Numerical Methods of Partial Differential Equations

From lessons by Alfio Quarteroni
Luigi Pagani

Polytechnic University of Milan
A.Y. 2023/2024

Disclaimer.
This document contains the lecture notes for the course on numerical methods for partial differential equations, taught by Professor Alfio Quarteroni at the Polytechnic University of Milan during the 2023/24 academic year, with minor additions and modifications. The intellectual property remains with the aforementioned professor, who has not reviewed this document. It is intended solely as a supplementary resource for the lectures, created by students for students, without any claims to replace official textbooks or attendance in the lectures. These notes are taken from the lectures and most of the illustrations are based on the professor's drawings and slides, to whom the intellectual property also belongs.
The widely cited references in these notes, in addition to the slides, are the following ones:

- Quarteroni, A.(2016). *Numerical Models for Differential Problems*. Springer.
- Quarteroni, A., & Valli, A. (1994). *Numerical Approximation of Partial Differential Equations*. Springer.
- Quarteroni, A., & Valli, A. (1999). *Domain Decomposition Methods for Partial Differential Equations*. Numerical Mathematics and Scientific Computation. Oxford University Press.
- Parolini, N. Computational Fluid Dynamics lecture notes.

Revised on June 22, 2024

The most up-to-date version of these notes is always available at this link.

Contents

1 Elliptic Equations **5**
- 1.1 Boundary value problems . . . 5
- 1.2 Weak formulation . . . 6
- 1.3 Theorem (Lax-Milgram) . . . 6
- 1.4 Problem 1.3 - Generalization of problem 1.2 . . . 7
- 1.5 Nečas Theorem . . . 7
- 1.6 Approximation . . . 8
- 1.7 Lax-Milgram hypothesis for the weak formulation of problem 1.2 . . . 11
- 1.8 Poincaré Inequality . . . 12
- 1.9 Stiffness Matrix . . . 13
- 1.10 Interpolation Error Estimate . . . 14
- 1.11 Finite Element Error Estimate . . . 14
- 1.12 A more general FE error estimate . . . 15

2 Parabolic equations **17**
- 2.1 Bilinear Forms . . . 18
- 2.2 Condition on the well-posedness of problem 2.5 . . . 19
- 2.3 Galerkin Approximation . . . 22
- 2.4 Algebraic Formulation . . . 22
- 2.5 Time Discretization . . . 22
- 2.6 A priori estimates . . . 24
- 2.7 Convergence analysis of the θ-method . . . 31
- 2.8 Parabolic ADR equation . . . 32
- 2.9 A semimplicit scheme . . . 32
- 2.10 Stability analysis of the semimplicit scheme . . . 32

3 Domain Decomposition Methods **35**
- 3.1 Classical Iterative DD Methods . . . 36
- 3.2 The Steklov-Poincaré interface Equation . . . 39
- 3.3 FEM: Multi-Domain Formulation . . . 42
- 3.4 The Schur Complement System . . . 45
- 3.5 Nonoverlapping Multiple Subdomains . . . 47

4 Navier Stokes **55**
- 4.1 Well-posedness . . . 56
- 4.2 Alternative formulations . . . 56
- 4.3 Weak formulation of Navier-Stokes equations . . . 57

4.4 Solution Uniqueness 59
4.5 The Reynolds number 60
4.6 Stokes equations and their approximation 61
4.7 Galerkin Approximation 62
4.8 Existence and Uniqueness 63
4.9 Algebraic formulation of the Stokes problem 64
4.10 Compatible couples of spaces 66
4.11 Time discretization of Navier-Stokes equations 68
4.12 Finite difference methods 69
4.13 Time dependent Generalized Stokes problem 71
4.14 Generalization of the Stokes problem to N-S 72

5 The ADR Boundary Value Problem **75**
5.1 Complete Case 75

6 Stokes Equation **79**

Chapter 1

Elliptic Equations

1.1 Boundary value problems

We are considering a second order differential equation of the form:

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

Where:

- Ω represents an open bounded domain in $\mathbb{R}^d$, with $d = 2, 3$.
- $\partial\Omega$ is the boundary of Ω.
- f is a given function.
- The Boundary Conditions (B.C.) are to be prescribed according to $\mathcal{L}$.
- $\mathcal{L}$ is a 2nd order differential operator.

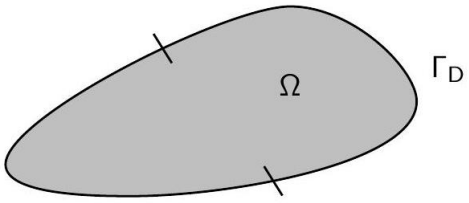
Examples of $\mathcal{L}$ include:

- Non-conservative form: $\mathcal{L}u = -\operatorname{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u$.
- Conservative form: $\mathcal{L}u = -\operatorname{div}(\mu\nabla u) + \operatorname{div}(\mathbf{b}u) + \sigma u$.

Example

$$\begin{cases} \mathcal{L}u = -\operatorname{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_{\mathrm{D}} \\ \mu\nabla u\cdot\mathbf{n} = g & \text{on } \Gamma_{\mathrm{N}} \end{cases} \tag{1.2}$$

$$g \in L^2(\Gamma_{\mathrm{N}}), \quad \partial\Omega = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}, \quad \Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$$



Γ_N

1.2 Weak formulation

Weak form of Eq. (1.2):
Idea: v = suitable test function $\rightarrow$ multiply (1.2) by $v : (\mathcal{L}u)v = fv$

$\rightarrow$ integrate in $\Omega \rightarrow$ apply integration by parts, product rule of divergence and divergence theorem to obtain:

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

whence:

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

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

1.3 Theorem (Lax-Milgram)

Assume that:

1. V is a Hilbert space with norm $\| \cdot \|$ and inner product $(\cdot, \cdot)$.
2. F is bounded, i.e., $F \in V'$ such that $|F(v)| \leq \|F\|_{V'} \|v\| \quad \forall v \in V$.
3. a is continuous, i.e., $\exists M > 0 : |a(u,v)| \leq M \|u\| \|v\| \quad \forall u, v \in V$.
4. a is coercive, i.e., $\exists \alpha > 0 : a(v,v) \geq \alpha \|v\|^2 \quad \forall v \in V$.

Then, there exists a unique solution u of (1.2).

Remark: V' is the dual space of V, which consists of linear and bounded (i.e., continuous) maps from V to $\mathbb{R}$ with norm:

$$\|F\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|F(v)|}{\|v\|_V}$$

Moreover, we have:

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

Hence, we can conclude that:

$$\|u\| \leq \frac{\|F\|_{V'}}{\alpha}$$

This leads to stability or continuous dependence on data.

1.4 Problem 1.3 - Generalization of problem 1.2

To formulate Nečas Theorem, consider a slightly more general problem than (1.2):

$$\begin{cases} \text{Find } u \in V \\ a(u,w) = \langle F, w\rangle \quad \forall w \in W \end{cases} \tag{1.3}$$

Note: $\langle F, w\rangle$ is a different (equivalent) notation for $F(w)$. $a : V \times W \to \mathbb{R}$ bilinear form, $F : W \to \mathbb{R}$ linear and continuous (that is $F \in W'$)

1.5 Nečas Theorem

Assume that $F \in W'$. Consider the following conditions.

i) a continuous: $\exists M > 0 : |a(u,w)| \leq M\|u\|_V\|w\|_w \ \forall u \in V, w \in W$.

ii) inf-sup condition: $\exists \alpha > 0 : \forall v \in V \sup_{w\in W\setminus\{0\}} \frac{a(v,w)}{\|w\|_W} \geq \alpha\|v\|_V$.

iii) $\forall w \in W,\ w \neq 0, \exists v \in V : a(v,w) \neq 0$.

These three conditions are necessary and sufficient for the existence and uniqueness of a solution of (1.3), for any $F \in W'$. Moreover (continuous dependence on data):

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{w'}$$

Remark

If $W = V$, then the previous theorem provides necessary and sufficient conditions for the existence and uniqueness of solutions of the weak formulation of the Example problem(1.2).
Note that in this case:

- Condition i) is equivalent to continuity.
- Condition ii) yields:

$$\exists \alpha > 0 : \forall v \in V \sup_{w\in V\setminus\{0\}} \frac{a(v,w)}{\|w\|_V} \geq \alpha\|v\|_V$$

- Condition iii) yields:

$$\forall w \in V, w \neq 0, \exists v \in V : a(v,w) \neq 0$$

Conditions ii) and iii are more general and weaker conditions than coercivity. Indeed, by taking $w = v$, Condition iv (coercitivity) of Lax Milgram implies Equations ii) and iii) here.

1.6 Approximation

Galerkin method (for problem (1.2))

Find $u_h \in V_h$ such that

$$a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h$$

This is the Petrov-Galerkin method for problem (3). We want to find $u_h \in V_h$ such that

$$a(u_h, w_h) = F(w_h) \quad \forall w_h \in W_h$$

where:

- $\{V_h, h > 0\}$ are finite dimensional subspaces of V,
- $\{W_h, h > 0\}$ are finite dimensional subspaces of W,
- $\dim V_h = \dim W_h = N_h < +\infty$.

Analysis of the Galerkin Problem

Existence and Uniqueness

This is a corollary of the Lax-Milgram Lemma, with V_h being a closed subspace of V.

Stability

We have a uniform bound with respect to h:

$$\|u_h\| \leq \frac{\|F\|_{V'}}{\alpha}$$

Consistency = Galerkin Orthogonality

This is equivalent to Galerkin orthogonality. By subtracting $a(u_h, v_h) = F(v_h)$ from $a(u, v_h) = F(v_h)$, we obtain:

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

The equation $a(u - u_h, v_h) = 0$ means that the error $u - u_h$ is orthogonal to the subspace V_h with respect to the bilinear form $a(\cdot, \cdot)$.

Convergence (Céa Lemma)

We have:

$$\begin{aligned}
\alpha \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) && (1.4)\\
&= a(u - u_h, u - v_h) + \underbrace{a(u - u_h, v_h - u_h)}_{=0, \text{ since } (v_h - u_h) \in V_h} && (1.5)\\
&\leq M \|u - u_h\| \|u - v_h\| \quad \forall v_h \in V_h && (1.6)
\end{aligned}$$

Hence:

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

Finally:

$$\|u - u_h\| \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|$$

Rate of Convergence

We start with the assumption of space saturation:

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

Then, we have convergence:

$$\lim_{h \to 0} \|u - u_h\| = 0$$

We consider the rate of convergence (for example, in Finite Elements). Let $\mathcal{T}_h = \bigcup K$ be a triangulation of Ω (actually of Ω_h). We define v_h as follows:

$$v_h = \left\{ v_h \in \mathcal{C}^0(\bar{\Omega}) : v_h|_K \in \mathbb{P}^r(K) \;\; \forall K \in \mathcal{T}_h, \, v_h|_{\Gamma_D} = 0 \right\}, \quad \text{for } r \geq 1$$

We have:

$$\inf_{v_h \in V_h} \|u - v_h\| \leq \|u - \bar{u}_h\|$$

where $\bar{u}_h$ is a suitable choice. For example, $\bar{u}_h = \Pi_h^r u$ is the Finite Element interpolant. Then, we have:

$$\|u - \bar{u}_h\| \leq C h^r |u|_{H^{r+1}(\Omega)}$$

provided $u \in V \cap H^{r+1}(\Omega)$.

We define v_h as a function that is continuous over the closure of the domain Ω, denoted as $\mathcal{C}^0(\bar{\Omega})$. For each element K in the triangulation $\mathcal{T}_h$, v_h takes the form of a polynomial of degree r, expressed as $v_h|_K \in \mathbb{P}^r(K)$ for all $K \in \mathcal{T}_h$. Here, r is an integer greater than or equal to 1, where a higher r provides a more refined approximation. Additionally, v_h satisfies the Dirichlet boundary conditions, as indicated by $v_h|_{\Gamma_D} = 0$.

Regarding error estimation, we consider the inequality $\inf_{v_h \in V_h} \|u - v_h\| \leq \|u - \bar{u}_h\|$. This compares the error between the true solution u and any function in the finite element space V_h with the error between u and a particular approximation $\bar{u}_h$, which is often chosen as the finite element interpolant $\Pi_h^r u$. The selection of $\bar{u}_h$ is critical for determining the error reduction rate.

Finally, the rate of convergence is described by $\|u - \bar{u}_h\| \leq C h^r |u|_{H^{r+1}(\Omega)}$. This inequality shows that the error is proportional to h^r, where h relates to the mesh size in the triangulation, and C is a constant. The condition $u \in V \cap H^{r+1}(\Omega)$ is essential, as it requires u to have sufficient smoothness (belonging to the Sobolev space $H^{r+1}(\Omega)$) for this error estimate to be valid.

Examples of Polynomial Spaces

- For 1D ($\mathbb{P}^r$), we have:

$$p(x) = \sum_{k=0}^{r} a_k x^k$$

- For 2D ($\mathbb{P}^r$), we have:

$$p\left(x_1, x_2\right) = \sum_{\substack{k=0,\ldots,r \\ m=0,\ldots,r \\ k+m\leq r}} a_{km} x_1^k x_2^m$$

- For 3D tetrahedra ($\mathbb{P}^r$), we have:

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

- For 3D hexahedra ($\mathbb{Q}^r$), we have:

$$p\left(x_1, x_2, x_3\right) = \sum_{\substack{k=0,\ldots,r \\ m=0,\ldots,r \\ n=0,\ldots,r}} a_{kmn} x_1^k x_2^m x_3^n$$

Basis functions

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

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

We will use the notation $\mathbf{v} = \left(v_1, \ldots, v_{N_h}\right)^T$ to denote a vector $\mathbf{v} \in \mathbb{R}^{N_h}$ collecting all the basis coefficients (also called degrees of freedom). A basis is called lagrangian if it satisfies the following property:

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

for a suitable collection of points $\{\mathbf{x}_j\}_{j=1}^{N_h}$ called nodes. When the basis is lagrangian, the following property holds:

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

We will now prove the conditions of Theorem 2 (Lax-Milgram).

1.7 Lax-Milgram hypothesis for the weak formulation of problem 1.2

Condition I

We have a Hilbert space, as it is a closed subspace of the Hilbert space $H^1(\Omega)$. The scalar product is defined as $(u,v) = \int_\Omega uv + \int_\Omega \nabla u \cdot \nabla v$ and the norm is defined as $\|v\| = \left(\int_\Omega v^2 + \int_\Omega |\nabla v|^2\right)^{1/2}$.

Condition II
We now check the condition $F \in V'$:

$$\begin{aligned}
|\langle F, v\rangle| &= \left|\int_\Omega fv + \int_{\Gamma_N} gv\right| \\
&\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_N)}\|v\|_{L^2(\Gamma_N)} \quad \text{(Cauchy-Schwarz inequality)} \\
&\leq \|f\|_{L^2(\Omega)}\|v\|_{H^1} + \|g\|_{L^2(\Gamma_N)} C_{\text{trace}}\, \|v\|_{H^1} \quad \text{(definition of } \|\cdot\| \text{ and trace ineq.)} \\
&= \left(\|f\|_{L^2(\Omega)} + C_{\text{trace}}\, \|g\|_{L^2(\Gamma_N)}\right) \|v\|_{H^1}
\end{aligned}$$

Therefore, we have $\|F\|_{V'} = \sup_{v\neq 0} \frac{|\langle F,v\rangle|}{\|v\|} \leq \|f\|_{L^2(\Omega)} + C_{\text{trace}}\, \|g\|_{L^2(\Gamma_N)} < +\infty$.

Condition III

We have:

$$\begin{aligned}
|a(u,v)| &\leq \|\mu\|_{L^\infty}\|\nabla u\|_{L^2}\|\nabla v\|_{L^2} + \|\mathbf{b}\|_{L^\infty}\|\nabla u\|_{L^2}\|v\|_{L^2} \\
&\quad + \|\sigma\|_{L^2}\|u\|_{L^4}\|v\|_{L^4} \\
&\leq \underbrace{\left(\|\mu\|_{L^\infty} + \|\mathbf{b}\|_{L^\infty} + \|\sigma\|_{L^2}\right)}_{M} \|u\|\|v\| \quad \forall u, v \in V
\end{aligned}$$

Note: By Sobolev embedding, we have $\|v\|_{L^4} \leq \|v\| \quad \forall v \in H^1$.

Estimation of the Term $\|\sigma\|_{L^2}\|u\|_{L^4}\|v\|_{L^4}$:

- *Hölder's Inequality*: Apply Hölder's inequality for $\sigma \in L^2(\Omega)$ and $uv \in L^2(\Omega)$, obtaining $\|\sigma uv\|_{L^1} \leq \|\sigma\|_{L^2}\|uv\|_{L^2}$.

- *Sobolev Embedding*: Use Sobolev embedding to assert $u, v \in H^1(\Omega) \Rightarrow u, v \in L^4(\Omega)$.

- *Algebraic Closure in L^p Spaces*: Leverage the closure property of L^p spaces under multiplication for $u, v \in L^4(\Omega)$, yielding $\|uv\|_{L^2} \leq \|u\|_{L^4}\|v\|_{L^4}$.

- *Final Estimation*: Combine the above to estimate $\|\sigma\|_{L^2}\|u\|_{L^4}\|v\|_{L^4}$ as part of the overall bound for $|a(u,v)|$.

Condition IV

This condition holds under the assumptions:

$$\text{Given: } \sigma - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0 \text{ in } \Omega, \quad \mathbf{b}\cdot\mathbf{n} \geq 0 \text{ on } \Gamma_{\mathrm{N}},$$

$$a(v,v) = \int_\Omega \mu|\nabla v|^2 + \underbrace{\int_\Omega \mathbf{b}\cdot\frac{1}{2}\nabla(v^2)}_{\text{Target Expression}} + \int_\Omega \sigma v^2,$$

$$\text{Target Expression}: \int_\Omega \mathbf{b}\cdot\frac{1}{2}\nabla(v^2) = \quad = \int_\Omega \mathbf{b}\cdot v\nabla v \quad \text{(product rule)}$$

Applying the Divergence Theorem:

$$\mathbf{F} = \frac{1}{2}\mathbf{b}(v^2) \text{ and } \nabla\cdot\mathbf{F} = \frac{1}{2}\nabla\cdot(\mathbf{b}(v^2)) = \frac{1}{2}v^2\nabla\cdot\mathbf{b} + \frac{1}{2}\mathbf{b}\cdot\nabla(v^2),$$

$$a(v,v) = \int_\Omega \mu|\nabla v|^2 + \int_\Omega -\frac{1}{2}\nabla\cdot\mathbf{b}v^2 + \frac{1}{2}\int_{\partial\Omega}\mathbf{b}\cdot\mathbf{n}v^2 + \int_\Omega \sigma v^2$$

$$= \int_\Omega \mu|\nabla v|^2 + \int_\Omega \left(\sigma - \frac{1}{2}\operatorname{div}\mathbf{b}\right)v^2 + \frac{1}{2}\int_{\Gamma_{\mathrm{N}}}\mathbf{b}\cdot\mathbf{n}v^2 \geq \mu_0\|\nabla v\|_{L^2}^2,$$

If $\mathbf{b}$ is constant, $\operatorname{div}\mathbf{b} = 0$, then $\sigma \geq 0$.

μ_0 is a positive lower bound for the coefficient μ throughout the domain Ω. This means that for all points in the domain, $\mu(x) \geq \mu_0 > 0$.

Note: if $\mathbf{b}$ is constant, then $\operatorname{div}\mathbf{b} = 0$, and the first term $\sigma - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0$ reduces to $\sigma \geq 0$.

1.8 Poincaré Inequality

The Poincaré inequality is a fundamental result in the theory of partial differential equations and the calculus of variations. It provides a relationship between the norm of a function and the norm of its derivatives.

If Γ_D is a set of positive measure (in 1D, it is sufficient that Γ_D contains a single point), then there exists a constant $C_{\mathrm{P}} > 0$ such that:

$$\|v\|_{L^2(\Omega)} \leq C_{\mathrm{P}}\|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H^1_{\Gamma_{\mathrm{D}}}(\Omega)$$

From this, we have:

$$\|v\|_V^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \leq \left(1 + C_{\mathrm{P}}^2\right)\|\nabla v\|_{L^2(\Omega)}^2$$

And hence:

$$\|\nabla v\|_{L^2(\Omega)}^2 \geq \left(1 + C_{\mathrm{P}}^2\right)^{-1}\|v\|^2$$

In conclusion (coercivity), we have:

$$a(v,v) \geq \frac{\mu_0}{1 + C_{\mathrm{P}}^2}\|v\|^2.$$

1.9 Stiffness Matrix

As a reminder, if A is symmetric positive definite (spd), then the condition number $K_2(A)$ is given by the ratio of the maximum to the minimum eigenvalues of A:

$$K_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

Proposition If the bilinear form $a(\cdot,\cdot)$ is symmetric and coercive, then the matrix A is symmetric positive definite.

Proof The symmetry of A is given by $A_{ij} = a(\phi_j, \phi_i) = a(\phi_i, \phi_j) = A_{ji}$.

For any vector $\mathbf{v} \in \mathbb{R}^{N_h}$, we have:

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

Hence, A is positive definite.

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

$$\|\mathbf{v}\|_A = \sqrt{(A\mathbf{v}, \mathbf{v})} = \sqrt{\sum_{i,j} a_{ij} v_i v_j}$$

Remark:
In the case the bilinear form a is symmetric, we have a stronger stability result. Since a is symmetric, the finite element solution can be viewed as the element belonging to V_h minimizing the distance to the exact solution u in the energy norm. Thus:

$$\alpha\|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) \leq a(u - v_h, u - v_h) \leq M\|u - v_h\|_V^2, \quad \forall v_h \in V_h$$

$$\implies \|u - u_h\|_V \leq \sqrt{\frac{M}{\alpha}} \inf_{v_h \in V_h} \|u - vh\|_V.$$

Conditioning of the Stiffness Matrix

We can prove that there exist constants $C_1, C_2 > 0$ such that for all eigenvalues λ_h of A:

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

From this, it follows that:

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

This is indeed an asymptotic estimate (also a lower bound, with a different constant). Then:

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

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

$$\left\|\mathbf{u}^{(k)} - \mathbf{u}\right\|_A \leq 2\left(\frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1}\right)^k \left\|\mathbf{u}^{(0)} - \mathbf{u}\right\|_A$$

The same holds for the gradient method, with $\sqrt{K_2(A)}$ replaced by $K_2(A)$. This leads to the need for preconditioners.

1.10 Interpolation Error Estimate

1. **Estimate Formula**:

$$|v - \Pi_h^r v|_{H^m(\Omega)} \leq C\left(\sum_{K\in\mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{H^{r+1}(K)}\right)^{\frac{1}{2}} \tag{1.7}$$

   This formula provides an upper bound on the error between the actual function v and its interpolant. The error is measured in the H^m norm.

