Lecture Notes: The Finite Element Method

Aurélien Larcher, Niyazi Cem Değirmenci

Fall 2013

Contents

Introduction

This document is a collection of short lecture notes written for the course "The Finite Element Method" (SF2561), at KTH, Royal Institute of Technology during Fall 2013. It is in no way intended as a comprehensive and rigorous introduction to Finite Element Methods but rather an attempt for providing a self-consistent overview in direction to students in Engineering without any prior knowlegde of Numerical Analysis.

Content

The course will go through the basic theory of the Finite Element Method during the first six lectures while the last three lectures will be devoted to some applications.

1. Introduction to PDEs, weak solution, variational formulation.
2. Ritz method for the approximation of solutions to elliptic PDEs
3. Galerkin method and well-posedness.
4. Construction of a Finite Element approximation space.
5. Polynomial approximation and error analysis.
6. Time dependent problems.
7. Adaptive control.
8. Stabilized finite element methods.
9. Mixed problems.

The course will attempt to introduce the practicals aspects of the methods without hiding the mathematical issues. There are indeed two side of the Finite Element Method: the Engineering approach and the Mathematical theory. Although any reasonable implementation of a Finite Element Method is likely to compute an approximate solution, usually the real challenge is to understand the properties of the obtained solution, which can be summarized in three main questions:

1. *Well-posedness*: Is the solution to the approximate problem unique ?
2. *Consistency*: Is the solution to the approximate problem close to the continuous solution (or at least "sufficiently" in a sense to determine) ?
3. *Stability, Maximum principle*: Is the solution to the approximate problem stable and/or satisfying physical bounds ?

Ultimately, the goal of designing numerical scheme is to combine these properties to ensure the convergence of the method to the unique solution of the continuous problem (if hopefully it exists) defined by the mathematical model. In a way, the main message of the course is that studying the mathematical properties of the continuous problem hints and deriving discrete counterparts of them (usually in terms of inequalities) is usually a good way to enforce stability and convergence.

Answering these questions requires some knowledge of elements of numerical analysis of PDEs which will be introduced throughout the document in a didactic manner. Nonetheless, while these difficulties will not be hidden, addressing some technical details is left to more serious and well-written works referenced in the bibliography.

Literature

The historical textbook used mainly for the exercises is *Computational Differential Equations* [6] which covers many examples from Engineering but is mainly limited to Galerkin method and in particular continuous Lagrange elements.

The two essential books in the list are *Theory and Practice of Finite Elements* [4] and *The Mathematical Theory of Finite Element Methods* [2]. The first work provides an extensive coverage of Finite Elements from a theoretical standpoint (including non-conforming Galerkin, Petrov-Galerkin, Discontinuous Galerkin) by expliciting the theoretical foundations and abstract framework in the first Part, then studying applications in the second Part and finally addressing more concrete questions about the implementation of the methods in a third Part. The Appendices are also quite valuable as they provide a toolset of results to be used for the numerical analysis of PDEs. The second work is written in a more theoretical fashion, providing to the Finite Element method in the first six Chapters which is suitable for a student with a good background in Mathematics. Section 2 about Ritz's method is based on the lecture notes [5] and Section 9.1 on the description of the Stokes problem in [7].

Two books listed in the bibliography are not concerned with Numerical Analysis but with the continuous setting. On the one hand, book *Functional Analysis, Sobolev Spaces and Partial Differential Equations* [3] is an excellent introduction to Functional Analysis, but has a steep learning curve without a solid background in Analyis. On the other hand, *Mathematical Tools for the Study of the Incompressible Navier–Stokes Equations and Related Models* [1], while retaining all the difficulties for the analysis of PDEs for fluid problems, possesses a really didactic approach in a clear and rigorous manner.

1 Weak formulation of Partial Differential Equations

1.1 Historical perspective

By "Finite Element Methods", we denote a family of approaches developed to compute an approximate solution to a partial differential equation (PDE). The physics of phenomena encoutere in engineering applicatios is often modelled under the form of a boundary value problem. Equations describing the evolution in time are called initial value problems and consist of the coupling of an ordinary differential equation (ODE) in time with a bounday value problem in space.

The study of equations involving derivatives of the unknown has led to rethinking the concept of derivation: from the idea of variation, then the study of the Cauchy problem, finally to the generalization of the notion of derivative with the Theory of Distributions.

1.2 Weak solution to the Dirichlet problem

Let us consider the Poisson problem posed in a domain Ω, an open bounded subset of $\mathbb{R}^d$, $d \geq 1$ supplemented with homogeneous Dirichlet boundary conditions:

$$-\boldsymbol{\Delta} u(\boldsymbol{x}) = f(\boldsymbol{x}) \tag{1a}$$

$$u(\boldsymbol{x}) = 0 \tag{1b}$$

with $f \in \mathrm{C}^0(\overline{\Omega})$ and the Laplace operator,

$$\boldsymbol{\Delta} = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \tag{2}$$

thus involving second order partial derivatives of the unknown u with respect to the space coordinates.

Definition 1.1 (Classical solution)**.** A classical solution (or strong solution) of Problem (1) is a function $u \in \mathrm{C}^2(\Omega)$ satisfying relations (1a) and (1b).

Problem (1) can be reformulated so as to look for a solution in the distributional sense by testing the equation against smooth functions. Reformulating the problem amounts to relaxing the pointwise regularity (*i.e.* continuity) required to ensure the existence of the classical derivative to the (weaker) existence of the distributional derivative which regularity is to be interpreted in term in terms of Lebesgue spaces: the obtained problem is a *weak formulation* and a solution to this problem (*i.e.* in the distributional sense) is called *weak solution.* Three properties of the weak formulation should be studied: firstly that a classical solution is a weak solution, secondly that such a weak solution is indeed a classical solution provided that it is regular enough and thirdly that the well-posedness of this reformulated problem, *i.e.* existence and uniqueness of the solution, is ensured.

1.2.1 Formal passage from classical solution to weak solution

Let $u \in \mathrm{C}^2(\overline{\Omega})$ be a classical solution to (1) and let us test Equation (1a) against any smooth function $\varphi \in \mathrm{C}_c^\infty(\Omega)$:

$$-\int_\Omega \boldsymbol{\Delta} u(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Since $u \in \mathrm{C}^2(\overline{\Omega})$, Δu is well defined. Integrating by parts, the left-hand side reads:

$$-\int_\Omega \boldsymbol{\Delta} u(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = -\int_{\partial\Omega} \boldsymbol{\nabla} u(\boldsymbol{x}) \cdot \boldsymbol{n}\varphi(\boldsymbol{x})\,\mathrm{d}s + \int_\Omega \boldsymbol{\nabla} u(\boldsymbol{x}) \cdot \boldsymbol{\nabla}\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

For simplicity, we recall the one-dimensional case:

$$-\int_0^1 \frac{\partial^2 u(x)}{\partial x^2}\varphi(x)\mathrm{d}x = -\left[\frac{\partial u(x)}{\partial x}\varphi(x)\right]_0^1 + \int_0^1 \frac{\partial u(x)}{\partial x}\frac{\partial \varphi(x)}{\partial x}\mathrm{d}x$$

Since φ has compact support in Ω, it vanishes on the boundary $\partial\Omega$, consequently the boundary integral is zero, thus the distributional formulation reads

$$\int_\Omega \boldsymbol{\nabla} u(\boldsymbol{x}) \cdot \boldsymbol{\nabla}\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad , \forall\, \varphi \in \mathrm{C}_c^\infty(\Omega)$$

and we are led to look for a solution u belonging to a functional space such that the previous relation makes sense.

A weak formulation of Problem (1) consists in solving:

$$\left|\begin{array}{l} \text{Find } u \in H \text{, given } f \in V' \text{, such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v\,\mathrm{d}\boldsymbol{x} = \int_\Omega f v\,\mathrm{d}\boldsymbol{x} \quad , \forall\, v \in V \end{array}\right. \tag{3}$$

in which H and V are a functional spaces yet to be defined, both satisfying regularity contraints and for H boundary condition constraints. The choice of the *solution space* H and the *test space* is described Section 1.3.

1.2.2 Formal passage from weak solution to classical solution

Provided that the weak solution to Problem (3) belongs to $\mathrm{C}^2(\overline{\Omega})$ then the second derivatives exist in the classical sense. Consequently the integration by parts can be performed the other way around and the weak solution is indeed a classical solution.

1.2.3 About the boundary conditions

Boundary condition	Expression on $\partial\Omega$	Property
Dirichlet	$u = u_D$	"essential" boundary condition
Neumann	$\boldsymbol{\nabla} u \cdot \boldsymbol{n} = 0$	"natural" boundary condition

Essential boundary conditions are embedded in the functional space, while natural boundary conditions appear in the weak formulation as linear forms.

1.3 Weak and variational formulations

1.3.1 Functional setting

Hilbert–Sobolev spaces H^s (Section C.4) are a natural choice to "measure" functions involved in the weak formulations of PDEs as the existence of the integrals relies on the fact that integrals of powers $|\cdot|^p$ of u and weak derivatives $\mathbf{D}^\alpha u$ for some $1 \leq p < +\infty$ exist:

$$\mathrm{H}^s(\Omega) = \left\{u \in \mathrm{L}^2(\Omega) \,:\, \mathbf{D}^\alpha u \in \mathrm{L}^2(\Omega)\,, 1 \leq \alpha \leq s\right\}$$

with the Lebesgue space of square integrable functions on Ω:

$$\mathrm{L}^2(\Omega) = \left\{u \,:\, \int_\Omega |u(\boldsymbol{x})|^2\,\mathrm{d}\boldsymbol{x} < +\infty\right\}$$

endowed with its natural scalar product

$$(u , v)_{\mathrm{L}^2(\Omega)} = \int_\Omega u\, v \,\mathrm{d}\boldsymbol{x}$$

Since Problem (3) involves first order derivatives according to relation,

$$\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x}$$

then we should consider a solution in $\mathrm{H}^1(\Omega)$.

$$\mathrm{H}^1(\Omega) = \left\{ u \in \mathrm{L}^2(\Omega) \;:\; \mathbf{D}u \in \mathrm{L}^2(\Omega) \right\}$$

with the weak derivative $\mathbf{D}u$ *i.e.* a function of $\mathrm{L}^2(\Omega)$ which identifies with th classical derivative (if it exists) "almost everywhere", and endowed with the norm,

$$\| \cdot \|_{\mathrm{H}^1(\Omega)} = (\cdot , \cdot)^{1/2}_{\mathrm{H}^1(\Omega)}$$

defined from the scalar product,

$$(u , v)_{\mathrm{H}^1(\Omega)} = \int_\Omega u\, v \,\mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x}$$

Moreover, the solution should satisfy the boundary condition of the strong form of the PDE problem. According to Section 1.2.3 the homogeneous Dirichlet condition is embedded in the functional space of the solution: u vanishing on the boundary $\partial\Omega$ yields that we should seek u in $\mathrm{H}^1_0(\Omega)$.

1.3.2 Determination of the solution space

We will now establish that any weak solution "lives" in $\mathrm{H}^1_0(\Omega)$.

Choice of test space: In order to give sense to the solution in a Hilbert–Sobolev space we need to choose the test function φ itself in the same kind of space. Indeed $\mathrm{C}^\infty_c(\Omega)$ is not equipped with a topology which allows us to work properly. If we chose $\varphi \in \mathrm{H}^1_0(\Omega)$ then by definition, w can construct a sequence $(\varphi^n)_{n\in\mathbb{N}}$ of functions in $\mathrm{C}^\infty_c(\Omega)$ converging in $\mathrm{H}^1_0(\Omega)$ to φ, *i.e.*

$$\|\varphi^n - \varphi\|_{\mathrm{H}^1(\Omega)} \to 0, \text{ as } n \to +\infty$$

For the sake of completeness, we show that we can pass to the limit in the formulation, term by term for any partial derivative:

$$\int_\Omega \partial_i u\, \partial_i \varphi^n \to \int_\Omega \partial_i u\, \partial_i \varphi$$

as $\partial_i \varphi^n \rightharpoonup \mathbf{D}_i \varphi$ in $\mathrm{L}^2(\Omega)$, which denotes the weak convergence *i.e.* tested on functions of the dual space (which, in case of $\mathrm{L}^2(\Omega)$, is $\mathrm{L}^2(\Omega)$ itself).

$$\int_\Omega f\, \varphi^n \to \int_\Omega f\, \varphi$$

as $\varphi^n \to \varphi$ in $\mathrm{L}^2(\Omega)$. Consequently, the weak formulation is satisfied if $\varphi \in \mathrm{H}^1_0(\Omega)$.

