_0^t \|f(s)\|_{L^2(\Omega)}\|u_0\|_{L^2(\Omega)}\, ds + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)}\, d\tau ds \\ = \left(\|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)}\, ds\right)^2\end{aligned}$$

Therefore we can obtain the following estimate

$$\begin{aligned}&\left(\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)}\, ds\right)^{\frac{1}{2}} \\ &\leq \|u_0\|^2_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)}\, ds \quad t > 0\end{aligned} \tag{5.19}$$

Now that we have found an estimate for (5.5), we can now estimate its discretization (5.6), like (5.15)

$$\|u_h(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)}\, ds \leq \|u_{0h}\|^2_{L^2(\Omega)} + \frac{C^2_\Omega}{\alpha}\int_0^t \|f(s)\|^2_{L^2(\Omega)}\, ds \tag{5.20}$$

We can proceed as before, simply taking for every $t > 0$, $v_h = u_h(t)$. Since $u_h(0) = u_{0h}$ we obtain the discrete counterparts of (5.17) and (5.19)

$$\|u_h(t)\|_{L^2(\Omega)} \leq \|u_{0h}(t)\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)}\, ds \quad t > 0 \tag{5.21}$$

and

$$\begin{aligned}&\left(\|u_h(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)}\, ds\right)^{\frac{1}{2}} \\ &\leq \|u_{0h}(t)\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)}\, ds \quad t > 0\end{aligned} \tag{5.22}$$

5.6 Convergence analysis

Theorem 5.2
There exists a constant $C > 0$ independent of both t and h s.t.

$$\begin{aligned}&\left\{\|u(t) - u_h(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u(s) - \nabla u_h(s)\|^2_{L^2(\Omega)}\, ds\right\}^{\frac{1}{2}} \\ &\leq Ch^r \left\{|u_0|^2_{H^r(\Omega)} + \int_0^t |u(s)|^2_{H^{r+1}(\Omega)}\, ds + \int_0^t \left|\frac{\partial u(s)}{\partial s}\right|^2_{H^{r+1}(\Omega)} ds\right\}^{\frac{1}{2}}\end{aligned}$$

5.7 Stability analysis of the θ-method

We can now analyze the stability of discretized problem. By applying the θ-method to the Galerkin problem (5.6) we obtain

$$\begin{aligned}&\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, v_h\right) + a(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h) \\ &= \theta F^{k+1}(v_h) + (1-\theta)F^k(v_h) \quad \forall\, v_h \in V_h\end{aligned} \tag{5.23}$$

for each $k \geq 0$, with F^k indicating the functional evaluated at the time k. For the analysis we will consider $F = 0$, and the case of implicit Euler $\theta = 1$.

$$\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, v_h\right) + a(u_h^{k+1}, v_h) = 0 \quad \forall\, v_h \in V_h$$

By choosing $v_h = u_h^{k+1}$, we obtain

$$(u_h^{k+1}, u_h^{k+1}) + \Delta t\, a(u_h^{k+1}, u_h^{k+1}) = (u_h^k, u_h^{k+1})$$

and then, thanks to these inequalities

$$a(u_h^{k+1}, u_h^{k+1}) \geq \alpha \left\| u_h^{k+1} \right\|_V^2$$
$$(u_h^k, u_h^{k+1}) \leq \frac{1}{2}\left\| u_h^k \right\|_{L^2(\Omega)}^2 + \frac{1}{2}\left\| u_h^{k+1} \right\|_{L^2(\Omega)}^2$$

which are derived from the coercivity of $a(\cdot,\cdot)$ and the CS inequality we obtain

$$\left\| u_h^{k+1} \right\|_{L^2(\Omega)}^2 + 2\alpha\Delta t \left\| u_h^{k+1} \right\|_V^2 \leq \left\| u_h^k \right\|_{L^2(\Omega)}^2 \tag{5.24}$$

Observing that $\left\| u_h^{k+1} \right\|_V \geq \left\| u_h^{k+1} \right\|_{L^2(\Omega)}$, we deduce from (5.24) that

$$(1 + 2\alpha\Delta t)\left\| u_h^{k+1} \right\|_{L^2(\Omega)}^2 \leq \left\| u_h^k \right\|_{L^2(\Omega)}^2$$

hence

$$\left\| u_h^{k+1} \right\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{1 + 2\alpha\Delta t}} \left\| u_h^k \right\|_{L^2(\Omega)}$$

which entails

$$\left\| u_h^k \right\|_{L^2(\Omega)} \leq \left(\frac{1}{\sqrt{1 + 2\alpha\Delta t}}\right)^k \|u_{0h}\|_{L^2(\Omega)}$$

and therefore

$$\lim_{k\to\infty} \left\| u_h^k \right\|_{L^2(\Omega)} = 0$$

so the backward Euler method is absolutely stable without any restriction on Δt.
Now we assume $f \neq 0$

$$\underbrace{\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, u_h^{k+1}\right)}_{(1)} + \underbrace{a(u_h^{k+1}, u_h^{k+1})}_{(2)} = \underbrace{\int_\Omega f^{k+1} u_h^{k+1}}_{(3)}$$

where

$$(1) \geq \frac{1}{2\Delta t}\left(\left\| u_h^{k+1} \right\|_{L^2(\Omega)}^2 - \left\| u_h^k \right\|_{L^2(\Omega)}^2\right)$$
$$(2) \geq \alpha \left\| u_h^{k+1} \right\|_{L^2(\Omega)}^2$$
$$(3) \underset{(\mathrm{CS})}{\leq} \left\| f^{k+1} \right\|_{L^2(\Omega)} \left\| u_h^{k+1} \right\|_V \underset{(\mathrm{Young})}{\leq} \frac{1}{2\alpha}\left\| f^{k+1} \right\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\left\| u_h^{k+1} \right\|_V^2$$

Then, after summation on k, for $k = 0, \ldots, n-1$:

$$\left\| u_h^2 \right\|_{L^2(\Omega)} + \underbrace{\alpha \sum_{k=1}^n \Delta t \left\| u_h^k \right\|_V^2}_{\simeq \alpha \int_0^{t_n} \|u_h(t)\|_V^2\, dt} \leq \|u_{0h}\|_{L^2(\Omega)}^2 + \underbrace{\frac{1}{\alpha}\sum_{k=1}^n \Delta t \left\| f^k \right\|_{L^2(\Omega)}^2}_{\simeq \frac{1}{\alpha}\int_0^{t_n} \|f^k\|_{L^2(\Omega)}^2\, dt}$$

meaning that we have unconditional stability.
Now, before analyzing the general case, where $0 \leq \theta \leq 1$, we need the following definition

Definition 5.4
We say that the scalar λ is an eigenvalue of the bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ and that $w \in V$ is its corresponding eigenfunction if

$$a(w, v) = \lambda(w, v) \quad \forall v \in V$$

If the bilinear form $a(\cdot,\cdot)$ is symmetric and coercive, it has positive and real eigenvalues forming an infinite sequence, moreover, its eigenfunctions form a basis of the space V. The eigenvalues and eigenfunctions of $a(\cdot,\cdot)$ can be approximated by finding the pairs $\lambda_h \in \mathbb{R}$ and $w_h \in V_h$, which satisy

$$a(w_h, v_h) = \lambda_h(w_h, v_h) \quad \forall\, v_h \in V_h \tag{5.25}$$

From an algebraic viewpoint, problem (5.25) can be formulated as

$$A\mathbf{w} = \lambda_h M\mathbf{w},$$

where A is the stiffness matrix and M the mass matrix. This is a generalized eigenvalue problem. Such eigenvalues are all positive and N_h (the usual dimension of V_h). After ordering them in ascending order we have

$$\lambda_h^{N_h} \quad \text{for } N_h \to \infty.$$

Moreover, the corresponding eigenfunctions form a basis for the subspace V_h and can be chosen to be orthonormal w.r.t. the scalar product of $L^2(\Omega)$. This means that, by denoting with w_h^i the eigenfunction corresponding to the eigenvalue λ_h^i, we have $(w_h^i, w_h^j) = \delta_{ij}$, $\forall\, i, j = 1, \dots, N_h$. Thus, each function $v_h \in V_h$ can be represented as follows

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

and, thanks to the eigenfunctions orthonormality

$$\|v_h\|^2_{L^2(\Omega)} = \sum_{j=1}^{N_h} v_j^2 \tag{5.26}$$

Let us know consider an arbitrary $\theta \in [0, 1]$ and assume that $a(\cdot,\cdot)$ is symmetric. Since $u_h^k \in V_h$ we can write

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

In this case, however, u_j^k no longer represents the nodal values of u_h^k.
If we now set $F = 0$ in (5.23) and take $v_h = w_h^i$, we find

$$\frac{1}{\Delta t}\sum_{k=1}^{N_h}[u_j^{k+1} - u_j^k](w_h^i, w_h^j) + \sum_{j=1}^{N_h} + \sum_{j=1}^{N_h}[\theta u_j^{k+1} + (1-\theta)u_j^k]a(w_h^j, w_h^i) = 0,$$

for each $i = 1, \dots, N_h$. For each pair $i, j = 1, \dots, N_h$ we have

$$a(w_h^j, w_h^i) = \lambda_h^j(w_h^j, w_h^i) = \lambda_h^i,$$

and thus, for each $i = 1, \dots, N_h$,

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} + [\theta u_i^{k+1} + (1-\theta)u_i^k]\lambda_h^i = 0.$$

Solving for u_i^{k+1}, we find:

$$u_i^{k+1} = u_i^k \frac{1 - (1-\theta)\lambda_h^i \Delta t}{1 + \theta\lambda_h^i \Delta t}.$$

Now, recalling (5.26), we can conclude that absolute stability comes from the following inequality

$$\left|\frac{1-(1-\theta)\lambda_h^i\Delta t}{1+\theta\lambda_h^i\Delta t}\right| < 1$$

that is

$$-1-\theta\lambda_h^i\Delta t < 1-(1-\theta)\lambda_h^i\theta < 1+\theta\lambda_h^i\theta.$$

Hence,

$$-\frac{2}{\lambda_h^i\Delta t}-\theta < \theta-1 < \theta.$$

While the second inequality is always verified, we can rewrite the first one as:

$$2\theta-1 > -\frac{2}{\lambda_h^i\Delta t}.$$

If $\theta \geq \frac{1}{2}$, the left-hand side is nonnegative, while the right-hand side is negative, so the relation always holds. In the case $\theta < \frac{1}{2}$, the inequality is satisfied if

$$\Delta t < \frac{2}{(1-2\theta)\lambda_h^i}. \tag{5.27}$$

So, we have that

- If $\theta \geq \frac{1}{2}$, the θ-method is unconditionally absolutely stable, for every value of Δt.
- If $\theta < \frac{1}{2}$ the θ-method is absolutely stable only if (5.27) is satisfied.

Thanks to the definition of eigenvalue (5.25) and the continuity property of $a(\cdot,\cdot)$, we deduce that

$$\lambda_h^{N_h} = \frac{a(w_{N_h}, w_{N_h})}{\|w_{N_h}\|^2_{L^2(\Omega)}} \leq \frac{M\|w_{N_h}\|^2_V}{\|w_{N_h}\|^2_{L^2(\Omega)}} \leq M(1+C^2h^{-2}).$$

The constant $C > 0$, which appears in the latter step derives from the following inverse inequality

$$\exists\, C > 0 : \|\nabla v_h\|_{L^2(\Omega)} \leq Ch^{-1}\|v_h\|_{L^2(\Omega)} \quad \forall\, v_h \in V_h$$

Hence, for h small enough, $\lambda_h^{N_h} \leq Ch^{-2}$. In fact, we can prove that $\lambda_h^{N_h}$ is indeed of the order h^-2, that is

$$\lambda_h^{N_h} = \max_i \lambda_h^i \simeq ch^{-2}.$$

Knowing that, we obtain the stability of the θ-method for $\theta < \frac{1}{2}$, which is

$$\Delta t \leq C(\theta)h^2 \tag{5.28}$$

where $C(\theta)$ denotes a positive constant depending on θ. We cannot choose Δt without keeping in mind h.

5.8 Convergence analysis

Theorem 5.3
Under the hypothesis that u_0, f and the exact solution are sufficiently regular, the following a priori estimate holds $\forall\, n \geq 1$:

$$\|u(t^n)-u_h^n\|^2_{L^2(\Omega)} + 2\alpha\Delta t\sum_{k=1}^{n}\left\|u(t^k)-u_n^k\right\|^2_V \leq C(u_0,f,u)(\Delta t^{p(\theta)}+h^{2r})$$

where $p(\theta)=2$ if $\theta \neq \frac{1}{2}$, $p(\frac{1}{2})=4$ and C depends on its arguments, but not on Δt or h.

5.9 Parabolic ADR equation

Consider the parabolic PDE, where $\Omega \subset \mathbb{R}^2$ is an open bounded domain

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu \Delta u + \beta \cdot \nabla u + \sigma u = f & \text{in } \Omega \times (0,T) \\ u = 0 & \text{on } \partial\Omega \times (0,T) \\ u(0) = u_0 & \text{in } \Omega \end{cases} \tag{5.29}$$

where μ, β, σ and f are regular functions, satisfying:

$$\begin{aligned} &0 < \mu_0 \leq \mu \leq \mu_1 && \text{a.e. in } \Omega \\ &|\beta| \leq b_1 && \text{a.e. in } \Omega \\ &0 < \sigma_0 \leq \sigma \leq \sigma_1 && \text{a.e. in } \Omega \end{aligned}$$

Then, introducing a finite element space $V_h \subset H_0^1(\Omega)$, the semi-discrete Galerkin formulation reads

$$\begin{aligned} \text{for all } t \in (0,T] \text{ find } u_h(t) \in V_h : \int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, dx + \int_\Omega \mu \nabla u_h(t) \cdot \nabla v_h + \int_\Omega \beta \cdot \nabla u_h(t) v_h \\ + \int_\Omega \sigma u_h(t) v_h = \int_\Omega f v_h \quad \forall\, v_h \in V_h \end{aligned} \tag{5.30}$$

And such that $u_h(0) = u_{0h}$, where u_{0h} is the projection of the initial condition into V_h.

5.9.1 A semimplicit scheme

Consider now a time-advancing scheme, where the diffusion and reaction term are treated implicitly, while the advection term is treated explicitedly. Let us denote $t_k = k\Delta t$, for $k = 0, \dots, N$, where $\Delta t = \frac{T}{N}$. Let now u_h^k be the approximation of $u(t_k)$. A fully discretized version of (5.30) reads:

$$\begin{cases} \left(\dfrac{u_h^{k+1} - u_H^k}{\Delta t}, v_h \right) + \left(\mu \nabla u_h^{k+1}, v_h \right) + \left(\beta \cdot \nabla u_h^k, v_h \right) \\ \qquad\qquad + \left(\sigma u_h^{k+1}, v_h \right) = (f, v_h) \quad \forall\, v_h \in V_h,\ k = 0, \dots, N-1 \\ u_h(0) = u_{0h} \end{cases} \tag{5.31}$$

Theorem 5.4
If the coefficients of the problem staisfy

$$b_1^2 < 4\mu_0\sigma_0 \tag{5.32}$$

then the semimplicit scheme (5.31) is absolutely stable for any chance of Δt. Consider now the case $\sigma = 0$. If the coefficients of the problem satisfy

$$b_1 < \frac{\mu_0}{C_p} \tag{5.33}$$

with C_p being the Poincaré constant, then the scheme is absolutely stable, for any choice of Δt.