2. **Interpretation**:

   - The error is influenced by the size of each mesh element (h_K) and the degree of the polynomial (r) used for interpolation.
   - The constant C is dependent on the polynomial degree r, the semi-norm m, and possibly other characteristics of the mesh (denoted as $\hat{k}$).

3. **Simplified Estimate**:

$$|v - \Pi_h^r v|_{H^m(\Omega)} \leq C h^{r+1-m} |v|_{H^{r+1}(\Omega)} \tag{1.8}$$

   In this simplified version, the estimate assumes $h_K \leq h$ for all elements.

4. **Special Case, m = 0**

   When $m = 0$, $H^0(\Omega) = L^2(\Omega)$, and the norm becomes the standard L^2 norm, which is the square root of the integral of the square of the function over the domain.

1.11 Finite Element Error Estimate

Recall that:

$$\|u - u_h\| = \|u - u_h\|_{H^1(\Omega)} \leq \frac{M}{\alpha} \inf_{v_h\in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \frac{M}{\alpha} \|u - \Pi_h^r u\|_{H^1(\Omega)} \tag{1.9}$$

Using (1.7):

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

Using (1.8):

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

1.12 A more general FE error estimate

Suppose $u \in H^{p+1}(\Omega)$, with $p \geq 0$.
Convergence rate for $\|u - u_h\|_v$:

	$p=0$	$p=1$	$p=2$	$p=3$	$p>3$
$r=1$	conv.	$\mathcal{O}(h)$	$\mathcal{O}(h)$	$\mathcal{O}(h)$	$\mathcal{O}(h)$
$r=2$	conv.	$\mathcal{O}(h)$	$\mathcal{O}\left(h^2\right)$	$\mathcal{O}\left(h^2\right)$	$\mathcal{O}\left(h^2\right)$
$r=3$	conv.	$\mathcal{O}(h)$	$\mathcal{O}\left(h^2\right)$	$\mathcal{O}(h)^3$	$\mathcal{O}\left(h^3\right)$

- **Optimal rate of convergence** for $p = r$. If the polynomial degree p of the function u matches the polynomial degree r of the finite element space, the convergence rate is optimal.
- **Waste of computational power** for $p < r$. If the polynomial degree p of the function u is less than the polynomial degree r of the finite element space, the convergence rate is suboptimal. This is because the finite element space has more degrees of freedom than necessary to accurately approximate u.
- **Suboptimal convergence rate** for $p > r$. If the polynomial degree p of the function u is greater than the polynomial degree r of the finite element space, the convergence rate is still optimal for the given finite element space (since the extra regularity of u does not penalize the convergence rate), but the finite element method does not fully exploit the higher regularity of u. This means that while the error still decreases at the optimal rate for r, it could potentially decrease faster if a higher degree finite element space were used. However, it is not a "waste of computational effort" because the computational resources are still effectively used to achieve the best possible accuracy for the chosen r.

Remark

For $\|u - u_h\|_{L^2}$, the convergence rates from the table should be increased by one order. This is due to the less stringent nature of the L^2 norm compared to other norms, such as the H^1 norm. The L^2 norm measures the error in an average sense over the domain and does not consider the derivatives of the function, making it more forgiving and resulting in higher apparent convergence rates.

Chapter 2

Parabolic equations

Parabolic Equations

We consider parabolic equations of the form

$$\frac{\partial u}{\partial t} + Lu = f, \quad \mathbf{x} \in \Omega, t > 0 \tag{2.1}$$

where:

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

When solved only for a bounded temporal interval, say for $0 < t < T$, the region $Q_T = \Omega \times (0, T)$ is called cylinder in the space $\mathbb{R}^d \times \mathbb{R}^+$.

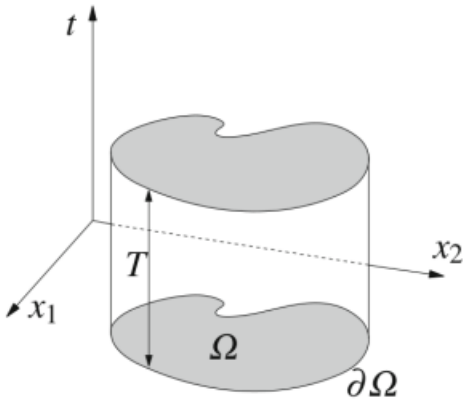


Figure 2.1: The cylinder $Q_T = \Omega \times (0, T), \Omega \subset \mathbb{R}^2$

In the case where $T = +\infty, Q = \{(\mathbf{x}, t) : \mathbf{x} \in \Omega, t > 0\}$ will be an infinite cylinder. Equation (2.1) must be completed by assigning an initial condition

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \tag{2.2}$$

together with boundary conditions, which can take the following form:

$$\begin{cases} u(\mathbf{x},t) = \varphi(\mathbf{x},t), & \mathbf{x} \in \Gamma_D \text{ and } t > 0 \\ \dfrac{\partial u(\mathbf{x},t)}{\partial n} = \psi(\mathbf{x},t), & \mathbf{x} \in \Gamma_N \text{ and } t > 0 \end{cases} \tag{2.3}$$

where u_0, φ and ψ are given functions and $\{\Gamma_D, \Gamma_N\}$ provides a boundary partition, that is $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\overset{\circ}{\Gamma}_D \cap \overset{\circ}{\Gamma}_N = \emptyset$. For obvious reasons, Γ_D is called Dirichlet boundary and Γ_N Neumann boundary. In the one-dimensional case, the problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = f, & 0 < x < d, \quad t > 0 \\ u(x,0) = u_0(x), & 0 < x < d \\ u(0,t) = u(d,t) = 0, & t > 0 \end{cases} \tag{2.4}$$

describes the evolution of the temperature $u(x,t)$ at point x and time t of a metal bar of length d occupying the interval $[0,d]$, whose thermal conductivity is ν and whose endpoints are kept at a constant temperature of zero degrees.

The function u_0 describes the initial temperature, while f represents the heat generated (per unit length) by the bar.

For this reason, (2.4) is called heat equation.

Weak Formulation and Its Approximation

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

$$\int_\Omega \frac{\partial u(t)}{\partial t} v d\Omega + a(u(t), v) = \int_\Omega f(t) v d\Omega \quad \forall v \in V \tag{2.5}$$

where

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

2.1 Bilinear Forms

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

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

yielding for $\lambda = 0$ the standard definition of coercivity.

Rationale for weak coercitivity

Before coming to the existence and uniqueness theorem, let us notice that, if we introduce the change of variable $u_\lambda(t,\mathbf{x}) := e^{-\lambda t}u(t,\mathbf{x})$, where u is the solution of (2.4), the new unknown u_λ satisfies

$$\frac{\partial u_\lambda}{\partial t} + Lu_\lambda + \lambda u_\lambda = e^{-\lambda t}f \quad \text{ in } Q_T.$$

If we have a weakly coercive bilinear form $a(w,v)$, the modified bilinear form $a_\lambda(w,v) := a(w,v)+\lambda(w,v)$ associated to this last problem is coercive, i.e., it satisfies (11.1.6) with $\lambda = 0$. Therefore, if we replace f with $e^{-\lambda t}f$ and L with $L+\lambda I, I$ being the identity operator, without loosing generality we can assume that the bilinear form associated to the initial-boundary value problem (2.4) satisfies the weak coercitivi with $\lambda = 0$.

This will be always assumed in the sequel of this chapter. However, it is worthy to notice that the estimates we will prove are valid for the auxiliary unknown $u_\lambda(t,\mathbf{x})$ (or its approximations), and that the corresponding estimates for the solution $u(t,\mathbf{x})$ show an extra multiplicative factor $e^{\lambda t}$.

Let us now prove the existence theorem. We notice that hereafter all norms refer to the space variables, i.e., $\|\cdot\|_k$ is the norm in the Sobolev space $H^k(\Omega)$ for $k \geq 0$.

2.2 Condition on the well-posedness of problem 2.5

Consider the following parabolic partial differential equation (PDE) which incorporates time-dependency:

$$\begin{cases} \partial_t u - \nabla\cdot(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \nabla\cdot(\mathbf{c}u) + \sigma u = f & \text{in } \Omega\times(0,T], \\ u = g_D & \text{on } \Gamma_D\times(0,T], \\ (\mu\nabla u - \mathbf{b}u)\cdot\mathbf{n} + \gamma u = g_N & \text{on } \Gamma_N\times(0,T], \\ u(\cdot,0) = u_0 & \text{in } \Omega. \end{cases}$$

Where u_0 is the initial condition and T is the final time.

To ensure the well-posedness of the problem, we assume the following regularity and compatibility conditions:

- $\mu \in \mathrm{L}^\infty(\Omega\times(0,T))$ with $\mu(\mathbf{x},t) \geq \mu_0 > 0$.
- $\sigma \in \mathrm{L}^\infty(\Omega\times(0,T))$.
- $\mathbf{b},\mathbf{c} \in [\mathrm{L}^\infty(\Omega\times(0,T))]^n$.
- $\mathrm{div}(\mathbf{b}-\mathbf{c}) \in \mathrm{L}^\infty(\Omega\times(0,T))$.
- $f \in \mathrm{L}^2(\Omega\times(0,T))$.
- $g_D \in \mathrm{H}^{1/2}(\partial\Omega\times(0,T))$.
- $g_N \in \mathrm{L}^2(\partial\Omega\times(0,T))$.
- $\gamma \in \mathrm{L}^\infty(\partial\Omega)$ is a constant.

- $u_0 \in L^2(\Omega)$.

1. **Weak Form:** The weak form for the parabolic problem can be written as:

$$\int_\Omega \partial_t \tilde{u} v \, d\Omega + a(\tilde{u}, v) = F(v) \quad \forall v \in V, \text{ for a.e. } t \in (0, T],$$

where $a(\tilde{u}, v)$ is the bilinear form from the elliptic case (See Chapters 5) and $F(v)$ is adjusted similarly:

$$F(v) = \int_\Omega f v \, d\Omega + \int_{\Gamma_N} g_N v \, d\Gamma - a(R_{g_D}, v) - \int_\Omega \partial_t R_{g_D} v \, d\Omega.$$

We consider the evolution of u over time. For the parabolic problem, we include a time derivative term and seek $u \in L^2(0, T; V)$ such that $\frac{\partial u}{\partial t} \in L^2(0, T; V')$ and the following equation holds for almost every $t \in (0, T)$ and for all $v \in V$:

$$\left(\frac{\partial \tilde{u}}{\partial t}, v \right) d\Omega + a(\tilde{u}(t), v) = f(v) - \left(\frac{\partial \tilde{R}_{gD}}{\partial t}, v \right), \quad \forall t \in (0, T)$$

where $f(v)$ represents external forces or source terms. The last term, is an integration and since we assumed that $g_D \in H^{1/2}(\Gamma_D \times (0, T)) \implies R_{gD} \in H^1(\Omega \times (0, T)) \implies$ the last term is well defined. The initial condition is given by:

$$u(x, 0) = u_0(x) \quad \text{for } x \in \Omega.$$

We are considering a bilinear form $a(\cdot, \cdot)$ associated with a boundary value problem, assuming it meets the criteria of being continuous and weakly coercive. With the given initial and boundary data where $u_0 \in L^2(\Omega)$ and $f \in L^2(\Omega \times (0, T))$, we establish that problem (2.5) admits a unique solution, denoted by u. The solution u possesses the following mathematical properties:

1. **Continuity in Time with L^2 Spatial Regularity**: The function u is continuous with respect to time when viewed in the $L^2(\Omega)$ space. This property is mathematically denoted as $u \in C^0(\mathbb{R}^+; L^2(\Omega))$, meaning that for any fixed time $t \in \mathbb{R}^+$, the function $u(t, \cdot)$ belongs to $L^2(\Omega)$, and the mapping from time to $L^2(\Omega)$ is continuous.

2. **Square Integrability in V Space**: Over the real positive time domain $\mathbb{R}^+$, the function u also resides in the space $L^2(\mathbb{R}^+; V)$. This implies that the integral of the square of the norm of $u(t, \cdot)$ in the space V over any finite interval of time is finite.

3. **Temporal Differentiability in a Weaker Sense**: The temporal derivative of u, denoted as $\frac{\partial u}{\partial t}$, exists in a weaker form. Specifically, it belongs to $L^2(\mathbb{R}^+; V')$, where V' is the dual space of V. This is further compactly expressed as $u \in H^1(\mathbb{R}^+; V, V')$, indicating that u has weak derivatives with respect to time that are square-integrable in V'.

These properties collectively suggest that the solution u not only adheres to the necessary regularity conditions in space but also behaves well in time, both in terms of continuity and derivative existence in an appropriate functional framework.

Now to prove weakly coercitivity of the bilinear forms we have two possible approaches:

1. We prove it for $\lambda = 0$, following the same approach in Chapter 5, imposing thus stricter conditions on the coefficients and functions.

2. We prove the weakly coercitivity: In this case we require $\mu(\mathbf{x}) \geq \mu_0 > 0, \sigma \in \mathrm{L}^\infty(\Omega), \gamma \in L^\infty(\partial\Omega)$, and $\mathbf{b}, \mathbf{c} \in [\mathrm{L}^\infty(\Omega)]^2, g_N \in L^2(\Gamma_N)$ and getting this weak coercitivity λ (QV NAPDE 11.1.1):

$$\lambda > \left(\|\mathbf{b} - \mathbf{c}\|_{L^\infty(\Omega)} \frac{1}{4\epsilon} + \|\sigma\|_{L^\infty(\Omega)} + C' \frac{1}{4\epsilon} \|\gamma\|_{L^\infty(\Omega)} \right)$$

Proof of weak coercitivity In the following step we will use the following notation:$\|v\|_{0,\Omega} = \|v\|_{L^2(\Omega)}$and $\|v\|_{1,\Omega} = \|v\|_{H^1(\Omega)}$.

$$\alpha(v,v) \geq \mu_0 \|\nabla v\|_0^2 - \|\mathbf{b}-\mathbf{c}\|_{L^\infty(\Omega)} \|\nabla v\|_0 \|v\|_0 - \|\sigma\|_{L^\infty(\Omega)}^2 \|v\|_0^2 - \|\gamma\|_{L^\infty(\Omega)} \|v\|_{0,\partial\Omega}^2$$

$$\alpha(v,v) + \|\mathbf{b}-\mathbf{c}\|_{L^\infty(\Omega)} \|\nabla v\|_0 \|v\|_0 + \|\sigma\|_{L^\infty(\Omega)}^2 \|v\|_0^2 + \|\gamma\|_{L^\infty(\Omega)} \|v\|_{0,\partial\Omega}^2 \geq \mu_0 \|\nabla v\|_0^2$$

Using the following inequalities:

$$\|\nabla v\|_0 \|v\|_0 \leq \epsilon \|\nabla v\|_0^2 + \frac{1}{4\epsilon} \|v\|_0^2$$

The first inequality of the following step comes from this reference, on Mathematics Stack Exchange: link

$$\|v\|_{L^2(\partial\Omega)}^2 \leq C' \|\nabla v\|_0 \|v\|_0 \leq C' \left(\epsilon' \|\nabla v\|_0^2 + \frac{1}{4\epsilon'} \|v\|_0^2 \right)$$

$$\begin{aligned} \alpha(v,v) + \|\mathbf{b} - \mathbf{c}\|_{L^\infty(\Omega)} \left(\epsilon \|\nabla v\|_0^2 + \frac{1}{4\epsilon} \|v\|_0^2 \right) + \|\sigma\|_{L^\infty(\Omega)}^2 \|v\|_0^2 + \\ + \|\gamma\|_{L^\infty(\Omega)} C' \left(\epsilon' \|\nabla v\|_0^2 + \frac{1}{4\epsilon'} \|v\|_0^2 \right) \|v\|_0^2 \geq \mu_0 \|\nabla v\|_0^2 \end{aligned}$$

Next steps, don't get discouraged:

$$\begin{aligned} \alpha(v,v) + \left(\|\mathbf{b} - \mathbf{c}\|_{L^\infty(\Omega)} \frac{1}{4\epsilon} + \|\sigma\|_{L^\infty(\Omega)} + C' \frac{1}{4\epsilon} \|\gamma\|_{L^\infty(\Omega)} \right) \|v\|_0^2 \geq \\ \left(\mu_0 - \epsilon \|\mathbf{b} - \mathbf{c}\|_{L^\infty(\Omega)} - C' \epsilon' \|\gamma\|_{L^\infty(\Omega)} \right) \|\nabla v\|_0^2, \end{aligned}$$

$$a(v,v) + \lambda \|v\|_0^2 \geq \frac{\left(\mu_0 - \epsilon \|\mathbf{b} - \mathbf{c}\|_{L^\infty(\Omega)} - C' \epsilon' \|\gamma\|_{L^\infty(\Omega)} \right)}{1 + C_\Omega^2} \|v\|_1^2.$$

Where we used in the last step the Poincaré inequality.

2.3 Galerkin Approximation

For each $t > 0$, find $u_h(t) \in V_h$ such that

$$\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h d\Omega + a\left(u_h(t), v_h\right) = \int_\Omega f(t) v_h d\Omega \quad \forall v_h \in v_h \tag{2.6}$$

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.

Such problem is called semi-discretization of (2.5), as the temporal variable has not yet been discretized.

2.4 Algebraic Formulation

We introduce a basis $\{\varphi_j\}$ for V_h and we observe that it suffices that (2.6) is verified for the basis functions in order to be satisfied by all the functions of the subspace.

Moreover, since for each $t > 0$ the solution to the Galerkin problem belongs to the subspace as well, we will have

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N_h} u_j(t) \varphi_j(\mathbf{x}) \tag{2.7}$$

where the coefficients $\{u_j(t)\}$ represent the unknowns of problem (2.6).

Denoting by $\dot{u}_j(t)$ the derivatives of the function $u_j(t)$ with respect to time, (2.6) becomes

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

that is

$$\sum_{j=1}^{N_h} \dot{u}_j(t) \underbrace{\int_\Omega \varphi_j \varphi_i d\Omega}_{m_{ij}} + \sum_{j=1}^{N_h} u_j(t) \underbrace{a\left(\varphi_j, \varphi_i\right)}_{a_{ij}} = \underbrace{\int_\Omega f(t) \phi_i d\Omega}_{f_i(t)}, \quad \forall i \leq N_h \tag{2.8}$$

If we define the vector of unknowns $\mathbf{u} = (u_1(t), u_2(t), \ldots, u_{N_h}(t))^T$, the mass matrix $\mathrm{M} = [m_{ij}]$, the stiffness matrix $\mathrm{A} = [a_{ij}]$ and the right-hand side vector $\mathbf{f} = (f_1(t), f_2(t), \ldots, f_{N_h}(t))^T$, the system (2.7) can be rewritten in matrix form as

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

2.5 Time Discretization

For the numerical solution of this ODE system, many finite difference methods are available. Here we limit ourselves to considering the so-called θ-method.

The latter discretizes the temporal derivative by a simple 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 $\theta (0 \leq \theta \leq 1)$,

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}\left[\theta \mathbf{u}^{k+1} + (1-\theta)\mathbf{u}^k\right] = \theta \mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k \tag{2.9}$$

The real positive parameter $\Delta t = t^{k+1} - t^k, k = 0, 1, \ldots$, denotes the discretization step (here assumed to be constant), while the superscript k indicates that the quantity under consideration refers to the time t^k. Let us see some particular cases of (2.8):

- For $\theta = 0$ we obtain the forward Euler (or explicit Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^k = \mathbf{f}^k \tag{2.10}$$

which is accurate to order one with respect to Δt;

- For $\theta = 1$ we have the backward Euler (or implicit Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^{k+1} = \mathbf{f}^{k+1}, \tag{2.11}$$

also of first order with respect to Δt;

- For $\theta = 1/2$ we have the Crank-Nicolson (or trapezoidal) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \frac{1}{2}\,\mathrm{A}\left(\mathbf{u}^{k+1} + \mathbf{u}^k\right) = \frac{1}{2}\left(\mathbf{f}^{k+1} + \mathbf{f}^k\right) \tag{2.12}$$

which is of second order in Δt. (More precisely, $\theta = 1/2$ is the only value for which we obtain a second-order method.) Let us consider the two extremal cases, $\theta = 0$ and $\theta = 1$. For both, we obtain a system of linear equations: In the θ-method for solving ordinary differential equations (ODEs), the matrix that precedes u^{k+1} in the system of linear equations depends on the choice of θ:

1. If $\theta = 0$, the system to solve has matrix $\frac{M}{\Delta t}$
2. If $\theta = 1$, the system to solve has matrix $\frac{\mathrm{M}}{\Delta t} + A$

We observe that the M matrix is invertible, being positive definite.

When $\theta = 0$, the scheme isn't unconditionally stable. For a subspace V_h of finite elements, the stability condition is given by: $\exists c > 0 : \Delta t \leq ch^2 \quad \forall h > 0$. This means that the time step Δt depends on the mesh size h; it can't be chosen independently.

If we make the matrix M diagonal, we decouple the system. This is achieved through 'lumping' of the mass matrix, which means transforming M into a diagonal matrix.

When $\theta > 0$, the system equation becomes: $K\mathbf{u}^{k+1} = \mathbf{g}$. Here, $\mathbf{g}$ is the source term, and $K = \frac{M}{\Delta t} + \theta A$. Since the spatial operator L and hence the matrix A are time-independent, and assuming the spatial mesh is constant, K can be factorized once at the start of the process.