Choice of solution space: The determination of the functional space is guided,

— firstly, by the regularity of the solution: if u is a classical solution then it belongs to $\mathrm{C}^2(\overline{\Omega})$ which involves that $u \in \mathrm{L}^2(\Omega)$ and $\partial_i u \in \mathrm{L}^2(\Omega)$, thus $u \in \mathrm{H}^1(\Omega)$,

— secondly by the boundary conditions: the space should satisfy the Dirichlet boundary condition on $\partial\Omega$. This constraint is satisfied thanks to the following trace theorem for the solution to the Dirichlet problem: since $Ker(\gamma) = \mathrm{H}^1_0(\Omega)$, we conclude $u \in \mathrm{H}^1_0(\Omega)$.

Lemma 1.2 (Trace Theorem)**.** *Let* Ω *be a bounded open subset of* $\mathbb{R}^d$ *with piecewise* C^1 *boundary, then there exists a linear application* $\gamma : \mathrm{H}^1(\Omega) \to \mathrm{L}^2(\partial\Omega)$ *continous on* $\mathrm{H}^1(\Omega)$ *such that* $\gamma(u) = 0 \Rightarrow u \in \mathrm{Ker}(\gamma)$*.*

The weak formulation of Problem (1) reads then:

$$\left| \begin{array}{l} \text{Find } u \in \mathrm{H}^1_0(\Omega), \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x} \quad , \forall\, v \in \mathrm{H}^1_0(\Omega) \end{array} \right. \tag{5}$$

1.4 Abstract problem

The study of mathematical properties of PDE problems is usually performed on a general formulation called *abstract problem* which reads in our case:

$$\left| \begin{array}{l} \text{Find } u \in V, \text{ such that:} \\ a(u,v) = L(v) \quad , \forall\, v \in V \end{array} \right. \tag{6}$$

with $a(\,\cdot\,,\,\cdot\,)$ a continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

Proposition 1.3 (Continuity)**.** *A bilinear form* $a(\,\cdot\,,\,\cdot\,)$ *is continuous on* $V \times \mathrm{W}$ *if there exists a positive constant real number* M *such that*

$$a(v,w) \leq M \left\|v\right\|_V \left\|w\right\|_{\mathrm{W}} \quad ,\forall\, (v,w) \in V \times \mathrm{W}$$

For example, in the previous section for Problem (5), the bilinear form reads

$$\begin{array}{rccl} a: & V \times V & \to & \mathbb{R} \\ & (u,v) & \mapsto & \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} \end{array}$$

and the linear form,

$$\begin{array}{rccl} L: & V & \to & \mathbb{R} \\ & v & \mapsto & \displaystyle\int_\Omega f\, v \,\mathrm{d}\boldsymbol{x} \end{array}$$

In the following chapters, we consider the case of elliptic PDEs, like the Poisson problem, for which the bilinear form $a(\,\cdot\,,\,\cdot\,)$ is coercive.

Proposition 1.4 (Coercivity)**.** *A bilinear form is said coercive in* V *if there exists a positive constant real number* α *such that for any* $v \in V$

$$a(v,v) \geq \alpha \left\|v\right\|^2_V$$

This property is also know as V–ellipticity.

1.5 Well-posedness

In the usual sense, a well-posed problem admits a unique solution which is bounded in the V-norm by the data (forcing term, boundary conditions). In this particular case of the Poisson problem the bilinear form $a(\,\cdot\,,\,\cdot\,)$ is the natural scalar product in $\mathrm{H}^1_0(\Omega)$, thus it defines a norm in $\mathrm{H}^1_0(\Omega)$ (but only a seminorm in $\mathrm{H}^1(\Omega)$ due to the lack of definiteness, not a norm !).

Theorem 1.5 (Riesz–Fréchet). *Let H be a Hilbert space and H' its topological dual, $\forall\ \Phi \in H'$, there exists a unique representant $u \in H$ such that for any $v \in H$,*

$$\Phi(v) = (\ u\ ,\ v\ )_H$$

and furthermore $\|u\|_H = \|\Phi\|_{H'}$

This result ensures directly the existence and uniqueness of a weak solution as soon as $a(\,\cdot\,,\,\cdot\,)$ is a scalar product and Φ is continuous for $\|\cdot\|_a$. Now that we have derived a variational problem for which there exists a unique solution with V infinite dimensional (*i.e.* for any point $x \in \Omega$), we need to construct an approximate problem which is also well-posed.

1.6 Exercises

Exercise 1.6 (Weak formulation — 1). Consider the following problem:

> Find $u \in C^2(\bar{\Omega}), \Omega = (0,1)$ such that:
>
> $$\begin{aligned} -\frac{d^2u}{dx^2}(x) &= 1+x \quad ,\forall\ x \in \Omega \\ u(0) = u(1) &= 0 \end{aligned} \tag{7}$$

1. Formulate the weak form of the problem.
2. Define the space where the solution will be searched for.
3. Formulate the bilinear and linear forms.

Exercise 1.7 (Weak formulation — 2). Solve Problem (7) with boundary conditions:

$$u(0) = 0,\ u(1) = 2$$

Exercise 1.8 (Weak formulation — 3). Solve Problem (7) with boundary conditions:

$$\frac{du}{dx}(0) = 1,\ u(1) = 2$$

Exercise 1.9 (Weak formulation + Regularity $\Rightarrow$ Strong formulation). For the first problem, show that a solution of the weak formulation u_w satisfies the original problem if it belongs to $C^2(\Omega)$.

Hint: Let us assume that

$$1 + x_0 - \frac{d^2u_w(x_0)}{dx^2} \neq 0$$

for some $x_0 \in \Omega$, use the test function

$$v(x) = \begin{cases} 0 & \text{if } x \notin (x_0 - \epsilon, x_0 + \epsilon) \\ (x - (x_0 - \epsilon))^2(x - (x_0 + \epsilon))^2 & \text{otherwise} \end{cases}$$

to show contradiction with the fact that u_w is a weak solution.

2 Ritz and Galerkin methods for elliptic problems

In Section 1 we have reformulated the Dirichlet problem to seek weak solutions and we showed its well-posedness. The problem being infinite dimensional, it is not computable. It should also be noted that computability is not the only reason to build a solution with the Galerkin method, but also to prove the existence of a solution by the method of approximate problems.

Question: How can we construct an approximation to Problem (1) which is also well-posed and how does the solution to this problem compare to the solution of the original problem ?

2.1 Approximate problem

In the previous section we showed how a classical PDE problem such as Problem (1) can be reformulated as a weak problem. The abstract problem for this class of PDE reads then:

$$\left|\begin{array}{l}\text{Find } u \in V\text{, such that:}\\[1ex] a(u,v) = L(v) \quad ,\ \forall\ v \in V\end{array}\right. \tag{8}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

Since in the case of the Poisson problem the bilinear form is continuous, coercive and symmetric, the well-posedness follows directly from Riesz–Fréchet representation Theorem. If the bilinear form is still coercive but not symmetric then we will see that the well-posedness is proven by the Lax–Milgram Theorem.

But for the moment, let us focus on the symmetric case: provided that the well-posedness holds, we want now to construct an approximate solution u_n to the Problem (8).

2.2 Ritz method for symmetric bilinear forms

2.2.1 Formulation

Ritz's method is based on replacing the solution space V (which is infinite dimensional) by a finite dimensional subspace $V_n \subset V$, $\dim(V_n) = n$.

Problem (9) is the approximate weak problem by Ritz's method:

$$\left|\begin{array}{l}\text{Find } u_n \in V_n,\ V_n \subset V\text{, such that:}\\[1ex] a(u_n,v_n) = L(v_n) \quad ,\forall\ v_n \in V_n\end{array}\right. \tag{9}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive symmetric continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

Provided that the bilinear form is symmetric, Problem (10) is the equivalent approximate variational problem under minimisation form:

$$\left|\begin{array}{l}\text{Find } u_n \in V_n,\ V_n \subset V\text{, such that:}\\[1ex] J(u_n) \leq J(v_n) \quad ,\forall\ v_n \in V_n\\[1ex] \text{with } J(v_n) = \frac{1}{2}a(v_n,v_n) - L(v_n)\end{array}\right. \tag{10}$$

Solution

$$u_n = \sum_{j=1}^{n} u_j \varphi_j \tag{11}$$

where ${u_j}_{1\leq j\leq n}$ is a family of real numbers and $\mathcal{B} = (\varphi_1, \ldots, \varphi_n)$ a basis of V_n.

2.2.2 Well-posedness

Theorem 2.1 (Well-posedness)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V, $\dim(V_n) = n$, Problem* (9) *admits a unique solution u_n.*

Proof. The proof can either use directly the Lax–Milgram Theorem or show that there exists a unique solution to the equivalent minimisation problem (10) by explicitly constructing an approximation $u_n \in V_n$ decomposed on a basis $(\varphi_1, \cdots, \varphi_n)$ of V_n:

$$u_n = \sum_{j=1}^{n} u_j \; \varphi_j$$

In so doing, the constructive approach paves the way to the Finite Element Method and is thus chosen as a prequel to establishing the Galerkin method.

Writing the minimisation functional for u_n reads:

$$\begin{aligned} J(u_n) &= \frac{1}{2} a(u_n, u_n) - L(u_n) \\ &= \frac{1}{2} a(\sum_{j=1}^{n} u_j\varphi_j, \sum_{i=1}^{n} u_i\varphi_i) - L(\sum_{j=1}^{n} u_i\varphi_i) \\ &= \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} a(u_j\varphi_j, u_i\varphi_i) - \sum_{j=1}^{n} L(u_i\varphi_i) \\ &= \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} u_j u_i a(\varphi_j, \varphi_i) - \sum_{j=1}^{n} u_i L(\varphi_i) \end{aligned}$$

Collecting the entries by index i, the functional can be rewritten under algebraic form:

$$\mathrm{J}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^t \mathrm{A}\mathbf{u} - \mathbf{u}^t\mathbf{b}$$

where $\mathbf{u}$ is the unknow vector:

$$\mathbf{u}^t = (u_1, \ldots, u_n)$$

and A, $\mathbf{b}$ are respectively the stiffnes matrix and load vector:

$$\mathrm{A}_{ij} = a(\varphi_j, \varphi_i), \mathbf{b}_i = L(\varphi_i)$$

Owing to Proposition 2.2,J is a strictly convex quadratic form, then there exists a unique $\mathbf{u} \in \mathbb{R}^n \; : \; \mathrm{J}(\mathbf{u}) \leq \mathrm{J}(\mathbf{v}), \forall \; \mathbf{v} \in \mathbb{R}^n$, which in turns proves the existence and uniqueness of $u_n \in V_n$.

The minimum is achieved with $\mathbf{u}$ satisfying $\mathrm{A}\mathbf{u} = \mathbf{b}$ which corresponds to the Euler condition $J'(u_n) = 0$ □

Proposition 2.2 (Convexity of a quadratic form)**.**

$$\mathrm{J}(\mathbf{u}) = \mathbf{u}^t\mathrm{K}\mathbf{u} - \mathbf{u}^t\mathrm{G} + \mathrm{F}$$

is a strictly convex quadratic functional iff K *symmetric positive definite non-singular.*

2.2.3 Convergence

The question in this section is: considering a sequence of discrete solutions $(u_n)_{n\in\mathbb{N}}$, with each u_n belonging to V_n, can we prove that $u_n \to u$ in V as $n \to \infty$?