Proof. Let us choose $v_h = u_h^{k+1}$. We have

$$\left(\mu \nabla u_h^{k+1}, \nabla u_h^{k+1} \right) \geq \mu_0 \left\| \nabla u_h^{k+1} \right\|^2_{L^2(\Omega)}$$
$$\left(\sigma u_h^{k+1}, u_h^{k+1} \right) \geq \sigma_0 \left\| u_h^{k+1} \right\|^2_{L^2(\Omega)}$$

which entails, for every k

$$\left\| u_h^{k+1} \right\|^2_{L^2(\Omega)} + \Delta t \mu_0 \left\| \nabla u_h^{k+1} \right\|^2_{L^2(\Omega)} + \Delta t \sigma_0 \left\| u_h^{k+1} \right\|^2_{L^2(\Omega)} \leq \left| \left(u_h^k, u_h^{k+1} \right) \right| + \Delta t \left| \left(\beta \cdot \nabla u_h^k, u_h^{k+1} \right) \right|$$

The two right-hand side terms can be bounded by combining (5.12) and (5.11):

$$\left|\left(u_h^k, u_h^{k+1}\right)\right| \leq \frac{1}{2\eta_1}\left\|u_h^k\right\|^2_{L^2(\Omega)} + \frac{\eta_1}{2}\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)}$$
$$\left|\left(\beta\cdot\nabla u_h^k, u_h^{k+1}\right)\right| \leq \frac{b_1}{2\eta_2}\left\|\nabla u_h^k\right\|^2_{L^2(\Omega)} + \frac{\eta_2 b_1}{2}\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)}$$

where the positive constant η_1 and η_2 will be later fixed accordingly. Now we end up with the following inequality

$$\underbrace{\left[1+\Delta t\sigma_0 - \frac{\eta_1}{2} - \frac{\Delta t\eta_2 b_1}{2}\right]}_{A}\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)} + \underbrace{\Delta t\mu_0}_{B}\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)}$$
$$\leq \underbrace{\frac{1}{2\eta_1}}_{A'}\left\|u_h^k\right\|^2_{L^2(\Omega)} + \underbrace{\frac{\Delta t b_1}{2\eta_2}}_{B'}\left\|\nabla u_h^k\right\|^2_{L^2(\Omega)}$$

In order to prove stability we need $A > A'$ and $B > B'$. In fact, if this were true, we would end up with

$$A\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)} + B\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)} \leq \max\left(\frac{A'}{A}, \frac{B'}{B}\right)\left[A\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)} + B\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)}\right],$$

which is the seeked stability in the norm $\|\cdot\|_{A,B} := \left(A\|\cdot\|^2_{L^2(\Omega)} + B\|\nabla\cdot\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}$, equivalent to the standard V_h norm.
Therefore, we look for a suitable choice of η_1 and η_2 that ensures $A > A'$ and $B > B'$. The second inequality is satisfied if and only if

$$\eta_2 = \frac{b_1+\varepsilon}{2\mu_0}$$

for some $\varepsilon > 0$.
Hence, the first inequality reads

$$1+\Delta t\sigma_0 - \frac{\Delta t b_1(b_1+\varepsilon)}{4\mu_0} > \frac{1}{2\eta_1} + \frac{\eta_1}{2}$$

The rigth hand side is minimized for $\eta_1 = 1$, thus leading to the condition

$$4\frac{\mu_0\sigma_0}{b_1(b_1+\varepsilon)} > 1 \tag{5.34}$$

Clearly, it is possible to find $\varepsilon > 0$ such that (5.34) holds if and only if

$$b_1^2 \leq 4\mu_0\sigma_0 \tag{5.35}$$

In conclusion, whenever the coefficients of the problem satisfy the condition (5.35), the scheme (5.31) is absolutely stable for any choice of Δt.
Let us now consider $\sigma = 0$. Proceeding as before, we have:

$$\left[1 - \frac{\eta_1}{2} - \frac{\Delta t\eta_2 b_1}{2}\right]\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)} + \Delta t\mu_0\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)}$$
$$\leq \frac{1}{2\eta_1}\left\|u_h^k\right\|^2_{L^2(\Omega)} + \frac{\Delta t b_1}{2\eta_2}\left\|\nabla u_h^k\right\|^2_{L^2(\Omega)}$$

Introducing now a constant $\omega \in (0,1)$, we have (thanks to (5.13)):

$$\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)} = (1-\omega)\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)} + \omega\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)}$$
$$\geq \frac{1-\omega}{C_p}\left\|u_h^{k+1}\right\|^2_{L^2(\Omega)} + \omega\left\|\nabla u_h^{k+1}\right\|^2_{L^2(\Omega)}$$

Combining the latter inequalities, we obtain

$$\underbrace{\left[1-\frac{\eta_1}{2}-\frac{\Delta t\eta_2 b_1}{2}-\frac{(1-\omega)\Delta t\mu_0}{C_p^2}\right]}_{A}\left\|u_h^{k+1}\right\|_{L^2(\Omega)}^2+\underbrace{\omega\Delta t\mu_0}_{B}\left\|\nabla u_h^{k+1}\right\|_{L^2(\Omega)}^2$$
$$\leq \underbrace{\frac{1}{2\eta_1}}_{A'}\left\|u_h^{k}\right\|_{L^2(\Omega)}^2+\underbrace{\frac{\Delta t b_1}{2\eta_2}}_{B'}\left\|\nabla u_h^{k}\right\|_{L^2(\Omega)}^2$$

As before, we look for conditions such that $A > A'$ and $B > B'$. The second inequality is satisfied if and only if

$$\eta_2 = \frac{b_1+\varepsilon}{2\omega\mu_0}$$

for some $\varepsilon > 0$. Then, the first inequality reads

$$1-\frac{\Delta t b_1(b_1+\varepsilon)}{4\omega\mu_0}+\frac{(1-\omega)\Delta t\mu_0}{C_p^2} > \frac{1}{2\eta_1}+\frac{\eta_1}{2}$$

The right-hand side is minimized for $\eta_1 = 1$. Rearranging the term, we get

$$-\omega^2+\omega-\frac{b_1(b_1+\varepsilon)C_p^2}{4\mu_0^2} > 0 \tag{5.36}$$

Real solutions $\omega \in (0,1)$ exists whenever the discriminant is positive, that is

$$b_1(b_1+\varepsilon)C_p^2 \leq \mu_0^2$$

The latter condition can be satisfied (by suitable choosing of ε) if and only if

$$b_1 < \frac{\mu_0}{C_p} \tag{5.37}$$

In conclusion, if (5.37) is satisfied, the scheme is absolutely stable for any choice of Δt. ★

6 Navier-Stokes equations

6.1 Introduction

Navier-Stokes equations describe the motion of a fluid with constant density ρ in a domain $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$. They read as follows

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \operatorname{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right)\right] + (\mathbf{u} \cdot \nabla)\,\mathbf{u} + \nabla p = \mathbf{f} & \mathbf{x} \in \Omega,\ t > 0 \\ \operatorname{div}\mathbf{u} = 0 & \mathbf{x} \in \Omega,\ t > 0 \end{cases} \tag{6.1}$$

where

- $\mathbf{u}$ is the fluid's velocity
- p is the pressure divided by the density (which will be simply called "pressure")
- ν is the kinematic viscosity
- $\mathbf{f} \in L^2(\mathbb{R}^+; [L^2(\Omega)]^d)$ is a forcing term per unit of mass

The first equation represents conservation of linear momentum, while the second one the conservation of mass. In fact, the term $(\mathbf{u} \cdot \nabla)\,\mathbf{u}$ describes the process of convective transport, while the term $-\operatorname{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^t\right)\right]$ describe the process of molecular diffusion.
When ν is constant, from the continuity equation we obtain

$$\operatorname{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right)\right] = \nu\left(\Delta \mathbf{u} + \nabla \operatorname{div} u\right) = \nu \Delta \mathbf{u}$$

and system (6.1) can be rewritten as

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\,\mathbf{u} + \nabla p = \mathbf{f} & \mathbf{x} \in \Omega,\ t > 0 \\ \operatorname{div}\mathbf{u} = 0 & \mathbf{x} \in \Omega,\ t > 0 \end{cases} \tag{6.2}$$

The (6.2) are often called incompressible Navier-Stokes equations. More in general, when $\operatorname{div}\mathbf{u} = 0$, fluids are said to be incompressible.
In order for (6.2) to be well posed, it is necessary to assign the initial condition

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega \tag{6.3}$$

where $\mathbf{u}_0$ is a given divergence-free vector field. Then we would need suitable BC

$$\begin{cases} \mathbf{u}(\mathbf{x}, t) = \boldsymbol{\varphi}(\mathbf{x}, t) & \forall\, \mathbf{x} \in \Gamma_D,\ t > 0 \\ \left(\nu \dfrac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n}\right)(\mathbf{x}, t) = \boldsymbol{\psi}(\mathbf{x}, t) & \forall\, \mathbf{x} \in \Gamma_N,\ t > 0 \end{cases} \tag{6.4}$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are given vector functions, while Γ_D and Γ_N provide a partition of the boundary, such that $\Gamma_D \cup \Gamma_N = \partial\Omega, \Gamma_D^{\circ} \cap \Gamma_N^{\circ} = \varnothing$. Finally, as usual $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$.
If we denote with $u_i, i = 1, \ldots, d$ the components of $\mathbf{u}$ w.r.t. a Cartesian frame, and, likewise, with f_i components of $\mathbf{f}$, we obtain, from (6.2)

$$\begin{cases} \dfrac{\partial u_i}{\partial t} - \nu \nabla u_i + \displaystyle\sum_{j=1}^{d} u_j \frac{\partial u_i}{\partial x_j} + \frac{\partial p}{\partial x_j} = f_i & i = 1, \ldots, d \\ \displaystyle\sum_{j=1}^{d} \frac{\partial u_j}{\partial x_j} = 0 \end{cases}$$

Remark 6.1
The Navier-Stokes equations have been written in terms of the primitive variable $\mathbf{u}$ and p, but other sets of variables may be used too. For instance, in the two-dimensional case it is common to see the vorticity ω and the streamfunction ψ that are related to the velocity

$$\omega = \operatorname{rot}\mathbf{u} = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}, \quad \mathbf{u} = \begin{bmatrix} \frac{\partial \psi}{\partial x_2} \\ -\frac{\partial \psi}{\partial x_1} \end{bmatrix}.$$

The various formulations are equivalent from a mathematical standpoint, but may give rise to different numerical methods.

6.2 Weak formulation

A weak formulation of the problem can be obtained as usual, by multiplying a test function $\mathbf{v}$, belonging to a suitable space V, and integrate in Ω

$$\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\Omega - \int_\Omega \nu \Delta \mathbf{u} \cdot \mathbf{v}\, d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\, \mathbf{u}] \cdot \mathbf{v}\, d\Omega + \int_\Omega \nabla p \cdot \mathbf{v}\, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega \tag{6.5}$$

Using Green's formula we find

$$\begin{aligned} -\int_\Omega \nu \Delta \mathbf{u} \cdot \mathbf{v}\, d\Omega &= \int_\Omega \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, d\Omega - \int_{\partial\Omega} \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \mathbf{v}\, d\Omega \\ \int_\Omega \nabla p \cdot \mathbf{v}\, d\Omega &= -\int_\Omega p \operatorname{div}\mathbf{v}\, d\Omega + \int_{\partial\Omega} p\mathbf{v} \cdot \mathbf{n}\, d\gamma \end{aligned}$$

Using these relations in the first of (6.2) we obtain

$$\begin{aligned} \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\Omega + \int_\Omega \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\, \mathbf{u}] \cdot \mathbf{v}\, d\Omega \\ -\int_\Omega p \operatorname{div}\mathbf{v}\, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega + \int_{\partial\Omega} \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} \right) \cdot \mathbf{v}\, d\gamma \quad \forall\, \mathbf{v} \in V \end{aligned} \tag{6.6}$$

Similarly, by multiplying the second equation of (6.2) by a test function q, belonging to a suitable space Q to be specified, then integrating in Ω it follows that

$$\int_\Omega q \operatorname{div}\mathbf{u}\, d\Omega = 0 \quad \forall\, q \in Q \tag{6.7}$$

Usually, V is chosen so that the test functions vanish on the boundary portion where a Dirichlet data is prescribed on $\mathbf{u}$

$$V = \left[H^1_{\Gamma_D}\right]^d = \left\{ \mathbf{v} \in [H^1(\Omega)]^d : \mathbf{v}|_{\Gamma_D} = \mathbf{0} \right\} \tag{6.8}$$

It will coincide with $\left[H^1_0(\Omega)\right]^d$ if $\Gamma_D = \partial\Omega$.
If Γ_N has positive measure, we can choose $Q = L^2(\Omega)$. If $\Gamma_D = \partial\Omega$, then the pressure space should be L^2_0 to ensure uniqueness for the pressure p.
Moreover, if $t > 0$, then $\mathbf{u}(t) \in \left[H^1(\Omega)\right]^d$, with $\mathbf{u}(t) = \boldsymbol{\varphi}(t)$ on Γ_D, $\mathbf{u}(0) = \mathbf{u}_0$ and $p(t) \in Q$. Having chosen these functional spaces, we can note first of all that

$$\int_{\partial\Omega} \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} \right) \cdot \mathbf{v}\, d\gamma = \int_{\Gamma_N} \boldsymbol{\psi} \cdot \mathbf{v}\, d\gamma \quad \forall\, \mathbf{v} \in V$$

Notation 6.1
For every function $\mathbf{v} \in \mathbf{H}^1(\Omega)$, we denote by

$$\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} = \left(\sum_{k=1}^d \|v_k\|^2_{H^1(\Omega)} \right)^{\frac{1}{2}}$$

its norm, and by

$$|\mathbf{v}|_{\mathbf{H}^1(\Omega)} = \left(\sum_{k=1}^{d} |v_k|^2_{H^1\Omega} \right)^{\frac{1}{2}}$$

its seminorm. Thanks to Poincaré's inequality, $|\mathbf{v}|_{\mathbf{H}^1(\Omega)}$ is equivalent to $\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}$ for all functions belonging to V, provided that the Dirichlet boundary has a positive measure.

All the integrals involving bilinear terms are finite. To be more precise, by using the vector notation $\mathbf{H}^k(\Omega) = [H^k(\Omega)]^d, \mathbf{L}^p(\Omega) = [L^p(\Omega)]^d, k \geq 1, 1 \leq p < \infty$, we find

$$\begin{aligned}
\left| \nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} \, d\Omega \right| &\leq \nu \, |\mathbf{u}|_{\mathbf{H}^1(\Omega)} \, |\mathbf{v}|_{\mathbf{H}^1(\Omega)} \\
\left| \int_\Omega p \operatorname{div} \mathbf{v} \, d\Omega \right| &\leq \|p\|_{L^2(\Omega)} \, |\mathbf{v}|_{\mathbf{H}^1(\Omega)} \\
\left| \int_\Omega q \nabla \mathbf{u} \, d\Omega \right| &\leq \|q\|_{L^2(\Omega)} \, |\mathbf{u}|_{\mathbf{H}^1(\Omega)}
\end{aligned}$$

Also the integral involving the trilinear term is finite. Before we see how let's recall the following result: if $d \leq 3$

$$\forall \, \mathbf{v} \in \mathbf{H}^1(\Omega), \text{ then } \mathbf{v} \in \mathbf{L}^4(\Omega) \text{ and } \exists \, C > 0 : \|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \leq C \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}.$$

Then, using the following three-term Hölder inequality

$$\left| \int_\Omega fgh \, d\Omega \right| \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)} \|h\|_{L^r(\Omega)},$$

valid for all $p, q, r > 1$ such that $p^{-1} + q^{-1} + r^{-1} = 1$, we conclude that

$$\left| \int_\Omega [(\mathbf{u} \cdot \nabla) \, \mathbf{u}] \cdot \mathbf{v} \, d\Omega \right| \leq \|\nabla \mathbf{u}\|_{\mathbf{L}^2(\Omega)} \|\mathbf{u}\|_{\mathbf{L}^4(\Omega)} \|\mathbf{u}\|_{\mathbf{L}^4(\Omega)} \leq C^2 \|\mathbf{u}\|_{\mathbf{H}^1(\Omega)} \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}.$$

6.3 Solution uniqueness

As for the solution's uniqueness, let us consider again (6.2). If $\Gamma_D = \partial\Omega$, when only boundary conditions of Dirichlet type are imposed, the pressure merely appears in terms of its gradient. In that case if we call $(\mathbf{u}, p)$ a solution, then for any constant c the couple $(\mathbf{u}, p + c)$ is a solution too since $\nabla (p + c) = \nabla p$.
To avoid that, one can fix a priori p at a given point $\mathbf{x}_0$ of the domain Ω such that $p(\mathbf{x}_0) = p_0$, or, alternatively, require the pressure average to be null, $\int_\Omega p \, d\Omega$. This condition requires to prescribe a pointwise value for the pressure, but this is inconsistent with the hypotesis that $p \in L^2$. For this reason, the pressure space will be considered as

$$Q = L^2_0(\Omega) = \left\{ p \in L^2(\Omega) : \int_\Omega p \, d\Omega = 0 \right\}.$$

Then, we observe that if $\Gamma_D = \partial\Omega$, the prescribed Dirichlet data $\boldsymbol{\varphi}$ must be compatible with the incompressiblility constant

$$\int_\Omega \boldsymbol{\varphi} \cdot \mathbf{n} \, d\gamma = \int_\Omega \operatorname{div} \mathbf{u} \, d\Omega = 0.$$

If Γ_N is not empty, we are ok with using $L^2(\Omega)$ for our pressure space. So:

$$Q = L^2(\Omega) \quad \text{if } \Gamma_N \neq \varnothing, \quad Q = L^2_0(\Omega) \quad \text{if } \Gamma_N = \varnothing \tag{6.9}$$

The weak formulation of (6.2), (6.3), (6.4) is:

$$\text{find } \mathbf{u} \in L^2(\mathbb{R}^+; \left[H^1(\Omega)\right]^d) \cap \mathcal{C}^0(\mathbb{R}^+; \left[L^2(\Omega)\right]^d), p \in L^2(\mathbb{R}^+; Q) :$$
$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\Omega + \nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, d\Omega + \int_\Omega \left[(\mathbf{u} \cdot \nabla)\, \mathbf{u}\right] \cdot \mathbf{v}\, d\Omega \\ \displaystyle\quad - \int_\Omega p \operatorname{div} \mathbf{v}\, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega + \int_{\Gamma_N} \boldsymbol{\psi} \cdot \mathbf{v}\, d\gamma \quad \forall\, \mathbf{v} \in V \\ \displaystyle\int_\Omega q \operatorname{div} \mathbf{u}\, d\Omega = 0 \quad \forall\, q \in Q \end{cases} \tag{6.10}$$

with $\mathbf{u}|_{\Gamma_D} = \boldsymbol{\psi}_D$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$. The space V is the one in (6.8), while Q the one in (6.9).