If both M and A are symmetric, then K is also symmetric. This allows the use of Cholesky factorization: $K = HH^T$, where H is a lower triangular matrix. At each time step, two triangular systems need to be solved with N_h unknowns:

- $H\mathbf{y} = \mathbf{g}$
- $H^T\mathbf{u}^{k+1} = \mathbf{y}$

2.6 A priori estimates

A priori estimates

Let us consider problem (2.5); since the corresponding equations must hold for each $v \in V$, it will be legitimate to set $v = u(t)$ (t being given), solution of the problem itself, yielding

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t) d\Omega + a(u(t), u(t)) = \int_\Omega f(t) u(t) d\Omega \quad \forall t > 0 \tag{2.13}$$

Considering the individual terms, we have

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t) d\Omega = \frac{1}{2}\frac{d}{dt}\int_\Omega |u(t)|^2 d\Omega = \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} \tag{2.14}$$

If we assume for simplicity that the bilinear form is coercive (with coercivity constant equal to α), we obtain:

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

while thanks to the Cauchy-Schwarz inequality, we find

$$(f(t), u(t)) \leq \|f(t)\|_{\mathrm{L}^2(\Omega)} \|u(t)\|_{\mathrm{L}^2(\Omega)} \tag{2.15}$$

In the remainder, we will often use Young's inequality

$$\forall a, b \in \mathbb{R}, \quad ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \quad \forall \varepsilon > 0$$

which descends from the elementary inequality

$$\left(\sqrt{\varepsilon} a - \frac{1}{2\sqrt{\varepsilon}} b\right)^2 \geq 0$$

Using first Poincaré inequality and Young's inequality, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha \|\nabla u(t)\|^2_{L^2(\Omega)} &\leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \\ &\leq \frac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.16}$$

The detailed steps are the following ones:

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha \|\nabla u(t)\|^2_{L^2(\Omega)} &\overset{\text{A}}{=} \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha \|u(t)\|^2_V \\ &\leq \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + a(u(t), u(t)) \\ &= \int_\Omega f(t) u(t) d\Omega \\ &\leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \\ &\leq \frac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \frac{1}{4C_\Omega^2/(2\alpha)}\|u(t)\|^2_{L^2(\Omega)} \\ &\overset{\text{B}}{\leq} \frac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}. \end{aligned}$$

B is Poincaré, but $V = H_0^1$, so for A we take $\|v\|_V := \|\nabla v\|_{L^2(\Omega)}$ which is equivalent to the H^1 norm on that space H_0^1(that is, both upper and lower bounds). Then, by integrating in time we obtain, for all $t > 0$,

$$\|u(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds \leq \|u_0\|^2_{L^2(\Omega)} + \frac{C_\Omega^2}{\alpha} \int_0^t \|f(s)\|^2_{L^2(\Omega)} ds. \quad (2.17)$$

This is an a priori energy estimate. Different kinds of a priori estimates can be obtained as follows. Note that:

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

Then from (5.14), using (2.14) and (2.15) we obtain (still using the Poincaré inequality)

$$\|u(t)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)} + \frac{\alpha}{C_\Omega}\|u(t)\|_{L^2(\Omega)}\|\nabla u(t)\|_{L^2(\Omega)} \leq \|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)}, \quad t > 0$$

The detailed steps are the following: Starting with the energy equality, applying the time derivative property, coercivity, the Cauchy-Schwarz inequality, and the Poincaré inequality, we derive the final inequality as follows:

$$\|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)} \geq \int_\Omega f(t)u(t)\, d\Omega = \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + a(u(t), u(t))$$

$$\geq \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)} + \alpha\|u(t)\|^2_V \geq \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)} + \frac{\alpha}{C_\Omega^2}\|u(t)\|^2_{L^2(\Omega)}$$

$$\Rightarrow \|u(t)\|_{L^2(\Omega)}\frac{d}{dt}\|u(t)\|_{L^2(\Omega)} + \frac{\alpha}{C_\Omega^2}\|u(t)\|^2_{L^2(\Omega)} \leq \|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)},$$

Remember that in this step we assume the space $V = H_0^1(\Omega)$, and so $||v||_V = ||v||_{H_0^1(\Omega)} = ||\nabla v||_{L^2_\Omega}$.

If $\|u(t)\|_{L^2(\Omega)} \neq 0$ (otherwise we should proceed differently, even though the final result is still true) we can divide by $\|u(t)\|_{L^2(\Omega)}$ and integrate in time to obtain:

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \quad t > 0 \quad (2.18)$$

This is a further a priori estimate. Let us now use the first inequality in (2.16) and integrate in time to yield:

$$\begin{aligned}
&\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)}\, ds \\
&\qquad \leq \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)}\|u(s)\|_{L^2(\Omega)}\, ds \\
&\qquad \leq_{\text{(equation 2.18)}} \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \left(\|u_0\|_{L^2(\Omega)} + \int_0^s \|f(\tau)\|_{L^2(\Omega)}\, d\tau\right) ds \\
&\qquad = \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)}\|u_0\|_{L^2(\Omega)}\, ds + \int_0^t 2\|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)}\, d\tau\, ds \\
&\qquad = \left(\|u_0\|_{L^2(\Omega)} + \int_0^t \|f(\tau)\|_{L^2(\Omega)}\, ds\right)^2
\end{aligned}$$

The latter equality follows upon noticing that:

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

Therefore if we define $F(s) = \left(\int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau \right)^2$, from the Fundamental Theorem of Calculus we can derive that:

$$\int_0^t \frac{d}{ds} F(s) ds = F(t) - F(0) = F(0).$$

We finally conclude with the additional a priori estimate:

$$\left(\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds \right)^{\frac{1}{2}} \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \quad t > 0. \tag{2.19}$$

We have seen that we can formulate the Galerkin problem (2.6) for problem (2.5) and that the latter, under suitable hypotheses, admits a unique solution. Similarly to what we did for problem (2.5) we can prove the following a priori (stability) estimates for the solution to problem (2.6):

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

For its proof we can take, for every $t > 0, v_h = u_h(t)$ and proceed as we did to obtain (2.16). Then, by recalling that the initial data is $u_h(0) = u_{0h}$, we can deduce the following discrete counterparts of (2.18 NMDP) and (2.19):

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

and

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

Convergence Analysis of the Semi-Discrete Problem

Theorem 1. *There exists a constant $C > 0$ independent of both t and h such that:*

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

Stability Analysis of the θ-Method

We now analyze the stability of the fully discretized problem. Applying the θ-method to the Galerkin problem (2.6) we obtain

$$\begin{aligned}\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h\right) \\ = \theta F^{k+1}(v_h) + (1-\theta)F^k(v_h) \quad \forall v_h \in v_h,\end{aligned} \tag{2.20}$$

for each $k \geq 0$, with $u_h^0 = u_{0h}$. F^k indicates that the functional is evaluated at time t^k. We will limit ourselves to the case where $F = 0$ and start to consider the case of the implicit Euler method ($\theta = 1$) that is

$$\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(u_h^{k+1}, v_h\right) = 0 \quad \forall v_h \in v_h \tag{2.21}$$

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

$$\left(u_h^{k+1}, u_h^{k+1}\right) + \Delta t\ a\left(u_h^{k+1}, u_h^{k+1}\right) = \left(u_h^k, u_h^{k+1}\right). \tag{2.22}$$

By exploiting the following inequalities:

$$a\left(u_h^{k+1}, u_h^{k+1}\right) \geq \alpha \left\|u_h^{k+1}\right\|_V^2 \tag{2.23}$$

$$\left(u_h^k, u_h^{k+1}\right) \leq \frac{1}{2}\left\|u_h^k\right\|_{\mathrm{L}^2(\Omega)}^2 + \frac{1}{2}\left\|u_h^{k+1}\right\|_{\mathrm{L}^2(\Omega)}^2, \tag{2.24}$$

the former deriving from the coercivity of the bilinear form $a(\cdot,\cdot)$, and the latter from the Cauchy-Schwarz and Young inequalities, we obtain

$$\left\|u_h^{k+1}\right\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \left\|u_h^{k+1}\right\|_V^2 \leq \left\|u_h^k\right\|_{\mathrm{L}^2(\Omega)}^2. \tag{2.25}$$

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

$$(1 + 2\alpha\Delta t)\left\|u_h^{k+1}\right\|_{\mathrm{L}^2(\Omega)}^2 \leq \left\|u_h^k\right\|_{\mathrm{L}^2(\Omega)}^2 \tag{2.26}$$

hence:

$$\left\|u_h^{k+1}\right\|_{\mathrm{L}^2(\Omega)} \leq \frac{1}{\sqrt{1+2\alpha\Delta t}}\left\|u_h^k\right\|_{\mathrm{L}^2(\Omega)} \tag{2.27}$$

which entails:

$$\left\|u_h^k\right\|_{\mathrm{L}^2(\Omega)} \leq \left(\frac{1}{\sqrt{1+2\alpha\Delta t}}\right)^k \|u_{0h}\|_{\mathrm{L}^2(\Omega)} \tag{2.28}$$

and therefore

$$\lim_{k\to\infty}\left\|u_h^k\right\|_{L^2(\Omega)} = 0 \tag{2.29}$$

that is the backward Euler method is absolutely stable without any restriction on the time step Δt. Assume now $f \neq 0$. We have

$$\underbrace{\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, u_h^{k+1}\right)}_{(I)} + \underbrace{a\left(u_h^{k+1}, u_h^{k+1}\right)}_{(II)} = \underbrace{\int_\Omega f^{k+1} u_h^{k+1}}_{(III)} \tag{2.30}$$

We can derive the following inequalities:

$$(I) \geq \frac{1}{2\Delta t}\left(\left\|u_h^{k+1}\right\|_{L^2}^2 - \left\|u_h^k\right\|_{L^2}^2\right)^1 \tag{2.31}$$

$$(II) \geq \alpha \left\|u_h^{k+1}\right\|_V^2 \quad \text{(coercivity of } a(\cdot,\cdot)) \tag{2.32}$$

$$(III) \overset{\text{(C.S.)}}{\leq} \left\|f^{k+1}\right\|_{L^2}\left\|u_h^{k+1}\right\|_V \overset{\text{(Young)}}{\leq} \frac{1}{2\alpha}\left\|f^{k+1}\right\|_{L^2}^2 + \frac{\alpha}{2}\left\|u_h^{k+1}\right\|_V^2 \tag{2.33}$$

The first inequality derives from the fact that $(a-b,a) \geq \frac{1}{2}\left(\|a\|^2 - \|b\|^2\right) \ \forall a, b$. Then, after summation on k, for $k = 0, \ldots, n-1$, we obtain:

$$\|u_h^n\|_{L^2}^2 + \underbrace{\alpha \sum_{k=1}^{n} \Delta t \left\|u_h^k\right\|_V^2}_{\simeq \alpha \int_0^{t^n} \|u_h(t)\|_V^2 dt} \leq \|u_{0,h}\|_{L^2}^2 + \underbrace{\frac{1}{\alpha}\sum_{k=1}^{n} \Delta t \left\|f^k\right\|_{L^2}^2}_{\simeq \frac{1}{\alpha}\int_0^{t^n} \|f(t)\|_{L^2}^2 dt} \tag{2.34}$$

This leads to unconditional stability (no restriction on Δt).

Before analyzing the general case where θ is an arbitrary parameter ranging between 0 and 1, we introduce the following definition.

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

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

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

$$a(w_h, v_h) = \lambda_h (w_h, v_h) \quad \forall v_h \in V_h. \tag{2.35}$$

From an algebraic viewpoint, problem (2.35) can be formulated as follows:

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

where A is the stiffness matrix and M the mass matrix. We are therefore dealing with a generalized eigenvalue problem. Such eigenvalues are all positive and N_h in number (N_h being as usual the dimension of the subspace V_h); after ordering them in ascending order, $\lambda_h^1 \leq \lambda_h^2 \leq \ldots \leq \lambda_h^{N_h}$, we have:

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

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

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

and, thanks to the eigenfunction orthonormality:

$$\|v_h\|^2_{\mathrm{L}^2(\Omega)} = \sum_{j=1}^{N_h} v_j^2 \tag{2.36}$$

Let us consider an arbitrary $\theta \in [0, 1]$ and let us limit ourselves to the case where the bilinear form $a(\cdot, \cdot)$ is symmetric (otherwise, although the final stability result holds in general, the following proof would not work, as the eigenfunctions would not necessarily form a basis).

Since $u_h^k \in V_h$, we can write

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

We observe that in this modal expansion, the u_j^k no longer represent the nodal values of u_h^k. If we now set $F = 0$ in (2.20) and take $v_h = w_h^i$, we find:

1. Original Equation (2.20):

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

   Here, u_h^k and u_h^{k+1} are the approximate solutions at time steps k and $k+1$, v_h is a test function from the finite element space V_h, and $a(\cdot, \cdot)$ is a bilinear form.

2. Discretization of u_h: The solution u_h is approximated as a linear combination of basis functions w_h^j in the finite element space:

$$u_h = \sum_{j=1}^{N_h} u_j w_h^j$$

   where N_h is the number of basis functions, and u_j are the coefficients.

3. Substituting in Equation (2.20):

   The term $\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, v_h\right)$ in (2.20) becomes:

$$\left(\frac{1}{\Delta t}\sum_{j=1}^{N_h}(u_j^{k+1} - u_j^k)w_h^j, v_h\right)$$

By choosing $v_h = w_h^i$, this term translates to:

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

where (w_h^j, w_h^i) represents the inner product of the basis functions.
Similarly, the bilinear form $a(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h)$ becomes:

$$a\left(\sum_{j=1}^{N_h}(\theta u_j^{k+1} + (1-\theta)u_j^k)w_h^j, w_h^i\right) = \sum_{j=1}^{N_h}(\theta u_j^{k+1} + (1-\theta)u_j^k)a(w_h^j, w_h^i)$$

4. Thus, the substitution leads to the equation:

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

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

for each $i = 1, \ldots, N_h$.

For each pair $i, j = 1, \ldots, N_h$ we have:

$$a\left(w_h^j, w_h^i\right) = \lambda_h^j\left(w_h^j, w_h^i\right) = \lambda_h^j\delta_{ij}$$

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

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

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

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

Recalling (2.36), we can conclude that for the method to be absolutely stable, we must impose the inequality

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

that is:

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

Hence,

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

The second inequality is always verified, while the first one can be rewritten as:

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

If $\theta \geq 1/2$, the left-hand side is non-negative, while the right-hand side is negative, so the inequality holds for each Δt. Instead, if $\theta < 1/2$, the inequality is satisfied (hence the method is stable) only if:

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

As such relation must hold for all the eigenvalues λ_h^i of the bilinear form, it will suffice to require that it holds for the largest among them, which we have supposed to be $\lambda_h^{N_h}$.
To summarize, we have:

- if $\theta \geq 1/2$, the θ-method is unconditionally absolutely stable, i.e. it is absolutely stable for each Δt;
- if $\theta < 1/2$, the θ-method is absolutely stable only for $\Delta t \leq \frac{2}{(1-2\theta)\lambda_h^{N_h}}$.

Thanks to the definition of eigenvalue (2.35) and to the continuity property of $a(\cdot,\cdot)$, we deduce

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

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

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

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

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

Keeping this into account, we obtain that for $\theta < 1/2$ the method is absolutely stable only if

$$\Delta t \leq C(\theta) h^2$$

where $C(\theta)$ denotes a positive constant depending on θ. The latter relation implies that for $\theta < 1/2, \Delta t$ cannot be chosen arbitrarily but is bound to the choice of h.

2.7 Convergence analysis of the θ-method

Theorem 2. *Under the hypothesis that u_0, f and the exact solution are sufficiently regular, the following a priori error estimate holds: $\forall n \geq 1$,*

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

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

2.8 Parabolic ADR equation

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

$$\begin{cases} \frac{\partial u}{\partial t} - \mu \Delta u + \boldsymbol{\beta} \cdot \nabla u + \sigma u = f & \text{in } \Omega \times (0,\ \mathrm{T}) \\ u = 0 & \text{on } \partial\Omega \times (0,\ \mathrm{T}), \\ u(0) = u_0 & \text{in } \Omega, \end{cases} \tag{2.37}$$

and where $\mu, \boldsymbol{\beta}, \sigma$ and f are regular functions, satisfying:

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

Introducing a finite dimensional space $V_h \subset H_0^1(\Omega)$, the semi-discrete Galerkin formulation reads: for all $t \in (0, T]$ find $u_h(t) \in V_h$ such that:

$$\left\{ \begin{array}{l} \int_\Omega \frac{\partial u_h(t)}{\partial t} v_h dx + \int_\Omega \mu \nabla u_h(t) \cdot \nabla v_h + \int_\Omega \beta \cdot \nabla u_h(t) v_h + \int_\Omega \sigma u_h(t) v_h \\ \qquad = \int_\Omega f v_h \quad \forall v_h \in v_h, \end{array} \right. \tag{2.38}$$

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

2.9 A semimplicit scheme

We consider a time-advancing scheme, where the diffusion and reaction terms are treated implicitly, while the advection term is treated explicitly. Let us denote $t_k = k\Delta t$, for $k = 0, \ldots, N$, where $\Delta t = T/N$. Let $u_h^{(k)}$ be the approximation of $u(t_k)$. A fully discretized version of (2.37) reads:

$$\left\{ \begin{array}{l} \left(\frac{u_h^{(k+1)} - u_h^{(k)}}{\Delta t}, v_h \right) + \left(\mu \nabla u_h^{(k+1)}, \nabla v_h \right) + \left(\boldsymbol{\beta} \cdot \nabla u_h^{(k)}, v_h \right) \\ \quad + \left(\sigma u_h^{(k+1)}, v_h \right) = (f, v_h) \quad \forall v_h \in v_h, \quad k = 0, \ldots, N-1 \\ u_h^{(0)} = u_{0,h} \end{array} \right. \tag{2.39}$$

where $(\cdot, \cdot)$ denotes the $L^2(\Omega)$ scalar product.

2.10 Stability analysis of the semimplicit scheme

Theorem 3. *In the following derivations, the assumption of $f = 0$ will be made. If the coefficients of the problem satisfy:*

$$b_1^2 < 4\mu_0\sigma_0,$$

then the semimplicit scheme (2.39) is absolutely stable for any choice of Δt. Consider now the case $\sigma = 0$. If the coefficients of the problem satisfy (C_p being the Poincaré constant):

$$b_1 < \mu_0 / C_p$$

then the scheme is absolutely stable for any choice of Δt.

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

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

which entails, assuming $f = 0$, for every k

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

The two right-hand side terms can be bounded by combining the Cauchy-Schwarz and the Young inequalities:

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

where the positive constants η_1 and η_2 will be later fixed according to our best convenience. We end up with the following inequality:

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

In order to prove stability, we need $A > A'$ and $B > B'$. Indeed, if this were true, then we would have:

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

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

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

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

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

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

$$4 \frac{\mu_0 \sigma_0}{b_1 (b_1 + \epsilon)} > 1 \tag{2.40}$$

Clearly, it is possible to find $\epsilon > 0$ such that this holds if and only if:

$$b_1^2 < 4\mu_0\sigma_0 \tag{2.41}$$

In conclusion, whenever the coefficients of the problem satisfy the condition (2.41), the scheme (2.39) is absolutely stable, for any choice of Δt. Let us consider now the case $\sigma = 0$. Proceeding as before, we have:

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

Let us introduce a constant $\omega \in (0,1)$ (to be fixed later). By the Poincaré inequality, we have:

$$\begin{aligned}\left\|\nabla u_h^{(k+1)}\right\|^2 &= (1-\omega)\left\|\nabla u_h^{(k+1)}\right\|^2 + \omega \left\|\nabla u_h^{(k+1)}\right\|^2 \\ &\geq \frac{1-\omega}{C_p^2}\left\|u_h^{(k+1)}\right\|^2 + \omega \left\|\nabla u_h^{(k+1)}\right\|^2\end{aligned}$$

Combining the latter inequalities, we have

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

As in the previous point, we look for conditions on the coefficients such that $A > A'$ and $B > B'$. The second inequality is satifsied if and only if

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

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

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

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

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

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

$$b_1 (b_1 + \epsilon) C_p^2 < \mu_0^2$$

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

$$b_1 < \mu_0 / C_p \tag{2.42}$$

In conclusion, if (2.42) is satisfied, the scheme is absolutely stable for any choice of Δt.

Chapter 3

Domain Decomposition Methods

An elementary introduction in 1D Consider the 1D elliptic BVP:

$$\begin{cases} \mathcal{L}u := -u'' = f \quad a < x < b \\ u(a) = u(b) = 0 \end{cases} \tag{3.1}$$

This is equivalent to:

$$\begin{cases} \text{Find } u \in V = H_0^1(a,b) \text{ s.t.} \\ a(u,v) = (f,v) \end{cases} \qquad \forall v \in V$$

with

$$a(u,v) = \int_a^b u'v', \quad (f,v) = \int_a^b fv$$

Consider the domain splitting: $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2, \Omega_1 \cap \Omega_2 = \emptyset$

$$a < \gamma < b$$

Correspondingly, consider the following splitting of the space V :

$$V = V_1 \oplus H_\gamma \oplus V_2$$

where

$$v_i = \tilde{H}_0^1(\Omega_i) := \left\{ v \in H_0^1(\Omega) : v|_{\Omega_i} \in H_0^1(\Omega_i), v|_{\Omega \setminus \Omega_i} = 0 \right\}, i = 1,2$$

H_γ = space of harmonic extensions of functions whose values at γ is given = $\left\{ v_H \in V : v(\gamma) = v_\gamma, v|_{\Omega_i} = v_i, v_i'' = 0, i = 1,2 \quad \forall v_\gamma \in \mathbb{R} \right\}$

Proposition 1. *u is a solution to (3.1) if and only if:*

$$\begin{cases} -u_1'' = f & \text{in } \Omega_1 & \\ -u_2'' = f & \text{in } \Omega_2 & \\ u_1 = u_2 & \text{on } \gamma & \textit{Transmission condition (D)} \\ u_1' = u_2' & \text{on } \gamma & \textit{Transmission condition (N)} \\ u_1(a) = 0, u_2(b) = 0 & & \textit{B.C.} \end{cases} \tag{3.2}$$

The converse is also true, that is: if u_1, u_2 are solutions to (3.1), then, setting u such that $u|_{\Omega_1} = u_1$, $u|_{\Omega_2} = u_2$, u is a solution to (3.2). Therefore, Problem (3.1) is equivalent to Problem (3.2). Similar results hold in $\mathbb{R}^d$:

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \iff \begin{cases} -\Delta u_1 = f & \text{in } \Omega_1 \\ -\Delta u_2 = f & \text{in } \Omega_2 \\ u_1 = u_2 & \text{on } \Gamma \textit{(Transmission cond. (D))} \\ \frac{\partial u_1}{\partial \mathbf{n}} = \frac{\partial u_2}{\partial \mathbf{n}} & \text{on } \Gamma \textit{(Transmission cond. (N))} \\ u_1 = 0 & \text{on } \partial\Omega_1 \backslash \Gamma \\ u_2 = 0 & \text{on } \partial\Omega_2 \backslash \Gamma \end{cases}$$

Motivation

Domain Decomposition (DD) can be applied within any discretization method for PDEs, such as Finite Element Method (FEM), Finite Volume (FV), Finite Difference (FD), and Spectral Element Method (SEM), to enhance the efficiency of their algebraic solutions on parallel computing platforms. **DD methods** facilitate the division of a boundary-value problem across subdivided computational domains. This approach offers a *particularly advantageous framework for solving heterogeneous or multiphysics problems*, that is, those involving different types of differential equations in various sections of the computational domain.

The Idea

The computational domain Ω, where the boundary value problem (BVP) is established, is partitioned into two or more subdomains. In these subdomains, problems of reduced dimension are solved. Parallel solution algorithms are applicable here. There are two methods for dividing the computational domain: using either disjoint or overlapping subdomains.

3.1 Classical Iterative DD Methods

Model problem

Consider the model problem: Find $u : \Omega \to \mathbb{R}$ s.t.

$$\begin{cases} Lu = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

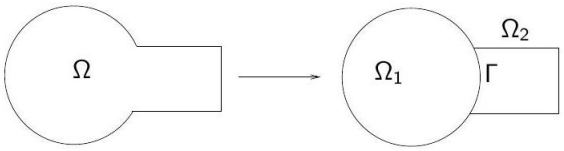
L is a generic second order elliptic operator. The weak formulation reads:

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

where $a(\cdot, \cdot)$ is the bilinear form associated with L.