Lemma 2.3 (Estimate in the energy norm)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V. We denote by $u \in V$, $u_n \in V_n$ respectively the solution to Problem* (8) *and the solution to approximate Problem* (9)*. Let us define the energy norm $\|\cdot\|_a = a(\cdot\,,\cdot)^{1/2}$, then the following inequality holds:*

$$\|u - u_n\|_a \leq \|u - v_n\|_a \quad , \forall\, v_n \in V_n$$

Proof. Using the coercivity and the continuity of the bilinear form, we have:

$$\alpha\,\|u\|_V^2 \leq \|u\|_a^2 \leq \|u\|_V^2$$

then $\|u\|_a$ is norm equivalent to $\|u\|_V$, thus $(V, \|\cdot\|_a)$ is a Hilbert space.

$$a(u - \mathrm{P}_{V_n} u, v_n) = 0 \quad ,\forall\, v_n \in V_n$$

by definition of P_{V_n} as the orthogonal projection of u onto V_n with respect to the scalar product defined by the bilinear form a.

$$\|u - u_n\|_a^2 = a(u - u_n, u - v_n) + a(u - u_n, v_n - u_n) \quad ,\forall\, v_n \in V_n$$

Since the second term of the right-hand side cancels due to the consistency of the approximation, we deduce $u_n = \mathrm{P}_{V_n} u$, then u_n minimizes the distance from u to V_n:

$$\|u - u_n\|_a^2 \leq \|u - v_n\|_a^2 \quad ,\forall\, v_n \in V_n$$

which means that the error estimate is *optimal* in the energy norm. □

Lemma 2.4 (Céa's Lemma)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V. we denote by $u \in V$, $u_n \in V_n$ respectively the solution to Problem* (8) *and the solution to approximate Problem* (9) *, then the following inequality holds:*

$$\|u - u_n\|_V \leq \sqrt{\frac{M}{\alpha}}\,\|u - v_n\|_V \quad , \forall\, v_n \in V_n$$

with $M > 0$ the continuity constant and $\alpha > 0$ the coercivity constant.

Proof. Using the coercivity and continuity of the bilinear form, we bound the left-hand side of the estimate (2.3) from below and its right-hand side from above:

$$\alpha\,\|u - u_n\|_V^2 \leq M\,\|u - v_n\|_V^2 \quad \forall\, v_n \in V_n$$

Consequently:

$$\|u - u_n\|_V \leq \sqrt{\frac{M}{\alpha}}\,\|u - v_n\|_V \quad , \forall\, v_n \in V_n$$

□

Lemma (2.4) gives a control on the discretisation error $e_n = u - u_n$ which is *quasi-optimal* in the V-norm (*i.e.* bound multiplied by a constant).

Lemma 2.5 (Stability)**.** *Any solution $u_n \in V_n$ to Problem* (9) *satisfies:*

$$\|u_n\|_V \leq \frac{\|L\|_{V'}}{\alpha}$$

Proof. Direct using the coercivity and the dual norm. □

2.2.4 Method

Algorithm 2.6 (Ritz's method)**.** *The following procedure applies:*

1. *Chose an approximation space* V_n
2. *Construct a basis* $\mathcal{B} = (\varphi_1, \ldots, \varphi_n)$
3. *Assemble stiffness matrix* A *and load vector* $\mathbf{b}$
4. *Solve* $\mathrm{A}\mathbf{u} = \mathbf{b}$ *as a minimisation problem*

2.3 Galerkin method

2.3.1 Formulation

We use a similar approach as for Ritz's method, except that the abstract problem does not require the symmetry of the bilinear form. Therefore we cannot endow V with a norm defined from the scalar product based on $a(\,\cdot\,,\,\cdot\,)$.

Problem (12) is the approximate weak problem by Galerkin's method:

$$\left|\begin{array}{l}\text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:}\\[4pt] a(u_n, v_n) = L(v_n) \quad , \forall\, v_n \in V_n\end{array}\right. \tag{12}$$

with $a(\,\cdot\,,\,\cdot\,)$ a coercive continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ a continuous linear form on V.

2.3.2 Convergence

The following property is merely a consequence of the consistency, as the continuous solution u is solution to the discrete problem (*i.e.* the bilinear form is the "same"), but it is quite useful to derive error estimates in Section 5. Consequently, whenever needed we will refer to the following proposition:

Proposition 2.7 (Galerkin orthogonality)**.** *Let* $u \in V$*,* $u_n \in V_n$ *respectively the solution to Problem* (8) *and the solution to approximate Problem* (12)*, then:*

$$a(\ u - u_n, v_n\ ) = 0 \quad , \ \forall\, v_n \in V_n$$

Proof. Direct consequence of the consistency of the method. □

Lemma 2.8 (Consistency)**.** *Let* V *be a Hilbert space and* V_n *a finite dimensional subspace of* V*. we denote by* $u \in V$*,* $u_n \in V_n$ *respectively the solution to Problem* (8) *and the solution to approximate Problem* (12)*, then the following inequality holds:*

$$\|u - u_n\|_V \leq \frac{M}{\alpha} \|u - v_n\|_V \quad , \ \forall\, v_n \in V_n$$

with $M > 0$ *the continuity constant and* $\alpha > 0$ *the coercivity constant.*

Proof. Using the coercivity:

$$\begin{aligned}
\alpha \|u - u_n\|_V^2 &\leq a(u - u_n, u - u_n)\\
&\leq a(u - u_n, u - v_n) + \underbrace{a(u - u_n, v_n - u_n)}_{0}\\
&\leq a(u - u_n, u - v_n)\\
&\leq M \|u - u_n\|_V \|u - v_n\|_V\\
\|u - u_n\|_V &\leq \frac{M}{\alpha} \|u - v_n\|_V
\end{aligned}$$

□

The only difference with the symmetric case is that the constant is squared due to the loss of the symmetry.

2.3.3 Well-posedness

Theorem 2.9 (Lax–Milgram). *Let V be a Hilbert space. Provided that $a(\,\cdot\,,\,\cdot\,)$ is a coercive continuous bilinear form on $V \times V$ and $L(\,\cdot\,)$ is a continuous linear form on V, Problem* (6) *admits a unique solution $u \in V$.*

Proof. □

2.3.4 Method

Algorithm 2.10 (Galerkin's method). *The following procedure applies:*

1. *Chose an approximation space V_n*
2. *Construct a basis $\mathcal{B} = (\varphi_1, \dots, \varphi_n)$*
3. *Assemble stiffness matrix* A *and load vector* $\mathbf{b}$
4. *Solve* $\mathrm{A}\mathbf{u} = \mathbf{b}$

2.4 Exercises

Exercise 2.11 (Ritz Galerkin method — 1). Let us consider Problem (7) from the previous chapter:

1. Formulate the approximation to this problem using the (finite dimensional) space of continuous functions, piecewise linear on intervals

$$\left\{ \left[0, \frac{1}{4}\right], \left[\frac{1}{4}, \frac{1}{2}\right], \left[\frac{1}{2}, \frac{3}{4}\right], \left[\frac{3}{4}, 1\right] \right\}$$

2. Write the basis functions.
3. Construct the linear system explicitely.

Exercise 2.12 (Ritz Galerkin method — 2). Repeat the previous exercise with boundary condition $u(1) = 2$.

Exercise 2.13 (Lax Milgram). Consider the following problem posed on $\Omega \subset \mathbb{R}^2$:

$$\left|\begin{array}{rcll} \multicolumn{4}{l}{\text{Find } u \in C^2(\bar{\Omega}) \text{ such that:}} \\ -\boldsymbol{\nabla}\cdot(k(\boldsymbol{x}) \cdot \boldsymbol{\nabla} u(\boldsymbol{x})) + r(\boldsymbol{x})u & = & f(\boldsymbol{x}) & , \forall\, \boldsymbol{x} \in \Omega \\ u & = & 0 & , \forall\, \boldsymbol{x} \in \partial\Omega \end{array}\right. \tag{13}$$

with the source term $f \in L^2(\Omega)$, the diffusive coefficient $0 < \alpha \leq k(x) \leq \beta$, and the reaction coefficient $0 < \alpha \leq r(x) \leq \beta$.

1. Formulate the weak problem.
2. Show that assumptions of the Lax-Milgram theorem hold.

3 Finite Element spaces

In the previous lectures we have studied the properties of coercive problems in an abstract setting and described Ritz and Galerkin methods for the approximation of the solution to a PDE, respectively in the case of symmetric and non-symmetric bilinear forms.

The abstract setting reads:

$$\left| \begin{array}{l} \text{Find } u_h \in V_h \subset \mathrm{H} \text{ such that:} \\ a(u_h, v_h) = L(v_h) \quad , \; \forall \; v_h \in V_h \end{array} \right.$$

such that:

- V_h is a finite dimensional approximation space characterized by a discretization parameter h,
- $a(\,\cdot\,,\,\cdot\,)$ is a continuous bilinear form on $V_h \times V_h$, coercive w.r.t $\|\cdot\|_V$,
- $L(\,\cdot\,)$ is a continuous linear form.

Under these assumptions existence and uniqueness of a solution to the approximate problem holds owing to the Lax–Milgram Theorem and u_h is called discrete solution. Provided this abstract framework which allows us to seek approximate solutions to PDEs, we need to chose the approximate space V_h and construct a basis $\mathcal{B} = (\varphi_1, \cdots, \varphi_N)$ of V_h on which the discrete solution is decomposed:

$$u_h = \sum_{j=1}^{N_{V_h}} u_j \; \varphi_j$$

with $N_{V_h} = \dim(V_h)$, $\{u_j\}$ a family of N_{V_h} real numbers called *global degrees of freedom* and $\{\varphi_j\}$ a family of N_{V_h} elements of V_h called *global shape functions.*

To construct the approximate space V_h, we need two ingredients:

1. An admissible mesh $\mathcal{T}_h$ generated by a tesselation of domain Ω.
2. A reference finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ to construct a basis of V_h.

3.1 Admissible mesh

Definition 3.1 (Mesh)**.** Let Ω be polygonal ($d = 2$) or polyhedral ($d = 3$) subset of $\mathbb{R}^d$, we define $\mathcal{T}_h$ (a triangulation in the simplicial case) as a finite family $\{K_i\}$ of disjoints non-empty subsets of Ω named cells. Moreover $\mathcal{N}_h = \{\mathcal{N}_i\}$ denotes the set a vertices of $\mathcal{T}_h$ and $\varepsilon_h = \{\sigma_{KL} = K \cap L\}$ denotes the set of edges.

Definition 3.2 (Mesh size)**.**

$$h_{\mathcal{T}} = \max_{K \in \mathcal{T}_h} (\mathrm{diam}(K))$$

Definition 3.3 (Geometrically conforming mesh)**.** A mesh is said geometrically conforming if two neighbouring cells share either exactly one vertex, exactly one edge, or in the case $d = 3$ exactly one facet.

The meaning of the previous condition is that there should not be any “hanging node” on a facet. Moreover some theoretical results require that the mesh satisfies some regularity condition: for example, bounded ratio of equivalent ball diameter, Delaunay condition on the angles of a triangle, . . .

3.2 Reference Finite Element

Definition 3.4 (Finite Element – [4] page 19, [2] page 69)**.** A Finite Element consists of a triple $(K, \mathcal{P}, \Sigma)$, such that

- K is a compact, connected subset of $\mathbb{R}^d$ with non-empty interior and with regular boundary (typically Lipshitz continuous),
- $\mathcal{P}$ is a finite dimensional vector space, $\dim(\mathcal{P}) = N$, of functions $p : K \to \mathbb{R}$, which is the space of shape functions,
- Σ is a set $\{\sigma\}_j$ of linear forms,

  $$\begin{array}{rccl} \sigma_j : & \mathcal{P} & \to & \mathbb{R} \\ & p & \mapsto & p_j = \sigma_j(p) \end{array} \quad , \forall\, j \in [\![1, N]\!]$$

  which is a basis of $\mathcal{L}(\mathcal{P}, \mathbb{R})$, the dual of $\mathcal{P}$.

Practically, the definition constructs first the Finite Element on a cell K which can be an interval ($d = 1$), a polygon ($d = 2$) or a polyhedron ($d = 3$) (Example: triangle, quadrangle, tetrahedron, hexahedron). Then an approximation space $\mathcal{P}$ (Example: polynomial space) and the local degrees of freedom Σ are chosen (Example: value at N geometrical nodes $\{a_i\}$, $\sigma_i(\varphi_j) = \varphi_j(a_i)$). The local shape functions $\{\varphi_i\}$ are then constructed so as to ensure unisolvence.