6.4 The Reynolds number

Let us define the Reynolds number,

$$Re = \frac{|\mathbf{U}|L}{\nu}$$

where $\mathbf{U}$ is a representative length of the domain Ω (e.g. the length of the channel in which the fluid flows), $\mathbf{U}$ a representative fluid velocity and ν the kinematic viscosity.
This number measures the extent to which convection dominates over diffusion. When $Re \ll 1$, the convective term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ can be omitted, reducing the equations to the so-called Stokes equation. On the other hand, for large values of Re problems may arise, concerning the uniqueness of the solution, the existence of stationary and stable solutions, the possibility of strange attractors and the transition towards turbulent flows.

6.5 Divergence free formulation of Navier-Stokes equations

By eliminating the pressure, the Navier-Stokes equations can be rewritten in a reduced form, with the sole variable $\mathbf{u}$. Introducing the following subspaces of $\left[H^1(\Omega)\right]^d$:

$$V_{div} = \left\{\mathbf{v} \in \left[H^1(\Omega)\right]^d : \operatorname{div}\mathbf{v} = 0\right\}$$
$$V_{div}^0 = \{\mathbf{v} \in V_{div} : \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}$$

If the test function $\mathbf{v}$ belongs to the space V_{div}, the term associated with the pressure gradient vanishes, whence we find the following reduced problem for the velocity

$$\text{find } \mathbf{u} \in L^2(\mathbb{R}^+; V_{div}) \cap \mathcal{C}^0(\mathbb{R}^+; \left[L^2(\Omega)\right]^d) :$$
$$\begin{aligned} \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\Omega + \nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, d\Omega + \int_\Omega \left[(\mathbf{u} \cdot \nabla)\, \mathbf{u}\right] \cdot \mathbf{v}\, d\Omega \\ = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega + \int_{\Gamma_N} \boldsymbol{\psi} \cdot \mathbf{v}\, d\gamma \quad \forall\, \mathbf{v} \in V_{div}, \end{aligned} \tag{6.11}$$

with $\mathbf{u}|_{\Gamma_D} = \boldsymbol{\psi}_D$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$.
Since we are dealing with a nonlinear parabolic problem, we can carry an analysis by using techniques similar to those applied in parabolic problems. Clearly a solution of (6.10) will be a suitable solution of (6.11), while, for the converse, we have the following theorem

Theorem 6.1
Let $\Omega \subset \mathbb{R}^d$ be a domain with Lipschitz-continuous boundary $\partial\Omega$. Let $\mathbf{u}$ be a solution of the reduced problem (6.11). Then exists a unique function $p \in L^2(\mathbb{R}^+; Q)$ such that $(\mathbf{u}, p)$ is a solution of the full problem (6.10)

In practice, however, the results of this theorem, are quite unsuitable from a numerical viewpoint, since it requires the construction of the V_{div} subspaces, of the divergence-free velocity functions, etc... Moreover, the result of the above theorem is not constructive, as it does not provide a way to build the solution pressure p.

6.6 Stokes equations and their approximation

In this section we will consider the generalized Stokes problem with homogeneous Dirichlet BC

$$\begin{cases} \sigma \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \\ \text{div}\mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega \end{cases} \tag{6.12}$$

for $\sigma \geq 0$. This is the motion of an incompressible viscous flow in which the convective term has been neglected, since $Re \ll 1$. Moreover, one can generate a problem (6.12) also while using an implicit temporal discretization of the Navier-Stokes equations and by neglecting the convective term.
We have indeed the following scheme, where k denotes the temporal index

$$\begin{cases} \frac{\mathbf{u}^k - \mathbf{u}^{k-1}}{\Delta t} - \nu \Delta \mathbf{u}^k + \nabla p^k = \mathbf{f}(t^k) & \mathbf{x} \in \Omega,\ t > 0 \\ \text{div}\mathbf{u}^k = 0 & \mathbf{x} \in \Omega,\ t > 0 \\ +B.C. \end{cases}$$

Hence, at each time step t^k we need to solve the following Stokes-like system of equations

$$\begin{cases} \sigma \mathbf{u}^k - \nu \Delta \mathbf{u}^k + \nabla p^k = \tilde{\mathbf{f}}^k & \text{in } \Omega \\ \text{div}\mathbf{u}^k = 0 & \text{in } \Omega \\ +B.C. \end{cases} \tag{6.13}$$

where $\sigma = (\Delta t)^{-1}$ and $\tilde{\mathbf{f}}^k = \tilde{\mathbf{f}}(t^k) + \frac{\mathbf{u}^{k-1}}{\Delta t}$. The weak formulation of problem (6.12) reads:

$$\text{find } \mathbf{u} \in V \text{ and } p \in Q : \begin{cases} \displaystyle\int_\Omega (\sigma \mathbf{u} \cdot \mathbf{v} + \nu \nabla u \cdot \nabla v)\, d\Omega - \int_\Omega p \text{div}\mathbf{v}\, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega & \forall\, \mathbf{v} \in V, \\ \displaystyle\int_\Omega q \text{div}\mathbf{u}\, d\Omega = 0 & \forall\, q \in Q, \end{cases} \tag{6.14}$$

where $V = \left[H_0^1(\Omega)\right]^d$ and $Q = L_0^2(\Omega)$. Now with the bilinear forms $a : V \times V \to \mathbb{R}$ and $b : V \times Q \to \mathbb{R}$:

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_\Omega (\sigma \mathbf{u} \cdot \mathbf{v} + \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v})\, d\Omega, \\ b(\mathbf{u}, q) &= -\int_\Omega q \text{div}\mathbf{u}\, d\Omega. \end{aligned} \tag{6.15}$$

Using these notations, problem (6.14) becomes

$$\text{find } (\mathbf{u}, p) \in V \times Q : \begin{cases} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) & \forall\, \mathbf{v} \in V \\ b(\mathbf{u}, q) = 0 & \forall\, q \in Q \end{cases} \tag{6.16}$$

where $(\mathbf{f}, v) = \sum_{i=1}^d \int_\Omega f_i v_i\, d\Omega$.
Considering non-homogeneous BC like in (6.4), the weak formulation of the Stokes problem becomes

$$\text{find } (\overset{\circ}{\mathbf{u}}, p) \in V \times Q : \begin{cases} a(\overset{\circ}{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, p) = \mathbf{F}(\mathbf{v}) & \forall\, \mathbf{v} in V \\ b(\overset{\circ}{\mathbf{u}}, q) = G(q) & \forall\, q \in Q \end{cases} \tag{6.17}$$

where V and Q are the spaces (6.8) and (6.9), respectively. Having denoted with $\mathbf{R}\boldsymbol{\varphi} \in \left[H^1(\Omega)\right]^d$ a lifting of the boundary datum $\boldsymbol{\varphi}$, we have set $\overset{\circ}{\mathbf{u}} = \mathbf{u} - \mathbf{R}\boldsymbol{\varphi}$, while the new terms on the right-hand side have the following expression

$$\begin{aligned} \mathbf{F}(\mathbf{v}) &= (\mathbf{f}, \mathbf{v}) + \int_{\Gamma_N} \boldsymbol{\psi} \mathbf{v}\, d\gamma - a(\mathbf{R}\boldsymbol{\varphi}, \mathbf{v}), \\ G(q) &= -b(\mathbf{R}\boldsymbol{\varphi}, q) \end{aligned} \tag{6.18}$$

Theorem 6.2
The couple $(\mathbf{u}, p)$ solves the problem (6.16) if and only if it is a saddle point of the Lagrangian functional

$$\mathcal{L}(\mathbf{v}, q) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) + b(\mathbf{v}, q) - (\mathbf{f}, \mathbf{v})$$

or equivalently

$$\mathcal{L}(\mathbf{u}, p) = \min_{\mathbf{v}\in V} \max_{q\in Q} \mathcal{L}(\mathbf{v}, q)$$

The pressure q hence plays the role of Lagrange multiplier associated to the divergence-free constraint. We need to define the Fréchet derivative.

Definition 6.1 (Fréchet derivative)
Let $F : X \to Y$ with X, Y two normed vector spaces. F is differentiable at $x \in X$ if $\exists\ \mathcal{L}_x : X \to Y$ linear and bounded such that

$$\forall\ \varepsilon > 0\ \exists\ \delta > 0\ \|F(x + h) - F(x) - \mathcal{L}_x h\|_Y \leq \varepsilon\|h\|_X \quad \forall\ h \in X : \|h\|_X < \delta$$

$F'(x) := \mathcal{L}_x$ is called Fréchet derivative of F at point x.

By formally taking the Fréchet derivative of the Lagrangian with respect to the two variables (thanks to the symmetry of $a(\cdot, \cdot)$):

$$\left\langle \frac{\partial \mathcal{L}(\mathbf{u}, p)}{\partial \mathbf{u}}, \mathbf{v} \right\rangle = a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) - (\mathbf{f}, \mathbf{v}) = 0 \quad \forall\ \mathbf{v} \in V$$
$$\left\langle \frac{\partial \mathcal{L}(\mathbf{u}, p)}{\partial p}, q \right\rangle = b(\mathbf{u}, q) = 0 \quad \forall\ q \in Q$$

6.7 Galerkin approximation

The Galerkin approximation of problem (6.16) has the following form:

$$\text{find } (\mathbf{u}_h, p_h) \in V_h \times Q_h : \begin{cases} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h) & \forall\ \mathbf{v}_h \in V_h \\ b(\mathbf{u}, q_h) = 0 & \forall\ q_h \in Q_h \end{cases} \tag{6.19}$$

If, instead we consider problem (6.17)-(6.18), we need to add $\mathbf{F}(\mathbf{v}_h)$ and $G(q_h)$. These new functionals can be obtained from (6.18) by replacing $\mathbf{R}\boldsymbol{\varphi}$ with the interpolant of $\boldsymbol{\varphi}$ at the nodes of Γ_D, and replacing $\boldsymbol{\varphi}$ with its interpolant at the nodes on Γ_N.
The existence and uniqueness is guaranteed by the theorem

Theorem 6.3
The Galerkin approximation (6.19) admits one and only one solution if the following conditions hold

- The bilinear form $a(\cdot, \cdot)$ is:
  (1) coercive, that is $\exists\ \alpha > 0$ (possibly depending on h) such that:
  $$a(\mathbf{v}_h, \mathbf{v}_h) \geq \alpha\|\mathbf{v}_h\|_V^2 \quad \forall\ \mathbf{v}_h \in V_h^*,$$
  where $V_h^* = \{\mathbf{v}_h \in V : b(\mathbf{v}_h, q_h) = 0 \quad \forall\ q_h \in Q_h\}$
  (2) continuous, that is $\exists\ \gamma > 0$ such that
  $$|a(\mathbf{u}_h, \mathbf{v}_h)| \leq \gamma\|\mathbf{u}_h\|_V\|\mathbf{v}_h\|_V \quad \forall\ \mathbf{u}_h, \mathbf{v}_h \in V_h$$
- The bilinear form $b(\cdot, \cdot)$ is continuous, that is $\exists\ \delta > 0$ such that
  $$|b(\mathbf{v}_h, q_h)| \leq \delta\|\mathbf{v}_h\|_V\|q_h\|_Q \quad \forall \mathbf{v}_h \in V_h, q_h \in Q_h$$

- Finally, there exist a positive constant β (possibly depending on h) such that:

$$\forall\, q_h \in Q_h,\ \exists\, \mathbf{v}_h \in V_h : b(\mathbf{v}_h, q_h) \geq \beta \|\mathbf{v}_h\|_{\mathbf{H}^1(\Omega)} \|q_h\|_{L^2(\Omega)} \tag{6.20}$$

Under the previous assumptions the discrete solution fullfills the following a priori estimates:

$$\begin{aligned} \|\mathbf{u}_h\|_V &\leq \frac{1}{\alpha}\|\mathbf{f}\|_{V'} \\ \|p_h\|_Q &\leq \frac{1}{\beta}\left(1+\frac{\gamma}{\alpha}\right)\|\mathbf{f}\|_{V'}, \end{aligned}$$

where V' is the dual space of V.

Moreover, the following convergence results hold:

$$\begin{aligned} \|\mathbf{u}-\mathbf{u}_h\|_V \leq{}& \left(1+\frac{\delta}{\beta}\right)\left(1+\frac{\gamma}{\alpha}\right) \inf_{\mathbf{v}_h \in V_h} \|\mathbf{u}-\mathbf{v}_h\|_V \\ &+ \frac{\delta}{\alpha} \inf_{q_h \in Q_h} \|p-q_h\|_Q, \\ \|p-p_h\|_Q \leq{}& \frac{\gamma}{\beta}\left(1+\frac{\gamma}{\alpha}\right)\left(1+\frac{\delta}{\beta}\right) \inf_{\mathbf{v}_h \in V_h} \|\mathbf{u}-\mathbf{v}_h\|_V \\ &+ \left(1+\frac{\delta}{\beta}+\frac{\delta\gamma}{\alpha\beta}\right) \inf_{q_h \in Q_h} \|p-q_h\|_Q \end{aligned}$$

It is worth noticing that condition (6.20) is equivalent to the existence of a positive constant β such that

$$\inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{\mathbf{v}_h \in V_h \\ \mathbf{v}_h \neq 0}} \frac{b(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{\mathbf{H}^1(\Omega)} \|q_h\|_{L^2(\Omega)}} \geq \beta \tag{6.21}$$

Which is often called inf-sup condition.

6.8 A general saddle-point problem

Let X and M be two Hilbert spaces endowed with norms $\|\cdot\|_X$ and $\|\cdot\|_M$. Denoting their dual spaces as X', M', we introduce the bilinear forms

- $a(\cdot,\cdot) : X \times X \to \mathbb{R}$
- $b(\cdot,\cdot) : M \times M \to \mathbb{R}$

that we suppose to be continuous, meaning there exist two constants $\gamma, \delta > 0$ such that for all $w, v \in X$ and $\mu \in M$

$$\begin{aligned} |a(w,v)| &\leq \gamma \|w\|_X \|v\|_X, \\ |b(w,\mu)| &\leq \|w\|_X \|\mu\|_M \end{aligned} \tag{6.22}$$

Consider now the following constrained problem

$$\text{find } (u,\eta) \in X \times M : \begin{cases} a(u,v) + b(v,\eta) = \langle I, v\rangle & \forall\, v \in X, \\ b(u,\mu) = \langle \sigma, \mu \rangle & \forall\, \mu \in M \end{cases} \tag{6.23}$$

where $I \in X'$ and $\sigma \in M'$ are two assigned linear functionals, while $\langle \cdot,\cdot \rangle$ denotes the pairing between X and X' or M and M'.

Formulation (6.23) is general enough to include formulation (6.16) of the Stokes problem. In order to analyze (6.23), we introduce the affine manifold

$$X^\sigma = \{v \in X : b(v,\mu) = \langle \sigma, \mu \rangle\ \forall\, \mu \in M\} \tag{6.24}$$

The space X^0 denotes the kernel of b, that is

$$X^0 = \{v \in X : b(v, \mu) = 0 \ \forall \ \mu \in M\}$$

This is a closed subspace of X. We can, therefore, associate (6.23) with the following reduced problem:

$$\text{find } u \in X^\sigma : a(u, v) = \langle I, v \rangle \quad \forall \ v \in X^0 \tag{6.25}$$

If (u, η) is a solution of (6.23), then it is a solution to (6.25). In the following, we will introduce suitable condititions that allow the converse to hold too. Also, if we are able to prove existence and uniqueness for (6.25), then this would allow us to obtain a result for (6.23).

Theorem 6.4 (Existence, uniqueness and stability)
Let the bilinear form $a(\cdot, \cdot)$ satisfy the continuity condition (6.22) and be coercive on X^0, so

$$\exists \ \alpha > 0 : a(v, v) \geq \alpha \|v\|_X^2 \quad \forall \ v \in X^0 \tag{6.26}$$

Suppose also that the bilinear form $b(\cdot, \cdot)$ satisfies the same continuity condition as well as the following: there exists $\beta^* > 0$ such that

$$\forall \ \mu \in M \exists \ v \in X, \ v \neq 0 : b(v, \mu) \geq \beta^* \|v\|_X \|\mu\|_M. \tag{6.27}$$

Then, for every $I \in X'$ and $\sigma \in M'$, there exists a unique solution $(u, \eta) \in X \times M$ to the saddle point problem (6.23).
Moreover, the map $(I, \sigma) \mapsto (u, \eta)$ is an isomorphism from $X' \times M'$ onto $X \times M$ and the following a priori estimates hold

$$\begin{aligned} \|u\|_X &\leq \frac{1}{\alpha} \left[\|I\|_{X'} + \frac{\alpha + \gamma}{\beta^*} \|\sigma\|_{M'} \right], \\ \|\eta\|_M &\leq \frac{1}{\beta} \left[\left(1 + \frac{\gamma}{\alpha} \|I\|_{X'} + \frac{\gamma(\alpha + \gamma)}{\alpha \beta^*} \|\sigma\|_{M'} \right) \right]. \end{aligned} \tag{6.28}$$

Galerkin approximation

To introduce a Galerkin approximation of the saddle-point problem (6.23), we consider two subspaces $X_h \subset X$ and $M_h \subset M$, respectively. They can be either finite element piecewise polynomial subspaces or spectral element subspaces. We look for a solution for the problem

$$\text{given } I \in X' \text{ and } \sigma \in M' \text{ find } (u_h, \eta_h) \in X_h \times M_h : \begin{cases} a(u_h, v_h) + b(v_h, \eta_h) = \langle I, v_h \rangle & \forall \ v_h \in X_h \\ b(u_h, \mu_h) = \langle \sigma, \mu_h \rangle & \forall \ \mu_h \in M_h \end{cases} \tag{6.29}$$

We introduce the subspace

$$X_h^\sigma = \{v_h \in X_h : b(v_h, \mu_h) = \langle \sigma, \mu_h \rangle \ \forall \ \mu_h \in M_h\} \tag{6.30}$$

which allow us to introduce the following reduced formulation

$$\text{find } u_h \in X_h^\sigma : a(u_h, v_h) = \langle I, v_h \rangle \quad \forall \ v_h \in X_h^0 \tag{6.31}$$

Since, in general, M_h is different from M, the space (6.30) is not necessarily a subspace of X^σ.
Clearly, every solution of (u_h, η_h) of (6.29) yields a solution u_h for the reduced problem (6.31). Now we look for a solution that allows us to prove that the converse is true.