Non Overlapping Decomposition

We partition now the domain Ω in two disjoint subdomains:



The following equivalence result holds.

Theorem 4. *The solution u of the model problem is such that $u_{|_{\Omega_i}} = u_i$ for $i = 1, 2$, where u_i is the solution to the problem*

$$\begin{cases} Lu_i = f & \text{in } \Omega_i \\ u_i = 0 & \text{on } \partial\Omega_i \backslash \Gamma \end{cases}$$

with interface conditions:

$$u_1 = u_2 \quad \text{and} \quad \frac{\partial u_1}{\partial n_L} = \frac{\partial u_2}{\partial n_L} \quad \text{on } \Gamma$$

where $\partial/\partial n_L$ is the conormal derivative.

Dirichlet-Neumann Method

Given $u_2^{(0)}$ on Γ, for $k \geq 1$ solve the problems:

$$\begin{cases} Lu_1^{(k)} = f & \text{in } \Omega_1 \\ u_1^{(k)} = u_2^{(k-1)} & \text{on } \Gamma \\ u_1^{(k)} = 0 & \text{on } \partial\Omega_1 \backslash \Gamma \end{cases}$$

$$\begin{cases} Lu_2^{(k)} = f & \text{in } \Omega_2 \\ \frac{\partial u_2^{(k)}}{\partial n_L} = \frac{\partial u_1^{(k)}}{\partial n_L} & \text{on } \Gamma \\ u_2^{(k)} = 0 & \text{on } \partial\Omega_2 \backslash \Gamma \end{cases}$$

The equivalence theorem ensures that if the sequences $\{u_1^{(k)}\}$ and $\{u_2^{(k)}\}$ converge, their limits will necessarily be the solution to the original problem. Thus, the Dirichlet-Neumann (DN) algorithm is consistent. However, it's important to note that convergence of this algorithm is not always assured.

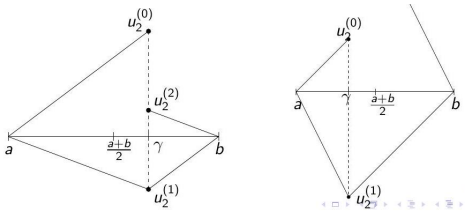
Example

Let $\Omega = (a, b), \gamma \in (a, b), L = -d^2/dx^2$ and $f = 0$. At every $k \geq 1$ the DN algorithm generates the two subproblems:

$$\begin{cases} -\left(u_1^{(k)}\right)'' = 0 & a < x < \gamma \\ u_1^{(k)} = u_2^{(k-1)} & x = \gamma \\ u_1^{(k)} = 0 & x = a \end{cases}$$

$$\begin{cases} -\left(u_2^{(k)}\right)'' = 0 & \gamma < x < b \\ \left(u_2^{(k)}\right)' = \left(u_1^{(k)}\right)' & x = \gamma \\ u_2^{(k)} = 0 & x = b. \end{cases}$$

The two sequences converge only if $\gamma > (a+b)/2$:



A variant of the Dirichlet-Neumann (DN) algorithm can be implemented by modifying the Dirichlet condition in the first subdomain to

$$u_1^{(k)} = \theta u_2^{(k-1)} + (1-\theta) u_1^{(k-1)} \text{ on } \Gamma$$

using a relaxation parameter $\theta > 0$. This approach allows for a reduction in the error between two subsequent iterates. In the aforementioned example, it can be easily verified that by selecting

$$\theta_{\text{opt}} = -\frac{u_1^{(k-1)}}{u_2^{(k-1)} - u_1^{(k-1)}}$$

the algorithm converges to the exact solution in a single iteration.

More generally, there exists an appropriate value for $0 < \theta_{\max} < 1$ such that the DN algorithm converges for any choice of the relaxation parameter θ in the interval $(0, \theta_{\max})$.

Neumann-Neumann Algorithm

Consider again a partition of Ω into two disjoint subdomains and denote by λ the (unknown) value of the solution u on their interface Γ :

$$\lambda = u_i \text{ on } \Gamma \quad (i = 1, 2)$$

Consider the following iterative algorithm: for any given $\lambda^{(0)}$ on Γ, for $k \geq 0$ and $i = 1, 2$, solve the following problems:

$$\begin{cases} L u_i^{(k+1)} = f & \text{in } \Omega_i \\ u_i^{(k+1)} = \lambda^{(k)} & \text{on } \Gamma \\ u_i^{(k+1)} = 0 & \text{on } \partial\Omega_i \backslash \Gamma \end{cases}$$

$$\begin{cases} L \psi_i^{(k+1)} = 0 & \text{in } \Omega_i \\ \dfrac{\partial \psi_i^{(k+1)}}{\partial n} = \dfrac{\partial u_1^{(k+1)}}{\partial n} - \dfrac{\partial u_2^{(k+1)}}{\partial n} & \text{on } \Gamma \\ \psi_i^{(k+1)} = 0 & \text{on } \partial\Omega_i \backslash \Gamma \end{cases}$$

with

$$\lambda^{(k+1)} = \lambda^{(k)} - \theta \left(\sigma_1 \psi_{1|\Gamma}^{(k+1)} - \sigma_2 \psi_{2|\Gamma}^{(k+1)} \right)$$

where θ is a positive acceleration parameter, while σ_1 and σ_2 are two positive coefficients.

3.2 The Steklov-Poincaré interface Equation

Multi-Domain Formulation of Poisson Problem and Interface Conditions

We consider now the model problem:

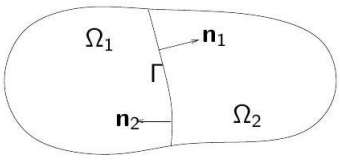
$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

For a domain partitioned into two disjoint subdomains, we can write the equivalent multi-domain formulation $\left(u_i = u_{|\Omega_i}, i = 1, 2\right)$:

$$\begin{cases} -\triangle u_1 = f & \text{in } \Omega_1 \\ u_1 = 0 & \text{on } \partial\Omega_1 \backslash \Gamma \\ -\triangle u_2 = f & \text{in } \Omega_2 \\ u_2 = 0 & \text{on } \partial\Omega_2 \backslash \Gamma \\ u_1 = u_2 & \text{on } \Gamma \\ \frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} & \text{on } \Gamma \end{cases}$$

Remark

- On the interface Γ we have the normal unit vectors $\mathbf{n}_1$ and $\mathbf{n}_2$:



- There holds: $\mathbf{n}_1 = -\mathbf{n}_2$ on Γ.
- We denote $\mathbf{n} = \mathbf{n}_1$ so that

$$\frac{\partial}{\partial n} = \frac{\partial}{\partial n_1} = -\frac{\partial}{\partial n_2} \quad \text{on } \Gamma.$$

The Steklov-Poincaré Operator

Let λ be the unknown value of the solution u on the interface Γ :

$$\lambda = u_{|r}$$

Should we know a priori the value λ on Γ, we could solve the following two independent boundary-value problems with Dirichlet condition on Γ $(i = 1, 2)$:

$$\begin{cases} -\triangle w_i = f & \text{in } \Omega_i \\ w_i = 0 & \text{on } \partial\Omega_i \backslash \Gamma \\ w_i = \lambda & \text{on } \Gamma. \end{cases}$$

With the aim of obtaining the value λ on Γ, let us split w_i as follows

$$w_i = w_i^* + u_i^0$$

where w_i^* and u_i^0 represent the solutions of the following problems $(i = 1, 2)$:

$$\begin{cases} -\triangle w_i^* = f & \text{in } \Omega_i \\ w_i^* = 0 & \text{on } \partial\Omega_i \cap \partial\Omega \\ w_i^* = 0 & \text{on } \Gamma \end{cases}$$

and

$$\begin{cases} -\triangle u_i^0 = 0 & \text{in } \Omega_i \\ u_i^0 = 0 & \text{on } \partial\Omega_i \cap \partial\Omega \\ u_i^0 = \lambda & \text{on } \Gamma. \end{cases}$$

The functions w_i^* depend solely on the source data f, thus we can express this as

$$w_i^* = G_i f$$

where G_i is a linear continuous operator.

Furthermore, u_i^0 depend solely on the value λ on Γ, leading to the expression

$$u_i^0 = H_i \lambda$$

where H_i is the so-called harmonic extension operator of λ on the domain Ω_i.
We have that:

$$\begin{aligned} & w_i = w_i^* + u_i^0 \quad (i = 1, 2) \\ \Leftrightarrow & \frac{\partial w_1}{\partial n} = \frac{\partial w_2}{\partial n} \quad \text{on } \Gamma \\ \Leftrightarrow & \frac{\partial}{\partial n}\left(w_1^* + u_1^0\right) = \frac{\partial}{\partial n}\left(w_2^* + u_2^0\right) \quad \text{on } \Gamma \\ \Leftrightarrow & \frac{\partial}{\partial n}\left(G_1 f + H_1 \lambda\right) = \frac{\partial}{\partial n}\left(G_2 f + H_2 \lambda\right) \quad \text{on } \Gamma \\ \Leftrightarrow & \left(\frac{\partial H_1}{\partial n} - \frac{\partial H_2}{\partial n}\right)\lambda = \left(\frac{\partial G_2}{\partial n} - \frac{\partial G_1}{\partial n}\right) f \quad \text{on } \Gamma. \end{aligned}$$

The equivalence between the decomposition $w_i = w_i^* + u_i^0$ and the equality of normal derivatives is not a direct mathematical derivation but rather a condition for the physical and mathematical consistency of the problem setup and its solution. We have obtained the Steklov-Poincaré equation for the unknown λ on the interface Γ :

$$S\lambda = \chi \quad \text{on } \Gamma$$

- S is the Steklov-Poincaré pseudo-differential operator:

$$S\mu = \frac{\partial}{\partial n} H_1\mu - \frac{\partial}{\partial n} H_2\mu = \sum_{i=1}^{2} \frac{\partial}{\partial n_i} H_i\mu$$

- χ is a linear functional which depends on f :

$$\chi = \frac{\partial}{\partial n} G_2 f - \frac{\partial}{\partial n} G_1 f = -\sum_{i=1}^{2} \frac{\partial}{\partial n_i} G_i f$$

- The operator

$$S_i : \mu \to S_i\mu = \left.\frac{\partial}{\partial n_i}\left(H_i\mu\right)\right|_{\Gamma} \quad i = 1, 2$$

is called local Steklov-Poincaré operator (Dirichlet-to-Neumann) which operates between the trace space

$$\Lambda = \left\{\mu : \exists v \in V \text{ s.t. } \mu = v_{|\Gamma}\right\} = H_{00}^{1/2}(\Gamma)$$

and its dual Λ'.

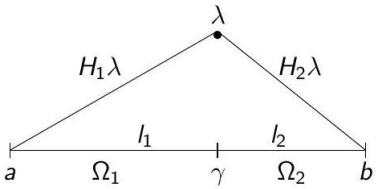
Example

To provide an example of the operator S, we consider a simple 1D problem.

Let $\Omega = (a, b) \subset \mathbb{R}$ as illustrated below. We split Ω in two nonoverlapping subdomains. In this case the interface Γ reduces to the point $\gamma \in (a, b)$ and the Steklov-Poincaré operator S becomes

$$S\lambda = \left(\frac{dH_1}{dx} - \frac{dH_2}{dx}\right)\lambda = \left(\frac{1}{l_1} + \frac{1}{l_2}\right)\lambda$$

where $I_1 = \gamma - a$ et $I_2 = b - \gamma$.



Equivalence Between the DD Schemes and Classical Iterative Methods

The preconditioned Richardson method is an iterative technique for solving linear systems $Ax = b$, enhanced with a preconditioning operator P. It is formulated as:

$$P(x^{(k+1)} - x^{(k)}) = \theta(b - Ax^{(k)}) \tag{3.3}$$

where $x^{(k)}$ is the k-th iteration approximation, and θ is a relaxation parameter.

In the context of numerical methods for partial differential equations, specifically domain decomposition (DD) methods, this concept is adapted as follows:

- **DN Method:** The Dirichlet-Neumann (DN) method uses a preconditioning operator $P_{DN} = S_2 = \frac{\partial(H_2\mu)}{\partial n_2}$. The method's iterative scheme can be seen as a variant of the preconditioned Richardson method:

$$P_{DN}(\lambda^{(k)} - \lambda^{(k-1)}) = \theta(\chi - S\lambda^{(k-1)}) \tag{3.4}$$

  where $\lambda^{(k)}$ represents the interface variable in the k-th iteration.

- **NN Method:** The Neumann-Neumann (NN) method is characterized by a different preconditioning operator $P_{NN} = (\sigma_1 S_1^{-1} + \sigma_2 S_2^{-1})^{-1}$. Its equivalence to the preconditioned Richardson method is reflected in its iterative formula:

$$P_{NN}(\lambda^{(k)} - \lambda^{(k-1)}) = \theta(\chi - S\lambda^{(k-1)}) \tag{3.5}$$

These formulations illustrate that both DN and NN methods in domain decomposition can be viewed as specific applications of the preconditioned Richardson method.

3.3 FEM: Multi-Domain Formulation

Consider the Poisson problem:

$$\begin{cases} -\triangle u = f \text{ in } \Omega \\ \quad u = 0 \text{ on } \partial\Omega \end{cases}$$

Its weak formulation reads

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

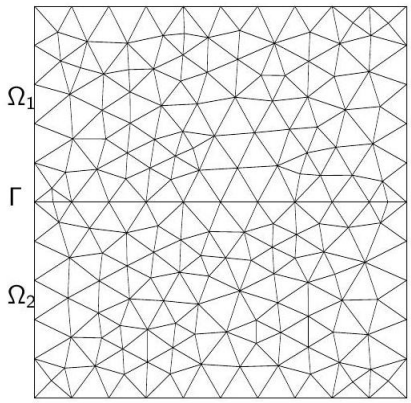
where $V = H_0^1(\Omega)$,

$$a(v,w) = \int_\Omega \nabla v \cdot \nabla w \quad \forall v, w \in V$$

and

$$F(v) = \int_\Omega fv \quad \forall v \in V$$

Suppose that Ω is split into two nonoverlapping subdomains and consider a uniform triangulation $\mathcal{T}_h$ of Ω, conforming on Γ :



The Galerkin finite element approximation of the Poisson problem reads:

$$\text{find } u_h \in V_h : \quad a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h \tag{3.6}$$

where

$$V_h = \left\{ v_h \in C^0(\bar{\Omega}) : v_{h|_K} \in \mathbb{P}_r \quad r \geq 1, \forall K \in \mathcal{T}_h, v_h = 0 \text{ on } \partial\Omega \right\}$$

is the space of finite element functions of degree r with basis $\{\varphi_i\}_{i=1}^{N_h}$. The Galerkin approximation (1) is equivalent to:

$$\text{find } u_h \in V_h : \quad a(u_h, \varphi_i) = F(\varphi_i) \quad \forall i = 1, \ldots, N_h. \tag{3.7}$$

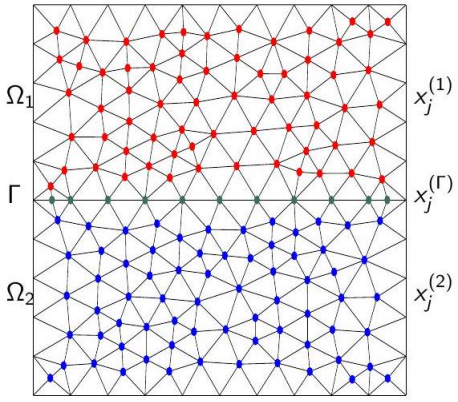
We partition the nodes of the triangulation as follows:

1. $\left\{ x_j^{(1)}, 1 \leq j \leq N_1 \right\}$ nodes in subdomain Ω_1
2. $\left\{ x_j^{(2)}, 1 \leq j \leq N_2 \right\}$ nodes in subdomain Ω_2
3. $\left\{ x_j^{(\Gamma)}, 1 \leq j \leq N_\Gamma \right\}$ nodes on the interface Γ,

and we split the basis functions accordingly:

1. $\varphi_j^{(1)}$ functions associated to the nodes $x_j^{(1)}$
2. $\varphi_j^{(2)}$ functions associated to the nodes $x_j^{(2)}$
3. $\varphi_j^{(\Gamma)}$functions associated to the nodes $x_j^{(\Gamma)}$ on the interface.

Example



We can reformulate a generic elliptic BVP in this way:

$$\text{find } u_h \in V_h : \begin{cases} a\left(u_h, \varphi_i^{(1)}\right) = F\left(\varphi_i^{(1)}\right) & \forall i = 1, \ldots, N_1 \\ a\left(u_h, \varphi_j^{(2)}\right) = F\left(\varphi_j^{(2)}\right) & \forall j = 1, \ldots, N_2 \\ a\left(u_h, \varphi_k^{(\Gamma)}\right) = F\left(\varphi_k^{(\Gamma)}\right) & \forall k = 1, \ldots, N_\Gamma \end{cases} \tag{3.8}$$

We introduce the bilinear form on Ω_i :

$$a_i(v,w) = \int_{\Omega_i} \nabla v \cdot \nabla w \quad \forall v, w \in V, i = 1,2$$

the space

$$V_h^i = \left\{ v \in H^1(\Omega_i) : v = 0 \text{ on } \partial\Omega_i \backslash \Gamma \right\} \quad (i = 1,2)$$

and let $u_h^{(i)} = u_{h|\Omega_i} (i = 1,2)$. Then, problem (3) can be written equivalently as: find $u_h^{(1)} \in V_h^1, u_h^{(2)} \in V_h^2$ s.t.

$$\begin{cases} a_1\left(u_h^{(1)}, \varphi_i^{(1)}\right) = F_1\left(\varphi_i^{(1)}\right), & \forall i = 1, \ldots, N_1 \\ a_2\left(u_h^{(2)}, \varphi_j^{(2)}\right) = F_2\left(\varphi_j^{(2)}\right), & \forall j = 1, \ldots, N_2 \\ a_1\left(u_h^{(1)}, \varphi_k^{(\Gamma)} \mid \Omega_1\right) + a_2\left(u_h^{(2)}, \varphi_k^{(\Gamma)} \mid \Omega_2\right) = & \\ \quad = F_1\left(\varphi_k^{(\Gamma)} \mid \Omega_1\right) + F_2\left(\varphi_k^{(\Gamma)} \mid \Omega_2\right), & \forall k = 1, \ldots, N_\Gamma \end{cases} \tag{3.9}$$

Remark

- Problem (3.8) corresponds to the finite element approximation of the multi-domain formulation of the Poisson problem:

$$\begin{cases} -\triangle u_1 = f & \text{in } \Omega_1 \\ u_1 = 0 & \text{on } \partial\Omega_1 \backslash \Gamma \\ -\triangle u_2 = f & \text{in } \Omega_2 \\ u_2 = 0 & \text{on } \partial\Omega_2 \backslash \Gamma \\ u_1 = u_2 & \text{on } \Gamma \\ \frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} & \text{on } \Gamma. \end{cases}$$

- The condition $u_1 = u_2$ on Γ is satisfied by definition of $u_h^{(i)}$.

We can write:

$$\begin{aligned} u_h(x) = & \sum_{j=1}^{N_1} u_h\left(x_j^{(1)}\right) \varphi_j^{(1)}(x) + \sum_{j=1}^{N_2} u_h\left(x_j^{(2)}\right) \varphi_j^{(2)}(x) \\ & + \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right) \varphi_j^{(\Gamma)}(x) \end{aligned}$$

and we substitute this expression in (3.7) to obtain:

$$\begin{cases} \sum_{j=1}^{N_1} u_h\left(x_j^{(1)}\right) a_1\left(\varphi_j^{(1)}, \varphi_i^{(1)}\right) + \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right) a_1\left(\varphi_j^{(\Gamma)}, \varphi_i^{(1)}\right) = F_1\left(\varphi_i^{(1)}\right) & \forall i = 1, \ldots, N_1 \\ \sum_{j=1}^{N_2} u_h\left(x_j^{(2)}\right) a_2\left(\varphi_j^{(2)}, \varphi_i^{(2)}\right) + \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right) a_2\left(\varphi_j^{(\Gamma)}, \varphi_i^{(2)}\right) = F_2\left(\varphi_i^{(2)}\right) & \forall i = 1, \ldots, N_2 \\ \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right) \left[a_1\left(\varphi_j^{(\Gamma)}, \varphi_i^{(\Gamma)}\right) + a_2\left(\varphi_j^{(\Gamma)}, \varphi_i^{(\Gamma)}\right)\right] & \\ \quad + \sum_{j=1}^{N_1} u_h\left(x_j^{(1)}\right) a_1\left(\varphi_j^{(1)}, \varphi_i^{(\Gamma)}\right) + \sum_{j=1}^{N_2} u_h\left(x_j^{(2)}\right) a_2\left(\varphi_j^{(2)}, \varphi_i^{(\Gamma)}\right) & \\ \quad = F_1\left(\varphi_i^{(\Gamma)}\Big|_{\Omega_1}\right) + F_2\left(\varphi_i^{(\Gamma)}\Big|_{\Omega_2}\right) & \forall i = 1, \ldots, N_\Gamma. \end{cases}$$

A bit of algebra...

$$\begin{cases} \sum_{j=1}^{N_1} u_h\left(x_j^{(1)}\right)(A_{11})_{ij} + \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right)(A_{1\Gamma})_{ij} = F_1\left(\varphi_i^{(1)}\right) & \forall i = 1,\dots,N_1 \\ \sum_{j=1}^{N_2} u_h\left(x_j^{(2)}\right)(A_{22})_{ij} + \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right)(A_{2\Gamma})_{ij} = F_2\left(\varphi_i^{(2)}\right) & \forall i = 1,\dots,N_2 \\ \sum_{j=1}^{N_\Gamma} u_h\left(x_j^{(\Gamma)}\right)\left[\left(A_{\Gamma\Gamma}^{(1)}\right)_{ij} + \left(A_{\Gamma\Gamma}^{(2)}\right)_{ij}\right] \\ \quad + \sum_{j=1}^{N_1} u_h\left(x_j^{(1)}\right)(A_{\Gamma 1})_{ij} + \sum_{j=1}^{N_2} u_h\left(x_j^{(2)}\right)(A_{\Gamma 2})_{ij} \\ \quad = F_1\left(\varphi_i^{(\Gamma)}{}_{|\Omega_1}\right) + F_2\left(\varphi_i^{(\Gamma)}{}_{|\Omega_2}\right) & \forall i = 1,\dots,N_\Gamma. \end{cases}$$

A bit of algebra... [continued]

$$\begin{cases} \sum_{j=1}^{N_1} \mathbf{u}_1 (A_{11})_{ij} + \sum_{j=1}^{N_\Gamma} \mathbf{u}_\Gamma (A_{1\Gamma})_{ij} = \mathbf{f}_1 \forall i = 1,\dots,N_1 \\ \sum_{j=1}^{N_2} \mathbf{u}_2 (A_{22})_{ij} + \sum_{j=1}^{N_\Gamma} \mathbf{u}_\Gamma (A_{2\Gamma})_{ij} = \mathbf{f}_2 \forall i = 1,\dots,N_2 \\ \quad \sum_{j=1}^{N_\Gamma} \mathbf{u}_\Gamma \left[\left(A_{\Gamma\Gamma}^{(1)}\right)_{ij} + \left(A_{\Gamma\Gamma}^{(2)}\right)_{ij}\right] \\ \quad + \sum_{j=1}^{N_1} \mathbf{u}_1 (A_{\Gamma 1})_{ij} + \sum_{j=1}^{N_2} \mathbf{u}_2 (A_{\Gamma 2})_{ij} \\ \quad = \mathbf{f}_1^{(\Gamma)} + \mathbf{f}_2^{(\Gamma)} \forall i = 1,\dots,N_\Gamma. \end{cases}$$

... so that we obtain the algebraic form:

$$\begin{cases} A_{11}\mathbf{u}_1 + A_{1\Gamma}\boldsymbol{\lambda} = \mathbf{f}_1 \\ A_{22}\mathbf{u}_2 + A_{2\Gamma}\boldsymbol{\lambda} = \mathbf{f}_2 \\ A_{\Gamma 1}\mathbf{u}_1 + A_{\Gamma 2}\mathbf{u}_2 + \left(A_{\Gamma\Gamma}^{(1)} + A_{\Gamma\Gamma}^{(2)}\right)\boldsymbol{\lambda} = \mathbf{f}_1^{(\Gamma)} + \mathbf{f}_2^{(\Gamma)} \end{cases}$$

or

$$\begin{pmatrix} A_{11} & 0 & A_{1\Gamma} \\ 0 & A_{22} & A_{2\Gamma} \\ A_{\Gamma 1} & A_{\Gamma 2} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_\Gamma \end{pmatrix}$$

where we denoted $A_{\Gamma\Gamma} = \left(A_{\Gamma\Gamma}^{(1)} + A_{\Gamma\Gamma}^{(2)}\right)$ and $\mathbf{f}_\Gamma = \mathbf{f}_1^{(\Gamma)} + \mathbf{f}_2^{(\Gamma)}$.