Proposition 3.5 (Determination of the local shape functions)**.** *Let $\{\sigma_i\}_{1 \leq i \leq N}$ be the set of local degrees of freedoms, the local shape functions are defined as $\{\varphi_i\}_{1 \leq i \leq N}$ a basis of $\mathcal{P}$ such that,*

$$\sigma_i(\varphi_j) = \delta_{ij} \quad , \forall\, i, j \in [\![1, N]\!]$$

Definition 3.6 (Unisolvence)**.** A Finite Element is said unisolvent if for any vector $(\alpha_1, \cdots, \alpha_N) \in \mathbb{R}^N$ there exists a unique representant $p \in \mathcal{P}$ such that $\sigma_i(p) = \alpha_i$, $\forall \ \in [\![1, N]\!]$.

The unisolvence property of a Finite Element is equivalent to construct Σ as dual basis of $\mathcal{P}$, thus we can express any function $p \in \mathcal{P}$ as

$$p = \sum_{j=1}^{N} \sigma_j(p)\ \varphi_j$$

the unique decomposition on $\{\varphi_j\}$, with $p_j = \sigma_j(p)$ the j-th degree of freedom. In other words, the choice of $\Sigma = \{\sigma_j\}$ ensures that the vector of degree of freedoms $(p_1, \cdots, p_N)$ uniquely defines a function of $\mathcal{P}$. Defining Σ as dual basis of $\mathcal{P}$ is equivalent to:

$$\dim(\mathcal{P}) = \mathrm{card}(\Sigma) = N \tag{14a}$$

$$\forall\, p \in \mathcal{P},\ (\sigma_i(p) = 0, 1 \leq i \leq N) \Rightarrow (p = 0) \tag{14b}$$

in which Property (14a) ensures that Σ generates $\mathcal{L}(\mathcal{P}, \mathbb{R})$ and Property (14b) that $\{\sigma_i\}$ are linearly independent.

Usually the unisolvence is part of the definition of a Finite Element since chosing the shape functions such that $\sigma_i(\varphi_j) = \delta_{ij}$ is equivalent.

Definition 3.7 (Local interpolation operator – [4] page 20)**.**

$$\begin{array}{rccl} \pi_K : & V(K) & \to & \mathcal{P} \\ & v & \mapsto & \displaystyle\sum_{j=1}^{N} \sigma_j(v)\,\varphi_j \end{array}$$

Remark 3.8. The notation using the dual basis can be confusing but with the relation $\sigma_i(p) = p(a_i)$ in the nodal Finite Element case it is easier to understand that the set Σ of linear forms defines how the interpolated function $\pi_h\, u$ "represents" its infinite dimensional counterpart u through the definition of the degrees of freedom. In the introduction, we defined simply $u_i = \sigma_i(u)$ without expliciting it. A natural choice is the pointwise representation $u_i = u(a_i)$ at geometrical nodes $\{a_i\}$, which is the case of Lagrange elements, but it is not the only possible choice ! For example, σ_i can be:

- a mean flux trough each facet of the element (Raviart–Thomas)

$$\sigma_i(v) = \int_\xi v \cdot \boldsymbol{n}_\xi \,\mathrm{d}s$$

- a mean value over each facet of the element (Crouzeix–Raviart)

$$\sigma_i(v) = \int_\xi v \,\mathrm{d}s$$

- a mean value of the tangential component over each facet of the element (Nédelec)

$$\sigma_i(v) = \int_\xi v \cdot \tau_\xi \,\mathbf{ds}$$

A specific choice of linear form allows a control on a certain quantity: divergence for the first two examples, and curl for the third. The approximations will then not only be H^s-conformal but also include the divergence or the curl in the space.

3.3 Transport of the Finite Element

In practice to avoid the construction of shape functions for any Finite Element $(K, \mathcal{P}, \Sigma)$, $K \in \mathcal{T}_h$, the local shape functions are evaluated for a *reference Finite Element* $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ defined on a *reference cell* $\hat{K}$ and then transported onto any cell K of the mesh. For example, in the case of simplicial meshes the reference cell in one dimension is the unit interval $[0, 1]$, in two dimension the unit triangle with vertices $\{(0, 0), (0, 1), (1, 0)\}$. In so doing, we can generate any Finite Element $(K, \mathcal{P}, \Sigma)$ on the mesh from $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ provided that we can construct a mapping such that $(K, \mathcal{P}, \Sigma)$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ are equivalents.

Definition 3.9 (Equivalent Finite Elements)**.** Two Finite Elements $(K, \mathcal{P}, \Sigma)$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ are said *equivalent* if there exists a bijection T_K from $\hat{K}$ onto K such that:

$$\forall\, p \in \mathcal{P},\ p \circ T_K \in \hat{\mathcal{P}}$$

and

$$\Sigma = T_K(\hat{\Sigma})$$

By collecting the local shape functions and local degrees of freedom from all the generated $(K, \mathcal{P}, \Sigma)$ on the mesh, we then construct *global shape functions* and *global degrees of freedom* and thus the approximation space V_h.

For Lagrange elements the transformation used to transport the Finite Element on the mesh is an *affine mapping*, but this is not suitable in general !

3.4 Numerical integration

The contributions are integrated numerically, usually using quadrature rules.

3.5 Method

Algorithm 3.10 (Finite Element Method)**.** *Solving a problem by a Finite Element Method is defined by the following procedure:*

1. *Choose a reference finite element* $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$.

2. *Construct an admissible mesh* $\mathcal{T}_h$ *such that any cell* $K \in \mathcal{T}_h$ *is in bijection with the reference cell* $\hat{K}$.

3. *Define a mapping to transport the reference finite element defined on* $\hat{K}$ *onto any* $K \in \mathcal{T}_h$ *to* $(K, \mathcal{P}, \Sigma)$.

4. *Construct a basis for* V_h *by collecting all the finite element basis of finite elements* $\{(K, \mathcal{P}, \Sigma)\}_{K \in \mathcal{T}_h}$ *sharing the same degree of freedom.*

Remark 3.11. The Finite Element approximation is said H-conformal if $V_h \subset$ H and is said non-conformal is $V_h \not\subset$ H. In this latter case the approximate problem can be constructed by building an approximate bilinear form

$$a_h(\,\cdot\,,\,\cdot\,) = a(\,\cdot\,,\,\cdot\,) + s(\,\cdot\,,\,\cdot\,)$$

as described, for instance, in the case of stabilized methods for advection-dominated problems in Section 8.3.

3.6 Exercises

3.6.1 Reference element, affine mapping

Consider the triangular element K with vertices having coordinates,$v_1 = (0,0), v_2 = (0.2, 0.2), v_3 = (0.1, 0.6)$ and piecewice linear basis functions Φ_1, Φ_2, Φ_3 where $\Phi_i(v_j) = \delta_{ij}$

Compute the integral $\int_K \nabla\Phi_3.\nabla.\Phi_3 dx$ by first finding the formula for Φ_3.

Consider the reference element $\hat{K}$ with vertices having coordinates $(0,0),(1,0),(0,1)$. Find an isoparametric mapping to map points in $\hat{K}$ to K. Compute the same integral using this mapping.

3.6.2 Local element matrix, local load vector

Compute a local element matrix and a local load vector for one cell for the first exercise Ritz-Galerkin methods chapter.

4 Simplicial Lagrange Finite Element

4.1 Polynomial interpolation in one dimension

Let $\mathbb{P}^k([a,b])$ be the space of polynomials $p = \sum_{i=0}^{k} \alpha_i x^i$ of degre lower or equal to k on the interval $[a,b]$, with $c_i x^i$ the monomial of order i, c_i a real number.

A natural basis of $\mathbb{P}^k([a,b])$ consists of the set of monomials $\left\{1, x, x^2, \cdots, x^k\right\}$. We can verify that its elements are linearly independent. But in the frame of Finite Elements we can chose another basis which is the Lagrange basis $\left\{\mathcal{L}_i^k\right\}_{0\leq i\leq k}$ of degree k defined on a set of $k+1$ points $\{\xi_i\}_{0\leq i\leq k}$ which are called *Lagrange nodes.*

Definition 4.1 (Lagrange polynomials – [4] page 21, [6] page 76)**.** The Lagrange polynomial of degree k associated with node ξ_m reads:

$$\mathcal{L}_m^k(x) = \frac{\prod_{\substack{i=0\\i\neq m}}^{k} (x - \xi_i)}{\prod_{\substack{i=0\\i\neq m}}^{k} (\xi_m - \xi_i)}$$

Proposition 4.2 (Nodal basis – [4])**.** *Lagrange polynomials form a nodal basis i.e.*

$$\mathcal{L}_i^k(\xi_j) = \delta_{ij} \quad , \ 0 \leq i,j \leq k$$

The following result gives a pointwise control of the interpolation error:

Theorem 4.3 (Pointwise interpolation inequality – [6] page 79)**.** *Let* $u \in \mathrm{C}^{k+1}([a,b])$ *and* $\pi_k\, u \in \mathbb{P}^k([a,b])$ *its Lagrange interpolate of order k, with Lagrange nodes* $\{\xi_i\}_{0\leq i\leq k}$*, then* $\forall\ x \in [a,b]$*:*

$$|u(x) - \ \pi_k\, u(x)| \leq \left| \frac{\prod_{i=0}^{k}(x-\xi_i)}{(k+1)!} \right| \max_{[a,b]} \left|\partial^{k+1} u\right|$$

4.2 A nodal element

Let us take $\{\xi_1, \cdots, \xi_N\}$ a family of points of K such that $\sigma_i(p) = p(\xi_i)$, $1 \leq i \leq N$:

- $\{\xi\}_{1\leq i\leq N}$ is the set of *geometrical nodes*,
- $\{\varphi_i\}_{1\leq i\leq N}$ is a *nodal basis* of $\mathcal{P}$, *i.e.* $\varphi_i(\xi_j) = \delta_{ij}$.

We can verify, for any $p \in \mathcal{P}$ that:

$$p(\xi_j) = \sum_{i=1}^{N} \sigma_i(p)\ \underbrace{\varphi_i(\xi_i)}_{\delta_{ij}} \quad , \ 1 \leq i,j \leq N$$

which reduces to:

$$p(\xi_j) = \sigma_i(p)$$

Remark 4.4 (Support of shape functions)**.** The polynomial basis being defined such that

$$\begin{cases} \varphi_i(\xi_i) = 1 \\ \varphi_i(\xi_j) = 0 \ , \ i \neq j \end{cases}$$

then any shape function φ_i is compactly supported on the union of cells containing the node ξ_i.

The Lagrange polynomials (4.1) are used to build directly the one-dimensional shape functions, while in higher dimensions the expression of the shape functions is reformulated in term of barycentric coordinates:

$$\begin{aligned}\lambda_i : \quad \mathbb{R}^d &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto \lambda_i(\boldsymbol{x}) = 1 - \frac{(\boldsymbol{x} - \xi_i) \cdot \boldsymbol{n}_i}{(\xi_f - \xi_i) \cdot \boldsymbol{n}_i}\end{aligned}$$

with $\boldsymbol{n}_i$ the unit outward normal to the facet opposite to ξ_i, and ξ_f a node belonging to this facet.

Example 4.5 (Lagrange elements of polynomial degree $k = 1, 2$ on triangle)**.** The shape functions are given by:

$$\begin{array}{llll} k = 1, & \varphi_i & = & \lambda_i & , \ 0 \leq i \leq d \\ k = 2, & \varphi_i & = & \lambda_i(2\lambda_i - 1) & , \ 0 \leq i \leq d \\ & \varphi_i & = & 4\lambda_i\lambda_j & , \ 0 \leq i \leq d \end{array}$$

4.3 Reference Finite Element

4.3.1 Examples in one dimension

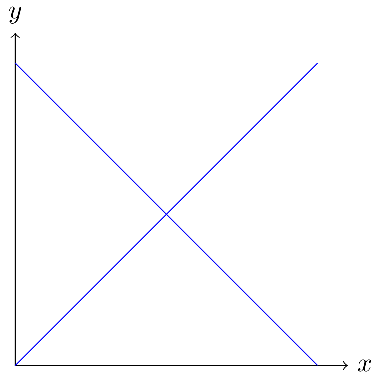

Figure 1: Lagrange $\mathbb{P}^1$ on the unit interval.