Theorem 6.5 (Existence, uniqueness and stability)
Let the bilinear form $a(\cdot, \cdot)$ satisfy the continuity condition (6.22) and be coercive on X^0, so

$$\exists \ \alpha > 0 : a(v_h, v_h) \geq \alpha_h \|v_h\|_X^2 \quad \forall \ v_h \in X_h^0 \tag{6.32}$$

Suppose also that the bilinear form $b(\cdot,\cdot)$ satisfies the same continuity condition as well as the following: there exists $\beta_h^* > 0$ such that

$$\forall\, \mu_h \in M_h \exists\, v_h \in X_h,\ v_h \neq 0 : b(v_h, \mu_h) \geq \beta_h^* \|v_h\|_X \|\mu_h\|_M. \tag{6.33}$$

Then, for every $I \in X'$ and $\sigma \in M'$, there exists a unique solution $(u_h, \eta_h) \in X_h \times M_h$ to the saddle point problem (6.29).
Moreover, the solution satisfies the following conditions:

$$\|u_h\|_X \leq \frac{1}{\alpha_h}\left[\|I\|_{X'} + \frac{\alpha_h + \gamma}{\beta_h^*}\|\sigma\|_{M'}\right], \tag{6.34}$$

$$\|\eta_h\|_M \leq \frac{1}{\beta_h}\left[\left(1 + \frac{\gamma}{\alpha_h}\|I\|_{X'} + \frac{\gamma(\alpha_h + \gamma)}{\alpha_h \beta_h}\|\sigma\|_{M'}\right)\right]. \tag{6.35}$$

The coercivity condition (6.26) does not necessarily guarantee (6.32), as $X_h^0 \not\subset X^0$, nor does the compatibility condition (6.27) in general imply the discrete compatibility condition (6.33) due to the fact that X_h is a proper subspace of X. (6.33) is called inf-sup condition.

Theorem 6.6
Let the assumptions of existence and uniqueness theorems (6.4) and (6.5) be satisfied. Then, the solutions (u, η) and (u_h, η_h) of problems (6.23) and (6.29), respectively, satisfy the following error estimates:

$$\|u - u_h\|_X \leq \left(1 + \frac{\gamma}{\alpha_h}\right) \int_{v_h^* \in X_h^\sigma} \|u - v_h\|_X + \frac{\delta}{\alpha_h} \inf_{\mu_h \in M_h} \|\eta - \mu_h\|_M \tag{6.36}$$

$$\|\eta - \eta_h\|_M \leq \frac{\gamma}{\beta_h}\left(1 + \frac{\gamma}{\alpha_h}\right) \inf_{v_h^* \in X_h^\sigma} \|u - v_h^*\|_X + \left(1 + \frac{\delta}{\beta_h} + \frac{\gamma\beta}{\alpha_h \beta_h}\right) \inf_{\mu_h \in M_h} \|\eta - \mu_h\|_M \tag{6.37}$$

where γ, δ, α_h and β_h are respectively defined by (6.22), (6.32) and (6.33). Also, the following estimate holds

$$\inf_{v_h^* \in X_h^\sigma} \|u - v_h\|_X \leq \left(1 + \frac{\delta}{\beta_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_X \tag{6.38}$$

The inequalities (6.36) and (6.37) yield error estimates with optimal convergence rate, provided that the constants α_h and β_h in (6.32) and (6.33) are bounded from below by two constants α and β indepedent of h. Let us also remark that (6.36) holds even if (6.27) and (6.33) are not satisfied.

Remark 6.2 (Spurious pressure modes)
The compatibility condition (6.33) is essential to guarantee the uniqueness of the η_h-component of the solution. Indeed, if (6.33) does not hold, then

$$\exists\, \mu_h^* \in M_h,\ \mu_h^* \neq 0 : b(v_h, \mu_h^*) = 0 \quad \forall M v_h \in X_h.$$

Consequently, if (u_h, η_h) is a solution to the problem (6.29), then $(u_h, \eta_h + \tau\mu_h^*)$, for all $\tau \in \mathbb{R}$, is a solution too.
Any such function μ_h^* is called spurious mode.

Algebraic form of Stokes problem

Let us investigate the structure of the algebraic system associated to the Galerkin approximation (6.19) to the Stokes problem (or, more generally, to a discrete saddle-point problem like (6.29)). Denote with

$$\{\boldsymbol{\varphi}_j \in V_h\}, \quad \{\phi_k \in Q_h\},$$

the basis functions of the spaces V_h and Q_h, respectively. Let us expand the discrete solutions $\mathbf{u}_h$ and p_h w.r.t. such bases

$$\begin{aligned} \mathbf{u}_h(\mathbf{x}) &= \sum_{j=1}^{N} u_j \boldsymbol{\varphi}_j(\mathbf{x}), \\ p_h(\mathbf{x}) &= \sum_{k=1}^{M} p_k \phi_k(\mathbf{x}), \end{aligned} \tag{6.39}$$

having set $N = \dim V_h$ and $M = \dim Q_h$.
By choosing as test functions in (6.19) the same basis functions we obtain the following block linear system

$$\begin{cases} A\mathbf{U} + B^T\mathbf{P} = \mathbf{F}, \\ B\mathbf{U} = \mathbf{0} \end{cases} \tag{6.40}$$

where $A \in \mathbb{R}^{N\times N}$ and $B \in \mathbb{R}^{M\times N}$ are the matrices related respectively to the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, whose elements are given by

$$\begin{aligned} A &= [a_{ij}] = [a(\boldsymbol{\varphi_j}, \boldsymbol{\varphi_i})], \\ B &= [b_{km}] = [b(\boldsymbol{\varphi_m}, \phi_k)], \end{aligned}$$

while $\mathbf{U}$ and $\mathbf{P}$ are the vectors of the unknowns,

$$\begin{aligned} \mathbf{U} &= [u_j], \\ \mathbf{P} &= [p_j]. \end{aligned}$$

The $(N+M)\times(N+M)$ matrix

$$S = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \tag{6.41}$$

is block-symmetric (as A is symmetric) and indefinite, featuring real eigenvalues with variable sign. S is non-singular if and only if it is null, a property that follows from the inf-sup condition (6.21).
To prove the latter statement we proceed as follows.
Since A is non-singular, because it is associated to the coercive bilinear form $a(\cdot,\cdot)$, from the first of (6.40), we can formally obtain $\mathbf{U}$ as

$$\mathbf{U} = A^{-1}(\mathbf{F} - B^T\mathbf{P}) \tag{6.42}$$

Using (6.42) in the second equation of (6.40) yields

$$R\mathbf{P} = BA^{-1}\mathbf{F}, \text{ where } R = BA^{-1}B^T \tag{6.43}$$

This corresponds to having carried out a block Gaussian elimination on system (6.41). In this way we obtain a reduced system for the sole unknown $\mathbf{P}$, which admits a unique solution in case R is non-singular and positive definite, we want to prove that the latter condition is satisfied if and only if B^T has a null kernel, that is

$$\ker(B^T) = \{\mathbf{0}\} \tag{6.44}$$

where $\ker(B^T) = \left\{\mathbf{x} \in \mathbb{R}^M : B^T\mathbf{x} = \mathbf{0}\right\}$.
We proceed as follows:

$$R\mathbf{p} = \mathbf{0} \Rightarrow \mathbf{p} = \mathbf{0}$$

that is

$$\langle BA^{-1}B^T\mathbf{p}, \mathbf{q}\rangle = \mathbf{0} \ \forall\ \mathbf{q} \Rightarrow \mathbf{p} = \mathbf{0}$$

Let us take $\mathbf{q} = \mathbf{p}$. We require:

$$\langle A^{-1}B^T\mathbf{p}, B^T\mathbf{q}\rangle = \mathbf{0} \Rightarrow \mathbf{p} = \mathbf{0}$$

Set $\mathbf{w} = B^T\mathbf{p}$. Since A is spd, we have that $\langle A^{-1}\mathbf{w}, \mathbf{w}\rangle = \mathbf{0}$, which implies $\mathbf{w} = \mathbf{0}$. Finally,

$$(\mathbf{w} = B^T\mathbf{p} = \mathbf{0} \Rightarrow \mathbf{p} = \mathbf{0}) \iff \ker(B^T) = \{\mathbf{0}\}$$

Remark 6.3
Condition (6.44) is equivalent to the inf-sup condition (6.21).

On the other hand, since A is non-singular, from the existence and uniqueness $\mathbf{P}$ we infer that there exists a unique vector $\mathbf{U}$ which satisfies (6.42).
In conclusion, system (6.40) admits a unique solution $(\mathbf{U}, \mathbf{P})$ if and only if condition (6.44).

Remark 6.4
We remark that, for an arbitrary matrix, $B^T(N \times M)$, we have $\text{rank}(B^T) + \dim\ker(B^T) = \min(M, N)$. Then, condition (6.44) is equivalent to asking that B^T has a full rank, because $\text{rank}(B^T)$ is the maximum number of linearly indepedent row vectors of B^T.

Let us consider the remark (6.2) concerning the general saddle-point problem and suppose that the inf-sup condition (6.21) does not hold. Then,

$$\exists\, q_h^* \in Q_h : b(\mathbf{v}_h, q_h^*) = 0 \quad \forall\, \mathbf{v}_h \in V_h. \tag{6.45}$$

Consequently, if $(\mathbf{u}_h, p_h)$ is a solution to the Stokes problem (6.19), then $(\mathbf{u}_h, p_h + q_h)$ is a solution too:

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h + q_h) &= a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}_h, p_h) + \underbrace{b(\mathbf{v}_h, q_h^*)}_{=0} \quad \forall\, \mathbf{v}_h \in V_h \\ &= a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h) \end{aligned}$$

Functions q_h^* which fail to satisfy the inf-sup condition are invisible to the Galerkin problem (6.19). For this reason, as already observed, they are called spurious pressure modes, or parasitic modes. Their presence yields numerical instabilities.
Two strategies are generally adopted in order to guarantee well-posedness of the numerical problem:

- choose spaces V_h and Q_h that satisfy the inf-sup condition.
- Stabilize the finite dimensional problem by eliminating spurious modes.

6.9 Suitable spaces

Let us analyze the first type of strategy. To start with, we will consider the case of finite element spaces. To characterize Q_h and V_h it suffices to choose on every elemnt their degrees of freedom. Since the weak formulation does not require a continuous pressure, we will consider the case of a discontinuous pressure first.
As Stokes equation are of order one in p and two in $\mathbf{u}$ it makes sense to use polynomials of degree $k \geq 1$ for V_h and of degree $k-1$ for Q_h. When looking for a compatible couple of spaces, the larger the velocity space V_h, the more likely the inf-sup condition is satisfied.
Suitable choices of spaces that fullfill the inf sup pressure are, for example, the couple $(\mathbb{P}_2, \mathbb{P}_0)$, $(\mathbb{Q}_2, \mathbb{P}_0)$ or the so-called Crouzeix Raviart elements, in which the velocity components are piecewise quadratic functions with a bubble function on each element, while the pressure components are piecewise linear discontinuous element.
In the case of continuous pressure, examples of incompatible spaces could be piecewise linear elements on triangles for both velocity and pressure. More in general, not a good idea to use the same polynomial degree.
The smallest degree possible for stability is the pair $(\mathbb{P}_2, \mathbb{P}_1)$, which are called Tayor-Hood elements.

Spectral methods

If we use spectral methods using equal-order polynomial spaces for both velocity and pressure yields subspaces that violate the inf-sup condition. Compatible spaces can be the ones with polynomial degree $N \geq 2$ for the velocity and $N-2$ for the pressure, yielding the so called $(\mathbb{Q}_N, \mathbb{Q}_{N-2})$ approximation. The degrees of freedom for each velocity components are represented by the $(N+1)^2$ LGL nodes.
In the case of discontinuous pressure, at least two sets of interpolation nodes can be used: either the subset represented by the $(N-1)^2$ internal nodes of the choice above, or the set $(N-1)^2$ LGL

nodes. This choice stands at the base of the spectral-type approximation, such as collocation, G-NI or SEM-NI.

6.10 A stabilized problem

We have seen that finite element or spectral methods that make use of equal-degree polynomial do not fullfill the inf-sup condition and are therefore unstable. However, stabilizing them is possible by using SUPG or GLS techniques like those enconuntered in the approximation of ADR equation.
Here we will limit ourselves to the case of piecewise continuous linear finite elements $(\mathbb{P}_1, \mathbb{P}_1)$, stabilized using GLS.

$$V_h = [\overset{\circ}{X}{}_h^1]^2, \qquad Q_h = \left\{ q_h \in X_h^1 : \int_\Omega d\Omega = 0 \right\}.$$

This choice is urged by the need of keeping the global numer of degrees of freedom as low as possible, especially when dealing with three-dimensional problem. However, as it violates the inf-sup condition, it will be unstable.
So, we set $W_h = V_h \times Q_h$, and, instead of (6.19), we consider the problem

$$\text{find } (\mathbf{u}_h, p_h) \in W_h : A_h((\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)) = F_h(\mathbf{v}_h, q_h) \quad \forall\, (\mathbf{v}_h, q_h) \in W_h \tag{6.46}$$

We have set

$$\begin{aligned}
&A_h : W_h \times W_h \to \mathbb{R}, \\
&A_h((\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)) = a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) - b(\mathbf{u}_h, q_h) \\
&\qquad + \delta \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \int_{\mathcal{K}} (-\nu \Delta \mathbf{u}_h + \nabla p_h)\,(-\nu \Delta \mathbf{v}_h + \nabla q_h)\; d\mathcal{K}, \\
&F_h : W_h \to \mathbb{R}, \\
&F_h(\mathbf{v}_h, q_h) = (\mathbf{f}, \mathbf{v}_h) + \delta \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \int_{\mathcal{K}} \mathbf{f}(-\nu \Delta \mathbf{v}_h + \nabla q_h)\, d\mathcal{K}
\end{aligned}$$

and δ is a positive parameter chosen accordingly.
This is a strongly consistent approximation of problem (6.12). As a matter of fact, the additional term is null when calculated on the exact solution thanks to (6.14). Note that, since $k = 1, \Delta \mathbf{u}_h|_{\mathcal{K}} = \Delta \mathbf{v}_h|_{\mathcal{K}} = \mathbf{0}\ \forall\, \mathcal{K} \in \mathcal{T}_h$ as we are using piecewise linear finite elements functions.
We obtain the following stability inequality

$$\nu \|\nabla \mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} + \delta \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \|\nabla p_h\|^2_{\mathbf{L}^2(\mathcal{K})} \le C \|\mathbf{f}\|^2_{\mathbf{L}^2(\Omega)}, \tag{6.47}$$

C being a constant that depends on ν but not on h.
By applying Strang's Lemma (2.1) we can now show that the solution to the generalized Galerkin problem (6.46) satisfies the following estimate

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}^1(\Omega)} + \left(\delta \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \|\nabla p - \nabla p_h\|^2_{\mathbf{L}^2(\mathcal{K})} \right)^{\frac{1}{2}} \le Ch.$$

We can show that (6.46) admits the following matrix form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix}. \tag{6.48}$$

This system differs from (6.40) without stabilization because of the presence of the non-null block occupying the position $(2, 2)$ which is associated to the stabilization term:

$$C = [c_{km}],\ c_{km} = \delta \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \int_{\mathcal{K}} \nabla \phi_m \cdot \nabla \phi_k \, d\mathcal{K}, k, m = 1, \ldots, M$$

while the components of the right-hand side $\mathbf{G}$ are

$$g_k = -\delta \sum_{\mathcal{K}\in\mathcal{T}_h} h_{\mathcal{K}}^2 \int_{\mathcal{K}} \mathbf{f}\cdot\nabla\phi_k \, d\mathcal{K}, k = 1,\dots,M$$

Note that

$$\begin{aligned} A\mathbf{U} + B^T\mathbf{P} = \mathbf{F} &\Rightarrow \mathbf{U} = A^{-1}(\mathbf{F} - B^T\mathbf{P}), \\ B\mathbf{U} - C\mathbf{P} &\Rightarrow BA^{-1}\mathbf{F} - BA^{-1}B^T\mathbf{P} - C\mathbf{P} = \mathbf{G} \\ &\Rightarrow (BA^{-1}B^T + C)\mathbf{P} = BA^{-1}\mathbf{F} - \mathbf{G}. \end{aligned}$$

Then, in this case, the reduced system reads

$$R\mathbf{P} = BA^{-1}\mathbf{F} - \mathbf{G}.$$

In contrast to (6.43), this time the matrix R is non-singular as C is positive defined, but B^T is not full-rank.