3.4 The Schur Complement System

Since λ represents the unknown value of u on Γ, its finite element correspondent is the vector $\boldsymbol{\lambda}$ of the values of u_h at the interface nodes. By Gaussian elimination, we can obtain a new reduced system in the sole unknown $\boldsymbol{\lambda}$:

- Matrices A_{11} and A_{22} are invertible since they are associated with two homogeneous Dirichlet boundary-value problems for the Laplace operator:

$$\mathbf{u}_1 = A_{11}^{-1}(\mathbf{f}_1 - A_{1\Gamma}\boldsymbol{\lambda}) \quad \text{and} \quad \mathbf{u}_2 = A_{22}^{-1}(\mathbf{f}_2 - A_{2\Gamma}\boldsymbol{\lambda}). \tag{3.10}$$

- From the third equation we obtain:

$$\left[\left(A_{\Gamma\Gamma}^{(1)} - A_{\Gamma 1}A_{11}^{-1}A_{1\Gamma}\right) + \left(A_{\Gamma\Gamma}^{(2)} - A_{\Gamma 2}A_{22}^{-1}A_{2\Gamma}\right)\right]\boldsymbol{\lambda} = \mathbf{f}_\Gamma - A_{\Gamma 1}A_{11}^{-1}\mathbf{f}_1 - A_{\Gamma 2}A_{22}^{-1}\mathbf{f}_2.$$

Setting:

$$\Sigma = \Sigma_1 + \Sigma_2 \quad \text{with} \quad \Sigma_i = A_{\Gamma\Gamma}^{(i)} - A_{\Gamma i}A_{ii}^{-1}A_{i\Gamma} \quad (i = 1, 2),$$

and

$$\chi_\Gamma = \mathbf{f}_\Gamma - A_{\Gamma 1}A_{11}^{-1}\mathbf{f}_1 - A_{\Gamma 2}A_{22}^{-1}\mathbf{f}_2,$$

we obtain the Schur complement system:

$$\Sigma\lambda = \chi_\Gamma.$$

- Σ and χ_Γ approximate S and χ.
- Σ is the so-called Schur complement of A with respect to $\mathbf{u}_1$ and $\mathbf{u}_2$.
- Σ_i are the Schur complements related to the subdomains Ω_i $(i = 1, 2)$.

Remark

After solving the Schur complement system in $\boldsymbol{\lambda}$, thanks to (3.8) we can compute $\mathbf{u}_1$ and $\mathbf{u}_2$. This is equivalent to solving two Poisson problems in the subdomains Ω_1 and Ω_2 with the Dirichlet boundary condition ${u_h^{(i)}}_{|r} = \lambda_h (i = 1, 2)$ on the interface Γ.

Properties of the Schur Complement Σ

The Schur complement Σ inherits some of the properties of A :

- if A is singular, so is Σ;
- if A (respectively, A_{ii}) is symmetric, then Σ (respectively, Σ_i) is symmetric too;
- if A is positive definite, so is Σ.

Moreover, concerning the condition number, we have

- $\kappa(A) \simeq Ch^{-2}$
- $\kappa(\Sigma) \simeq Ch^{-1}$

Preconditioners for the Schur Complement System

The iterative methods that we have illustrated are equivalent to preconditioned Richardson methods for the Schur complement system with preconditioners:

- for the DN algorithm: $P_h = \Sigma_2$
- for the ND algorithm: $P_h = \Sigma_1$
- for the NN algorithm: $P_h = \left(\sigma_1 \Sigma_1^{-1} + \sigma_2 \Sigma_2^{-1}\right)^{-1}$
- for the RR algorithm: $P_h = (\gamma_1 + \gamma_2)^{-1} (\gamma_1 I + \Sigma_1) (\gamma_2 I + \Sigma_2)$

All these preconditioners are optimal in the sense of the following definition. We must anticipate that these optimality results do not hold in the case of multiple subdomains.

Definition

A preconditioner P is optimal for a matrix $A \in \mathbb{R}^{N \times N}$ if the condition number of $P^{-1}A$ is bounded uniformly with respect to the dimension N of A.

In particular, we have

- $\kappa\left(\Sigma_i^{-1}\Sigma\right) = O(1)$ (for $i = 1, 2$)
- $\kappa\left(\left(\sigma_1 \Sigma_1^{-1} + \sigma_2 \Sigma_2^{-1}\right)\Sigma\right) = O(1) \forall \sigma_1, \sigma_2 > 0$

Subdomain iterations and parallelism

With the exception of the Neumann–Neumann method, the different iterative procedures introduced thus far share the feature of generating at each step two boundary value problems, the former set in Ω_1, the latter in Ω_2, to be solved *sequentially*.

A simple modification of this procedure making it more interesting in view of parallel implementation is in order. As a matter of fact, when solving the boundary value problem in Ω_2 at the new step $k+1$ it is enough to use as data on Γ those generated by u_i^k (rather than u_i^{k+1}).

For instance, following this approach, the Dirichlet-Neumann method should be modified by simply replacing the Neumann condition on Γ by the new one

$$\frac{\partial u_2^{k+1}}{\partial n} = \frac{\partial u_1^k}{\partial n} \quad \text{on } \Gamma.$$

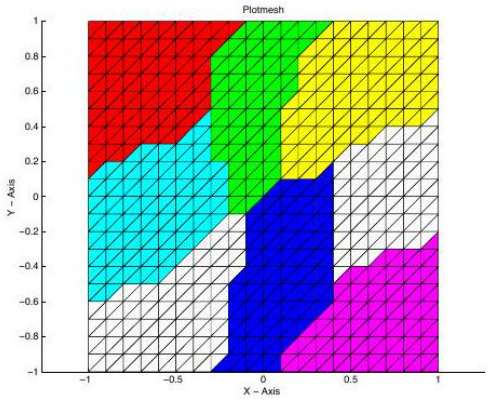
On the other hand, the issue of parallelism is relevant in the case of partitions of Ω using many (more than two) subdomains.

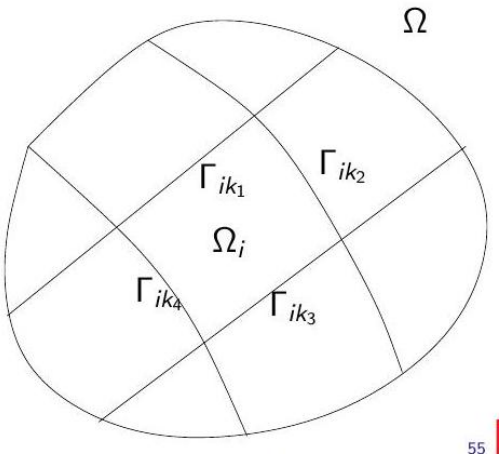
3.5 Nonoverlapping Multiple Subdomains

Multi-Domain Formulation for $M > 2$ Subdomains

We generalize now the nonoverlapping methods to the case of a domain Ω split into $M > 2$ subdomains:

- $\Omega_i (i = 1, \ldots, M)$ such that $\bigcup \bar{\Omega}_i = \bar{\Omega}$
- $\Gamma_i = \partial\Omega_i \backslash \partial\Omega$
- $\Gamma = \bigcup \Gamma_i$.





At the differential level, we have the equivalent multi-domain formulation:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \Leftrightarrow \begin{cases} -\triangle u_i = f & \text{in } \Omega_i \\ u_i = u_k & \text{on } \Gamma_{ik} \\ \frac{\partial u_i}{\partial n_i} = \frac{\partial u_k}{\partial n_i} & \text{on } \Gamma_{ik} \\ u_i = 0 & \text{on } \partial\Omega_i \cap \partial\Omega \end{cases}$$

where $\Gamma_{ik} = \partial\Omega_i \cap \partial\Omega_k \neq \emptyset$.

Finite Element Approximation

Considering a conforming finite element approximation, we obtain the linear system:

$$\begin{pmatrix} A_{II} & A_{I\Gamma} \\ A_{\Gamma I} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_I \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_I \\ \mathbf{f}_\Gamma \end{pmatrix} \tag{3.11}$$

where $\mathbf{u}_I$ is the vector of unknowns in the internal nodes and $\boldsymbol{\lambda}$ is the vector of unknowns on Γ: $\boldsymbol{\lambda} = \mathbf{u}_\Gamma$.
The submatrix A_{II} associated with the internal nodes is block-diagonal:

$$A_{II} = \begin{pmatrix} A_{11} & 0 & \ldots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & A_{MM} \end{pmatrix} \tag{3.12}$$

$A_{I\Gamma}$ is a banded matrix (interactions with local interfaces).

The Schur Complement System

$$\begin{aligned} A_{II}u_I + A_{I\Gamma}\lambda &= f_I \\ A_{\Gamma I}u_I + A_{\Gamma\Gamma}\lambda &= f_\Gamma \end{aligned}$$

These two equations imply:

$$\begin{aligned} u_I &= A_{II}^{-1}(f_I - A_{I\Gamma}\lambda) \\ A_{\Gamma I}(A_{II}^{-1}(f_I - A_{I\Gamma}\lambda)) + A_{\Gamma\Gamma}\lambda &= f_\Gamma \\ A_{\Gamma I}A_{II}^{-1}f_I - A_{\Gamma I}A_{II}^{-1}A_{I\Gamma}\lambda + A_{\Gamma\Gamma}\lambda &= f_\Gamma \\ (A_{\Gamma\Gamma} - A_{\Gamma I}A_{II}^{-1}A_{I\Gamma})\lambda &= f_\Gamma - A_{\Gamma I}A_{II}^{-1}f_I \end{aligned}$$

Denoting

$$\Sigma = A_{\Gamma\Gamma} - A_{\Gamma I}A_{II}^{-1}A_{I\Gamma}$$

and

$$\chi_\Gamma = \mathbf{f}_\Gamma - A_{\Gamma I}A_{II}^{-1}\mathbf{f}_I,$$

we obtain the Schur complement system in the multi-domain case:

$$\Sigma\boldsymbol{\lambda} = \chi_\Gamma.$$

Remarks

The local Schur complements are defined as:

$$\Sigma_i = A_{\Gamma_i\Gamma_i} - A_{\Gamma_i i}A_{ii}^{-1}A_{i\Gamma_i}$$

so that:

$$\Sigma = \Sigma_1 + \ldots + \Sigma_M$$

A Simple Algorithm

To compute a Finite Element (FE) approximation of the solution u of the Poisson problem:

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

we can follow these steps:

1. Solve the Schur complement system

   $$\Sigma\lambda = \chi_\Gamma$$

   to compute $\boldsymbol{\lambda}$ on the whole interface Γ;

2. Solve

   $$A_{II}\mathbf{u}_I = \mathbf{f}_I - A_{|\Gamma}\lambda$$

   i.e., M independent problems of reduced dimension

   $$A_{ii}\mathbf{u}_l^i = \mathbf{g}^i \quad (i = 1, \ldots, M)$$

   possibly in parallel.

Estimate of the Condition Number

The following estimate can be proved for the condition number of the Schur complement matrix Σ:

There exists a constant $C > 0$, independent of h and H, such that:

$$\kappa(\Sigma) \leq C \frac{H}{h H_{\min}^2}$$

where H and $H_{\min}$ are the maximal and minimal diameters of the subdomains, respectively.
To elaborate further, the diameter of each subdomain Ω_k is denoted as H_k, defined by:

$$H_k = \operatorname{diam}(\Omega_k)$$

We can also express the condition number of a matrix Σ as:

$$\operatorname{cond}(\Sigma) \simeq O\left(\frac{H_{\max}}{h H_{\min}^2}\right)$$

In the specific case where all H_k approximates to H, the condition number of Σ is approximated as:

$$\operatorname{cond}(\Sigma_1) \simeq O\left(\frac{1}{hH}\right)$$

Considering a scenario with two subdomains, we have:

$$H \cong \frac{1}{2} \operatorname{diam}(\Omega)$$
$$\operatorname{cond}(\Sigma) \simeq O\left(\frac{1}{h}\right)$$

When solving $\sum \vec{\lambda} = \vec{X}_\Gamma$ using the Richardson method without preconditioning, the convergence rate ρ is given by:

$$\rho = \frac{\operatorname{cond}(\Sigma) - 1}{\operatorname{cond}(\Sigma) + 1}$$

$$\Rightarrow \quad \rho = \rho\left(h^{-1}, H^{-1}\right)$$

This result indicates that the method is not optimal or scalable, as evidenced by the fact that increasing M leads to a decrease in H.

Therefore, to address these limitations, a parallel preconditioner is necessary. The ideal preconditioner should be both optimal and scalable.

Scalability

Definition A preconditioner P_h of Σ is said to be scalable if the condition number of the preconditioned matrix $P_h^{-1}\Sigma$ is independent of the number of subdomains.

Iterative methods using scalable preconditioners allow henceforth to achieve convergence rates independent of the subdomain number.

Dirichlet-Neumann Preconditioner

The Dirichlet-Neumann preconditioner is not widely used for many subdomains due to the necessity of a "black and white" subdivision. This subdivision requires alternating between Dirichlet (prescribed values) and Neumann (prescribed normal derivatives) boundary conditions on subdomain interfaces to ensure convergence. However, this strict subdivision can be challenging for complex geometries or numerous subdomains, leading to load imbalances and inefficiencies. As a result, alternative domain decomposition methods like the additive Schwarz or Balancing Neumann-Neumann techniques, which do not require this strict subdivision, are often preferred for problems with many subdomains or irregular decompositions. Anyway we can define the following preconditioning strategy, I will cite word for word the book by Quarteroni Valli Domain Decomposition Methods for Partial Differential Equations.

The Dirichlet-Neumann iterative substructuring method introduced for the differential boundary value problem can be formulated at the algebraic level as follows. We use a superscript B (black) and W (white) to denote the colour of the subdomain. Then the Schur complement matrix can be split as

$$\Sigma_h = \sum_{i \in I_B} \left(R_{\Gamma_i}^{(B)}\right)^T \Sigma_{i,h}^{(B)} R_{\Gamma_i}^{(B)} + \sum_{i \in I_W} \left(R_{\Gamma_i}^{(W)}\right)^T \Sigma_{i,h}^{(W)} R_{\Gamma_i}^{(W)}$$

and the Dirichlet-Neumann preconditioner is defined through

$$\left(P_h^{\mathrm{DN}}\right)^{-1} := \sum_{i \in I_W} \left(R_{\Gamma_i}^{(W)}\right)^T \left(\Sigma_{i,h}^{(W)}\right)^{-1} R_{\Gamma_i}^{(W)}$$

Also for this preconditioner it is possible to include a global coarse problem.

Neumann-Neumann Preconditioner

The Neumann-Neumann preconditioner for more subdomains reads:

$$\left(P_h^{NN}\right)^{-1} = \sum_{i=1}^{M} R_{\Gamma_i}^T D_i \Sigma_i^* D_i R_{\Gamma_i}$$

where Σ_i^* is either Σ_i^{-1} or an approximation of Σ_i^{-1}. D_i is a diagonal matrix of positive weights

$$D_i = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}$$

d_j is the number of subdomains that share the j-th node. We have the following estimate:

$$\kappa\left(\left(P_h^{NN}\right)^{-1}\Sigma\right) \leq CH^{-2}\left(1 + \log\frac{H}{h}\right)^2$$

The presence of D_i and R_Γ only entails matrix-matrix multiplications. On the other hand, if $\Sigma_i^* = \Sigma_i^{-1}$, applying Σ_i^{-1} to a given vector can be reconducted to the use

of local inverses. As a matter of fact, let $\mathbf{q}$ be a vector whose components are the nodal values on the local interface Γ_i; then

$$\Sigma_i^{-1}\mathbf{q} = [0, I]A_i^{-1}[0, I]^T\mathbf{q}.$$

In particular, $[0, I]^T\mathbf{q} = [0, \mathbf{q}]^T$, and the matrix-vector product

$$\underbrace{\left[\begin{array}{c|c} \text{internal} & \\ \text{nodes} & \\ & \end{array}\right]}_{A_i^{-1}} \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline \mathbf{q} \end{array}\right]$$

corresponds to the solution on Ω_i of the Neumann boundary-value problem:

$$\begin{cases} -\Delta w_i = 0 & \text{in } \Omega_i \\ \frac{\partial w_i}{\partial n} = q & \text{on } \Gamma_i \end{cases}$$

In essence, the operation $\Sigma_i^{-1}q$ (applying the inverse to vector q) can be interpreted as solving a Neumann boundary value problem in the subdomain Ω_i where the boundary condition on Γ_i is given by the vector q. This provides a way to apply the preconditioner by solving local PDE problems rather than explicitly inverting the matrix Σ_i.

Balanced Neumann-Neumann Preconditioner

The Neumann-Neumann preconditioner of the Schur complement system is not scalable. A substantial improvement can be achieved by adding a coarse grid correction:

$$\left(P_h^{BNN}\right)^{-1} = \Sigma_H^{-1} + \left(I - \Sigma_H^{-1}\Sigma\right)\left(P_h^{NN}\right)^{-1}\left(I - \Sigma\Sigma_H^{-1}\right).$$

where $\Sigma_H^{-1} = R_\Gamma^T A_H^{-1} R_\Gamma$. This is called balanced Neumann-Neumann preconditioner. We can prove that:

$$\kappa\left(\left(P_h^{BNN}\right)^{-1}\Sigma\right) \leq C\left(1 + \log\frac{H}{h}\right)^2.$$

If the original matrix was symmetric positive definite this is an almost (except for a logarithmic term) scalable and optimal preconditioner, which can be used also with the conjugate gradient method if the original matrix satisfy the applicability conditions.

Convergence Properties

For the two-subdomain cases of the Laplace problem with homogeneous Dirichlet boundary conditions, both Dirichlet–Neumann and Neumann–Neumann iterations converge at a rate independent of h. The Neumann–Neumann algorithm has no particular advantage over the Dirichlet–Neumann algorithm (and actually the former requires more subdomain solves per iteration than the latter). The situation can, however, be different in the case of many subdomains.

The need of pseud-inverses in multiple-subdomains decomposition

In a domain decomposition method, when the computational domain is divided into only two subdomains, each subdomain will indeed inherit part of the original problem's boundary. Typically, the original problem has Dirichlet boundary conditions (fixed values) on at least part of its boundary. Therefore, in a two-subdomain decomposition, each subdomain ends up with a portion of the boundary where Dirichlet conditions apply and the interface between the two subdomains, where Neumann conditions (or, more precisely, continuity conditions that can be mathematically similar to Neumann conditions) are enforced. This mixture of boundary conditions (Dirichlet on the outer boundary and Neumann-like on the interface) helps ensure that each subdomain problem remains well-posed, meaning that it has a unique solution.

However, when a domain is decomposed into more than two subdomains, especially in complex geometries or in 3D, some subdomains may end up without any part of the original boundary. This means that they are entirely enclosed by interfaces with other subdomains and thus are only subject to Neumann-like conditions derived from those interfaces. Such subdomains can lead to ill-posed problems because Neumann problems require additional conditions (like a fixed integral value of the solution over the domain) to ensure uniqueness of the solution. Without any Dirichlet boundary conditions (which help pin down the solution), these purely Neumann subdomains can suffer from indeterminacy issues (e.g., the solution might be determined only up to an additive constant).

The need for a pseudoinverse, rather than a straightforward matrix inverse, in these cases stems from this potential lack of well-posedness. So in this context, waht we indicated as inverse Σ_H^{-1} must be interpreted as a regularized inverse of the origimal matrix, and not the strict inverse which might not even exist.

Concluding Remarks

From the numerical results that we have presented, we can conclude with the following remarks:

- Even if better conditioned with respect to A, Σ is ill-conditioned, and therefore a suitable preconditioner must be applied.
- The Neumann-Neumann preconditioner can be satisfactorily applied using a moderate number of subdomains, while for larger M, $\kappa\left(\left(P_h^{NN}\right)^{-1}\Sigma\right) > \kappa(\Sigma)$.
- The balancing Neumann-Neumann preconditioner is almost optimally scalable and therefore recommended for partitions with a large number of subdomains.

Chapter 4

Navier Stokes

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

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} - \operatorname{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right)\right] + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}, & \mathbf{x} \in \Omega, t > 0 \\ \operatorname{div} \mathbf{u} = 0, & \mathbf{x} \in \Omega, t > 0 \end{cases} \tag{4.1}$$

where:

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

The first equation is that of conservation of linear momentum, the second one that of conservation of mass, which is also called the continuity equation.

- The term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ describes the process of convective transport.
- The term $-\operatorname{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right)\right]$ describes the molecular diffusion process.

When ν is constant, from the continuity equation we obtain:

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

whence system (4.1) can be written in the equivalent form:

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}, & \mathbf{x} \in \Omega, t > 0 \\ \operatorname{div} \mathbf{u} = 0, & \mathbf{x} \in \Omega, t > 0 \end{cases} \tag{4.2}$$

which is the one that we will consider in the following.