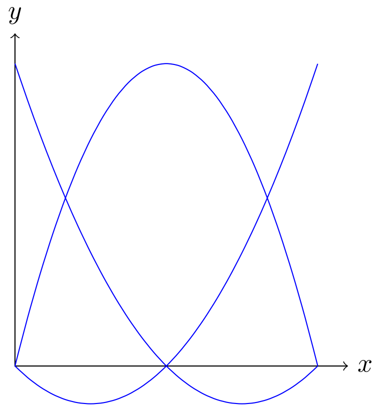

Figure 2: Lagrange $\mathbb{P}^2$ on the unit interval.

4.4 Formulation of the Poisson problem

The approximate problem of Problem (5) by Lagrange $\mathbb{P}^1$ elements reads:

$$\left|\begin{array}{l} \text{Find } u \in V_h\text{, given } f \in \mathrm{L}^2(\Omega)\text{, such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, \mathrm{d}\boldsymbol{x} = \int_\Omega f v \, \mathrm{d}\boldsymbol{x} \quad , \forall\, v \in V_h \end{array}\right. \tag{15a}$$

with the approximation space V_h chosen as:

$$V_h = \left\{ v \in \mathrm{C}^0(\Omega) \cap \mathrm{H}^1_0(\Omega) : v|_K \in \mathbb{P}^1(K), \forall\, K \in \mathcal{T}_h \right\} \tag{15b}$$

4.5 Exercises

5 Error analysis

The goal of this section is to bound the discretisation error $e_h = u - u_h$ in a Sobolev or Lebesgue norm. To this purpose we have already two ingredients:

— on the one hand, i in the analysis of Ritz's and Galerkin's methods, consistency estimates like Cea's Lemma gives a control on the discretisation error in the solution space V in term of "distance" between the solution space and the approximation space:

$$\|u - u_h\|_V \leq C \, \|u - v_h\|_V \quad , \forall \, v_h \in V_h$$

with $C > 0$ a constant real number,

— on the other hand, the pointwise interpolation inequality of Theorem (4.3) gives a control on the interpolation error $e_\pi = u - \pi_k \, u$.

Question: Can we control the discretization error by bounding the right-hand side of the consistency inequality using interpolation properties ?

5.1 *A priori* error estimate with Lagrange $\mathbb{P}^1$

Theorem 5.1 (Interpolation inequality in $\mathrm{H}^1_0(\Omega)$ and $\mathrm{L}^2(\Omega)$)**.** *Let $v \in \mathrm{H}^2(\Omega)$, $\exists C_1 > 0$ such that*

$$|v - v_h|_{\mathrm{H}^1(\Omega)} \leq C_1 h_{\mathcal{T}} \; |v|_{\mathrm{H}^2(\Omega)} \tag{16}$$

and $\exists C_0 > 0$ such that

$$\|v - v_h\|_{\mathrm{L}^2(\Omega)} \leq C_0 h_{\mathcal{T}}^2 \, |v|_{\mathrm{H}^2(\Omega)} \tag{17}$$

with $h_{\mathcal{T}} = \max_{K \in \mathcal{T}_h}(h_K)$.

Proof. The proof is based on the Mean-Value Theorem and a decomposition of the error per element. The global interpolation error is then recovered by summing over the cells. This makes sense since the polynomial estimate is defined pointwise: this is then a local property. In the same spirit the stability of the interpolation operator is also a local property, defined element-wise. □

The discretization error being bounded in $O(h_{\mathcal{T}})$ the method is first order in $\mathrm{H}^1_0(\Omega)$.

Remark 5.2 (Convergence order in $\mathrm{H}^1(\Omega)$)**.** On the other hand we know that $\exists C_I > 0$ such that, $\forall \, v \in \mathrm{H}^2(\Omega)$:

$$\|v - \, \pi_1 \, v\|_{\mathrm{L}^2(\Omega)} \leq C_I \; h_{\mathcal{T}}^2 \, |v|_{\mathrm{H}^2(\Omega)}$$

Using the definition of the norm

$$\|v - \, \pi_1 \, v\|^2_{\mathrm{H}^1(\Omega)} = \|v - \, \pi_1 \, v\|^2_{\mathrm{L}^2(\Omega)} + |v - \, \pi_1 \, v|^2_{\mathrm{H}^1(\Omega)}$$

we get:

$$\|v - \, \pi_1 \, v\|^2_{\mathrm{H}^1(\Omega)} \leq C_I^2 \; (h_{\mathcal{T}}^4 \, |v|^2_{\mathrm{H}^2(\Omega)} + h_{\mathcal{T}}^2 \, |v|^2_{\mathrm{H}^2(\Omega)})$$

$$\|v - \, \pi_1 \, v\|_{\mathrm{H}^1(\Omega)} \leq C_I \; h_{\mathcal{T}} \; (1 + h_{\mathcal{T}}^2)^{1/2} \, |v|_{\mathrm{H}^2(\Omega)}$$

Thus, we verify that the approximation is also first order in $\mathrm{H}^1(\Omega)$.

5.2 Superconvergence

The following result shows the the convergence properties of the method is not only limited by the interpolation inequality. Indeed, using a result by Aubin–Nitsche, we show that even if the approximation is not H^2-conformal, we can improve the error estimate by one order: the convergence order in $\mathrm{L}^2(\Omega)$ becomes then two.

Theorem 5.3 (Superconvergence)**.** *Let Ω be a convex polygonal subset of $\mathbb{R}^d$, $d \geq 1$, $f \in \mathrm{L}^2(\Omega)$, u solution to the Dirichlet Problem* (1) *and u_h approximate solution, $h_{\mathcal{T}} = \max_{K \in \mathcal{T}_h}(h_K)$:*

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq C_1\ h_{\mathcal{T}} \quad \textit{and} \quad \|u - u_h\|_{\mathrm{L}^2(\Omega)} \leq C_0\ h_{\mathcal{T}}^2$$

Proof. If u is solution to the Poisson problen then $u \in \mathrm{H}^1_0(\Omega)$, then by regularity Theorem (**??**) (by density of $\mathrm{H}^2(\Omega)$ in $\mathrm{H}^1(\Omega)$), $u \in \mathrm{H}^2(\Omega)$, thus $\exists C_1 > 0$ such that:

$$\|u\|_{\mathrm{H}^2(\Omega)} \leq C_1\ \|f\|_{\mathrm{L}^2(\Omega)}$$

Thus replacing the H^2-seminorm in the right-hand side of the error estimate, we have

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq C_1\ h_{\mathcal{T}}\ \|f\|_{\mathrm{L}^2(\Omega)} \tag{18}$$

Let us introduce the following auxiliary problem:

$$-\boldsymbol{\Delta}\varphi(\boldsymbol{x}) = e_h(\boldsymbol{x}) \quad , \ x \in \Omega \tag{19a}$$

$$\varphi(\boldsymbol{x}) = 0 \quad , \ x \in \partial\Omega \tag{19b}$$

and its weak formulation:

$$\left| \begin{array}{l} \text{Find } \varphi \in \mathrm{H}^1_0(\Omega), \text{ given } e_h \in \mathrm{L}^2(\Omega), \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\phi \, \mathrm{d}\boldsymbol{x} = \int_\Omega e_h \phi \, \mathrm{d}\boldsymbol{x} \quad , \ \forall\ \phi \in \mathrm{H}^1_0(\Omega) \end{array} \right. \tag{20}$$

Since e_h is bounded in $\mathrm{L}^2(\Omega)$ then the same regularity result holds for the auxiliary Problem (19), $\exists C_2 > 0$ such that:

$$\|\varphi\|_{\mathrm{H}^2(\Omega)} \leq C_2\ \|e_h\|_{\mathrm{L}^2(\Omega)}$$

and thus: estimate, we have

$$\|\varphi - \varphi_h\|_{\mathrm{H}^1(\Omega)} \leq C_2\ h_{\mathcal{T}}\ \|e_h\|_{\mathrm{L}^2(\Omega)} \tag{21}$$

Let us try to bound the L^2-norm of the discretization error by noticing that its amounts to take $\phi = e_h$ in (20):

$$\|e_h\|_{\mathrm{L}^2(\Omega)} = \int_\Omega |e_h|^2 \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}e_h \, \mathrm{d}\boldsymbol{x}$$

If we consider the approximate of Problem (20) by Galerkin's method, with $\varphi_h \in V_h$ its solution, then the Galerkin orthogonality reads:

$$\int_\Omega \boldsymbol{\nabla}\varphi_h \cdot \boldsymbol{\nabla}e_h \, \mathrm{d}\boldsymbol{x} = 0$$

Thus we can substract and add this latter to the previous expression:

$$\|e_h\|_{\mathrm{L}^2(\Omega)} = \int_\Omega \boldsymbol{\nabla}(\varphi - \varphi_h) \cdot \boldsymbol{\nabla}e_h \, \mathrm{d}\boldsymbol{x} + \underbrace{\int_\Omega \boldsymbol{\nabla}\varphi_h \cdot \boldsymbol{\nabla}e_h \, \mathrm{d}\boldsymbol{x}}_{0}$$

First we use Cauchy-Schwarz and make the H^1-norm of the discretization errors appear since we control them by Equation (18) and (21):

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq \|\varphi - \varphi_h\|_{\mathrm{H}^1(\Omega)} \, \|e_h\|_{\mathrm{H}^1(\Omega)}$$

Replacing by the bounds from the interpolation inequalities we get:

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_1 \; C_2 \; h_{\mathcal{T}}^2 \;\; \|f\|_{\mathrm{L}^2(\Omega)}$$

which concludes the proof. We have then a second order error estimate in L^2. □

5.3 Exercises

5.3.1 A priori error estimation

Consider the following differential equation

$$\begin{aligned} -u''(x) + u(x) = f(x), & x \in (0,1) \\ u(0) = u(1) = 0 & \end{aligned}$$

Derive a weak formulation and galerkin discretization with appropriate functional spaces, prove a priori error estimates in the energy norm $\|v\|_E$ where $\|v\|_E^2 = \|v'\|_{L_2(\Omega)}^2 + \|v\|_{L_2(\Omega)}^2$

6 Time-dependent problems

The objective of this section is to introduce the *a priori* stability analysis of time-dependent problems on several examples to obtain estimates similar to Lemma (**??**).

