# Numerical Analysis for Partial Differential Equations

Andrea Bonifacio

June 1, 2023

# 1 Boundary Value Problems

## 1.1 Weak Formulation

Let's consider a problem

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

- $\Omega$: open bounded domain in $\mathbb{R}^d$, with $d = 2, 3$
- $\partial\Omega$: boundary of $\Omega$
- $f$: given
- B.C. accordingly to $\mathcal{L}$
- $\mathcal{L}$: $2^{\text{nd}}$ order operator, like:
  (1) $\mathcal{L}u = -\text{div}(\mu\nabla u) + \mathbf{b} \cdot \nabla u + \sigma u$ (non-conservative form)
  (2) $\mathcal{L}u = -\text{div}(\mu\nabla u) + \text{div}(\mathbf{b}u) + \sigma u$ (conservative form)
  - $\mu \in L^\infty(\Omega), \quad \mu(\mathbf{x}) \geq \mu_0 > 0$ uniformly bounded from below
  - $\mathbf{b} \in (L^\infty(\Omega))^d$ transport term
  - $\sigma \in L^2(\Omega)$ reaction term
  - $f \in L^2(\Omega)$ can be less regular

### General elliptic problems

Consider

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

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 $\Omega$. Then, given a test function $v$, we multiply the equation by $v$, and integrate on the domain $\Omega$

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

By applying Green's formula

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

So the weak formulation of the problem is

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

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

**Theorem 1.1** (Lax-Milgram)
Assume that

- $V$ Hilbert space with $\|\cdot\|$ and inner product $(\cdot, \cdot)$
- $F \in V^* : |F(v)| \leq \|F\|_{V^*}\|v\| \;\forall\, v \in V$

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

Then, there exists a unique solution $u$ of 1.3

Moreover

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

where $\alpha$ is the coercivity costant. Hence

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

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

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

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

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

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

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

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

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

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

What could be our choice of $V$? Given that

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

and

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

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

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

.
Knowing that a Sobolev space

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

we can say $V_{\Gamma_D} = \left\{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\right\}$, 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 $H^1(), a(\cdot, \cdot)$ being a continuous and coercive bilinear forrm from $V \times V \to \mathbb{R}$, $F(\cdot)$ being a continuous linear functional from $V \to \mathbb{R}$.
Let $V_h \subset V$ be a family of spaces that depends on a parameter $h > 0$, such that $\dim V_h = N_h < \infty$. We can rewrite the weak formulation

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

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

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

Since $u_h \in V_h$

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

where $u_j$ are unknown coefficients. Then

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

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

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

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

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

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

and

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

The correspondence

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

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

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

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

**Existence and uniqueness**

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

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

### Stability

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

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

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

### Convergence

**Lemma 1.1** (Galerkin orthogonality)
The solution $u_h$ of the Galerkin method satisfies

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

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

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

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

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

from which the claim follows. ★

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

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

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

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

hence

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

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

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

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

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

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

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

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

## 1.3 Finite Element Method

### Partitions

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

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

and set $h = \max_j h_j$.

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

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

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

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

The condition (1.13) excludes very deformed triangles.

**Definition 1.1** (Seminorms)
A seminorm is defined as

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

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:

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

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

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

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

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

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

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

### The space $X_h^2$

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

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

### The space $V_h$

This space is generated by

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

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

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

A basis is called Lagrangian if it satisfies the following properties

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

and then the following property holds:

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

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

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

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

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

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

## 1.4 Advection Diffusion Reaction Problem

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

Lax-Milgram tells us that if $\sigma - \frac{1}{2}\text{div}\mathbf{b} \geq \gamma > 0$ 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):
$$\exists\, \alpha, \lambda : a(v,v) \geq \alpha\|v\|^2 - \|v\|^2_{L^2(\Omega)} \quad \forall\, v \in V$$
- 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: $A_{ij} = a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j) = A_{ji}$
$\forall\, \mathbf{v} \in \mathbb{R}^{N_h}$:

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

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

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

Since $A$ is positive defined $\Rightarrow \text{Re}(\lambda_k(A)) \Rightarrow \lambda_k(A) \neq 0$. 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 $A$:

$$\alpha C_1 h^d \leq \lambda_h \leq MC_2h^{d-2} \qquad d = 1,2,3$$

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.

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

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

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

Recall also that

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

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 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 $(f, g) = \int_{-1}^{1} f g\, dx$ 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

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

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

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

### 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 a $L_N(x)$ 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 $\{\varphi_0, \dots, \varphi_N\}$ characteristics 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 estabilished 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

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

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

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

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

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

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

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

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

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

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

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

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

#### General ideas

The idea proposed until now are the following:

| | | | |
|---|---|---|---|
| (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

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

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

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

### 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

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

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

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

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 of all, 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

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

that has the usual weak formulation

$$\text{find } u \in V = H^1_0(-1,1) : a(u,v) = F(v), \forall\, v \in V$$

The GNI formulation follows

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

Note that, thanks to the exactness of LGL quadrature formula

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

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 $\Omega$. 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 $j = 0, \ldots, p$ 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 $\exists\, c > 0 : \forall\, f \in H^r(\Omega), r \geq 1, p \geq 1 : \forall\, v_\delta \in X_\delta$:

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

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

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

## 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

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

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

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

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

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

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

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

So $u_\delta^{\text{GNI}}$ 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 soltion 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 know 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

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

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

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

This way we obtain this formula

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

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))

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

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 $3^{rd}$ term because $[\![u'(x_k)]\!] = 0$
- add symmetrization term ($= 0$ if $u$ is the exact solution)

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

  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

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

#### 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-homogeneus 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 homogeneus 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 $\alpha$ and $\beta$ 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 $f$, $\alpha$ and $\beta$.
Note that in (3.11), if we define $[\![u_h(x_0)]\!]$ and $[\![u_h(x_N)]\!]$ as $[\![v_h(x_0)]\!]$ and $[\![v_h(x_N)]\!]$

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

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

$$+\sum_{k=1}^{N-1} \gamma [\![u_h(x_k)]\!] [\![v_h(x_k)]\!] + \gamma u_h(x_0^+)v_h(x_0^+) + \gamma u_h(x_N^-)v_h(x_N^-)$$
$$+\sum_{k=0}^{N} \gamma [\![u_h(x_k)]\!] [\![v_h(x_k)]\!]$$

## 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 $F \in \mathcal{F}_h^I$ 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:

$$\sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\partial\mathcal{K}} \boldsymbol{\tau} \cdot \mathbf{n}_{\mathcal{K}} v = \sum_{F\in\mathcal{F}_h} \int_F \{\!\{\boldsymbol{\tau}\}\!\} \cdot [\![v]\!] + \sum_{F\in\mathcal{F}_h^I} \int_F [\![\boldsymbol{\tau}]\!] \{\!\{v\}\!\} \tag{3.15}$$

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 fullfilled 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 $u\in H^2(\Omega)\cap H_0^1(\Omega)$) 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 $\nabla_h$ is the elementwise 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 $\gamma$ is a stabilization function.
Now we can define the DG discrete space

$$V_h^p = \left\{v_h\in L^2(\Omega) : v_h|_{\mathcal{K}}\in\mathcal{P}^{p_{\mathcal{K}}}(\mathcal{K})\ \forall\ \mathcal{K}\in\mathcal{T}_h\right\}\not\subseteq H_0^1(\Omega)$$

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 homogeneus Dirichlet BC, but in the case of non-homogeneus 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 $\gamma$**

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

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 elementwise 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 penaly parameter $\alpha$ should be large enough.

- Strong-consistency (Galerkin orthogonality):

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

- 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 $\|u - u_h\|_{DG} \lesssim |||u - \prod_h^p|||_{DG}$.
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 $\Omega$ 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

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

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

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

$$a(u,v) = \begin{cases} \displaystyle\int_\Omega (\mu\nabla u + \mathbf{b}u)\cdot\nabla v + \int_\Omega \sigma uv & \text{conservative form}\\[2ex] \displaystyle\int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega \mathbf{b}\cdot\nabla uv + \int_\Omega \sigma uv & \text{non-conservative form}\end{cases}$$

Let's verify the uniqueness of the solution:

**Coercivity**

Sufficient conditions for coercivity:

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

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 tradeoff 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

- $\varphi(v_h)=\mathcal{L}v_h\rightarrow$ GLS - Galerkin Least Squares method
- $\varphi(v_h)=\mathcal{L}_{ss}v_h\rightarrow$ 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

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

Indeed we can see

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

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

**Remark 4.1**
If $\mathrm{div}\mathbf{b} = 0$, 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

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

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

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

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

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

In the end

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

★

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

$$\sum_{\mathcal{K}\in\mathcal{T}_h} h^2_\mathcal{K}\int_\mathcal{K}|\Delta v_h|^2\,d\mathcal{K} \le C_0\|\nabla v_h\|^2_{L^2(\Omega)} \quad \forall\, v_h \in X^r_h \tag{4.2}$$

**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.

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

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 $0 < \delta \le 2C_0^{-1}$.
Then, as long as $u \in H^{r+1}(\Omega)$, the following super-optimal estimate holds:

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

***Proof.*** First of all, 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, $V_h = X^r_h \cap H^1_0(\Omega)$, we can choose $\hat{u}_h = \prod^r_h u$ 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:

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

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

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

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

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

So we can rewrite everything bounded as

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

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

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

It's time to estimate $\mathcal{E}(\eta)$, by bounding each of it's 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 $\gamma_1$, the reaction constant $\sigma$ or the norm $\|\mathbf{b}\|_{L^\infty(\Omega)}$, the stabilization parameter $\delta$. Then we have

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

For the fouth term we have

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

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

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

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

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

that is

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

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

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

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:

- $\Omega$ 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 $u_0, \varphi$ and $\psi$ are given funcion and $\{\Gamma_D, \Gamma_N\}$ provides a boundary partition that is $\Gamma_D \cup \Gamma_N = \partial\Omega, {\Gamma_D}^\circ \cap {\Gamma_N}^\circ = \varnothing$. For obvious reasons $\Gamma_D$ is the Dirichlet boundary, while $\Gamma_N$ 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 $\nu$ 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 ($u = e^{\lambda t}$)

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

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

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

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)

- $\theta = 0$ wThe 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 $\Delta t$.

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 $\Delta t$ 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 from $K\mathbf{u}^{k+1} = \mathbf{g}$, 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

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