Equations (4.2) are often called incompressible Navier-Stokes equations. More in general, fluids satisfying the incompressibility condition $\operatorname{div} \mathbf{u} = 0$ are said to be incompressible.

Constant density fluids necessarily satisfy this condition, however there exist incompressible fluids featuring variable density (e.g., stratified fluids) that are governed by a different system of equations in which the density ρ explicitly shows up.

In order for problem (4.2) to be well posed it is necessary to assign the initial condition:

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega \tag{4.3}$$

where $\mathbf{u}_0$ is a given divergence-free vector field, together with suitable boundary conditions, such as, e.g., $\forall t > 0$,

$$\begin{cases} \mathbf{u}(\mathbf{x}, t) = \varphi(\mathbf{x}, t) \quad \forall \mathbf{x} \in \Gamma_D, \\ \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n}\right)(\mathbf{x}, t) = \psi(\mathbf{x}, t) \quad \forall \mathbf{x} \in \Gamma_N, \end{cases} \tag{4.4}$$

where φ and ψ are given vector functions, while Γ_D and Γ_N provide a partition of the domain boundary $\partial\Omega$, that is $\Gamma_D \cup \Gamma_N = \partial\Omega, \Gamma_D^\circ \cap \Gamma_N^\circ = \emptyset$.

Finally, as usual $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$.

If we use the alternative formulation (4.1), the second equation in (4.4) must be replaced by:

$$\left[\nu \left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right) \mathbf{n} - p\mathbf{n}\right](\mathbf{x}, t) = \psi(\mathbf{x}, t) \quad \forall \mathbf{x} \in \Gamma_N$$

Denoting with $u_i, i = 1, \ldots, d$ the components of the vector $\mathbf{u}$ with respect to a Cartesian frame, and with f_i the components of $\mathbf{f}$, system (4.2) can be written componentwise as

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

4.1 Well-posedness

In the two-dimensional case the Navier-Stokes equations with the boundary conditions previously indicated yield well-posed problems. This means that if all data (initial condition, forcing term, boundary data) are smooth enough, then the solution is continuous together with its derivatives and does not develop singularities in time.

Things may go differently in three dimensions, where existence and uniqueness of classical solutions have been proven only locally in time (that is for a sufficiently small time interval). In the following slides we will introduce the weak formulation of the Navier-Stokes equations, for which existence of a solution has been proven for all times. However, the issue of uniqueness (which is related to that of regularity) is still open, and is actually the central issue of Navier-Stokes theory.

4.2 Alternative formulations

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

are related to the velocity as follows:

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

The various formulations are in fact equivalent from a mathematical standpoint, although they give rise to different numerical methods.

References

- B.F. Smith, P.E. Bjørstad, W.D. Gropp (1996) *Domain Decomposition.* Cambridge University Press, Cambridge.
- A. Quarteroni and A. Valli (1999) *Domain Decomposition Methods for Partial Differential Equations.* Oxford Science Publications, Oxford.
- A. Toselli and O.B. Widlund (2005) *Domain Decomposition Methods – Algorithms and Theory.* Springer-Verlag, Berlin and Heidelberg.

4.3 Weak formulation of Navier-Stokes equations

A weak formulation of problem (4.2) can be obtained by proceeding formally, as follows. Let us multiply the first equation of (4.2) by a test function $\mathbf{v}$ belonging to a suitable space V that will be specified later on, and integrate in Ω

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

Using Green's formulae [ref. NMDP, 3.16 & 3.17] we find:

$$-\int_\Omega \nu \Delta \mathbf{u} \cdot \mathbf{v} d\Omega = \int_\Omega \nu \nabla \mathbf{u} \colon \nabla \mathbf{v} d\Omega - \int_{\partial\Omega} \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \mathbf{v} d\gamma$$

$$\int_\Omega \nabla p \cdot \mathbf{v} d\Omega = -\int_\Omega p \,\text{div}\, \mathbf{v} d\Omega + \int_{\partial\Omega} p \mathbf{v} \cdot \mathbf{n} d\gamma$$

Using these relations in the first of (4.2), we obtain:

$$\begin{aligned} &\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} d\Omega + \int_\Omega \nu \nabla \mathbf{u} \colon \nabla \mathbf{v} d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v} d\Omega \\ &- \int_\Omega p \,\text{div}\, \mathbf{v} d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} d\Omega + \int_{\partial\Omega} \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} \right) \cdot \mathbf{v} d\gamma \quad \forall \mathbf{v} \in V. \end{aligned} \tag{4.5}$$

Remark
All boundary integrals should indeed be regarded as duality pairings between V' and V.
Similarly, by multiplying the second equation of (4.2) by a test function q, belonging to a suitable space Q to be specified, then integrating on Ω it follows:

$$\int_\Omega q \,\text{div}\, \mathbf{u} \, d\Omega = 0 \quad \forall q \in Q. \tag{4.6}$$

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

$$V = \left[\mathrm{H}^1_{\Gamma_D}(\Omega)\right]^d = \left\{\mathbf{v} \in \left[\mathrm{H}^1(\Omega)\right]^d : \mathbf{v}|_{\Gamma_D} = \mathbf{0}\right\} \tag{4.7}$$

It will coincide with $\left[\mathrm{H}^1_0(\Omega)\right]^d$ if $\Gamma_D = \partial\Omega$. If Γ_N has positive measure, we can choose $Q = \mathrm{L}^2(\Omega)$. If $\Gamma_D = \partial\Omega$, then the pressure space should be L^2_0 to ensure uniqueness for the pressure p.

Moreover, if $t > 0$, then $\mathbf{u}(t) \in \left[\mathrm{H}^1(\Omega)\right]^d$, with $\mathbf{u}(t) = \varphi(t)$ on Γ_D, $\mathbf{u}(0) = \mathbf{u}_0$ and $p(t) \in Q$.

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

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

its norm and by

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

its seminorm.

The notation $\|\mathbf{v}\|_{L^p(\Omega)}, 1 \le p < \infty$, has a similar meaning. The same symbols will be used in case of tensor functions. Thanks to Poincaré's inequality, $|\mathbf{v}|_{\mathbf{H}^1(\Omega)}$ is equivalent to the norm $\|\mathbf{v}\|_V$ for all functions belonging to V, provided that the Dirichlet boundary has a positive measure.

Having chosen these functional spaces, we can note first of all that:

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

All the integrals involving bilinear terms are finite. More precisely, by using the vector notation $\mathbf{H}^k(\Omega) = \left[\mathrm{H}^k(\Omega)\right]^d, \mathbf{L}^p(\Omega) = \left[\mathrm{L}^p(\Omega)\right]^d, k \ge 1, 1 \le p < \infty$, we find:

$$\left|\nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\Omega\right| \le \nu |\mathbf{u}|_{\mathbf{H}^1(\Omega)} |\mathbf{v}|_{\mathbf{H}^1(\Omega)},$$
$$\left|\int_\Omega p \nabla \cdot \mathbf{v} d\Omega\right| \le \|p\|_{\mathrm{L}^2(\Omega)} |\mathbf{v}|_{\mathbf{H}^1(\Omega)},$$
$$\left|\int_\Omega q \nabla \mathbf{u} d\Omega\right| \le \|q\|_{\mathrm{L}^2(\Omega)} |\mathbf{u}|_{\mathbf{H}^1(\Omega)}.$$

Also the integral involving the trilinear term is finite. To see how, let us start by recalling the following result.

Proposition 2. *If* $d \leq 3$, $\forall \mathbf{v} \in \mathbf{H}^1(\Omega)$, *then* $\mathbf{v} \in \mathbf{L}^4(\Omega)$ *and* $\exists C > 0$ *s.t.* $\|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \leq C\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}$.

Using the following three-term Hölder inequality:

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

valid for all $p, q, r > 1$ such that $p^{-1} + q^{-1} + r^{-1} = 1$, we conclude by applying the usual inclusion theorem that:

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

4.4 Solution Uniqueness

As for the solution's uniqueness, let us consider again the Navier-Stokes equations in strong form (4.2) (similar considerations can be made on the weak form).

If $\Gamma_D = \partial\Omega$, when only boundary conditions of Dirichlet type are imposed, the pressure appears merely in terms of its gradient; in such a case, if we call $(\mathbf{u}, p)$ a solution of (4.2), for any possible constant c the couple $(\mathbf{u}, p + c)$ is a solution too, since $\nabla(p + c) = \nabla p$.

To avoid such indeterminacy one can fix a priori the value of p at one given point $\mathbf{x}_0$ of the domain Ω, that is set $p(\mathbf{x}_0) = p_0$, or, alternatively, require the pressure to have null average, i.e., $\int_\Omega p d\Omega = 0$.

The former condition requires to prescribe a pointwise value for the pressure, but this is inconsistent with our ansatz that $p \in \mathrm{L}^2(\Omega)$. (We anticipate, however, that this is admissible at the numerical level when we look for a continuous finite-dimensional pressure).

Prescribing a pointwise value for the pressure is therefore inconsistent with the assumption that p is an L^2 function. This kind of pointwise condition is more appropriate for functions in spaces that include continuity (like continuous functions or, in a numerical setting, continuous finite-dimensional approximations).

For this reason we assume from now on that the pressure is average-free. More specifically, we will consider the following pressure space:

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

Further, we observe that if $\Gamma_D = \partial\Omega$, the prescribed Dirichlet data φ must be compatible with the incompressibility constraint; indeed,

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

If Γ_N is not empty, i.e. in presence of either Neumann or mixed Dirichlet-Neumann boundary conditions, the problem of pressure indeterminacy (up to an additive constant) no longer exists. In this case we can take $\mathrm{Q} = \mathrm{L}^2(\Omega)$.

In conclusion, from now on we shall implicitly assume:

$$Q = \mathrm{L}^2(\Omega) \quad \text{if} \quad \Gamma_N \neq \emptyset, \quad Q = \mathrm{L}^2_0(\Omega) \quad \text{if } \Gamma_N = \emptyset \tag{4.8}$$

The weak formulation of the system (4.2), is therefore:
find $\mathbf{u} \in \mathrm{L}^2\left(\mathbb{R}^+; \left[\mathrm{H}^1(\Omega)\right]^d\right) \cap C^0\left(\mathbb{R}^+; \left[\mathrm{L}^2(\Omega)\right]^d\right), p \in \mathrm{L}^2\left(\mathbb{R}^+; Q\right)$ such that:

$$\begin{cases} \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} d\Omega + \nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v} d\Omega \\ - \int_\Omega p \operatorname{div} \mathbf{v} d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} d\Omega + \int_{\Gamma_N} \psi \cdot \mathbf{v} d\gamma \quad \forall \mathbf{v} \in V, \\ \int_\Omega q \operatorname{div} \mathbf{u} d\Omega = 0 \quad \forall q \in Q, \end{cases} \tag{4.9}$$

with $\mathbf{u}|_{\Gamma_D} = \varphi_D$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$. The space V is the one in (4.7) while Q is the space introduced in (4.8). Let's also keep in mind that we are requiring continuity in time.

As we have already anticipated, existence of solutions can be proven for this problem for both dimensions $d = 2$ and $d = 3$, whereas uniqueness has been proven only in the case $d = 2$ for sufficiently small data.

4.5 The Reynolds number

Let us define the Reynolds number:

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

where L is a representative length of the domain Ω (e.g. the length of a channel where the fluid's flow is studied) and $\mathbf{U}$ a representative fluid velocity.

The Reynolds number measures the extent to which convection dominates over diffusion.

- When $\mathrm{Re} \ll 1$ the convective term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ can be omitted, and the Navier-Stokes equations reduce to the so-called Stokes equations, that will be investigated later.

- On the other hand, if *Re* is large, problems may arise concerning uniqueness of the solution, the existence of stationary and stable solutions, the possible existence of strange attractors, the transition towards turbulent flows.

4.6 Stokes equations and their approximation

In this section we will consider the following generalized Stokes problem with homogeneous Dirichlet boundary conditions:

$$\begin{cases} \sigma\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \\ \operatorname{div}\mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega \end{cases} \tag{4.10}$$

for a given coefficient $\sigma \geq 0$.

This problem describes the motion of an incompressible viscous flow in which the (quadratic) convective term has been neglected, a simplification that is acceptable when $\mathrm{Re} \ll 1$.

Moreover, one can generate a problem like (4.10) also while using an implicit temporal discretization of the Navier-Stokes equations and by neglecting the convective term (i.e. $(\mathbf{u}\cdot\nabla)\mathbf{u}$).

We have indeed the following scheme, where k denotes the temporal index:

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

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

$$\begin{cases} \sigma\mathbf{u}^k - \nu\Delta\mathbf{u}^k + \nabla p^k = \tilde{\mathbf{f}}^k & \text{in } \Omega, \\ \operatorname{div}\mathbf{u}^k = 0 & \text{in } \Omega \\ + \text{ B.C.} \end{cases} \tag{4.11}$$

where $\sigma = (\Delta t)^{-1}$ and $\tilde{\mathbf{f}}^k = \mathbf{f}\left(t^k\right) + \frac{\mathbf{u}^{k-1}}{\Delta t}$.

The weak formulation of problem (4.10) reads:

Find $\mathbf{u} \in V$ and $p \in Q$ such that

$$\begin{cases} \int_\Omega (\sigma\mathbf{u}\cdot\mathbf{v} + \nu\nabla\mathbf{u}\cdot\nabla\mathbf{v})d\Omega - \int_\Omega p \operatorname{div}\mathbf{v} d\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v} d\Omega & \forall \mathbf{v} \in V, \\ \int_\Omega q \operatorname{div}\mathbf{u} d\Omega = 0 & \forall q \in Q, \end{cases} \tag{4.12}$$

where $V = \left[\mathrm{H}_0^1(\Omega)\right]^d$ and $Q = \mathrm{L}_0^2(\Omega)$.

In the weak formulation of fluid dynamics problems like the Stokes problem, a key transformation occurs with the term $\nabla p \cdot \mathbf{v}$ from the strong form. This term is modified to $p\nabla\cdot\mathbf{v}$ in the weak form through integration by parts. The process is as follows:

Integrating the term $\nabla p \cdot \mathbf{v}$ over the domain Ω. This step involves applying integration by parts, which transforms the integral into a different form. The integration

by parts formula states that $\int_\Omega \nabla p \cdot \mathbf{v}\, d\Omega = -\int_\Omega p\, \nabla \cdot \mathbf{v}\, d\Omega + \int_{\partial\Omega} p\, \mathbf{v} \cdot \mathbf{n}\, dS$, where $\partial\Omega$ denotes the boundary of Ω and $\mathbf{n}$ is the outward unit normal on the boundary. However, due to Dirichlet boundary conditions, the boundary term $\int_{\partial\Omega} p\, \mathbf{v} \cdot \mathbf{n}\, dS$ vanishes. Thus, the result simplifies to $\int_\Omega \nabla p \cdot \mathbf{v}\, d\Omega = -\int_\Omega p\, \nabla \cdot \mathbf{v}\, d\Omega$.

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_\Omega (\sigma \mathbf{u} \cdot \mathbf{v} + \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}) d\Omega, \\ b(\mathbf{u}, q) &= -\int_\Omega q \,\mathrm{div}\, \mathbf{u} d\Omega. \end{aligned} \tag{4.13}$$

With these notations, problem (4.12) becomes: find $(\mathbf{u}, p) \in V \times Q$ such that:

$$\begin{cases} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) & \forall \mathbf{v} \in V, \\ b(\mathbf{u}, q) = 0 & \forall q \in Q, \end{cases} \tag{4.14}$$

where $(\mathbf{f}, \mathbf{v}) = \sum_{i=1}^d \int_\Omega f_i v_i d\Omega$.

If we consider non-homogeneous boundary conditions, as indicated in (4.4), the weak formulation of the Stokes problem becomes: find $\left(\overset{\circ}{\mathbf{u}}, p\right) \in V \times Q$ such that:

$$\begin{cases} a(\overset{\circ}{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, p) = \mathbf{F}(\mathbf{v}) & \forall \mathbf{v} \in V, \\ b(\overset{\circ}{\mathbf{u}}, q) = G(q) & \forall q \in Q, \end{cases} \tag{4.15}$$

where V and Q are the spaces introduced in (4.7) and (4.8), respectively.

Having denoted with $\mathbf{R}\varphi \in \left[\mathrm{H}^1(\Omega)\right]^d$ a lifting of the boundary datum φ, we have set $\overset{\circ}{\mathbf{u}} = \mathbf{u} - \mathbf{R}\varphi$, while the new terms on the right-hand side have the following expression:

$$\mathbf{F}(\mathbf{v}) = (\mathbf{f}, \mathbf{v}) + \int_{\Gamma_N} \psi \mathbf{v} d\gamma - a(\mathbf{R}\varphi, \mathbf{v}), \quad G(q) = -b(\mathbf{R}\varphi, q) \tag{4.16}$$

4.7 Galerkin Approximation

The Galerkin approximation of problem (4.14) has the following form: find $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ such that

$$\begin{cases} a\left(\mathbf{u}_h, \mathbf{v}_h\right) + b\left(\mathbf{v}_h, p_h\right) = (\mathbf{f}, \mathbf{v}_h) & \forall \mathbf{v}_h \in V_h \\ b\left(\mathbf{u}_h, q_h\right) = 0 & \forall q_h \in Q_h \end{cases} \tag{4.17}$$

where $\{V_h \subset V\}$ and $\{Q_h \subset Q\}$ represent two families of finite-dimensional subspaces depending on a real discretization parameter h.

If, instead, we consider problem corresponding to non-homogeneous boundary data, the above formulation needs to be modified by using $\mathbf{F}\left(\mathbf{v}_h\right)$ on the right-hand side of the first equation and $G\left(q_h\right)$ on that of the second equation. These new functionals can be obtained from (4.16) by replacing $\mathbf{R}\varphi$ with $\mathbf{R}_h\varphi$, and replacing $\boldsymbol{\psi}$ with its interpolant at the nodes sitting on Γ_N, ψ_h.

4.8 Existence and Uniqueness

The following celebrated theorem is due to F. Brezzi, and guarantees uniqueness and existence for problem (4.17):

Theorem 5 (Existence and Uniqueness)**.** *The Galerkin approximation (4.17) admits one and only one solution if the following conditions hold:*

- *The bilinear form $a(\cdot,\cdot)$ is:*

  a) coercive, that is $\exists \alpha > 0$ (possibly depending on h) such that

$$a(\mathbf{v}_h, \mathbf{v}_h) \geq \alpha \|\mathbf{v}_h\|_V^2 \quad \forall \mathbf{v}_h \in V_h^*,$$

  where $V_h^* = \{\mathbf{v}_h \in V_h : b(\mathbf{v}_h, q_h) = 0 \ \forall q_h \in Q_h\}$;
  b) continuous, that is $\exists \gamma > 0$ such that

$$|a(\mathbf{u}_h, \mathbf{v}_h)| \leq \gamma \|\mathbf{u}_h\| \, v \, \|\mathbf{v}_h\| \, v \quad \forall \mathbf{u}_h, \mathbf{v}_h \in V_h$$

- *The bilinear form $b(\cdot,\cdot)$ is continuous, that is $\exists \delta > 0$ such that*

$$|b(\mathbf{v}_h, q_h)| \leq \delta \|\mathbf{v}_h\| \, v \, \|q_h\|_Q \quad \forall \mathbf{v}_h \in V_h, q_h \in Q_h$$

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

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

Under the previous assumptions the discrete solution fulfills the following a-priori estimates:

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

where V' is the dual space of V.
Moreover, the following convergence results hold:

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

It is worth noticing that condition on β is equivalent to the existence of a positive constant β such that:

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

For such a reason it is often called the inf-sup condition.

4.9 Algebraic formulation of the Stokes problem

Let us investigate the structure of the algebraic system associated to the Galerkin approximation (4.17) to the Stokes problem. Denote with:

$$\{\varphi_j \in V_h\}, \quad \{\phi_k \in Q_h\}$$

the basis functions of the spaces V_h and Q_h, respectively.

Let us expand the discrete solutions $\mathbf{u}_h$ and p_h with respect to such bases,

$$\mathbf{u}_h(\mathbf{x}) = \sum_{j=1}^{N} u_j \varphi_j(\mathbf{x}), \quad p_h(\mathbf{x}) = \sum_{k=1}^{M} p_k \phi_k(\mathbf{x}) \tag{4.18}$$

having set $N = \dim V_h$ and $M = \dim Q_h$.

By choosing as test functions in (4.18) the same basis functions we obtain the following block linear system:

$$\begin{cases} \mathrm{A}\mathbf{U} + \mathrm{B}^T\mathbf{P} = \mathbf{F} \\ \mathrm{B}\mathbf{U} = \mathbf{0} \end{cases} \tag{4.19}$$

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

$$\mathrm{A} = [a_{ij}] = [a(\varphi_j, \varphi_i)], \quad \mathrm{B} = [b_{km}] = [b(\varphi_m, \phi_k)]$$

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

$$\mathbf{U} = [u_j], \quad \mathbf{P} = [p_j]$$

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

$$\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & 0 \end{bmatrix} \tag{4.20}$$

is block symmetric (as $\mathbf{A}$ is symmetric) and indefinite, featuring real eigenvalues with variable sign (either positive and negative).

Proposition 3. $\mathbf{S}$ *is non-singular if and only if no eigenvalue is null, iff* $\ker(\mathbf{B}^T) = 0$.

Proof. Let's remember that $\mathbf{A}$ is non-singular, since it is associated with the coercive bilinear form $a(\cdot,\cdot)$. From the first of (4.19) we can formally obtain $\mathbf{U}$ as:

$$\mathbf{U} = \mathbf{A}^{-1}(\mathbf{F} - \mathbf{B}^\top\mathbf{P}) \tag{4.21}$$

Using (4.21) in the second equation of (4.19) yields:

$$\mathbf{R}\mathbf{P} = \mathbf{B}\mathbf{A}^{-1}\mathbf{F}, \quad \text{where} \quad \mathbf{R} = \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$$

This corresponds to having carried out a block Gaussian elimination on system (4.20).

In this way, we obtain a reduced system for the sole unknown $\mathbf{P}$ (the pressure), which admits a unique solution in case $\mathbf{R}$ is non-singular. Since $\mathbf{A}$ is non-singular and positive definite, we want to prove that the latter condition is satisfied if and only if $\mathbf{B}^\top$ has a null kernel, that is

$$\ker(\mathbf{B}^\top) = \{\mathbf{0}\} \tag{4.22}$$

where $\ker(\mathbf{B}^\top) = \left\{\mathbf{x} \in \mathbb{R}^M : \mathbf{B}^\top \mathbf{x} = \mathbf{0}\right\}$.

We proceed as follows:

$$\mathbf{Rp} = 0 \implies p = 0 \quad \text{(uniqueness)}$$

that is

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

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

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

Set $\mathbf{w} = \mathbf{B}^T\mathbf{p}$. Since $\mathbf{A}$ is spd, we have

$$\langle \mathbf{A}^{-1}\mathbf{w}, \mathbf{w}\rangle = 0 \implies w = 0$$

Finally,

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

□

Proposition 4. *Equivalency between inf-sup condition and full rank condition of* B^T
Condition (4.22) is equivalent to the inf-sup condition.