6.11 Time discretization

We consider the following semi discretized formulation

$$\begin{cases} M\dfrac{d\mathbf{u}(t)}{dt} + A\mathbf{u}(t) + C(\mathbf{u}(t))\mathbf{u}(t) + B^T\mathbf{p}(t) = \mathbf{f}(t), \\ B\mathbf{u}(t) = \mathbf{0}, \end{cases} \tag{6.49}$$

with $\mathbf{u}(0) = \mathbf{u}_0$. $C(\mathbf{u}(t))$ is in fact a matrix depending on $\mathbf{u}(t)$, whose generic coefficient is $c_{ij}(t) = c(\mathbf{u}(t), \varphi_j, \varphi_i)$.
For the temporal discretization of this system let us use the θ-method. By setting

$$\begin{aligned} &\mathbf{u}_\theta^{n+1} = \theta\mathbf{u}^{n+1} + (1-\theta)\mathbf{u}^n, \\ &\mathbf{p}_\theta^{n+1} = \theta\mathbf{p}^{n+1} + (1-\theta)\mathbf{p}^n, \\ &\mathbf{f}_\theta^{n+1} = \theta\mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n, \\ &C_\theta(\mathbf{u}^{n+1,n})\mathbf{u}^{n+1,n} = \theta C(\mathbf{u}^{n+1})\mathbf{u}^{n+1} + (1-\theta)C(\mathbf{u}^n)\mathbf{u}^n \end{aligned}$$

we obtain

$$\begin{cases} M\dfrac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + A\mathbf{u}_\theta^{n+1} + C_\theta(\mathbf{u}^{n+1,n})\mathbf{u}^{n+1,n} + B^T\mathbf{p}_\theta^{n+1} = \mathbf{f}_\theta^{n+1} \\ B\mathbf{u}^{n+1} = \mathbf{0} \end{cases} \tag{6.50}$$

Since, except for the forward Euler method, the solution of this system is quite involved, a possible alternative is a semi-implicit scheme in which the linear part of the equation is advanced implicitly, while the nonlinear ones explicitly. In this way, for $\theta \geq \frac{1}{2}$, the resulting scheme is unconditionally stable, whereas, in all other cases, it must obey a stability restriction on the timestep Δt.

Finite difference methods

We consider an explicit temporal discretization of the first equation in (6.49), corresponding to $\theta = 0$ in (6.50). If all the quantities are known at t^n, then we can write

$$\begin{cases} M\mathbf{u}^{n+1} = H(\mathbf{u}^n, \mathbf{p}^n, \mathbf{f}^n) \\ B\mathbf{u}^{n+1} = \mathbf{0} \end{cases}$$

with M mass matrix

$$m_{ij} = \int_\Omega \boldsymbol{\varphi}_i \boldsymbol{\varphi}_j$$

This system does not allow the determination of the pressure $\mathbf{p}^{n+1}$. In particular, there is no way to enforce the divergence-free constraints on $\mathbf{u}^{n+1}$. However, if we replace $\mathbf{p}^n$ by $\mathbf{p}^{n+1}$ we obtain

$$\begin{cases} \frac{1}{\Delta t} M\mathbf{u}^{n+1} + B^T\mathbf{p}^{n+1} = \mathbf{G}, \\ B\mathbf{u}^{n+1} = \mathbf{0}, \end{cases} \tag{6.51}$$

with $\mathbf{G}$ being a suitable vector. This system corresponds to a semi-explicit discretization of (6.49). Since M is symmetric and positive definite, if condition (6.44) is satisfied, then the reduced system $BM^{-1}B^T\mathbf{p}^{n+1} = BM^{-1}\mathbf{G}$ is non-singular. Once solved, the velocity vector $\mathbf{u}^{n+1}$ can be recovered from the first equation of (6.51). This discretization method is temporally stable if

$$\Delta t \leq C \min\left(\frac{h^2}{\nu}, \frac{h}{\max_{\mathbf{x}\in\Omega} |\mathbf{u}^n(\mathbf{x})|}\right).$$

Let us now consider an implicit discretization of (6.49), for instance the backward Euler method. As already seen, this is an unconditionally stable method. Its algebraic system can be regarded as the finite element space approximation of the Navier-Stokes problem

$$\begin{cases} -\nu\Delta u^{n+1} + (\mathbf{u}^{n+1}\cdot\nabla)\mathbf{u}^{n+1} + \nabla p^{n+1} + \frac{\mathbf{u}^{n+1}}{\Delta t} = \tilde{\mathbf{f}}, \\ \operatorname{div}\mathbf{u}^{n+1} = 0. \end{cases}$$

The solution of such nonlinear algebraic system can be achieved by Newton-Krylov techniques, that is using a Krylov method to solve, at each timestep, a system obtained by a Newton method.

Fractional step methods

Let us consider an abstract time dependent problem,

$$\frac{\partial w}{\partial t} + \mathcal{L}w = f,$$

where $\mathcal{L}$ is a differential operator that splits into the sum of two operators $\mathcal{L}_1$ and $\mathcal{L}_2$ such that

$$\mathcal{L}v = \mathcal{L}_1 v + \mathcal{L}_2 v$$

We perform an advancement in time only on the first operator, and then correct the solution by performing the advancement on the second operator. The idea is to separate a complex problem into smaller simpler ones, for example separating diffusion from transport.

Chorin Temam method

This strategy also involves two operators: $\mathcal{L}_1(\mathbf{w}) = -\nu\Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{w}$ whereas $\mathcal{L}_2$ is associated to the remaining terms of (6.2). Thanks to this solution we can split the main difficult of the Navier-Stokes equations, the nonlinear part from the incompressiblility constraint.
The corresponding scheme reads

(1) Solve the diffusion-transport equation for the velocity $\tilde{\mathbf{u}}^{n+1}$

$$\begin{cases} \dfrac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\tilde{\mathbf{u}}^{n+1} + (\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \tilde{\mathbf{u}}^{n+1} = \mathbf{0} & \text{on } \partial\Omega \end{cases} \tag{6.52}$$

(2) solve the following problem for $\mathbf{u}^{n+1}$ and p^{n+1}

$$\begin{cases} \dfrac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\tilde{\mathbf{u}}^{n+1} + (\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \operatorname{div}\mathbf{u}^{n+1} = 0 & \text{in } \Omega, \\ \mathbf{u}^{n+1}\cdot\mathbf{n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{6.53}$$

where $\mathbf{u}^*$ and $\mathbf{u}^{**}$ can be either $\tilde{\mathbf{u}}^{n+1}$ or $\mathbf{u}^n$ depending on whether the nonlinear terms are treated. In such a way, in the first step the velocity is calculated as $\tilde{\mathbf{u}}^{n+1}$, and then it is corrected in the second step to satisfy the constraints.
While the first step is a classic advection-diffusion problem, the second step is a bit more tricky. First we need to apply the divergence operator to the first equation of (6.53)

$$\operatorname{div}\frac{\mathbf{u}^{n+1}}{\Delta t} - \operatorname{div}\frac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} + \Delta p^{n+1} = 0,$$

that is an elliptic boundary value problem with Neumann boundary conditions

$$\begin{cases} -\Delta p^{n+1} = -\operatorname{div}\dfrac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} & \text{in } \Omega, \\ \dfrac{\partial p^{n+1}}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.54}$$

The Neumann condition follows from $\mathbf{u}^{n+1} \cdot \mathbf{n} = 0$ on $\partial\Omega$.
From the solution of (6.54) we obtain p^{n+1} and $\mathbf{u}^{n+1}$ by using the first equation on (6.53),

$$\mathbf{u}^{n+1} = \tilde{\mathbf{u}}^{n+1} - \Delta t \nabla p^{n+1} \tag{6.55}$$

This is precisely the correction to operate on the velocity field in order to fullfill the divergence-free constraint.
To conclude, the full algorithm reads:

(1) Solve the elliptic system (6.52) to obtain $\tilde{\mathbf{u}}^{n+1}$

$$\begin{cases} \dfrac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\tilde{\mathbf{u}}^{n+1} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \tilde{\mathbf{u}}^{n+1} = \mathbf{0} & \text{on } \partial\Omega \end{cases} \tag{6.56}$$

(2) solve the scalar elliptic problem (6.54) to obtain p^{n+1}

$$\begin{cases} -\Delta p^{n+1} = -\operatorname{div}\dfrac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} & \text{in } \Omega, \\ \dfrac{\partial p^{n+1}}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.57}$$

(3) obtain the corrected velocity $\mathbf{u}^{n+1}$ thanks to the correction equation (6.55)

$$\mathbf{u}^{n+1} = \tilde{\mathbf{u}}^{n+1} - \Delta t \nabla p^{n+1} \tag{6.58}$$

As with many explicit schemes, if we take $\mathbf{u}^* = \mathbf{u}^{**} = \mathbf{u}^n$ we obtain a method that has a stability restriction on the timestep like

$$\Delta t \leq C \frac{h}{\max_{\mathbf{x}\in\Omega} |\mathbf{u}^n(\mathbf{x})|}.$$

On the other hand, this system splits into independent systems of smaller size, one for each spatial component of the velocity fields.
If we decided to use $\mathbf{u}^* = \mathbf{u}^{**} = \mathbf{u}^{n+1}$ we obtain an unconditionally stable method, but the downside is that the spatial components aren't separated due to the nonlinear convective term. Its algebraic formulation is of the form

$$\left(\frac{1}{\Delta t}M + A\right)\tilde{\mathbf{u}}^{n+1} = \tilde{\mathbf{f}}^{n+1}$$

On the second step, we generate a divergence free solution, but there is a failure to satisfy the physical condition in the case the pressure boundary layer has a width greater than

$$\sqrt{\nu\Delta t}$$

7 Hyperbolic equations

7.1 Introduction

Let us consider the following scalar hyperbolic problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 & x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{7.1}$$

where $a \in \mathbb{R}\setminus\{0\}$. The solution of such problem is a wave travelling at a velocity a, in the (x,t) plane, given by

$$u(x,t) = u_0(x - at),\ t \geq 0.$$

We consider the curves $x(t)$ in the plane (x,t), solutions of the following ordinary differential equation

$$\begin{cases} \dfrac{dx}{dt} = a, & t > 0, x(0) = x_0, \end{cases}$$

for varying values of $x_0 \in \mathbb{R}$. They read $x(t) = x_0 + at$ and are called characteristic lines. The solution of (7.1) along these lines remains constant, for

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0.$$

In the case of the more general problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x}a_0 u = f, & x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{7.2}$$

where a, a_0 and f are given functions of (x,t), the characteristic lines $x(t)$ are the solution of the Cauchy problem

$$\begin{cases} \dfrac{dx}{dt} = a(x,t), & t > 0, \\ x(0) = x_0. \end{cases}$$

In this case, solutions of (7.2) satisfy

$$\frac{d}{dt}u(x(t),t) = f(x(t),t) - a_0(x(t),t)u(x(t),t).$$

Therefore, it is possible to extract the solution u by solving an ordinary differential equation on each characteristic curve. If, for instance $a_0 = 0$, we find

$$u(x,t) = u_0(x - at) + \int_\Omega \int_0^t f(x - a(t-s), s)\, ds,\ t > 0.$$

Let us now consider problem (7.1) in a bounded interval. For instance, let us $x \in [0,1]$ and $a > 0$. As u is constant on the characteristics, from the figure we deduce that the value of the solution at a point P coincides with the value of u_0 at the foot P_0 of the characteristic outgoing from P. Instead, a characteristic outgoing from Q intersects $x = 0$ for $t > 0$, so $x = 0$ is an inflow point. Same thing for $x = 1$ and $a < 0$.

In the case of u_0 discontinuous in x_0, such discontinuity would propagate along the characteristic outgoing from x_0. In order to regularize it, one could approximate the initial datum u_0 with a series of regular functions $u_0^\varepsilon(x)$, $\varepsilon > 0$. But is an effective solution to a linear problem. In the case of nonlinear hyperbolic problems can indeed develop discontinuities for regular initial data. In such cases, the best approach is to regularize the differential equation itself, rather than the initial datum, obtaining, for example, in the case of diffusion-transport equation

$$\frac{\partial u^\varepsilon}{\partial t} + a\frac{\partial u^\varepsilon}{\partial x} = \varepsilon\frac{\partial^2 u^\varepsilon}{\partial x^2},\ x \in \mathbb{R},\ t > 0,$$

which, for small values of ε can be regarded as a parabolic regularization of (7.1).

7.2 A priori estimate

Let us now return to the problem (7.2) on a bounded interval

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} + a_0 u = f, & x \in (\alpha, \beta),\ t > 0, \\ u(x,0) = u_0(x), & x \in (\alpha, \beta), \\ u(\alpha, t) = \varphi(t) & t > 0, \end{cases} \tag{7.3}$$

where $a(x), f(x,t)$ and $\varphi(t)$ are assigned functions. We have made the assumptions that $a(x) > 0$, so that the inflow point is α and β the outflow point. By multiplying the first equation of (7.3) by u, integrating in x by parts, we obtain

$$\frac{1}{2}\frac{d}{dt}\int_\alpha^\beta u^2\,dx + \int_\alpha^\beta \left(a_0 - \frac{1}{2}a_x\right) u^2\,dx + \frac{1}{2}(au^2)(\beta) - \frac{1}{2}(au^2)(\alpha) = \int_\alpha^\beta fu\,dx.$$

By supposing that there exists a $\mu_0 \geq 0$ such that

$$a_0 - \frac{1}{2}a_x \geq \mu_0 \quad \forall\, x \in [\alpha, \beta],$$

we find

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\alpha,\beta)} + \mu_0\|u(t)\|^2_{L^2(\alpha,\beta)} + \frac{1}{2}(au^2)(\beta) \leq \int_\alpha^\beta fu\,dx + \frac{1}{2}a(\alpha)\varphi^2(t).$$

In the case f and φ are null, then

$$\|u(t)\|_{L^2(\alpha,\beta)} \leq \|u_0\|_{L^2(\alpha,\beta)} \quad \forall\, t > 0.$$

In the case of the more general problem (7.2), if we suppose that $\mu_0 > 0$, thanks to the Cauchy-Schwartz and Young inequalities, we have

$$\int_\alpha^\beta fu\,dx \leq \|f\|_{L^2(\alpha,\beta)}\|u\|_{L^2(\alpha,\beta)} \leq \frac{\mu_0}{2}\|u\|^2_{L^2(\alpha,\beta)} + \frac{1}{2\mu_0}\|f\|^2_{L^2(\alpha,\beta)}.$$

Then, integrating over time, we obtain

$$\begin{aligned} &\|u(t)\|^2_{L^2(\alpha,\beta)} + \mu_0\int_0^t \|u(s)\|^2_{L^2(\alpha,\beta)}\,ds + a(\beta)\int_0^t u^2(\beta,s)\,ds \\ &\leq \|u_0\|^2_{L^2(\alpha,\beta)} + a(\alpha)\int_0^t \varphi^2(s)\,ds + \frac{1}{\mu_0}\int_0^t \|f\|^2_{L^2(\alpha,\beta)}. \end{aligned}$$

An alternative estimate could be obtained without requiring the differentiability of $a(x)$, by using the hypothesis that $a_0 \leq a(x) \leq a_1$ and multiplying the equation for a^{-1},

$$a^{-1}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = a^{-1}f.$$

By multiplying by u and integrating

$$\frac{1}{2}\frac{d}{dt}\int_\alpha^\beta a^{-1}(x)u^2(x,t)\,dx\frac{1}{2}u^2(\beta,t) = \int_\alpha^\beta a^{-1}(x)f(x,t)u(x,t)\,dx + \frac{1}{2}\varphi^2(t).$$

We can define

$$\|v\|_a = \left(\int_\alpha^\beta a^{-1}(x)v(x)\,dx\right)^{\frac{1}{2}}.$$

If $f = 0$ we obtain

$$\|u(t)\|^2_a + \int_0^t u^2(\beta,s)\,ds = \|u_0\|^2_a + \int_0^t \varphi^2(s)\,ds, \quad t > 0.$$

Thanks to the lower and upper bounds of a^{-1}, $\|v\|_a$ is equivalent to the norm of $L^2(\alpha,\beta)$. On the other hand, if f is non-null we have

$$\|u(t)\|_a^2 + \int_0^t u^2(\beta,s)\,ds = \|u_0\|_a^2 + \int_0^t \varphi^2(s)\,ds + \int_0^t \|f\|_a^2\,ds + \int_0^t \|u(s)\|_a^2\,ds,$$

thanks to Cauchy-Schwartz. Then, applying Gronwall's lemma we obtain, for each $t>0$,

$$\|u(t)\|_a^2 + \int_0^t u^2(\beta,s)\,ds \le e^t\left(\|u_0\|_a^2 + \int_0^t \varphi^2(s)\,ds + \int_0^t \|f\|_a^2\,ds\right).$$

7.3 The finite difference method

Going back to (7.1) out of simplicity. To solve it numerically, we can use spatio-temporal discretization based on the finite difference method. In this case, the half-plane $\{t>0\}$ is discretized choosing a temporal step Δt, a spatial discretization h and defining the gridpoint (x_j,t^n) in the following way

$$\begin{aligned} x_j &= jh, && j\in\mathbb{Z},\\ t^n &= n\Delta t, && n\in\mathbb{N}. \end{aligned}$$

Set $\lambda=\frac{\Delta t}{h}$, and define

$$x_{j+\frac{1}{2}} = x_j + \frac{h}{2}$$

Usually hyperbolic initial value problems are discretized using explicit methods. Of course this imposes restriction on λ. Any explicit finite difference method can be written as

$$u_j^{n+1} = u_j^n - \lambda(H^n_{j+\frac{1}{2}} - H^2_{j-\frac{1}{2}}), \tag{7.4}$$

where $H^n_{j+\frac{1}{2}} = H(u_j^n, u_{j+1}^n)$ for a suitable function $H(\cdot,\cdot)$ called numerical flux. The numerical scheme (7.4) is basically the outcome of the following consideration. Suppose that a is a constant and let us write (7.1) in conservation form

$$\frac{\partial u}{\partial t} + \frac{\partial(au)}{\partial x} = 0,$$

with au being the flux associated to the equation. By integrating in space we obtain

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \frac{\partial u}{\partial t}\,dx + [au]_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} = 0, \quad j\in\mathbb{Z},$$

that is

$$\frac{\partial}{\partial t}U_j + \frac{(au)(x_{j+\frac{1}{2}}) - (au)(x_{j-\frac{1}{2}})}{h},$$

where

$$U_j = h^{-1}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\,dx.$$

Equation (7.4) can now be interpreted as an approximation where the temporal derivative is discretized using the forward Euler finite difference scheme, U_j is replaced by u_j and $H_{j+\frac{1}{2}}$ is an approximation of $(au)(x_{j+\frac{1}{2}})$.