6.1 Time marching schemes

6.2 *A priori* stability estimates

6.2.1 Heat equation

Firstly, let us derive the energy estimate for the heat equation in the by recalling the weak form and then taking the test function to be the unknown u:

$$\begin{aligned}
\int_\Omega \partial_t u\, v \,\mathrm{d}\boldsymbol{x} + \kappa \int_\Omega \boldsymbol{\nabla} u . \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} &= \int_\Omega f\, v \,\mathrm{d}\boldsymbol{x} \\
\int_\Omega \partial_t u\, u \,\mathrm{d}\boldsymbol{x} + \kappa \int_\Omega |\boldsymbol{\nabla} u|^2 \,\mathrm{d}\boldsymbol{x} &= \int_\Omega f\, u \,\mathrm{d}\boldsymbol{x} \\
\frac{1}{2} \int_\Omega \partial_t |u|^2 \,\mathrm{d}\boldsymbol{x} + \kappa \int_\Omega |\boldsymbol{\nabla} u|^2 \,\mathrm{d}\boldsymbol{x} &= \int_\Omega f\, u \,\mathrm{d}\boldsymbol{x} \\
\frac{1}{2} \frac{\mathrm{d}}{\mathrm{dt}} \int_\Omega |u|^2 \,\mathrm{d}\boldsymbol{x} + \kappa \int_\Omega |\boldsymbol{\nabla} u|^2 \,\mathrm{d}\boldsymbol{x} &= \int_\Omega f\, u \,\mathrm{d}\boldsymbol{x} \\
\frac{1}{2} \frac{\mathrm{d}}{\mathrm{dt}} \|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa \, |u|^2_{\mathrm{H}^1(\Omega)} \,\mathrm{d}\boldsymbol{x} &= \int_\Omega f\, u \,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

In the case of an homogeneous equation, the latest relation is directly the instantaneous conservation of energy

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{dt}} \|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa \, |u|^2_{\mathrm{H}^1(\Omega)} = 0 \tag{22}$$

with the first term being the variation of kinetic energy and the second term being the dissipation of energy with diffusion coefficient κ. Integrating over the time interval, we get the energy budget over $[0, T]$:

$$\frac{1}{2} \|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa \int_0^T |u|^2_{\mathrm{H}^1(\Omega)} \,\mathrm{d}t = 0 \tag{23}$$

Let us consider now a non-zero source term f, then using the Cauchy–Schwarz inequality yields the following relation:

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{dt}} \|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa \, |u|^2_{\mathrm{H}^1(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)} \, \|u\|_{\mathrm{L}^2(\Omega)} \tag{24}$$

Since the bound should depend only on the data, the name of the game is to absorb any term involving the unknown in the left-hand side. To this purpose, inequalities like Hölder, Korn, Sobolev injections are to be used in order to get a power of the proper L^p or H^s norm of the unknown. In the case of coercive problems, the diffusion term giving directly the H^1 seminorm (to a factor depending on the diffusive coefficient), we should try to make it pop from the right-hand side. Using first the Poincaré inequality (Lemma D.8) and then the Young inequality (Lemma D.3), we can bound the right-hand side by the data and the H^1 seminorm,

$$\|f\|_{\mathrm{L}^2(\Omega)} \, \|u\|_{\mathrm{L}^2(\Omega)} \leq \frac{1}{2\gamma^2 c_P^2} \|f\|^2_{\mathrm{L}^2(\Omega)} + \frac{\gamma^2}{2} |u|^2_{\mathrm{H}^1(\Omega)} \tag{25}$$

with γ a positive real number which can be chosen arbitrarily. Therefore, as soon as we choose $\gamma < \sqrt{2\kappa}$, it is possible to substract the second term of (25) to the left-hand side of the estimate, given that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{dt}}\|u\|^2_{\mathrm{L}^2(\Omega)} + \frac{2\kappa-\gamma^2}{2}|u|^2_{\mathrm{H}^1(\Omega)} \leq \frac{1}{2\gamma^2 c_P^2}\|f\|^2_{\mathrm{L}^2(\Omega)} \tag{26}$$

Consequently, taking $\gamma = \sqrt{\kappa}$ there exists a constant $C > 0$ depending on the Poincaré constant, such that

$$\frac{\mathrm{d}}{\mathrm{dt}}\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa\,|u|^2_{\mathrm{H}^1(\Omega)} \leq C(c_P)\,\|f\|^2_{\mathrm{L}^2(\Omega)} \tag{27}$$

This inequality yields a control of the L^2 norm and H^1 seminorn of the solution at any time t of the time interval $[0,T]$. Similarly to Equation (23), if we integrate over the time, we get

$$\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa\int_0^T |u|^2_{\mathrm{H}^1(\Omega)}\,\mathrm{d}t \leq C(c_P)\int_0^T \|f\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}t$$

which, by defining,

$$\|v\|_{\mathrm{L}^r(0,T;\mathrm{L}^p(\Omega))} = \left(\int_0^T \|v\|^r_{\mathrm{L}^p(\Omega)}\,\mathrm{d}t\right)^{1/r} \tag{28}$$

can be rewritten as

$$\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa\,\|u\|^2_{\mathrm{L}^2(0,T;\mathrm{H}^1_0(\Omega))} \leq C(c_P)\,\|f\|^2_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))}$$

The solution is said to be bounded in $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$, *i.e.* $u \in \mathrm{L}^2(\Omega)$ for almost every $t \in [0,T]$, and is it also bounded in $\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ by the data (provided that $f \in \mathrm{L}^2(0,T;\mathrm{L}^2(\Omega))$ of course).

Now, if we turn to the discrete case the estimate is not different aside from the the discrete time derivative. The term for the discrete time derivative in the case of backward Euler reads

$$\frac{1}{\delta t}\int_\Omega (u-u^*)\,u\,\mathrm{d}\boldsymbol{x}$$

with δt the current time step, u and u^* respectively the solution at the current and previous time stepping.

6.2.2 Wave equation

6.3 Well-posedness

6.4 Exercises

Exercise 6.1. Consider the following initial-boundary value problem

$$\begin{array}{rcll} \dot{u}(\boldsymbol{x}) - \boldsymbol{\Delta}u(\boldsymbol{x}) & = & 0 & ,\ (\boldsymbol{x},t) \in \Omega\times I \\ u & = & 0 & ,\ (\boldsymbol{x},t) \in \partial\Omega\times I \\ u(\boldsymbol{x},0) & = & u_0(\boldsymbol{x}) & ,\ \boldsymbol{x}\in\Omega \end{array}$$

1. Show the stability estimates

$$\|u(t)\|^2_{L_2(\Omega)} + \int_0^t \|\nabla u(s)\|^2_{L_2(\Omega)}\,ds \leq \|u(0)\|^2_{L_2(\Omega)}$$

$$\|\nabla u(t)\|^2_{L_2(\Omega)} + \int_0^t \|\triangle u(s)\|^2_{L_2(\Omega)}\,ds \leq \|\nabla u(0)\|^2_{L_2(\Omega)}$$

7 Adaptive error control

In Section 5. we derived *a priori* error estimates which give a control of the discretization error for any approximate solution. The order of convergence given by the exponent $O(h_{\mathcal{T}}^{\alpha})$ is an indication on "how close" to the continuous solution any approximate solution is expected to be. Provided that we are able to compute an approximate solution u_h, we want now to evaluate the "quality" of this solution in the sense of the residual of the equation: such an estimate is thus called *a posteriori* as it gives a quality measure of a computed solution.

Question: How can we evaluate the quality of a computed approximate solution in the sense of the residual of the equation ?

7.1 *A posteriori* error estimate

Let u and u_h be respectively the solutions to Problem (5) and Problem (15)

$$\begin{aligned}
|e_h|^2_{\mathrm{H}^1(\Omega)} &= \int_\Omega \boldsymbol{\nabla} e_h \cdot \boldsymbol{\nabla} e_h \,\mathrm{d}\boldsymbol{x} \\
&= \int_\Omega \boldsymbol{\nabla} e_h \cdot \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} \\
&= \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} \\
&= \int_\Omega f\; (e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

To obtain the residual $\mathcal{R}(u_h)$ we need to consider the equation element-wise, then integrating by part on any cell $K \in \mathcal{T}_h$, we obtain

$$\int_K \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} = \int_{\partial K} \boldsymbol{\nabla} u_h \cdot \boldsymbol{n}\; (e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_K \Delta u_h\; (e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x}$$

with

$$\mathcal{R}_K(u_h) = (f + \Delta u_h)|_K$$

Summing again over the domain yields

$$|e_h|^2_{\mathrm{H}^1(\Omega)} = \sum_{K \in \mathcal{T}_h} \left[\int_K \mathcal{R}_K(u_h)\; \boldsymbol{\nabla}(e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} + \int_{\partial K} \boldsymbol{\nabla} u_h \cdot \boldsymbol{n}\; (e_h - \pi_h\, e_h) \,\mathrm{d}\boldsymbol{x} \right]$$

noting that in the case of continuous elements the boundary term cancels. Using first the Cauchy–Schwarz inequality

$$|e_h|^2_{\mathrm{H}^1(\Omega)} \le \|\mathcal{R}(u_h)\|_{\mathrm{L}^2(\Omega)}\, \|e_h - \pi_h\, e_h\|_{\mathrm{L}^2(\Omega)}$$

then the interpolation inequality with constant C_I

$$|e_h|^2_{\mathrm{H}^1(\Omega)} \le C_I\, \|\mathcal{R}(u_h)\|_{\mathrm{L}^2(\Omega)}\, |h e_h|_{\mathrm{H}^1(\Omega)}$$

Consequenlty, we conclude

$$|e_h|_{\mathrm{H}^1(\Omega)} \le C_I h_{\mathcal{T}}\, \|\mathcal{R}(u_h)\|_{\mathrm{L}^2(\Omega)}$$

7.2 Dual weighted residual estimate

7.2.1 Adjoint operator

Definition 7.1 (Adjoint operator)**.** Let us define $\mathcal{A}^\star$, the adjoint operator of $\mathcal{A}$ as:

$$(\mathcal{A}u \,,\, v) = (u \,,\, \mathcal{A}^\star v)$$

Example 7.2 (Matrix of $\mathcal{M}_N(\mathbb{R})$)**.** Let $\mathcal{A} = \mathrm{A}$ be a real square matrix of dimension $N \times N$ and $x, y \in \mathbb{R}^N$:

$$(\mathcal{A}x \,,\, y) = (\mathrm{A}x \,,\, y) = (x \,,\, \mathrm{A}^t y) = (x \,,\, \mathcal{A}^\star y)$$

with $(\cdot \,,\, \cdot)$ the scalar product of $\mathbb{R}^N$, then $\mathcal{A}^\star = \mathrm{A}^t$.

Example 7.3 (Weak derivative)**.** Let $\mathcal{A} = \mathrm{D}_x$ and $u, v \in \mathrm{L}^2(\Omega)$, with compact support on Ω:

$$(\mathcal{A}u \,,\, v) = (\mathrm{D}_x u \,,\, v) = -(u \,,\, \mathrm{D}_x v) = (u \,,\, \mathcal{A}^\star v)$$

with $(\cdot \,,\, \cdot)$ the scalar product of $\mathrm{L}^2(\Omega)$ (to simplify), then $\mathcal{A}^\star = -\mathrm{D}_x$.

Example 7.4 (Laplace operator)**.** Let $\mathcal{A} = -\boldsymbol{\Delta}$ and $u, v \in \mathrm{H}^1_0(\Omega)$:

$$(\mathcal{A}u \,,\, v) = (-\boldsymbol{\Delta} u \,,\, v) = (\boldsymbol{\nabla} u \,,\, \boldsymbol{\nabla} v) = (u \,,\, -\boldsymbol{\Delta} v) = (u \,,\, \mathcal{A}^\star v)$$

with $(\cdot \,,\, \cdot)$ the scalar product of $\mathrm{L}^2(\Omega)$, then $\mathcal{A}^\star = -\boldsymbol{\Delta}$. The Laplace operator is said *self-adjoint*.