Proof. Note that condition (4.22) is violated iff $\exists \mathbf{p}^* \neq \mathbf{0}$ with $\mathbf{p}^* \in \mathbb{R}^M$ such that $\mathrm{B}^T\mathbf{p}^* = \mathbf{0}$ or, equivalently, if $\exists p_h^* \in \mathbb{Q}_h$ such that $b(\varphi_n, p_h^*) = 0 \ \forall n = 1, \ldots, N$. This is equivalent to $b(\mathbf{v}_h, p_h^*) = 0 \quad \forall \mathbf{v}_h \in V_h$, which in turn is equivalent to violating the inf-sup condition. Indeed:

$$\exists \beta_h > 0 \ \forall \mathbf{q}_h \in Q_h \ \exists \mathbf{v}_h \text{ in } X_h : \frac{b(\mathbf{v}_h, \mathbf{q}_h)}{\|\mathbf{v}_h\| \, v \, \|\mathbf{q}_h\|_Q} \geq \beta_h$$

is violated if

$$\forall \beta_h > 0 \ \exists \mathbf{p}_h^* \in Q_h \ \forall \mathbf{v}_h \text{ in } X_h : \frac{b(\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\| \, v \, \|\mathbf{p}_h^*\|_Q} < \beta_h$$

Take now $-\mathbf{v}_h$:

$$\frac{b(-\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\| \, v \, \|\mathbf{p}_h^*\|_Q} = -\frac{b(\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\| \, v \, \|\mathbf{p}_h^*\|_Q} < \beta_h$$

Then

$$-\beta_h < \frac{b\left(\mathbf{v}_h, \mathbf{p}_h^*\right)}{\|\mathbf{v}_h\| \, v \, \left\|\mathbf{p}_h^*\right\|_Q} < \beta_h$$

Because of the arbitrariness of β_h we conclude that $b\left(\mathbf{v}_h, \mathbf{p}_h^*\right) = 0 \; \forall \mathbf{v}_h \in X_h$. On the other hand, since A is non-singular, from the existence and uniqueness of $\mathbf{P}$ we infer that there exists a unique vector $\mathbf{U}$ which satisfies (4.21), and so the inf-sup condition does hold. In conclusion, system (4.22) admits a unique solution (U, P) if and only if condition (4.22) holds.

□

Remark We recall that, for an arbitrary matrix $B^T \in \mathbb{R}^{N \times M}$, we have rank $\left(B^T\right)$ + dim ker $\left(B^T\right) = \min(M, N)$.

Then, condition (4.22) is equivalent to asking that B^T (and consequently B) has full rank, i.e. that rank $\left(B^T\right) = \min(N, M)$, because rank $\left(B^T\right)$ is the maximum number of linearly independent row vectors (or, equivalently, column vectors) of B^T.

Let us consider again the remark about spurious pressure modes concerning the general saddle-point problem and suppose that the inf-sup condition does not hold. Then:

$$\exists q_h^* \in Q_h : \quad b\left(\mathbf{v}_h, q_h^*\right) = 0 \quad \forall \mathbf{v}_h \in V_h. \tag{4.23}$$

Consequently, if $(\mathbf{u}_h, p_h)$ is a solution to the Stokes problem (4.17), then $(\mathbf{u}_h, p_h + q_h^*)$ is a solution too, as:

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

Functions q_h^* which fail to satisfy the inf-sup condition are invisible to the Galerkin problem(4.17). For this reason, as already observed, they are called spurious pressure modes, or even parasitic modes. Their presence inhibits the pressure solution from being unique, yielding numerical instabilities. For this reason, those finite-dimensional subspaces that violate the compatibility condition are said to be unstable, or incompatible.

Two strategies are generally adopted in order to guarantee well-posedness of the numerical problem:

- choose spaces V_h and Q_h that satisfy the inf-sup condition;
- stabilize (either a priori or a posteriori) the finite dimensional problem by eliminating the spurious modes.

4.10 Compatible couples of spaces

Let us analyze the first type of strategy. To start with, we will consider the case of finite element spaces. To characterize Q_h and V_h it suffices to choose on every element

of the triangulation their degrees of freedom. Since the weak formulation does not require a continuous pressure, we will consider first the case of discontinuous pressures.

As Stokes equations are of order one in p and order two in $\mathbf{u}$, generally speaking it makes sense to use piecewise polynomials of degree $k \geq 1$ for the velocity space V_h and of degree $k-1$ for the space Q_h.

In particular, we might want to use piecewise linear finite elements $\mathbb{P}_1$ for each velocity component, and piecewise constant finite elements $\mathbb{P}_0$ for the pressure. In fact, this choice, although being quite natural, does not pass the inf-sup test (4.19).

When looking for a compatible couple of spaces, the larger the velocity space V_h, the more likely the inf-sup condition is satisfied. Otherwise said, the space V_h should be "rich" enough compared to the space Q_h.

In the following pictures, by means of the symbol $\square$ we indicate the degrees of freedom for the pressure, whereas the symbol $\bullet$ identifies those for each velocity component.

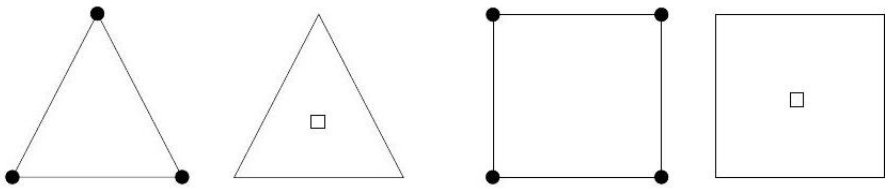


Figure 4.1: Case of discontinuous pressure: choices that do not satisfy the inf-sup condition, on triangles (left), and on quadrilaterals (right)

In this figure we report three different choices of spaces that fulfill the inf-sup condition, still in the case of continuous velocity and discontinuous pressure. Choice (left) is made by $\mathbb{P}_2 - \mathbb{P}_0$ elements, (center) by $\mathbb{Q}_2 - \mathbb{P}_0$ elements, while choice (right) by piecewise linear discontinuous elements for the pressure, while the velocity components are made by piecewise quadratic continuous elements enriched by a cubic bubble function on each triangle - these are the so-called Crouzeix-Raviart elements.

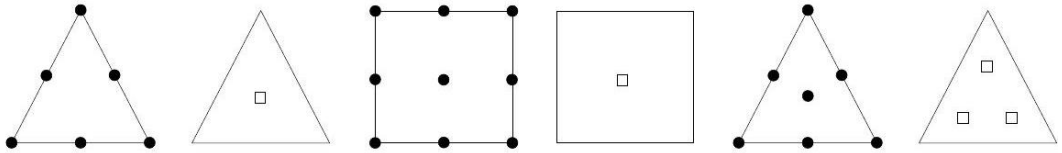


Figure 4.2: Case of discontinuous pressure: choices that do satisfy the inf-sup condition: on triangles, (left), and on quadrilaterals, (center). Also the couple (right), known as Crouzeix-Raviart elements, satisfies the inf-sup condition

In this figure, we report two choices of incompatible finite elements in the case of continuous pressure. They consist of piecewise linear elements on triangles (respectively, bilinear on quadrilaterals) for both velocity and pressure. More generally,

finite elements of the same polynomial degree $k \geq 1$ for both velocity and pressure are unstable (equal order interpolation).

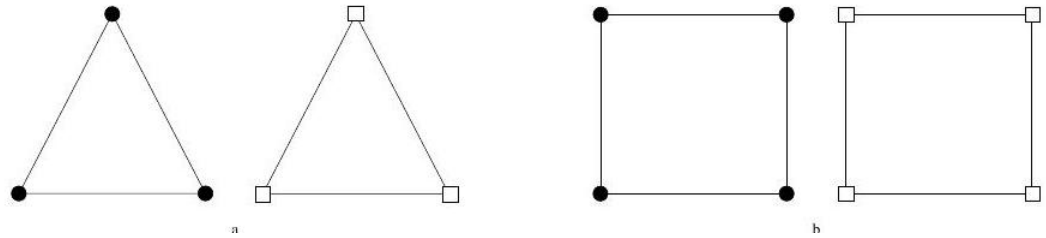


Figure 4.3: Case of continuous pressure: the couples displayed in this figure do not satisfy the inf-sup condition.

In this figure, the elements displayed are instead stable. In both cases, pressure is a piecewise linear continuous function, whereas velocities are piecewise linear polynomials on each of the four sub-triangles (left), or piecewise linear polynomials enriched by a cubic bubble function (right).

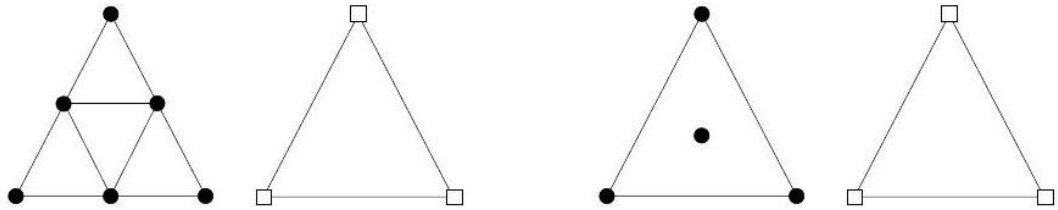


Figure 4.4: Case of continuous pressure: the elements used for the velocity components in (left) are known as $\mathbb{P}_1$-iso $\mathbb{P}_2$ finite elements, whereas couple (right) is called mini-element

The pair $\mathbb{P}_2 - \mathbb{P}_1$ (continuous piecewise quadratic velocities and continuous piecewise linear pressure) is stable. This is the smallest degree representative of the family of the so-called Taylor-Hood elements $\mathbb{P}_k - \mathbb{P}_{k-1}, k \geq 2$ (continuous velocities and continuous pressure), that are inf-sup stable. They yield the following error estimate:

$$\|\mathbf{u} - \mathbf{u}_h\|_{[H^1(\Omega)]^d} + \|p - p_h\|_{L^2(\Omega)} \leq Ch^k \left(|\mathbf{u}|_{\left[H^{k+1}(\Omega)\right]^d} + |p|_{\left[H^k(\Omega)\right]^d} \right)$$

4.11 Time discretization of Navier-Stokes equations

We consider the following semidiscretized formulation:

$$\begin{cases} \mathrm{M}\frac{d\mathbf{u}(t)}{dt} + \mathrm{A}\mathbf{u}(t) + \mathrm{C}(\mathbf{u}(t))\mathbf{u}(t) + \mathrm{B}^T\mathbf{p}(t) = \mathbf{f}(t) \\ \mathrm{Bu}(t) = \mathbf{0} \end{cases} \tag{4.24}$$

with $\mathbf{u}(0) = \mathbf{u}_0$. $C(\mathbf{u}(t))$ is in fact a matrix depending on $\mathbf{u}(t)$, whose generic coefficient is $c_{mi}(t) = c\left(\mathbf{u}(t), \boldsymbol{\varphi}_i, \boldsymbol{\varphi}_m\right)$.

For the temporal discretization of this system let us use, for instance, the θ-method, that was introduced for parabolic equations. By setting:

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

we obtain the following system of algebraic equations:

$$\begin{cases} \mathrm{M}\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + \mathrm{A}_\theta^{n+1} + \mathrm{C}_\theta\left(\mathbf{u}^{n+1,n}\right)\mathbf{u}^{n+1,n} + \mathrm{B}^T\mathbf{p}_\theta^{n+1} = \mathbf{f}_\theta^{n+1} \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0} \end{cases} \tag{4.25}$$

Except for the special case $\theta = 0$, which corresponds to the forward Euler method, the solution of this system is quite involved.

A possible alternative is to use a semi-implicit scheme, in which the linear part of the equation is advanced implicitly, while nonlinear terms explicitly.

By doing so, if $\theta \geq 1/2$, the resulting scheme is unconditionally stable, whereas it must obey a stability restriction on the time step Δt (depending on h and ν) in all other cases.

4.12 Finite difference methods

We consider at first an explicit temporal discretization of the first equation in (4.24), corresponding to the choice $\theta = 0$ in (4.25). If we suppose that all quantities are known at the time t^n, we can write the associated problem at time $t^n + 1$ as follows:

$$\begin{cases} \mathrm{M}\mathbf{u}^{n+1} = H\left(\mathbf{u}^n, \mathbf{p}^n, \mathbf{f}^n\right) \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0} \end{cases} \tag{4.26}$$

where M is the mass matrix whose entries are

$$m_{ij} = \int_\Omega \varphi_i \varphi_j d\Omega$$

This system does not allow the determination of the pressure $\mathbf{p}^{n+1}$. In particular, there is no way to enforce the divergence free constraint on $\mathbf{u}^{n+1}$.

However, if we replace $\mathbf{p}^n$ by $\mathbf{p}^{n+1}$ in the momentum equation, we obtain the new linear system

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

G being a suitable known vector.

This system corresponds to a semi-explicit discretization of (4.24). Since M is symmetric and positive definite, if condition (4.22) is satisfied, then the reduced system

$\mathrm{BM}^{-1}\,\mathrm{B}^T\mathbf{p}^{n+1} = \mathrm{BM}^{-1}\mathbf{G}$ is non-singular.

This discretization method is temporally stable provided the time step satisfies the following limitation (of parabolic type):

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

Let us now consider an implicit discretization of (4.24), for instance the backward Euler method, which corresponds to choosing $\theta = 1$ in (4.25). As already observed, this scheme is unconditionally stable. It yields a nonlinear algebraic system which can be regarded as the finite element space approximation to the steady Navier-Stokes problem

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

The solution of such nonlinear algebraic system can be achieved by Newton-Krylov techniques, that is by using a Krylov method (e.g. GMRES or BiCGStab) for the solution of the linear system that is obtained at each Newton iteration step.

We recall that Newton's method is based on the full linearization of the convective term, $\mathbf{u}_k^{n+1}\cdot\nabla\mathbf{u}_{k+1}^{n+1} + \mathbf{u}_{k+1}^{n+1}\cdot\nabla\mathbf{u}_k^{n+1}$.

A popular approach consists in starting Newton iterations after few Piccard iterations in which the convective term is evaluated as follows: $\mathbf{u}_k^{n+1}\cdot\nabla\mathbf{u}_{k+1}^{n+1}$.

This approach entails three nested cycles:

- temporal iteration: $t^n \to t^{n+1}$;
- Newton iteration: $\mathbf{x}_k^{n+1} \to \mathbf{x}_{k+1}^{n+1}$;
- Krylov iteration: $\left[\mathbf{x}_k^{n+1}\right]_j \to \left[\mathbf{x}_k^{n+1}\right]_{j+1}$;

for simplicity we have called $\mathbf{x}^n$ the couple $(\mathbf{u}^n, \mathbf{p}^n)$. Obviously, the goal is the following convergence result:

$$\lim_{k\to\infty}\lim_{j\to\infty}\left[\mathbf{x}_k^{n+1}\right]_j = \begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{p}^{n+1} \end{bmatrix}$$

Finally, let us operate a semi-implicit, temporal discretization, consisting in treating explicitly the nonlinear convective term. The following algebraic linear system, whose form is similar to (4.22), is obtained in this case

$$\begin{cases} \frac{1}{\Delta t}\mathrm{M}\mathrm{u}^{n+1} + \mathrm{A}\mathbf{u}^{n+1} + \mathrm{B}^T\mathbf{p}^{n+1} = \mathbf{G}, \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0}, \end{cases} \tag{4.27}$$

where $\mathbf{G}$ is a suitable known vector. In this case the stability restriction on the time step takes the following form:

$$\Delta t \leq C\frac{h}{\max_{\mathbf{x}\in\Omega}|\mathbf{u}^n(\mathbf{x})|} \tag{4.28}$$

In all cases, optimal error estimates can be proven.

4.13 Time dependent Generalized Stokes problem

Fully explicit discretization

The fully explicit discretization of the weak formulation of (Time Dependant Stokes) using (FE) with time step Δt is

$$\begin{cases} \int_\Omega \frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^n}{\Delta t}\mathbf{v}_h + a\left(\mathbf{u}_h^n, \mathbf{v}_h\right) + b\left(\mathbf{v}_h, p_h^n\right) = \mathbf{F}^n\left(\mathbf{v}_h\right) & \forall \mathbf{v}_h \in V_h, \\ b\left(\mathbf{u}_h^{n+1}, q_h\right) = 0 & \forall q_h \in Q_h, \\ \mathbf{u}^0 = \mathbf{u}_0. \end{cases}$$

Which can be put in the form:

$$\begin{cases} M\mathbf{u}^{n+1} = \Delta t G\left(\mathbf{u}^n, \mathbf{p}^n, \mathbf{f}^n\right), \\ B\mathbf{u}^{n+1} = \mathbf{0}. \end{cases}$$

The second equation is impossible to be satisfied, so this is not a functioning numerical method.

Semi-implicit discretization

The semi-implicit discretization using (FE) and (BE) is in the same form as the fully explicit one, except for the first equation, which is instead:

$$\int_\Omega \frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^n}{\Delta t}\mathbf{v}_h + a\left(\mathbf{u}_h^n, \mathbf{v}_h\right) + b\left(\mathbf{v}_h, p_h^{n+1}\right) = \mathbf{F}^{n+1}\left(\mathbf{v}_h\right).$$

The problem can be put in the form:

$$\begin{bmatrix} \frac{M}{\Delta T} & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{p}^{n+1} \end{bmatrix} = \begin{bmatrix} H(\mathbf{u}^n, \mathbf{f}^{n+1}) \\ 0 \end{bmatrix}$$

The stability condition is the following one:

$$\Delta T \leq \frac{Ch^2}{\nu}.$$

We can also illustrate the semi-implicit Θ method:

For every $n \geq 0$, find $u^n \in V_h$, $p^n \in Q_h$ such that $\forall v_h \in V_h$:

$$\frac{1}{\Delta t}(\mathbf{u}^{n+1} - \mathbf{u}^n, \mathbf{v}_h) + a(\mathbf{u}^n, \mathbf{v}_h) + b(\mathbf{v}_h, \theta p^{n+1} + (1-\theta)p^n) = (\theta \mathbf{F}^{n+1} + (1-\theta)\mathbf{F}^n, \mathbf{v}_h)$$

$$b(\theta \mathbf{u}^{n+1}, q_h) = \mathbf{G}^{n+1}((1-\theta)\mathbf{u}^n, q_h) \quad \forall q_h \in Q_h$$

Fully implicit discretization

The fully implicit discretization using (BE) is in the same form as the semi-implicit one, except for the first equation, which is instead:

$$\int_\Omega \frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^n}{\Delta t}\mathbf{v}_h + a\left(\mathbf{u}_h^{n+1},\mathbf{v}_h\right) + b\left(\mathbf{v}_h, p_h^{n+1}\right) = \mathbf{F}^{n+1}\left(\mathbf{v}_h\right).$$

$$\underbrace{\begin{bmatrix} \frac{1}{\Delta t}(M+A) & B^\top \\ B & 0 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{p}^{n+1} \end{bmatrix}}_{\mathbf{w}} = \begin{bmatrix} H(\mathbf{u}^n, \mathbf{f}^{n+1}) \\ 0 \end{bmatrix}$$

This method is unconditionally stable.

We can also illustrate the fully implicit Θ method:

For every $n \geq 0$, find $u^n \in V_h$, $p^n \in Q_h$ such that $\forall v_h \in V_h$:

$$\frac{1}{\Delta t}(\mathbf{u}^{n+1}-\mathbf{u}^n, \mathbf{v}_h)+a(\theta\mathbf{u}^{n+1}+(1-\theta)\mathbf{u}^n, \mathbf{v}_h)+b(\mathbf{v}_h, \theta p^{n+1}+(1-\theta)p^n) = (\theta\mathbf{F}^{n+1}+(1-\theta)\mathbf{F}^n, \mathbf{v}_h)$$

$$b(\mathbf{u}^{n+1}, q_h) = \mathbf{G}^{n+1}((1-\theta)\mathbf{u}^n, q_h) \quad \forall q_h \in Q_h$$

4.14 Generalization of the Stokes problem to N-S

In this section we will revise the concept already presented in section 4.12, starting from the generalization of the Stokes problem. Some of the information presented will be redundant. There is also an inconsistent notation since here, fully/semi or implicit/explicit refer just to the nonlinear c term. In particular in this section the approach followed in class will be presented, while section 4.12 followed the approach presented in the reference textbook.

The terms in the semi-discretized formulation of (NS) in common with that of (Time-GS) are discretized in time using fully implicit discretization, while the remaining term $c(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h)$ can be discretized in multiple ways. We start from the linear system arising from the fully implicit Stokes discretization, $S\mathbf{w} = \mathbf{G}$ which has the advantage of not having any stability restrictions. The cost would be similar to the one starting from the semi-implicit discretization, so this last one is preferred.

Fully Explicit Discretization

The discretized term c using Forward Euler can be put in the form $N\left(\mathbf{u}^n\right)\mathbf{u}^n$. The problem is still in the form $S\mathbf{w} = \tilde{\mathbf{G}}$, with $\tilde{\mathbf{G}} := \mathbf{G} - N\left(\mathbf{u}^n\right)\mathbf{u}^n$. This method has the following stability condition:

$$\Delta t \leq C\frac{h}{\max_{\mathbf{x}\in\Omega}|\mathbf{u}^n(\mathbf{x})|} \tag{4.29}$$

An explicit treatment of the nonlinear term leads to this kind of algorithm:

$$(\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = (\mathbf{u}^n \cdot \nabla)\mathbf{u}^n$$

corresponding to the following time discretization

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n+\nabla p^{n+1}=\mathbf{f}^{n+1},\quad \text{in } \Omega$$

$$\operatorname{div}\mathbf{u}^{n+1}=0.$$

The problem to solve at each time step reduces, in this case, to a generalized Stokes problem characterized by a symmetric matrix which does not change at each time step. This strong reduction of the computational complexity required at each time step is, however, balanced by poorer stability properties. In particular, due to the explicit treatment of the convective term, the time step should satisfy a CFL (Courant-Friedrichs-Lewy) condition given by

$$\Delta t\leq\frac{h}{\|\mathbf{u}^n\|_\infty},$$

which may become too penalizing in presence of high velocities.