7.4 Discretization of the scalar equation

The choice of H decide which numerical method will be use. Some of them are

- forward/Centered Euler (FE/C)
- Lax-Friedrichs (LF)
- Lax-Wendroff (LW)
- Upwind (U)

Forward/Centered Euler

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2} a(u_{j+1}^n - u_{j-1}^n), \tag{7.5}$$

that takes the form (7.4) provided that we define

$$H_{j+\frac{1}{2}} = \frac{1}{2} a(u_{j+1} + u_j). \tag{7.6}$$

Lax-Friedrichs

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\lambda}{2} a(u_{j+1}^n - u_{j-1}^n), \tag{7.7}$$

that takes the form (7.4) when

$$H_{j+\frac{1}{2}} = \frac{1}{2}\left[a(u_{j+1} + u_j) - \lambda^{-1}(u_{j+1} - u_j)\right]. \tag{7.8}$$

Lax-Wendroff

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2} a(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda^2}{2} a^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \tag{7.9}$$

that takes the form (7.4) with

$$H_{j+\frac{1}{2}} = \frac{1}{2}\left[a(u_{j+1} + u_j) - \lambda a^2(u_{j+1} - u_j)\right]. \tag{7.10}$$

Upwind

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2} a(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda}{2}|a|(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \tag{7.11}$$

that takes the form (7.4) if

$$H_{j+\frac{1}{2}} = \frac{1}{2}\left[a(u_{j+1} + u_j) - \lambda|a|(u_{j+1} - u_j)\right]. \tag{7.12}$$

The LF method is obtained from the FE/C by replacing the nodal value u_j^n in (7.5) with the average of u_{j-1}^n and u_{j+1}^n. The LW method derive from the Taylor expansion in time

$$u^{n+1} = u^n + (\partial_t u)^n \Delta t + (\partial_{tt} u)^n \frac{\Delta t^2}{2} + \mathcal{O}(\Delta t^3)$$

where $(\partial_t u)^n$ denotes the partial derivative of u at time t^n. Then, using the equation (7.1), we replace $\partial_t u$ by $-a\partial_x u$, and $\partial_{tt} u$ by $a^2 \partial_{xx} u$. Neglecting the remainder $\mathcal{O}(\Delta t^3)$ and approximating the spatial derivatives with centered finite differences, we get (7.9). Finally, the U method is obtained by discretizing the convective term $a\partial_x u$ of the equation with the upwind finite difference.
An example of an implicit method is the following

Backward/Centered Euler (BE/C)

$$u_j^{n+1} + \frac{\lambda}{2}(u_{j+1}^{n+1} - u_{j+1}^{n+1}) = u_j^n. \tag{7.13}$$

7.5 Boundary treatment

In case we want to discretize the hyperbolic equation (7.3) on a bounded interval, we will obviously need to use the inflow $x = \alpha$ to impose the boundary condition, say $u_0^{n+1} = \varphi(t^{n+1})$, while at all other nodes x_j, $1 \leq j \leq m$ (including the outflow $x_m = \beta$) we will write the finite difference scheme.
However, schemes using a centered discretization of the space derivative require a particular treatment at x_m. Indeed, they would require the value u_{m+1}, which is unavailable as it relates to the point with coordinates $\beta + h$, outside the integration interval. This can be solved in various ways. An option is to use only the upwind decentered discretization on the last node, as such discretization does not require knowing the datum in x_{m+1}, but this is a first order approach.
Alternatively, the value u_m^{n+1} can be obtained through extrapolation from the values available at the internal nodes, like characteristic lines for example.

7.6 Analysis of the finite difference method

Consistency and convergence

For a given numerical scheme, the local truncation error is the error generated by expecting the exact solution to verify the numerical scheme. For instance, in the case of FE/C, having denoted by u the solution to the exact problem (7.1), we can define the truncation error at the point (x_j, t^n) as

$$\tau_j^n = \frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\Delta t} + a\frac{u(x_{j+1}, t^n) - u(x_{j-1}, t^n)}{2h}.$$

If the truncation error

$$\tau(\Delta t, h) = \max_{j,n} \left|\tau_j^n\right|$$

tends to zero when Δt and h tend to zero, independently, then the numerical scheme will be said to be consistent.
Moreover, we will say that a numerical scheme is accurate to order p in time and to order q in space. If, for a sufficiently regular solution of the exact problem we have

$$\tau(\Delta t, h) = \mathcal{O}(\Delta t^p + h^q).$$

Using Taylor expansion suitably, we can then see that the truncation of the previously introduced methods is:

- Euler: $\mathcal{O}(\Delta t + h^2)$.
- Upwind: $\mathcal{O}(\Delta t + h)$.
- Lax-Friedrichs: $\mathcal{O}(\frac{h^2}{\Delta t} + \Delta t + h^2)$.
- Lax-Wendroff: $\mathcal{O}(\Delta t^2 + h^2 + h^2\Delta t)$.

Finally, we will say that a scheme is convergent (in the maximum norm) if

$$\lim_{\Delta t, h \to 0} (\max_{j,} \left|u(x_j, t^n) - u_j^n\right|) = 0.$$

Obviously, we can also consider weaker norms, such as $\|\cdot\|_{\Delta,1}$ and $\|\cdot\|_{\Delta,2}$, which we will introduce later.

Stability

We will say that a numerical method for a linear hyperbolic problem is stable if for each instant T there exists a constant $C_T > 0$ such that for each $h > 0$, there exists $\delta_0 > 0$ such that for each $0 < \Delta t < \delta_0$ we have

$$\|\mathbf{u}^n\|_\Delta \leq C_T \left\|\mathbf{u}^0\right\|_\Delta, \tag{7.14}$$

for each n such that $n\Delta t \leq T$, and for each initial datum $\mathbf{u}_0$. C_T should not depend on Δt or h.

Notation 7.1
$\|\cdot\|_\Delta$ denotes a suitable discrete norm, for example

$$\|\mathbf{v}\|_{\Delta,p} = \left(h \sum_{j=-\infty}^{\infty} |v_j|^p \right)^{\frac{1}{p}} \quad \text{for } p = 1, 2 \quad \|\mathbf{v}\|_{\Delta,\infty} = \sup_j |v_j|. \tag{7.15}$$

This is an approximation of the L^p norm for some p.

Theorem 7.1
The implicit backward/centered Euler scheme (7.13) is stable in the norm $\|\cdot\|_{\Delta,2}$ for any choice of Δt and h. Precisely,

$$\|\mathbf{u}\|_{\Delta,2} \leq \|\mathbf{u}_0\|_{\Delta,2} \ \forall\ \Delta t, h > 0.$$

A scheme is called strongly stable with respect to the norm $\|\cdot\|_\Delta$ if

$$\|\mathbf{u}^n\|_\Delta \leq \left\|\mathbf{u}^{n-1}\right\|_\Delta, \tag{7.16}$$

for each n such that $n\Delta t \leq T$, and for each initial datum $\mathbf{u}_0$, which implies that (7.14) is verified with $C^T = 1$.

Remark 7.1
In the contest of hyperbolic problems, one often wants long-time solutions. Such cases require a strongly stable scheme, as this guarantees that the numerical solution is bounded for each value of T.

As we will see a necessary condition for the stability of an explicit numerical scheme like (7.4) is that the temporal and spatial discretization steps satisfy

$$|a\lambda| \leq 1, \text{ or } \Delta t \leq \frac{h}{|a|}. \tag{7.17}$$

This is the celebrated CFL condition (from Courant, Friedrichs and Lewy). $a\lambda$ is commonly called CFL number and is a physically dimensionless quantity.

Remark 7.2
The CFL condition establishes, in particular, that there is no explicit finite difference scheme for hyperbolic initial value problems that is unconditionally stable and consistent. Indeed, suppose the CFL condition is violated. Then there exists at least a point x^* in the dependency domain that doesn't belong to the numerical dependency domain. Then, changing the initial datum to x^* will only modify the exact solution and not the numerical one.

Remark 7.3
In the case where $a = a(x,t)$ is no longer constant in (7.1), the CFL condition

$$\Delta t \leq \frac{h}{\sup_{\substack{x\in\mathbb{R}\\ t>0}} |a(x,t)|}.$$

If the spatial discretization step varies, we have

$$\Delta t \leq \frac{h}{\sup_{\substack{x\in(x_k,x_{k+1})\\ t>0}} |a(x,t)|}$$

with $h_k = x_{k+1} - x_k$.

Theorem 7.2
If the CFL condition (7.17) is satisfied, the upwind, Lax-Friedrichs and Lax-Wendroff schemes are

strongly stable $\|\cdot\|_{\Delta,1}$.

Theorem 7.3
If the CFL is verified, the upwind scheme satisfies

$$\|\mathbf{u}^n\|_{\Delta,\infty} \leq \|\mathbf{u}_0\|_{\Delta,\infty} \quad \forall\, n \geq 0. \tag{7.18}$$

This relation is called discrete maximum principle.

Theorem 7.4
The backward Euler BE/C is strongly stable in the norm $\|\cdot\|_{\Delta,2}$, with no restriction on Δt. The forward FE/C, instead, is never strongly stable. However, it is stable with constant $C_T = e^{\frac{T}{2}}$ provided that we assume that Δt satisfies the following condition (more restrictive than the CFL condition)

$$\Delta t \leq \left(\frac{h}{a}\right)^2 \tag{7.19}$$

7.7 Wave equation

Let us consider the following second order hyperbolic problem

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} = 0 & x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \\ \frac{\partial u}{\partial t} = v_0(x), & x \in \mathbb{R}, \end{cases} \tag{7.20}$$

where $c > 0$ is characteristic velocity.
The function u can represent the vertical displacement of a vibrating string elastic chord. The functions $u_0(x)$ and $v_0(x)$ describe the initial displacement and the velocity of the chord.
By performing, the change of variables $\mathbf{w} = (\partial_x u, \partial_t u)^T$, we can rewrite (7.20) as a first order system of PDEs, in the form

$$\begin{cases} \dfrac{\partial \mathbf{w}}{\partial t} + A \dfrac{\partial \mathbf{w}}{\partial t} = \mathbf{0}, & x \in \mathbb{R},\ t > 0, \\ \mathbf{w}(x,0) = \mathbf{w}_0, & x \in \mathbb{R}, \end{cases} \tag{7.21}$$

with

$$A = \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \end{pmatrix}, \quad \mathbf{w}_0 = \begin{pmatrix} \partial_x u_0(x) \\ v_0(x) \end{pmatrix}.$$

Since the eigenvalues of A are distinct real $\pm c$ (representing the wave propagation rates), system (7.21) is hyperbolic.

7.8 Finite Difference discretization

The half-plane $\{t > 0\}$ is discretized choosing a temporal step Δt, a spatial discretization step h and defining (x_j, t^n) in the following way

$$\begin{aligned} x_j &= jh, & j &\in \mathbb{Z}, \\ t^n &= n\Delta t, & n &\in \mathbb{N}. \end{aligned}$$

Moreover, we set $\lambda = \frac{\Delta t}{h}$ and seek discrete solutions $\mathbf{w}_j^n$ which approximate $\mathbf{w}(x_j, t^n)$. On one hand, FE/C reads

$$\begin{cases} \mathbf{w}_j^{n+1} = \mathbf{w}_j^n - \frac{\lambda}{2} A(\mathbf{w}_{j+1}^n - \mathbf{w}_{j-1}^n), & \text{for } j \in \mathbb{Z}, n \in \mathbb{N}, \\ \mathbf{w}_j^0 = \mathbf{w}_0(x_j), & \text{for } j \in \mathbb{Z}. \end{cases} \tag{7.22}$$

On the other hand, the BE/C scheme

$$\begin{cases} \mathbf{w}_j^{n+1} = \mathbf{w}_j^n - \frac{\lambda}{2} A(\mathbf{w}_{j+1}^{n+1} - \mathbf{w}_{j-1}^{n+1}), & \text{for } j \in \mathbb{Z}, n \in \mathbb{N}, \\ \mathbf{w}_j^0 = \mathbf{w}_0(x_j), & \text{for } j \in \mathbb{Z}. \end{cases} \tag{7.23}$$

Let us study the numerical stability of the FE/C and BE/C schemes. The matrix A can be diagonalized as

$$A = T\Lambda T^{-1},$$

where $\Lambda = \mathrm{diag}(c, -c)$ and

$$T = \begin{pmatrix} 1 & 1 \\ -c & c \end{pmatrix}.$$

Then:

$$\begin{aligned} \frac{\partial \mathbf{w}}{\partial t} + A\frac{\partial \mathbf{w}}{\partial x} &= \mathbf{0} \\ T^{-1}\frac{\partial \mathbf{w}}{\partial t} + T^{-1}A\frac{\partial \mathbf{w}}{\partial x} &= \mathbf{0} \\ \frac{\partial T^{-1}\mathbf{w}}{\partial t} + \underbrace{T^{-1}AT}_{\Lambda}\underbrace{T^{-1}\frac{\partial \mathbf{w}}{\partial x}}_{\frac{\partial T^{-1}\mathbf{w}}{\partial x}} &= \mathbf{0} \end{aligned}$$

We then introduce the so-called characteristic variables $\mathbf{z} = T^{-1}\mathbf{w}$. Specifically, we have

$$\begin{aligned} z_1 &= \frac{1}{2}\left(w_1 - \frac{w_2}{c}\right), \\ z_2 &= \frac{1}{2}\left(w_1 + \frac{w_2}{c}\right). \end{aligned} \tag{7.24}$$

Thus:

$$\frac{\partial \mathbf{z}}{\partial t} + \Lambda\frac{\partial \mathbf{z}}{\partial x} = \mathbf{0}.$$

Hence, left-multiplying (7.22) by T^{-1}, and by writing $A = ATT^{-1}$, we get

$$\mathbf{z}_n^{n+1} = \mathbf{z}_j^n - \frac{\lambda}{2}\underbrace{T^{-1}AT}_{\Lambda}(\mathbf{z}_{j+1}^n - \mathbf{z}_{j-1}^n). \tag{7.25}$$

Being Λ a diagonal matrix, the equations of the above system can be decoupled as follows

$$\begin{aligned} (z_1)_j^{n+1} &= (z_1)_j^n - \frac{\lambda}{2}c((z_1)_{j+1}^n - (z_1)_{j-1}^n), \\ (z_2)_j^{n+1} &= (z_2)_j^n + \frac{\lambda}{2}c((z_1)_{j+1}^n - (z_2)_{j-1}^n). \end{aligned}$$