7.2.2 Duality-based *a posteriori* error estimate

We define the dual problem as seeking η satisfying $\mathcal{A}^\star \eta = e_h$, which gives a control on the discretization error, using the definition of the adjoint operator $\mathcal{A}^\star$:

$$\begin{aligned}
\|e_h\|_{\mathrm{L}^2(\Omega)} &= (e_h \,,\, e_h) \\
&= (e_h \,,\, \mathcal{A}^\star \eta) \\
&= (\mathcal{A} e_h \,,\, \eta) \\
&= (\mathcal{A}u \,,\, \eta) - (\mathcal{A}u_h \,,\, \eta) \\
&= (f - \mathcal{A}u_h \,,\, \eta) \\
&= (\mathcal{R}(u_h) \,,\, \eta)
\end{aligned}$$

with $\mathcal{R}(u_h) = f - \mathcal{A}u_h$. Moreover, if the dual problem is stable then there exists a constant $\mathcal{S}$ such that the dual solution η is bounded:

$$\|\eta\|_{\mathrm{L}^2(\Omega)} \leq \mathcal{S} \, \|e_h\|_{\mathrm{L}^2(\Omega)}$$

with the stability factor $\mathcal{S}$ satisfying

$$\mathcal{S} = \max_{\theta \in \mathrm{L}^2(\Omega)} \frac{|\eta|_{\mathrm{H}^2(\Omega)}}{\|\theta\|_{\mathrm{L}^2(\Omega)}}$$

Thus we can obtain a bound of the form:

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq \mathcal{S} \, \|\mathcal{R}(u_h)\|_{\mathrm{L}^2(\Omega)}$$

Combining this estimate with an interpolation inequality in H^α, we can bound the discretization error in terms of the residual and the stability factor. For instance, if we control the second derivatives of the dual solution, *i.e.* $\alpha = 2$,

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_I \, \left\|h^2 \mathcal{R}(u_h)\right\|_{\mathrm{L}^2(\Omega)} \frac{|\eta|_{\mathrm{H}^2(\Omega)}}{\|e_h\|_{\mathrm{L}^2(\Omega)}}$$

Consequently,

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_I \, \mathcal{S} \left\|h^2 \mathcal{R}(u_h)\right\|_{\mathrm{L}^2(\Omega)}$$

7.3 Method

Definition 7.5 (h-adaptivity)**.** Given a tolerance parameter $\epsilon_{\mathrm{tol}} > 0$ defining a quality criterion for the computed solution u_h, adapt the mesh such that it satisfies:

$$\epsilon_{\mathcal{T}} = \sum_{K \in \mathcal{T}_h} \epsilon_K < \epsilon_{\mathrm{tol}}$$

Algorithm 7.6 (Adaptive mesh strategy)**.** *The following procedure applies:*

- *Generate an initial coarse mesh $\mathcal{T}_h^0$.*
- *Perform adaptive iterations for levels $\ell = 0, \cdots, \ell_{\max}$:*
  1. *Solve the primal problem with solution ${u_h}^0 \in V_h^\ell$.*
  2. *Compute the residual of the equation $\mathcal{R}({u_h}^\ell)$.*
  3. *If dual weighted, solve the dual problem with solution $\eta \in \mathrm{W}_h^\ell$.*
  4. *Compute error indicators ϵ_K, $\forall\, K \in \mathcal{T}_h^\ell$.*
  5. *If ($\epsilon_{\mathcal{T}} \geq \epsilon_{\mathrm{tol}}$) :*
     $\rightarrow$ Generate mesh $\mathcal{T}_h^{\ell+1}$ by refining cells with largest values of ϵ_K.
     Else :
     $\rightarrow$ Terminate adaptive iterations, $\ell_{\max} = \ell$.

7.4 Exercises

Exercise 7.7 (Diffusion–Reaction problem on the unit interval)**.** Consider the following one-dimensional problem:

$$-\partial_x\big(a(x)\,\partial_x u(x)\big) + c(x)\,u(x) = f(x) \quad , \forall\, x \in \Omega = [0,1]$$

with $a > 0$, $c \geq 0$, and supplemented with homogeneous Dirichlet boundary conditions

$$u(0) = u(1) = 0$$

1. Write the weak formulation for the given problem and its approximation by piecewise linear Lagrange elements.
2. Write the dual problem for unknown η.
3. Obtain the following estimate:

   $$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq \big\|h^2\mathcal{R}(u_h)\big\|_{\mathrm{L}^2(\Omega)}\,\big\|h^{-2}(\eta - \pi_1\,\eta)\big\|_{\mathrm{L}^2(\Omega)}$$

   with the discretization error $e_h = u - u_h$, the equation residual $\mathcal{R}(u_h) = f + \partial_x\big(a\,\partial_x u_h\big) - c\,u_h$ and the Lagrange $\mathbb{P}^1$ intepolation operator π_1 . First you should test the dual equation against e_h, then write the expression elementwise to be able to define the residual.
4. Conclude that the *a posteriori* error estimate holds

   $$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_I\,\mathcal{S}\,\big\|h^2\mathcal{R}(u_h)\big\|_{\mathrm{L}^2(\Omega)}$$

   with C_I the interpolation constant and $\mathcal{S}$ a stability factor that you will define.

7.4.1 A posteriori error estimation

Consider the following differential equation

$$-u''(x) + u'(x) + u(x) = f(x),x \in (0,1)$$
$$u(0) = u(1) = 0$$

Derive a weak formulation and Galerkin discretization with appropriate functional spaces,

Use the solution of the dual problem

$$-\theta''(x) - \theta'(x) + \theta(x) = ex \in (0,1)$$
$$\theta(0) = \theta(1) = 0$$

to derive the a posteriori error estimate in L_2 norm.

8 Stabilized methods for advection dominated problems

8.1 An advection–diffusion problem in one dimension

According to Problem 18.6 from [6], let us consider the following one-dimensional advection–diffusion problem:

$$-\partial_x\big(\nu(x)\,\partial_x u(x)\big) + \partial_x u(x) = f(x) \quad , \,\forall\, x \in \Omega = [0,1]$$

with viscosity $\nu > 0$, and supplemented with boundary conditions:

$$u(0) = 1 \,, \quad u(1) = 0$$

8.2 Coercivity loss

8.3 Stabilization of the Galerkin method

Galerkin	$(\mathcal{A}u \,,\, v)$			$=$	$(f \,,\, v)$
Galerkin–Least squares	$(\mathcal{A}u \,,\, v + \delta\mathcal{A}v)$			$=$	$(f \,,\, v + \delta\mathcal{A}v)$
	$(\mathcal{A}u \,,\, v)$	$+$	$(\mathcal{A}u \,,\, \delta\mathcal{A}v)$	$=$	$(f \,,\, v) + (f \,,\, \delta\mathcal{A}v)$
Streamline Diffusion	$(\mathcal{A}u \,,\, v + \delta\mathcal{A}v)$	$+$	$(\nu_h \boldsymbol{\nabla} u \,,\, \boldsymbol{\nabla} v)$	$=$	$(f \,,\, v + \delta\mathcal{A}v)$
Entropy viscosity	$(\mathcal{A}u \,,\, v)$	$+$	$(\nu_h \boldsymbol{\nabla} u \,,\, \boldsymbol{\nabla} v)$	$=$	$(f \,,\, v)$

8.4 Exercises

9 Mixed problems

This section is an opportunity to describe step by step the methodology described throughout the course by studying the Stokes problem and to give an overview of the difficulties arising in mixed problems.

QUESTION: In the case of a problem involving a pair of unknown $(\boldsymbol{u}, p)$, is there a criterion to chose the approximation spaces ?

9.1 The Stokes equations

9.1.1 Position of the problem

Let us consider the equations governing the velocity $\bar{\boldsymbol{u}}$ and pressure p of an incompressible creeping flow, subject to the gravity, in a domain Ω, open bounded subset of $\mathbb{R}^d$. As the flow is supposed to be sufficiently slow to neglect the advection compared to the diffusion, the momentum balance equation reduces to

$$-\boldsymbol{\nabla}\cdot \sigma(\boldsymbol{x}) = \varrho(\boldsymbol{x})\mathbf{g}(\boldsymbol{x}) \tag{29a}$$

with the stress tensor

$$\sigma = \tau - p\,\mathbb{I} \tag{29b}$$

consisting of a viscous stress tensor τ and a pressure term with $\mathbb{I}$ the identity matrix of $\mathcal{M}_d(\mathbb{R})$. The incompressibilty constraint

$$\boldsymbol{\nabla}\cdot \bar{\boldsymbol{u}}(\boldsymbol{x}) = 0 \tag{29c}$$

represents the mass conservation for an incompressible continuum. Moreover, the relations are supplemented with boundary conditions on $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$. Dirichlet boundary conditions are enforced on $\partial\Omega_D$

$$\bar{\boldsymbol{u}} = \boldsymbol{u}_D \tag{29d}$$

with u_D while Neumann boundary conditions on $\partial\Omega_N$

$$\sigma \cdot \boldsymbol{n} = \sigma_N \tag{29e}$$

with σ_N a surface force acting on $\partial\Omega_N$.

According to the method developed during the course, we would like first of all to derive a weak formulation by testing Equations (29a) and (29c) against smooth functions, such that we consider

$$-\int_\Omega \boldsymbol{\nabla}\cdot \tau \cdot \boldsymbol{v}\, \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{\nabla} p \cdot \boldsymbol{v}\, \mathrm{d}\boldsymbol{x} = \int_\Omega \varrho\mathbf{g} \cdot \boldsymbol{v}\, \mathrm{d}\boldsymbol{x} \quad , \forall\, \boldsymbol{v} \in \boldsymbol{V}$$

and

$$\int_\Omega \boldsymbol{\nabla}\cdot \bar{\boldsymbol{u}}\, q\, \mathrm{d}\boldsymbol{x} = 0 \quad , \forall\, q \in M$$

Integrating by parts to report the derivatives on the tests functions:

$$-\int_\Omega \boldsymbol{\nabla}\cdot \tau \cdot \boldsymbol{v}\, \mathrm{d}\boldsymbol{x} = -\int_\Omega \boldsymbol{\nabla}\cdot (\tau^t \cdot \boldsymbol{v})\, \mathrm{d}\boldsymbol{x} + \int_\Omega \tau : \boldsymbol{\nabla}\boldsymbol{v}\, \mathrm{d}\boldsymbol{x}$$

which uses the tensor identity, given under repeated indices form:

$$\partial_j\,(\tau_i j)\boldsymbol{v}_i = \partial_j\,(\tau_{ji}\boldsymbol{v}_i) - \tau_{ij}\partial_j\,\boldsymbol{v}_i$$

Owing to relation

$$-\int_\Omega \boldsymbol{\nabla}\cdot\tau\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} = -\int_{\partial\Omega}\tau\cdot\boldsymbol{n}\cdot\boldsymbol{v}\,\mathrm{d}s + \int_\Omega \tau : \boldsymbol{\nabla}\boldsymbol{v}\,\mathrm{d}\boldsymbol{x}$$

and

$$-\int_\Omega \boldsymbol{\nabla}p\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} = -\int_{\partial\Omega} p\boldsymbol{n}\cdot\boldsymbol{v}\,\mathrm{d}s + \int_\Omega p\,\boldsymbol{\nabla}\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x}$$

the weak formulation of Problem (29) reads:

> Find $(\bar{\boldsymbol{u}}, p) \in \boldsymbol{W} \times M$ such that:
>
> $$\begin{aligned} \int_\Omega \tau : \boldsymbol{\nabla}\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} - \int_\Omega p\,\boldsymbol{\nabla}\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} &= \int_\Omega \varrho\mathbf{g}\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega_N}\sigma_N\cdot\boldsymbol{n}\,\mathrm{d}s \quad ,\forall\,\boldsymbol{v}\in\boldsymbol{V} \\ \int_\Omega \boldsymbol{\nabla}\cdot\bar{\boldsymbol{u}}\,q\,\mathrm{d}\boldsymbol{x} &= 0 \quad ,\forall\,q\in M \end{aligned}$$

In the case of a Newtonian fluid the stress tensor reads

$$\sigma(\bar{\boldsymbol{u}}, p) = 2\nu\varepsilon(\bar{\boldsymbol{u}}) - p\mathbb{I}$$

with the strain rate tensor

$$\varepsilon(\bar{\boldsymbol{u}}) = \frac{1}{2}(\boldsymbol{\nabla}\bar{\boldsymbol{u}} + \boldsymbol{\nabla}^t\bar{\boldsymbol{u}})$$

which is symmetric.