Semi-implicit discretization

The term c can be discretized using (FE) and (BE) and put in two different forms, $c\left(\mathbf{u}_h^{n+1},\mathbf{u}_h^n,\mathbf{v}_h\right)$ and $c\left(\mathbf{u}_h^n,\mathbf{u}_h^{n+1},\mathbf{v}_h\right)$, resulting in two different matrices and linear systems:

$$\begin{bmatrix}\frac{1}{\Delta t}(M+A)+C_{1/2} & B^T\\ B & 0\end{bmatrix}\begin{bmatrix}\mathbf{u}^{n+1}\\ \mathbf{p}^{n+1}\end{bmatrix}=\begin{bmatrix}H(\mathbf{u}^n,\mathbf{f}^{n+1})\\ 0\end{bmatrix}$$

The only difference from the original algebraic version of the Stokes system is the fact that the resulting matrix is non symmetric. The stability conditions are the following:

- $\Delta t_n\leq Kh$ in the case of $C_1,\ c\left(\mathbf{u}_h^{n+1},\mathbf{u}_h^n,\mathbf{v}_h\right).$
- $\Delta t_n\leq\frac{h}{\max_{\mathbf{x}\in\Omega}|\mathbf{u}^n(\mathbf{x})|}$ in the case of $C_2,\ c\left(\mathbf{u}_h^n,\mathbf{u}_h^{n+1},\mathbf{v}_h\right)$.

We can have thus 2 discretization schemes:

$$(\mathbf{u}^*\cdot\nabla)\mathbf{u}^*=(\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1}.$$

or

$$(\mathbf{u}^*\cdot\nabla)\mathbf{u}^*=(\mathbf{u}^n+1\cdot\nabla)\mathbf{u}^n.$$

The resulting time-discretization of the Navier-Stokes problem reads

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1}+\nabla p^{n+1}=\mathbf{F}^{n+1},\quad \text{in } \Omega$$

$$\operatorname{div}\mathbf{u}^{n+1}=0.$$

or alternativly:

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^{n+1}\cdot\nabla)\mathbf{u}^n+\nabla p^{n+1}=\mathbf{F}^{n+1},\quad \text{in } \Omega$$

$$\operatorname{div}\mathbf{u}^{n+1}=0.$$

The advantage of this approach with respect to the fully implicit case relies on the fact that, in this case, the system to be solved at each time step is linear. However, the matrix depends on the solution u^n and needs to be recomputed each time and is non-symmetric. The computational cost for one time step is equivalent to one iteration of the fixed-point algorithm for steady state. (Parolini,N. Computational Fluid Dynamics).

Fully implicit discretization

The discretized term c using (BE) can be put in the form $N\left(\mathbf{u}^{n+1}\right)\mathbf{u}^{n+1}$, which is a nonlinear term. The problem is in the form:

$$S\mathbf{w} + \begin{pmatrix} N\left(\mathbf{u}^{n+1}\right)\mathbf{u}^{n+1} \\ \mathbf{0} \end{pmatrix} = \mathbf{G}$$

which is a nonlinear algebraic system. This method is unconditionally stable.

If we consider an implicit treatment for the nonlinear convective term, namely:

$$(\mathbf{u}^* \cdot \nabla)\mathbf{u}^* = (\mathbf{u}^{n+1} \cdot \nabla)\mathbf{u}^{n+1},$$

the resulting time-discretization reads:

$$\mathbf{u}^{n+1} - \mathbf{u}^n - \Delta t\nu\Delta\mathbf{u}^{n+1} + \Delta t(\mathbf{u}^{n+1} \cdot \nabla)\mathbf{u}^{n+1} + \Delta t\nabla p^{n+1} = \Delta t f^{n+1} \quad \text{in } \Omega$$

$$\operatorname{div}\mathbf{u}^{n+1} = 0.$$

In this case, the algebraic system to be solved at each time step is nonlinear, thus requires for its solution a fixed-point or Newton iteration. The computational cost of such approach is further incremented by the need of assembling matrix (and possibly preconditioner) at each time step. On the other hand, the main advantage of an implicit approach is that it is unconditionally stable regardless the time step used. (Parolini, Computational Fluid Dynamycs).

Error analysis of fully discretized problems

If Taylor-Hood elements with degree k and a time discretization method of order q are used, we have the following error estimate:

$$\forall t > 0, \|\mathbf{u}(t) - \mathbf{u}_h(t)\|_{H^1(\Omega)} + \|p(t) - p_h(t)\|_{L^2(\Omega)} \leq C\left(\Delta t^q + h^k\right) \left(|\mathbf{u}(t)|_{H^{k+1}(\Omega)} + |p(t)|_{H^k(\Omega)} + K\left(\partial_t\mathbf{u}(t), \partial_t p(t)\right)\right)$$

for a certain $C \in \mathbb{R}$, where the term $K\left(\partial_t\mathbf{u}(t), \partial_t p(t)\right)$. represents a measure of the time derivatives of the exact solutions at time t.

These prove, in particular, that the gradient of the discrete solution (as well as that of the weak solution u) could be as large as μ_0 is small.

Chapter 5

The ADR Boundary Value Problem

5.1 Complete Case

In this chapter we consider the following general formulation for an Elliptic Boundary value problem problems. By choosing some coefficient to be null, all the particular cases can be recovered.

$$\begin{cases} -\nabla \cdot (\mu \nabla u) + \nabla \cdot (\mathbf{b}u) + \mathbf{c} \cdot \nabla u + \sigma u = f & \text{in } \Omega, \\ u = g_D & \text{on } \Gamma_D, \\ (\mu \nabla u - \mathbf{b}u) \cdot \mathbf{n} + \gamma u = g_N & \text{on } \Gamma_N, \end{cases} \tag{5.1}$$

where μ, σ, f, $\mathbf{b}$ and $\mathbf{c}$ are given functions or constants. In the most general case, we will suppose that $\mu \in \mathrm{L}^\infty(\Omega)$ with $\mu(\mathbf{x}) \geq \mu_0 > 0, \sigma \in \mathrm{L}^\infty(\Omega)$, $\mathbf{b}, \mathbf{c} \in [\mathrm{L}^\infty(\Omega)]^2$ with $\mathrm{div}(\mathbf{b} - \mathbf{c}) \in \mathrm{L}^\infty(\Omega)$, and $f \in \mathrm{L}^2(\Omega)$. Moreover $\gamma \in L^\infty(\partial\Omega)$ is a constant and $g_D \in H^{1/2}(\partial\Omega)$ and $g_N \in L^2(\partial\Omega)$, in order to have the continuity of the functional on the right-hand side.

Weak problem formulation

We will expand the first terms using the Green Inequalities:

$$\nabla \cdot (\mu \nabla u v) = \nabla \cdot (\mu \nabla u) v + \mu \nabla u \cdot \nabla v$$

Applying Gauss theorem we get:

$$\int_\Omega -\nabla \cdot (\mu \nabla u) v d\Omega = \mu \nabla u \cdot \nabla v dx - \int_{\Gamma_N} \mu \nabla u \cdot \mathbf{n} d\Gamma$$

For the second term instead we get:

$$\nabla \cdot (\mathbf{b}uv) = u\mathbf{b} \cdot \nabla v + \nabla \cdot (\mathbf{b}u) v$$

Applying again Gauss theorem:

$$\int_\Omega \nabla \cdot (\mathbf{b}u) v = -\int_\Omega u\mathbf{b} \cdot \nabla v d\Omega + \int_{\Gamma_N} \mathbf{b} \cdot \mathbf{n} uv d\Gamma$$

So we will get on the right hand side:

$$\int_{\Gamma_N} -(\mu\nabla u - \mathbf{b}u)\cdot\mathbf{n}vd\Omega$$

which we will substitute with:

$$\int_{\Gamma_N} \gamma uv - g_N vd\Gamma$$

Let $V = \mathrm{H}^1_{\Gamma D}(\Omega)$ such that $V = \left\{v \in H^1(\Omega) : v_{|\Gamma_D} = g_d\right\}$. By introducing the bilinear form $a : V \times V \mapsto \mathbb{R}$:

$$a(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v d\Omega - \int_\Omega u\mathbf{b}\cdot\nabla v d\Omega + \int_\Omega v\mathbf{c}\cdot\nabla u d\Omega + \int_\Omega \sigma uv d\Omega + \int_{\Gamma_N}\gamma uv d\Gamma \quad (5.2)$$

Introduce the lifting function R_{g_D} such that $R_{g_D} \in H^1(\Omega)$ and $R_{g_D} = g_D$ on Γ_D. The new unknown function is $\tilde{u} = u - R_{g_D}$, which satisfies $\tilde{u} = 0$ on Γ_D.

The weak form simplifies to:

$$a(\tilde{u},v) = \int_\Omega fv\, d\Omega + \int_{\Gamma_N} g_N v\, d\Gamma \; - a(R_{g_D}, v)$$

where:

- $V = H^1_{\Gamma_D}(\Omega)$ as the space of test functions.
- The bilinear form $a(u,v)$ and the linear form $F(v)$ are defined by:

$$a(\tilde{u},v) = \int_\Omega \mu\nabla\tilde{u}\cdot\nabla v\, d\Omega - \int_\Omega \tilde{u}\mathbf{b}\cdot\nabla v\, d\Omega + \int_\Omega v\mathbf{c}\cdot\nabla\tilde{u}\, d\Omega + \int_\Omega \sigma\tilde{u}v\, d\Omega + \int_{\Gamma_N}\gamma\tilde{u}v\, d\Gamma$$

$$F(v) = \int_\Omega fv\, d\Omega + \int_{\Gamma_N} g_N v\, d\Gamma \; - a(R_{g_D}, v).$$

The weak formulation becomes :

Find $\tilde{u} \in V$ such that:

$$a(\tilde{u},v) = F(v) \quad \forall v \in V.$$

The actual solution u can be recovered by $u = \tilde{u} + R_{g_D}$.

In order to prove the existence and uniqueness of the solution of (5.12) we will put ourselves in the condition to apply the Lax-Milgram lemma.

To verify the coercivity of the bilinear form $a(\cdot,\cdot)$, we proceed separately on the single terms:

For the first term we have:

$$\int_\Omega \mu\nabla v\cdot\nabla v d\Omega \geq \mu_0\|\nabla v\|^2_{\mathrm{L}^2(\Omega)} \quad (5.3)$$

As $v \in \mathrm{H}^1_{\Gamma D}(\Omega)$, the Poincaré inequality holds (see (2.13 NMDP)); then

$$\|v\|^2_{\mathrm{H}^1(\Omega)} = \|v\|^2_{\mathrm{L}^2(\Omega)} + \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} \leq \left(1 + C^2_\Omega\right) \|\nabla v\|^2_{\mathrm{L}^2(\Omega)}$$

and therefore it follows that

$$\int_\Omega \mu \nabla v \cdot \nabla v d\Omega \geq \frac{\mu_0}{1 + C^2_\Omega} \|v\|^2_{\mathrm{H}^1(\Omega)}$$

We now move to the convective term. Using Green's formula (3.16 NMDE) yields

$$-\int_\Omega v(\mathbf{b} - \mathbf{c}) \cdot \nabla v d\Omega = -\frac{1}{2} \int_\Omega (\mathbf{b} - \mathbf{c}) \cdot \nabla \left(v^2\right) d\Omega = \frac{1}{2} \int_\Omega v^2 \nabla \cdot (\mathbf{b} - \mathbf{c}) d\Omega - \frac{1}{2} \int_{\Gamma_N} (\mathbf{b} - \mathbf{c}) \cdot \mathbf{n} v^2 d\gamma$$

Then we can conclude another two conditions for the coercitivity, taking into account also the term related to σ and γ:

$$\frac{1}{2} \nabla \cdot (\mathbf{b} - \mathbf{c}) + \sigma \geq 0 \quad \text{a.e. in } \Omega, \quad \gamma - \frac{1}{2} (\mathbf{b} - \mathbf{c}) \cdot \mathbf{n} \geq 0 \quad \text{a.e. in } \Gamma_N \tag{5.4}$$

Consequently, the bilinear form $a(\cdot, \cdot)$ is coercive, as

$$a(v, v) \geq \alpha \|v\|^2_V \quad \forall v \in V, \qquad \text{with} \qquad \alpha = \frac{\mu_0}{1 + C^2_\Omega} \tag{5.5}$$

To prove that the bilinear form $a(\cdot, \cdot)$ is continuous, that is it satisfies (2.6 NMDP), we bound the first term on the right-hand side of (5.11) as follows:

$$\left| \int_\Omega \mu \nabla u \cdot \nabla v d\Omega \right| \leq \|\mu\|_{\mathrm{L}^\infty(\Omega)} \|\nabla u\|_{\mathrm{L}^2(\Omega)} \|\nabla v\|_{\mathrm{L}^2(\Omega)} \leq \|\mu\|_{\mathrm{L}^\infty(\Omega)} \|u\|_{\mathrm{V}} \|v\|_{\mathrm{V}} \tag{5.6}$$

We have used the Hölder and Cauchy-Schwarz inequalities (see Sect. 2.5), as well as the inequality $\|\nabla w\|_{\mathrm{L}^2(\Omega)} \leq \|w\|_{\mathrm{V}(\Omega)}$ $\forall w \in V$. For the second term, proceeding in a similar way we find

$$\left| \int_\Omega u \mathbf{b} \cdot \nabla v d\Omega \right| \leq \|\mathbf{b}\|_{\mathrm{L}^\infty(\Omega)} \|u\|_{\mathrm{L}^2(\Omega)} \|\nabla v\|_{\mathrm{L}^2(\Omega)} \leq \|\mathbf{b}\|_{\mathrm{L}^\infty(\Omega)} \|v\|_V \|u\|_V \tag{5.7}$$

$$\left| \int_\Omega v \mathbf{c} \cdot \nabla u d\Omega \right| \leq \|\mathbf{c}\|_{\mathrm{L}^\infty(\Omega)} \|v\|_{\mathrm{L}^2(\Omega)} \|\nabla u\|_{\mathrm{L}^2(\Omega)} \leq \|\mathbf{c}\|_{\mathrm{L}^\infty(\Omega)} \|v\|_V \|u\|_V \tag{5.8}$$

For the last term we have, thanks again to the Cauchy-Schwarz inequality:

$$\left| \int_\Omega \sigma u v d\Omega \right| \leq \|\sigma\|_{\mathrm{L}^2(\Omega)} \|uv\|_{\mathrm{L}^2(\Omega)} \leq C^2 \|\sigma\|_{\mathrm{L}^2(\Omega)} \|u\|_V \|v\|_V \tag{5.9}$$

Indeed, $\|uv\|_{\mathrm{L}^2(\Omega)} \leq \|u\|_{\mathrm{L}^4(\Omega)} \|v\|_{\mathrm{L}^4(\Omega)} \leq C^2 \|u\|_{\mathrm{H}^1(\Omega)} \|v\|_{\mathrm{H}^1(\Omega)}$, having applied inequality (2.18 NMDP) and exploited inclusions (2.19), with C being the inclusion constant. Remember that in H^1_0 or equivalently in $H^1_{\Gamma D}$ the $||\cdot||_{H^1}$ is equivalent to the $||\cdot||_{H^1_0 / H^1_{\Gamma_D}}$ norms, due to the Poincaré inequality.

Finally, applying Cauchy Scharz on the boundary and the trace inequality we get:

$$\int_{\Gamma_N} \gamma u v \, d\Gamma \leq |\gamma| C'^2 \|u\|_V \|v\|_V$$

Summing all the terms we obtained, the continuity property (2.6 NMDP) follows by taking, e.g.,

$$M = \|\mu\|_{L^\infty(\Omega)} + \|\mathbf{b}\|_{L^\infty(\Omega)} + \|\mathbf{c}\|_{L^\infty(\Omega)} + C^2 \|\sigma\|_{L^2(\Omega)} + \gamma C'^2 \tag{5.10}$$

On the other hand, the right-hand side of (5.12) defines a bounded and linear functional thanks to the Cauchy-Schwarz inequality and to the Poincaré inequality (2.13 NMDP). This must be proved, let's do it. We need R_{gD}, the extension of g_D to the whole domain, to be in $H^1(\Omega)$, thus we require that $g_D \in H^{1/2}(\partial\Omega)$, and then we can apply the trace inequality:

$$\begin{aligned} |F(v)| &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g_N\|_{L^2(\Gamma_N)} \|v\|_{L^2(\Gamma_N)} + M \|R_{g_D}\|_V \|v\|_V \\ &\leq \left(\|f\|_{L^2(\Omega)} + C' \|g_N\|_{L^2(\Gamma_N)} + M' \|g_D\|_{H^{1/2}(\partial\Omega)} \right) \|v\|_V = K \|v\|_V . \end{aligned}$$

The Galerkin approximation of the problem is:

$$\text{find } \tilde{u_h} \in V_h : \quad a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h \tag{5.11}$$

where V_h is a suitable family of subspaces of $V = H^1_{\Gamma_D}$. By replicating the proof carried out above for the finite element formulation, the following estimates can be proved:

$$\|u_h\|_{V_h} \leq \frac{1}{\alpha} K, \quad \|\nabla u_h\|_{\mathrm{L}^2(\Omega)} \leq \frac{\sqrt{1 + C_\Omega^2}}{\mu_0} K$$

These prove, in particular, that the gradient of the discrete solution (as well as that of the weak solution u) could be as large as μ_0 is small.

Moreover, the Galerkin error inequality gives:

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V \tag{5.12}$$

By the definitions of α and M, the upper-bounding constant M/α becomes as large (and, correspondingly, the estimate (5.12) meaningless) as the ratio $\|\mathbf{b}\|_{L^\infty(\Omega)} / \|\mu\|_{L^\infty(\Omega)}$ (resp. the ratio $\|\sigma\|_{L^2(\Omega)} / \|\mu\|_{L^\infty(\Omega)}$) grows, which happens when the convective (resp. reactive) term dominates over diffusive one.

In such cases the Galerkin method can give inaccurate solutions, unless - as we will see - an extremely small discretization step h is used.

Chapter 6

Stokes Equation

Let $\Omega \subset \mathbb{R}^3$be a domain. Let us consider the stationary Stokes problem:

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{u}_{\text{in}} \quad \text{on } \Gamma_{\text{in}}, \\ \nu(\nabla\mathbf{u})\mathbf{n} - p\mathbf{n} &= -p_{\text{out}}\mathbf{n} \quad \text{on } \Gamma_N, \\ \mathbf{u} &= \mathbf{0} \quad \Gamma_D, \end{cases} \tag{6.1}$$

where $\mathbf{u} : \Omega \to \mathbb{R}^3$ and $p : \Omega \to \mathbb{R}$ are the velocity and pressure fields of a viscous, incompressible fluid.

Derive the weak formulation of the problem.

Solution. Let us introduce the function spaces

$$\begin{aligned} V &= \{\mathbf{v} \in H^1(\Omega)^3 : \mathbf{v} = \mathbf{u}_{\text{in}} \text{ on } \Gamma_{\text{in}}, \mathbf{v} = \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}, \\ V_0 &= \{v \in H^1(\Omega)^3 : \mathbf{v} = \mathbf{0} \text{ on } \Gamma_{\text{in}}, \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}, \\ Q &= L^2(\Omega). \end{aligned}$$

$$\int_\Omega \nabla p \cdot \mathbf{v}\, dx = -\int_\Omega p\nabla \cdot \mathbf{v}\, dx + \int_{\Gamma_N} p\mathbf{n} \cdot \mathbf{v}\, ds,$$

$$\int_\Omega \mathbf{v}\cdot\Delta\mathbf{u}\, dx = -\int_\Omega \nabla\mathbf{v} : \nabla\mathbf{u}\, dx + \int_{\Gamma_N} \mathbf{v}\cdot(\nabla\mathbf{u})\mathbf{n}\, ds = -\int_\Omega \nabla\mathbf{v} : \nabla\mathbf{u}\, dx + \int_{\Gamma_N} \mathbf{v}\cdot\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\, ds.$$

Thus, we get:

$$\int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\, dx - \int_\Omega p\nabla \cdot \mathbf{v}\, dx = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, dx + \int_{\Gamma_{\Gamma_N}} -p_{\text{out}}\mathbf{n} \cdot \mathbf{v}\, da. \tag{6.2}$$

We can write the velocity in terms of a lifting function for the boundary datum:

$$\mathbf{u} = \mathbf{u_0} + \mathbf{R}(\mathbf{u}_{\text{in}}) \tag{6.3}$$

Where $u_0 \in V_0$ and $\mathbf{R}(u_{\text{in}}) \in V$ such that $\mathbf{R}(u_{\text{in}}) = \mathbf{u}_{\text{in}}$ on Γ_{in} and $\nabla \cdot \mathbf{R}(u_{\text{in}}) = 0$. Inserting equation (6.3) into equation (6.2), we obtain:

$$a(\mathbf{u}_0, \mathbf{v}) + b(\mathbf{v}, p) = F(\mathbf{v}) - a(\mathbf{R}(\mathbf{u}_{\text{in}}), \mathbf{v}).$$

We proceed similarly for the second equation: we multiply by $q \in Q$ and integrate over Ω:

$$\int_\Omega \nabla \cdot q\mathbf{u}\, dx = 0,$$

$$\int_\Omega q\nabla \cdot \mathbf{u}_0\, dx = 0,$$

$$b(\mathbf{u}_0, q) = 0.$$

Therefore, the weak formulation reads: find $\mathbf{u}_0 \in V_0$ and $p \in Q$ such that, for all $\mathbf{v} \in V_0$ and $q \in Q$, the following holds:

$$a(\mathbf{u}_0, \mathbf{v}) + b(\mathbf{v}, p) = F(\mathbf{v}) - a(\mathbf{R}(\mathbf{u}_{\text{in}}), \mathbf{v}),$$

$$b(\mathbf{u_0}, q) = -b(\mathbf{R}(\mathbf{u}_{\text{in}}), q).$$

Galerkin Formulation

Let us introduce a mesh over Ω, and let $V_{0,h} = V_0 \cap (X_h^r(\Omega))^3$ and $Q_h = Q \cap X_h^p(\Omega)$. Then, the discrete weak formulation reads: find $\mathbf{u}_{0,h} \in V_{0,h}$ and $p_h \in Q_h$ such that, for all $\mathbf{v}_h \in V_{0,h}$ and $q_h \in Q_h$, there holds

$$a(\mathbf{u}_{0,h}, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = F(\mathbf{v}_h) - a(\mathbf{R_h}(\mathbf{u_{in}}), \mathbf{v}_h),$$

$$b(\mathbf{u}_h, q_h) = -b(\mathbf{R}_h(\mathbf{u}_{\text{in}}, q_h)).$$

Upon discretization, this leads to an algebraic system of the form

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{0} \end{bmatrix},$$

where, denoting by $\{\boldsymbol{\psi}_i\}_{i=1}^{N_u}$ the basis functions for the velocity space and by $\{\boldsymbol{\phi}_i\}_{i=1}^{N_p}$ those for the pressure space,

$$A_{ij} = \int_\Omega \nabla \boldsymbol{\psi}_i : \nabla \boldsymbol{\psi_j}\, dx,$$

$$B_{ij} = -\int_\Omega \boldsymbol{\psi}_i \cdot \boldsymbol{\phi}_j\, dx,$$

$$F_i = \int_\Omega \mathbf{f} \cdot \boldsymbol{\psi}_i\, dx + \int_{\Gamma_N} -p_{\text{out}}\mathbf{n} \cdot \boldsymbol{\psi}_i\, d\sigma.$$