In conclusion, applying FE/C scheme (7.22) to system (7.21) is equivalent to independently apply scalar FE/C to each characteristic variable. The same argument can be applied to the BE/C scheme. Therefore, the FE/C and BE/C numerical schemes (7.22) and (7.23) are stable if and only if the corresponding scalar schemes are stable for each characteristic variable. We conclude that the BE/C scheme is strongly stable in the norm $\|\cdot\|_{\Delta,2}$ for any choice of the parameters h and Δt. Conversely, the FE/C scheme is stable in the norm $\|\cdot\|_{\Delta,2}$ whenever

$$\Delta t \leq \left(\frac{h}{c}\right)^2. \tag{7.26}$$

7.9 Wave equation in a bounded domain

Consider now problem (7.20) in a bounded domain $(a, b) \subset \mathbb{R}$.

$$\begin{cases} \dfrac{\partial^2 u}{\partial u^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} = 0 & x \in (a, b),\ t > 0, \\ u(x, 0) = u_0(x), & x \in \mathbb{R}, \\ \dfrac{\partial u}{\partial t} = v_0(x), & x \in (a, b), +\text{B.C.} \end{cases} \tag{7.27}$$

Equation (7.27) can be rewritten as

$$\begin{cases} \dfrac{\partial \mathbf{w}}{\partial t} + A \dfrac{\partial \mathbf{w}}{\partial x} = \mathbf{0}, & x \in (a, b),\ t > 0 \\ \mathbf{w}(x, 0) = \mathbf{w}_0, & x \in (a, b), \\ +\text{B.C.} \end{cases} \tag{7.28}$$

Let us know study boundary conditions can be meaningful for (7.27).
Left-multiplying (7.28) by T^{-1} we get

$$\begin{cases} \frac{\partial \mathbf{z}}{\partial t} + \Lambda \frac{\partial \mathbf{z}}{\partial x} = 0, & x \in (a, b), t > 0, \\ \mathbf{z}(x, 0) = \mathbf{z}_0 = T^{-1} \mathbf{w}_0, & x \in (a, b). \end{cases} \tag{7.29}$$

The first line can be decoupled into two independent equations:

$$\begin{aligned} &\frac{\partial z_1}{\partial t} + c \frac{\partial z_1}{\partial x} = 0 && x \in (a, b),\ t > 0, \\ &\frac{\partial z_2}{\partial t} - c \frac{\partial z_2}{\partial x} = 0 && x \in (a, b),\ t > 0, \end{aligned}$$

The two characteristic variables z_1 and z_2 are associated with positive c and negative $-c$ velocities respectively. Therefore, a possible set of boundary conditions would consist in assigning the first characteristic variable on the left boundary ($z_1(a, t) = g_a(t)$) and the second one on the right boundary ($z_2(b, t) = g_b(t)$). Considering the primitive variable $u(x, t)$, these conditions read as follows

$$\begin{cases} \dfrac{\partial u}{\partial x} - \dfrac{1}{c} \dfrac{\partial u}{\partial t} = 2 g_a(t), & t > 0, \\ \dfrac{\partial u}{\partial x} + \dfrac{1}{c} \dfrac{\partial u}{\partial t} = 2 g_b(t), & t > 0. \end{cases} \tag{7.30}$$

In general terms, given a hyperbolic system in the form (7.21), the number of positive eigenvalues of A determines the number of boundary conditions to be assigned at $x = a$, while $x = b$ we will need to assign as many conditions as the number of negative eigenvalues.

7.10 Finite element method

We now illustrate how to apply Galerkin methods to the discretization of scalar hyperbolic equations. Let us consider the following model problem, where a is a positive constant

$$\begin{cases} \dfrac{\partial u}{\partial t} + a \dfrac{\partial u}{\partial x} = f, & x \in (0, 1),\ t > 0 \\ u(x, 0) = u_0(x), & x \in [0, 1], \\ u(0, t) = 0, & t > 0, \end{cases} \tag{7.31}$$

To start with, we proceed with a spatial discretization based on continuous finite element. We therefore attempt a semidiscretization of the following form

$$\begin{aligned} &\forall\, t > 0,\ \text{find } u_h = u_h(t) \in V_h : \\ &\qquad \left(\frac{\partial u_h}{\partial t}, v_h \right) + a \left(\frac{\partial u_h}{\partial x}, v_h \right) = (f, v_h) \quad \forall\, v_h \in V_h, \end{aligned} \tag{7.32}$$

with u_h^0 being the approximation of the initial datum. We have set

$$V_h = \{v_h \in X_h^r : v_h(0) = 0\}, \quad r \geq 1.$$

For the temporal discretization we will use finite difference schemes. In the case of BE/C we will have

$$\begin{aligned} &\forall\ > 0,\ \text{find } u_h^{n+1} \in V_h : \\ &\quad \frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v_h) + a\left(\frac{\partial u_h^{n+1}}{\partial x}, v_h\right) = (f^{n+1}, v_h) \quad \forall\, v_h \in V_h. \end{aligned} \tag{7.33}$$

By choosing $v_h = u_h^{n+1}$ with $f = 0$, we obtain, as in the case of finite differences, that the BE/C scheme is strongly with no restriction on Δt.
In case we use the FE/C scheme instead, the discrete problem becomes:

$$\begin{aligned} &\forall\ > 0,\ \text{find } u_h^{n+1} \in V_h : \\ &\quad \frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v_h) + a\left(\frac{\partial u_h^{n}}{\partial x}, v_h\right) = (f^{n+1}, v_h) \quad \forall\, v_h \in V_h. \end{aligned} \tag{7.34}$$

As in the case of finite differences, the FE/C is never strongly stable.
To make it stable, we use the other methods, LF, U or LW. By observing that the these schemes can be rewritten as

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} - \mu\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0. \tag{7.35}$$

The second term is the discretization via centered finite differences of the convective term $au_x(t^n)$, while the third one is a numerical diffusion term and corresponds to the discretization via finite differences of $-\mu u_{xx}(t^n)$. The numerical viscosity coefficient μ is given by

$$\mu = \begin{cases} \dfrac{h^2}{2\Delta t} & LF \\ \dfrac{a^2\Delta t}{2} & LW \\ \dfrac{ah}{2} & \end{cases} \tag{7.36}$$

Equation (7.35) suggests the following finite element version for the approximation of problem (7.31):

$$\begin{aligned} &\forall\ > 0,\ \text{find } u_h^{n+1} \in V_h : \\ &\quad \frac{1}{\Delta t}(u_h^{n+1} - u_n^h, v_h) + a\left(\frac{\partial u_n^h}{\partial x}, v_h\right) + \mu\left(\frac{\partial u_h^n}{\partial x}, \frac{\partial v_h}{\partial x}\right) \\ &\quad (f^n, v_h) \quad \forall\, v_h \in V_h, \end{aligned} \tag{7.37}$$

where the artificial viscosity μ is given in (7.36) by analogy with the finite difference case. We denote the resulting methods as LF-FE, LW-FE and U-FE, respectively.
As a matter of fact, it is possible to prove that the same stability results of finite differences holds for finite elements as well.

8 Domain Decomposition

Consider the elliptic BVP:

$$\begin{cases} \mathcal{L}u := -u'' = f & a < x < b, \\ u(a) = u(b) = 0. \end{cases} \tag{8.1}$$

Now consider the splitting of $\overline{\Omega}$ in $\overline{\Omega}_1 \cup \overline{\Omega}_2 = \varnothing$. And now consider the splitting of the space V:

$$V = V_1 \oplus H_\gamma \oplus V_2 \tag{8.2}$$

where

$$\begin{aligned} V_i = \tilde{H}_0^1 &:= \left\{ v \in H_0^1(\Omega) : v|_{\Omega_i} \in H_0^1(\Omega_i), v|_{\Omega \setminus \Omega_i} = 0 \right\}, \ i = 1, 2 \\ H_\gamma &= \text{ space of harmonic extensions of functions whose values at } \gamma \text{ is given} \\ &= \left\{ v_H \in V : v(\gamma) = v_\gamma, v|_{\Omega_i} = v_i, v_i'' = 0, i = 1, 2 \ \forall \ v_\gamma \in \mathbb{R} \right\} \end{aligned}$$

Remark 8.1
The splitting (8.2) is unique.

So we can say

Proposition 8.1
u is solution to the weak formulation of (8.1) if and only if:

$$\begin{cases} -u_1'' = f & \text{in } \Omega_1 \\ -u_2'' = f & \text{in } \Omega_2 \\ u_1 = u_2 & \text{Dirichlet transmission condition} \\ u_1' = u_2' & \text{Neumann transmission condition} \\ u_1(a) = u_2(b) = 0 & \text{B.C.} \end{cases} \tag{8.3}$$

The converse is also true, meaning that if $u1, u_2$ is the solution to (8.3), then $u : u|_{\Omega_1} = u_1, u|_{\Omega_2} = u_2$ is solution to the weak formulation of (8.1). So (8.1) and (8.3) are equivalent.

In the case of overlapping domains we have $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}, \Omega_1 \cap \Omega_2 = \Omega_{12}$ which has a non-null measure. Then we have

Proposition 8.2
This problem is equivalent to (8.3)

$$\begin{cases} -u_1'' = f & \text{in } \Omega_1 \\ -u_2'' = f & \text{in } \Omega_2 \\ u_1 = u_2 & \text{in } \Omega_{12} \\ u_1(a) = u_2(b) = 0 & txtB.C. \end{cases} \tag{8.4}$$

The condition on Ω_{12} is called continuity condition.

This method can be used in the discretization method of any PDEs, to make their algebraic solution more efficient on parallel computer systems. These methods allow for the reformulation of problems into multiple subdomains, which is very convenient for multi-physics problems, in which there are different equations in different subregions of the domain.

8.1 Iterative methods

Consider the problem

$$\text{find } u : \Omega \to \mathbb{R} \text{ s.t. } \begin{cases} \mathcal{L}u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{8.5}$$

The weak formulation reads:

$$\text{find } u \in V = H_0^1(\Omega) : a(u, v) = (f, v) \quad \forall \ v \in V,$$

Schwarz Method

Now consider a decomposition of Ω into subdomains with an overlap. We have Ω_1, Ω_2 and their overlap Ω_{12} which boundary can be divided into two subregions $\Gamma_1 \cup \Gamma_2 = \partial\partial\Omega_{12}$, where Γ_1 is the old boundary of Ω_1 and Γ_2 of Ω_2.
Then, given u_2^0 on Γ_1, for $k \geq 1$:

$$\text{solve} \begin{cases} \mathcal{L}u_1^{(k)} = f & \text{in } \Omega_1 \\ u_1^{(k)} = u_2^{(k-1)} & \text{on } \Gamma_1 \\ u_1^{(k)} = 0 & \text{on } \partial\Omega_1 \setminus \Gamma_1 \end{cases}$$

$$\text{solve} \begin{cases} \mathcal{L}u_2^{(k)} = f & \text{in } \Omega_2 \\ u_2^{(k)} = \begin{cases} u_1^{(k-1)} \\ u_1^{(k)} \end{cases} & \text{on } \Gamma_2 \\ u_2^{(k)} = 0 & \text{on } \partial\Omega_2 \setminus \Gamma_2 \end{cases}$$

The choice of $u_1^{(k)}$ implies that we are using a multiplicative Schwarz method, otherwise we are using an additive one.
The Schwarz method always converges with a rate that increase as the overlap of the two domain increases.

Remark 8.2
The Schwarz method requires that at each iteration the solution of the subproblems is of the same kind of the original problem, and that the boundary condition remain unchanged. Also, it is necessary to apply Dirichlet B.C. at the interfaces.

Dirichlet-Neumann method

In the case of non overlapping domain, so a common interface Γ, the following theorem holds

Theorem 8.1
The solution of the (8.5) is such that $u|_{\Omega_i} = u_i$ for $i = 1, 2$ where u_i is the solution to

$$\begin{cases} \mathcal{L}u_i = f & \text{in } \Omega_i \\ u_i = 0 & \text{on } \partial\Omega_i \setminus \Gamma \end{cases}$$

with interface conditions

$$u_1 = u_2 \quad \frac{\partial u_i}{\partial n_L} = \frac{\partial u_2}{\partial n_L} \quad \text{on } \Gamma$$

where $\frac{\partial}{\partial n_L}$ is the conormal derivative.

And we can use the Dirichlet-Neumann, which, given u_2^0 on Γ, for $k \geq 1$

$$\text{solve} \begin{cases} \mathcal{L}u_1^{(k)} = f & \text{in } \Omega_1 \\ u_1^{(k)} = u_2^{(k-1)} & \text{on } \Gamma \\ u_1^{(k)} = 0 & \text{on } \partial\Omega_1 \setminus \Gamma \end{cases}$$

$$\text{solve} \begin{cases} \mathcal{L}u_2^{(k)} = f & \text{in } \Omega_2 \\ \dfrac{\partial u_2^{(k)}}{\partial n_L} = \dfrac{\partial u_1^{(k)}}{\partial n_L} & \text{on } \Gamma \\ u_2^{(k)} = 0 & \text{on } \partial\Omega_2 \setminus \Gamma \end{cases}$$

Thanks to the theorem, when the two sequences $\left\{u_1^{(k)}\right\}$ and $\left\{u_2^{(k)}\right\}$ converge, their limit will be the solution to the exact problem. So the method is consistent, but its convergence is not guaranteed.

Other methods

Considering a partition of Ω into two disjoint subdomains and denote by λ the unknown value of the solution on their interface Γ we have $\lambda = u_i$ on Γ for $i = 1, 2$. Consider the following iterative algorithm (called Neumann-Neumann): for any given $\lambda^{(0)}$ on Γ, for $k \geq 0$ and $i = 1, 2$

$$\text{solve} \begin{cases} \mathcal{L}u_i^{(k+1)} = f & \text{in } \Omega_i \\ u_i^{(k+1)} = \lambda^{(k)} & \text{on } \Gamma \\ u_i^{(k+1)} = 0 & \text{on } \partial\Omega_i \setminus \Gamma \end{cases}$$

$$\text{solve} \begin{cases} \mathcal{L}\psi_i^{(k+1)} = 0 & \text{in } \Omega_i \\ \dfrac{\partial \psi_i^{(k+1)}}{\partial n} = \dfrac{\partial u_1^{(k+1)}}{\partial n} - \dfrac{\partial u_2^{(k+1)}}{\partial n} & \text{on } \Gamma \\ \psi_i^{(k+1)} = 0 & \text{on } \partial\Omega_i \setminus \Gamma \end{cases}$$

with

$$\lambda^{(k+1)} = \lambda^{(k)} - \theta \left(\sigma_1 \psi_{1|_\Gamma}^{(k+1)} - \sigma_2 \psi_{2|_\Gamma}^{(k+1)} \right)$$

Another method is the Robin-Robin algorithm: for every $k \geq 0$

$$\text{solve} \begin{cases} \mathcal{L}u_i^{(k+1)} = f & \text{in } \Omega_1 \\ \dfrac{\partial u^{(k+1)}}{\partial n} + \gamma_1 u_1^{(k+1)} = \dfrac{\partial u_2^{(k)}}{\partial n} + \gamma_1 u_2^{(2)} & \text{on } \Gamma \\ u_i^{(k+1)} = 0 & \text{on } \partial\Omega_1 \setminus \Gamma \end{cases}$$

$$\text{solve} \begin{cases} \mathcal{L}u_2^{(k+1)} = f & \text{in } \Omega_2 \\ \dfrac{\partial u_2^{(k+1)}}{\partial n} + \gamma_2 u_2^{(k+1)} = \dfrac{\partial u_1^{(k+1)}}{\partial n} - \gamma_2 u_1^{(k+1)} & \text{on } \Gamma \\ u_2^{(k+1)} = 0 & \text{on } \partial\Omega_2 \setminus \Gamma \end{cases}$$

where $u_2^{(0)}$ is assigned and γ_1, γ_2 are non-negative acceleration parameters that satisfy $\gamma_1 + \gamma_2 > 0$.

8.2 The Steklov-Poincaré interface equation

Let us change the model problem with a classical Poisson equation

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{8.6}$$

The equivalent multidomain formulation is ($u_i = u|_{\Omega_i}$)

$$\begin{cases} -\Delta u_1 = f & \text{in } \Omega_1 \\ -\Delta u_2 = f & \text{in } \Omega_2 \\ u_1 = 0 & \text{on } \partial\Omega_1 \setminus \Gamma \\ u_2 = 0 & \text{on } \partial\Omega_2 \setminus \Gamma \\ u_1 = u_2 & \text{on } \Gamma \\ \frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} & \text{on } \Gamma \end{cases} \tag{8.7}$$

Remark 8.3
On the interface we Γ we have the normal unit vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, with $\mathbf{n}_1 = -\mathbf{n}_2$. We can denote $\mathbf{n} = \mathbf{n}_1$ so that

$$\frac{\partial}{\partial n} = \frac{\partial}{\partial n_1} = -\frac{\partial}{\partial n_2} \quad \Gamma.$$

The Steklov Poincaré operator

Let λ be the unknown value of u on Γ:

$$\lambda = u|_\Gamma$$

If we know λ on Γ, we could solve the following two independent boundary-value problems with Dirichlet conditions on Γ and $i = 1, 2$:

$$\begin{cases} -\Delta w_i = f & \text{in } \Omega_i \\ w_i = \lambda & \text{on } \Gamma \\ w_i = 0 & \text{on } \partial\Omega_i \setminus \Gamma \end{cases}$$

With the aim of obtaining the value λ on Γ, let us split w_i in

$$w_i = w_i^* + u_i^0,$$

where w_i^* and u_i^0 are the solutions to ($i = 1, 2$):

$$\begin{cases} -\Delta w_i^* = f & \text{in } \Omega_i \\ w_i^* = 0 & \text{on } \partial\Omega_i \cap \partial\Omega \\ w_i^* = 0 & \text{on } \Gamma \end{cases}$$

$$\begin{cases} -\Delta u_i^0 = 0 & \text{in } \Omega_i \\ u_i^0 = 0 & \text{on } \partial\Omega_i \cap \partial\Omega \\ u_i^0 = 0 & \text{on } \Gamma \end{cases}$$

The functions w_i^* depend solely on the source data f ($\lambda = G_i f$), while u_i^0 depends on λ. We could write $u_i^0 = H_i\lambda$, where H_i is the harmonic extension operator of λ on the domain Ω_i.

$$\begin{aligned} u_i = w_i^* + u_i^0 &\iff \frac{\partial w_1}{\partial n} = \frac{\partial w_2}{\partial n} \quad \text{on } \Gamma \\ &\iff \frac{\partial}{\partial n}\big(w_i^* + u_i^0\big) = \frac{\partial}{\partial n}\big(w_2^* + u_2^0\big) \quad \text{on } \Gamma \\ &\iff \frac{\partial}{\partial G_1 f + H_1\lambda} = \frac{\partial}{\partial n}(G_2 f + H_2\lambda) \quad \text{on } \Gamma \\ &\iff \left(\frac{\partial H_1}{\partial n} - \frac{\partial H_2}{\partial n}\right)\lambda = \left(\frac{\partial G_2}{\partial n} - \frac{\partial G_1}{\partial n}\right) f \quad \text{on } \Gamma. \end{aligned}$$

which is the Steklov-Poincaré equation for λ on Γ

$$S\lambda = \chi \quad \text{on } \Gamma$$

where S is the Steklov-Poincaré operator

$$S\mu = \frac{\partial}{\partial n} H_1\mu - \frac{\partial}{\partial n} H_2\mu = \sum_{i=1}^{2} \frac{\partial}{\partial n_i} H_i\mu$$

χ is a linear functional

$$\chi = \frac{\partial}{\partial n} G_2 f - \frac{\partial}{\partial n} G_1 f = -\sum_{i=1}^{2} \frac{\partial}{\partial n_i} G_i f.$$

The operator

$$S_i : \mu \to S_i\mu = \frac{\partial}{\partial n_i}(H_i\mu)\Big|_\Gamma$$

is called local Steklov-Poincaré operator which operates between

$$\Lambda = \{\mu : \exists\, v \in V \text{s.t.} \mu = v|_\Gamma\} = H_{00}^{\frac{1}{2}}(\Gamma)$$

and its dual Λ'.

8.3 Finite Element Method: multi-domain formulation

Consider (8.6) and its weak formulation

$$\text{find } u \in V : a(u,v) = F(v) \forall\, v \in V$$

with $V = H_0^1(\Omega)$. Suppose Ω is split into two non-overlapping subdomains and consider a uniform triangulation $\mathcal{T}_h$ of Ω, conforming on Γ. The Galerkin finite element approximation of the Poisson problem reads:

$$\text{find } u_h \in V_h : a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h$$

where

$$V_h = \left\{ v_h \in \mathcal{C}^0(\overline{\Omega}) : v_h|_{\mathcal{K}} \in \mathbb{P}^r(\mathcal{K}), r \geq 1, \forall\, \mathcal{K} \in \mathcal{T}_h, v_h = 0, \text{ on } \partial\Omega \right\}$$

is the space of finite element functions of degree r with basis $\{\varphi_i\}_{i=1}^{N_h}$.
The Galerkin finite element approximation can be rewritten as

$$\text{find } u_h \in V_h : a(u_h, \varphi_i) = F(\varphi_i) \quad \forall\, i = 1, \dots, N_h$$

Given a partition of the nodes such as

- $\left\{ x_j^{(1)}, 1 \leq j \leq N_1 \right\}$ nodes in subdomain Ω_1
- $\left\{ x_j^{(2)}, 1 \leq j \leq N_2 \right\}$ nodes in subdomain Ω_2
- $\left\{ x_j^{(\Gamma)}, 1 \leq j \leq N_\Gamma \right\}$ nodes in the interface Γ

and a division of the basis function

- $\varphi_j^{(1)}$ functions associated to the nodes $x_j^{(1)}$
- $\varphi_j^{(2)}$ functions associated to the nodes $x_j^{(2)}$
- $\varphi_j^{(\Gamma)}$ functions associated to the nodes $x_j^{(\Gamma)}$ on the interface

So the Galerkin formulation is equivalent to

$$\text{find } u_h \in V_h : \begin{cases} a(u_h, \varphi_i^{(1)}) = F(\varphi_i^{(1)}) & \forall\, i = 1, \dots, N_1 \\ a(u_h, \varphi_j^{(2)}) = F(\varphi_j^{(2)}) & \forall\, j = 1, \dots, N_2 \\ a(u_h, \varphi_k^{(\Gamma)}) = F(\varphi_k^{(\Gamma)}) & \forall\, k = 1, \dots, N_\Gamma \end{cases}$$

Now we can introduce the bilinear form on Ω_i:

$$a_i(v,w) = \int_{\Omega_i} \nabla v \cdot \nabla w \quad \forall\, v, w \in V, i = 1, 2,$$

the space

$$V_h^i = \left\{ v \in H^1(\Omega_i) : v = 0 \text{ on } \partial\Omega_i \setminus \Gamma \right\} (i = 1, 2),$$

$u_h^{(i)} = u_h|_{\partial\Omega_i}$ and rewrite the formulation as

$$\text{find } u_h^{(1)} \in V_h^1, u_h^{(2)} \in V_h^2 : \begin{cases} a_1(u_h^{(1)}, \varphi_i^{(1)}) = F_1(\varphi_i^{(1)}) & \forall\, i = 1, \dots, N_1 \\ a_2(u_h^{(2)}, \varphi_j^{(2)}) = F_2(\varphi_j^{(2)}) & \forall\, j = 1, \dots, N_2 \\ a_1(u_h^{(1)}, \varphi_k^{(\Gamma)}) + a_2(u_h^{(2)}, \varphi_k^{(\Gamma)}) & \\ \quad = F_1(\varphi_k^{(\Gamma)}|_{\Omega_1}) + F_2(\varphi_k^{(\Gamma)}|_{\Omega_2}) & \forall\, k = 1, \dots, N_\Gamma \end{cases}$$

By writing

$$u_h(x) = \sum_{j=1}^{N_1} u_h(x_j^{(1)})\varphi_j^{(1)}(x) + \sum_{j=1}^{N_2} u_h(x_j^{(2)})\varphi_j^{(2)}(x) + \sum_{j=1}^{N_\Gamma} u_h(x_j^{(\Gamma)})\varphi_j^{(\Gamma)}(x)$$

and substituting it inside the weak formulation and doing a bit[80] of algebra we obtain the algebraic form:

$$\begin{cases} A_{11}\mathbf{u}_1 + A_{1\Gamma}\boldsymbol{\lambda} = \mathbf{f}_1 \\ A_{22}\mathbf{u}_2 + A_{2\Gamma}\boldsymbol{\lambda} = \mathbf{f}_2 \\ A_{\Gamma 1}\mathbf{u}_1 + A_{\Gamma 2}\mathbf{u}_2 + \left(A_{\Gamma\Gamma}^{(1)} + A_{\Gamma\Gamma}^{(2)}\right)\boldsymbol{\lambda} = \mathbf{f}_1^{(\Gamma)} + \mathbf{f}_2^{(\Gamma)} \end{cases}$$

or

$$\begin{pmatrix} A_{11} & 0 & A_{1\Gamma} \\ 0 & A_{22} & A_{2\Gamma} \\ A_{\Gamma 1} & A_{\Gamma 2} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_\Gamma \end{pmatrix}$$

where $A_{\Gamma\Gamma} = \left(A_{\Gamma\Gamma}^{(1)} + A_{\Gamma\Gamma}^{(2)}\right)$ and $\mathbf{f}_\Gamma = \mathbf{f}_1^{(\Gamma)} + \mathbf{f}_2^{(\Gamma)}$. Consider now M subdomains Ω_i $i = 1, \dots, M$. Its finite element approximation is defined by the linear system

$$\begin{pmatrix} A_{ii} & A_{i\Gamma} \\ A_{\Gamma i} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_i \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_i \\ \mathbf{f}_\Gamma \end{pmatrix}$$

where $\mathbf{u}_i$ is the vector of unknowns in the subdomain and $_\lambda$ the vector of the unknowns on Γ: $\mathbf{u}_\Gamma = \boldsymbol{\lambda}$. The submatrix associated to the internal nodes is block diagonal

$$A_{ii} = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & A_{MM} \end{pmatrix}$$

while $A_{i\Gamma}$ is a banded matrix.

Remark 8.4
On each subdomain Ω_i the matrix

$$A_i = \begin{pmatrix} A_i i & A_{i\Gamma} \\ A_{\Gamma i} & A_{\Gamma\Gamma} \end{pmatrix}$$

represents the local stiffness matrix associated to a Neumann problem.

8.4 Algebraic form of Schwarz iterative methods

Let $A\mathbf{u} = \mathbf{f}$ be the system associated to the finite element approximation of (8.6) and still assume that Ω is decomposed in two subdomains Ω_1 and Ω_2. The total number of internal nodes N_h is made by the nodes in the domains n_1 and n_2 such that $N_h \leq n_1 + n_2$, and the stiffness matrix contain two submatrices A_1 and A_2 which corresponds to the bilinear form of the problem. Consider the multiplicative method: given u_2^0 on Γ_1, for $k \geq 1$:

$$\text{solve} \begin{cases} \mathcal{L}u_1^{(k)} = f & \text{in } \Omega_1 \\ u_1^{(k)} = u_2^{(k-1)} & \text{on } \Gamma_1 \\ u_1^{(k)} = 0 & \text{on } \partial\Omega_1 \setminus \Gamma_1 \end{cases}$$

$$\text{solve} \begin{cases} \mathcal{L}u_2^{(k)} = f & \text{in } \Omega_2 \\ u_2^{(k)} = u_1^{(k)} & \text{on } \Gamma_2 \\ u_2^{(k)} = 0 & \text{on } \partial\Omega_2 \setminus \Gamma_2 \end{cases}$$

[80] A lot

We introduce the discrete spaces:

$$
\begin{aligned}
V_h &= \left\{ v_h \in \mathcal{C}^0(\Omega) : v_h|_{\mathcal{K}} \in \mathbb{P}^r \ \forall\, \mathcal{K} \in \mathcal{T}_h, v_h|_{\partial\Omega} = 0 \right\} \\
V_{i,h} &= \left\{ v_h \in \mathcal{C}^0(\Omega_i) : v_h|_{\mathcal{K}} \in \mathbb{P}^r \ \forall\, \mathcal{K} \in \mathcal{T}_h^i, v_h|_{\partial\Omega} = 0 \right\} \\
\overset{\circ}{V}_{i,h} &= \left\{ v_h \in \mathcal{C}^0(\Omega_i) : v_h|_{\mathcal{K}} \in \mathbb{P}^r \ \forall\, \mathcal{K} \in \mathcal{T}_h^i, v_h|_{\partial\Omega_i} = 0 \right\}
\end{aligned}
$$

with $i = 1, 2$. The Galerkin formulation reads

$$
\text{find } u_{1,h}^{(k)} \in V_{1,h} : \begin{cases} a_1(u_{1,h}^{(k)}, v_{1,h}) = \mathcal{L}_1(v_{1,h}) & \forall\, v_{1,h} \in \overset{\circ}{V}_{1,h} \\ u_{1,h}^{(k)} = u_{2,h}^{(k-1)} & \text{on } \Gamma_1 \end{cases}
$$

$$
\text{find } u_{2,h}^{(k)} \in V_{2,h} : \begin{cases} a_2(u_{2,h}^{(k)}, v_{2,h}) = \mathcal{L}_2(v_{2,h}) & \forall\, v_{2,h} \in \overset{\circ}{V}_{2,h} \\ u_{2,h}^{(k)} = u_{1,h}^{(k)} & \text{on } \Gamma_2 \end{cases}
$$

Then we can use a lifting strategy for the boundary datum, setting $\tilde{w}_{i,h}^{(k)}$ = extension by zero to $\Omega \setminus \Omega_i$ of $w_{i,h}^{(k)}$

$$
\text{find } w_{1,h}^{(k)} \in \overset{\circ}{V}_{1,h} : \begin{cases} a_1(w_{1,h}^{(k)}, v_{1,h}) = \mathcal{L}_1(v_{1,h}) - a_1(u_h^{(k)}, v_{1,h}) & \forall\, v_{1,h} \in \overset{\circ}{V}_{1,h} \\ u_h^{(k+\frac{1}{2})} = u_h^{(k)} + \tilde{w}_{2,h}^{(k)} & \text{on } \Gamma_1 \end{cases}
$$

$$
\text{find } w_{2,h}^{(k)} \in \overset{\circ}{V}_{2,h} : \begin{cases} a_2(w_{2,h}^{(k)}, v_{2,h}) = \mathcal{L}_2(v_{2,h}) - a_2(u^{(k)})_h, v_{2,h} & \forall\, v_{2,h} \in \overset{\circ}{V}_{2,h} \\ u_h^{(k+1)} = u_h^{(k+\frac{1}{2})} + \tilde{w}_{2,h}^{(k)} & \text{on } \Gamma_2 \end{cases}
$$

Now, thanks to the following matrices

$$
R_1^T = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ & 0 & \end{bmatrix}, \quad R_2^T \begin{bmatrix} & 0 & \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}
$$

So that R_1^T, R_2^T are prolongation matrices and R_1, R_2 are restriction matrices. Obviously, $R^T R = I(N_h \times N_h)$ and $A_1 = R_1 A R_1^T, A_2 = R_2 A R_2^T$.
We can now write the algebraic formulation

$$
A_1 \mathbf{w}^{(k)} = R_1 \mathbf{f} - A_1 R_1 \mathbf{u}^{(k)}
$$

that is the same as

$$
A_1 R_1 \tilde{\mathbf{w}}_1^{(k)} = R_1 \mathbf{f} - A_1 R_1 \mathbf{u}^{(k)}
$$

which entails, using $A_1 = R_1 A R_1^T$:

$$
\begin{aligned}
A_1 R_1 \left(\mathbf{u}^{(k+\frac{1}{2})} - \mathbf{u}^{(k)} \right) &= R_1 \mathbf{f} - R_1 A \overbrace{R_1^T R_1}^{I(N_h \times N_h)} \mathbf{u}^{(k)} = R_1 \left(\mathbf{f} - A \mathbf{u}^{(k)} \right) \\
R_1 \left(\mathbf{u}^{(k+\frac{1}{2})} - \mathbf{u}^{(k)} \right) &= A_1^{-1} R_1 \left(\mathbf{f} - A \mathbf{u}^{(k)} \right)
\end{aligned}
$$

Left-multiplying by R_1^T and noting that $R_1^T R_1 = I(N_h \times N_h)$ yields:

$$
\mathbf{u}^{(k+\frac{1}{2})} = \mathbf{u}^{(k)} + R_1^T A_1^{-1} R_1 \left(\mathbf{f} - A \mathbf{u}^{(k)} \right)
$$

By working similarly on the second step, we get:

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k+\frac{1}{2})} R_2^T (R_2 A R_2^T)^{-1} R_2 \left(\mathbf{f} - A\mathbf{u}^{(k+\frac{1}{2})}\right)$$

In conclusion, we obtain:

$$\begin{aligned}
\mathbf{u}^{(k+\frac{1}{2})} &= \mathbf{u}^{(k)} R_1^T (R_1 A R_1^T)^{-1} R_1 \left(\mathbf{f} - A\mathbf{u}^{(k)}\right) \\
\mathbf{u}^{(k+1)} &= \mathbf{u}^{(k+\frac{1}{2})} R_2^T (R_2 A R_2^T)^{-1} R_2 \left(\mathbf{f} - A\mathbf{u}^{(k+\frac{1}{2})}\right)
\end{aligned}$$

Define, for $i = 1, 2$

$$Q_i = R_i (R_i A R_i^T)^{-1} R_i$$

We get

$$\begin{aligned}
\mathbf{u}^{(k+\frac{1}{2})} &= \mathbf{u}^{(k)} + Q_1 \left(\mathbf{f} - A\mathbf{u}^{(k)}\right) \\
\mathbf{u}^{(k+1)} &= \mathbf{u}^{(k+\frac{1}{2})} Q_2 \left(\mathbf{f} - A\mathbf{u}^{(k+\frac{1}{2})}\right)
\end{aligned}$$

and

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + Q_1 \left(\mathbf{f} - A\mathbf{u}^{(k)}\right) + Q_2 \left[\mathbf{f} - A\left(\mathbf{u}^{(k)} + Q_1 \left(\mathbf{f} - A\mathbf{u}^{(k)}\right)\right)\right]$$

or equivalently,

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + (Q_1 Q_2 - Q_2 A Q_1)(\mathbf{f} - A\mathbf{u}^{(k)})$$