9.1.2 Abstract weak formulation

As a first step we can reformulate the previous problem as:

> Find $(\bar{\boldsymbol{u}}, p) \in \boldsymbol{W} \times M$ such that:
>
> $$\begin{aligned} a(\bar{\boldsymbol{u}}, \boldsymbol{v}) + b(\boldsymbol{v}, p) &= L(\boldsymbol{v}) \quad ,\forall\,\boldsymbol{v}\in\boldsymbol{V} \\ b(\bar{\boldsymbol{u}}, q) &= 0 \quad ,\forall\,q\in M \end{aligned}$$

defining $a(\,\cdot\,,\,\cdot\,)$ as the continuous bilinear form:

$$\begin{aligned} a:\ \boldsymbol{W}\times\boldsymbol{V} &\to \mathbb{R} \\ (\bar{\boldsymbol{u}}, \boldsymbol{v}) &\mapsto \int_\Omega \tau : \boldsymbol{\nabla}\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} \end{aligned}$$

$b(\,\cdot\,,\,\cdot\,)$ as the continuous bilinear form:

$$\begin{aligned} b:\ \boldsymbol{V}\times M &\to \mathbb{R} \\ (\boldsymbol{v}, p) &\mapsto -\int_\Omega p\,\boldsymbol{\nabla}\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} \end{aligned}$$

and $L(\,\cdot\,)$ as the continuous linear form:

$$\begin{aligned} L:\ \boldsymbol{V} &\to \mathbb{R} \\ \boldsymbol{v} &\mapsto \int_\Omega \varrho\mathbf{g}\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega_N}\sigma_N\cdot\boldsymbol{n}\,\mathrm{d}s \end{aligned}$$

Choice of the functional spaces: — Regularity: as in Section 1 we chose the test and solution space so that the integrals make sense. Owing to these requirements,

$\boldsymbol{W}$ and $\boldsymbol{V}$ should be subspaces of $\mathrm{H}^1(\Omega)^d$ and M should be a subspace of $\mathrm{L}^2(\Omega)$,
— Boundary conditions: the boundary condition on $\partial\Omega_N$ appears in the weak formulation as a linear form so that the solution will satisfy the constraint $\sigma \cdot \boldsymbol{n} = \sigma_N$, while the boundary condition is included in the definition of the functional space $\boldsymbol{W}$:

$$\boldsymbol{W} = \left\{\boldsymbol{v} \in \mathrm{H}^1(\Omega)^d : \bar{\boldsymbol{u}} = \boldsymbol{u}_D \text{ ,on } \partial\Omega_D\right\}$$

By homogenizing the Dirichlet boundary condition, we can lift the solution $\bar{\boldsymbol{u}}$ so that the problem is rewritten to seek a velocity $\boldsymbol{u}$ in $\boldsymbol{V}$.

The generalized Stokes problem reads then:

$$\left|\begin{array}{l} \text{Find } (\boldsymbol{u}, p) \in \boldsymbol{V} \times M \text{ such that:} \\ \qquad \begin{array}{rcll} a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) & = & L(\boldsymbol{v}) & ,\forall\, \boldsymbol{v} \in \boldsymbol{V} \\ b(\boldsymbol{u}, q) & = & \langle\, \Psi \,,\, p \,\rangle_{M',M} & ,\forall\, q \in M \end{array} \end{array}\right. \tag{30}$$

with $(\boldsymbol{V}, M)$ a pair of Hilbert spaces to be determined, $a(\,\cdot\,,\,\cdot\,)$ bilinear form continuous on $\boldsymbol{V} \times \boldsymbol{V}$, $L(\,\cdot\,)$ linear form continuous on $\boldsymbol{V}$ and Ψ a given continuity constraint in M'.

9.1.3 Well-posedness in the continuous setting

Let us change the space in which test functions are chosen to the space of divergence-free functions of $\boldsymbol{V}$ to satisfy the continuity constraint:

$$\boldsymbol{V}_0 = \{\boldsymbol{v} \in \boldsymbol{V} : b(\boldsymbol{v}, q) = 0 \text{ ,} \forall\, q \in M\}$$

The bilinear form b is continuous on $\boldsymbol{V}_0 \times M$, *i.e.* $b(\boldsymbol{v}, q) \leq \|\boldsymbol{v}\|_{\boldsymbol{V}_0} \|q\|_M$, thus $\mathrm{Im}(b)$ is closed and $\boldsymbol{V} = \boldsymbol{V}_0 \oplus \boldsymbol{V}_0^{\perp}$. The first relation of the Stokes problem becomes then:

$$a(\boldsymbol{u}, \boldsymbol{v}) + \underbrace{b(\boldsymbol{v}, p)}_{0} = L(\boldsymbol{v}) \quad ,\forall\, \boldsymbol{v} \in \boldsymbol{V}_0$$

Therefore, the new abstract problem with solenoidal test functions reads:

$$\left|\begin{array}{l} \text{Find } (\boldsymbol{u}, p) \in \boldsymbol{V} \times M \text{ such that:} \\ \qquad \begin{array}{rcll} a(\boldsymbol{u}, \boldsymbol{v}) & = & L(\boldsymbol{v}) & ,\forall\, \boldsymbol{v} \in \boldsymbol{V}_0 \\ b(\boldsymbol{u}, q) & = & \langle\, \Psi \,,\, p \,\rangle_{M',M} & ,\forall\, q \in M \end{array} \end{array}\right.$$

Theorem 9.1 (Well-posedness of constrained problem)**.** *Let us define the space*

$$\boldsymbol{V}_\Psi = \left\{\boldsymbol{v} \in \boldsymbol{V} : b(\boldsymbol{v}, q) = \langle\, \Psi \,,\, p \,\rangle_{M',M} \text{ ,} \forall\, q \in M\right\}$$

supposed non-empty and consider $a(\,\cdot\,,\,\cdot\,)$ a bilinear form coercive on V. The problem

$$\left|\begin{array}{l} \textit{Find } (\boldsymbol{u}, p) \in \boldsymbol{V}_\Psi \times M \textit{ such that:} \\ \qquad \begin{array}{rcll} a(\boldsymbol{u}, \boldsymbol{v}) & = & L(\boldsymbol{v}) & ,\forall\, \boldsymbol{v} \in \boldsymbol{V}_0 \end{array} \end{array}\right.$$

admits a unique solution.

Proof. The given problem satisfies the assumptions of the Lax–Milgram Theorem. □

We denote by $\mathcal{L}(V \times \mathrm{W};\mathbb{R})$, the space of bilinear form continuous on $V \times \mathrm{W}$ which is a Banach space for the operator norm

$$\|a\|_{V,\mathrm{W}} = \sup_{\substack{v\in V\\ w\in \mathrm{W}}} \frac{a(v,w)}{\|v\|_V \, \|w\|_\mathrm{W}}$$

Proposition 9.2 (Babuska–Necas–Brezzi condition)**.** *The bilinear form $a \in \mathcal{L}(V \times \mathrm{W};\mathbb{R})$ satisfies the (BNB) condition if there exists $\beta > 0$ such that*

$$\inf_{w\in \mathrm{W}} \sup_{v\in V} \frac{a(v,w)}{\|v\|_V \, \|w\|_\mathrm{W}} \geq \beta$$

Theorem 9.3 (Existence)**.** *If $\boldsymbol{V}_\Psi$ is non-empty, $a(\,\cdot\,,\,\cdot\,)$ is a bilinear form coercive on $\boldsymbol{V}$ with coercivity constant α, and the bilinear form $b(\,\cdot\,,\,\cdot\,)$ satisfies Proposition* (9.2)*, i.e.*

$$\exists \beta > 0 : \inf_{q\in \tilde{\mathrm{M}}} \sup_{\boldsymbol{v}\in \boldsymbol{V}} \frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\boldsymbol{V}} \, \|q\|_M} \geq \beta$$

then Problem (9.1.3) *admits solution pairs $(\boldsymbol{u},p) \in V \times M$ such that $\boldsymbol{u}$ is unique, satisfying*

$$\|\boldsymbol{u}\|_{\boldsymbol{V}} \leq \frac{1}{\alpha}\|L\|_{\boldsymbol{V}'} + \frac{1}{\alpha}\left(1 + \|a\|_{\boldsymbol{V},\boldsymbol{V}}\right)\|\Psi\|_{M'}$$

and any $p \in M$ can be written as $p = \tilde{p} + M_0$, $\tilde{p} \in M_0^\perp$

$$\|\tilde{p}\|_M \leq \left(1 + \frac{\|a\|_{\boldsymbol{V},\boldsymbol{V}}}{\alpha}\right)\left(\frac{1}{\beta}\|L\|_{\boldsymbol{V}'} + \frac{1}{\beta^2}\|a\|_{\boldsymbol{V},\boldsymbol{V}}\|\Psi\|_{M'}\right)$$

Indeed, p playing the role of a potential, it is defined up to a constant. Then we can interpret the space M_0 as the space of functions on which gradients are vanishing which is the space of constants on Ω, so that we seek $\tilde{p} \in \tilde{\mathrm{M}}$, with $\tilde{M} = M_0^\perp$ defined as the equivalent class: $\forall\, p,q \in M$, $p \equiv q \Leftrightarrow p = q + C : C \in \mathbb{R}$.

Consequently, we add the constraint that the pressure has a zero average on Ω and as a by-product we consider $(u,\tilde{p}) \in \boldsymbol{V}_h \times \tilde{\mathrm{M}}$ the unique solution pair to Problem (9.1.3).

Historically, the (BNB) condition relates to a well-known result characterizing the surjectivity of the divergence operator which can be generalized as:

Theorem 9.4 (De Rham – [4] page 492)**.** *The continuous bilinear forms on $\mathrm{W}^{1,p}(\Omega)$ which are zero on $\ker(\boldsymbol{\nabla}\cdot\,)$ are gradients of functions in $\mathrm{L}^{p'}_{\int=0}(\Omega)$.*

9.2 The discrete Inf-Sup condition

9.2.1 Results

Let us consider an approximation of Problem 30 by a Galerkin method:

> Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times \tilde{\mathrm{M}}_h$ such that:
>
> $$\begin{aligned} a(\boldsymbol{u}_h,\boldsymbol{v}_h) + b(\boldsymbol{v}_h,p_h) &= L(\boldsymbol{v}_h) && ,\forall\, \boldsymbol{v}_h \in \boldsymbol{V}_h \\ b(\boldsymbol{u}_h,q_h) &= \langle\, \Psi\,,\, p_h\,\rangle_{M',M} && ,\forall\, q_h \in M_h \end{aligned}$$

with $(\boldsymbol{V}_h, \tilde{\mathrm{M}}_h)$ a pair of approximation spaces to be chosen and the discrete divergence operator $\mathbf{B}_h$,

$$b(\boldsymbol{u}_h,q_h) = (\,\mathbf{B}_h\boldsymbol{u}\,,\,q_h\,)$$

Theorem 9.5 (Well-posedness)**.** *If* $\Psi_h \in \mathrm{Im}(\mathbf{B}_h)$ *then Problem* (9.2.1) *admits solutions* $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times \tilde{\mathrm{M}}_h$ *such that* $\boldsymbol{u}_h$ *is unique and the pressure can be written as* $p_h = \tilde{p}_h + \ker(\mathbf{B}_h^t)$ *with* $\tilde{p}_h \in \ker(\mathbf{B}_h^t)^\perp$ *unique.*

Theorem 9.6 (Convergence – [7] page 21)**.** *Let* $(\boldsymbol{u}, p) \in \boldsymbol{V} \times \tilde{\mathrm{M}}$ *be the solution of Problem* (29) *and* $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times \tilde{\mathrm{M}}_h$ *the solution of discrete Problem* (9.2.1) *and we denote by* α_h *the coercivity constant of* $a(\,\cdot\,,\,\cdot\,)$ *on* $\boldsymbol{V}_{0,h}$ *and by* β_h *the constant of the discrete Inf-Sup condition. If* $\Psi_h \in \mathrm{Im}(\mathbf{B}_h)$ *then the following two consistency estimates hold:*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}_h} \leq C_1 \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_{\boldsymbol{V}} + C_2 \inf_{q_h \in M_h} \|p - q_h\|_M$$

$$\|p_h - p_h\|_{\tilde{\mathrm{M}}_h} \leq C_3 \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_{\boldsymbol{V}} + C_4 \inf_{\boldsymbol{v}_h \in M_h} \|p - q_h\|_M$$

with constants

$$\begin{aligned} C_1 &= \left(1 + \frac{\|a\|_{\boldsymbol{V},\boldsymbol{V}}}{\alpha_h}\right)\left(1 + \frac{\|b\|_{\boldsymbol{V},M}}{\beta_h}\right) \\ C_2 &= \frac{\|b\|_{\boldsymbol{V},M}}{\alpha_h} \\ C_3 &= \frac{\|a\|_{\boldsymbol{V},M}}{\beta_h} C_1 \\ C_4 &= 1 + \frac{\|b\|_{\boldsymbol{V},M}}{\beta_h} + \frac{\|a\|_{\boldsymbol{V},M}}{\beta_h} C_2 \end{aligned}$$

The previous result shows then that satisfying the discrete Inf-Sup condition is crucial to ensure optimal convergence of the numerical scheme, *i.e.* the discretization error decreases with the mesh size $h_{\mathcal{T}}$. Indeed, if the parameter β_h is not bounded from below then it is clear that values tending to zero will degrade the consistency estimates.

9.2.2 Commonly used pairs of approximation spaces

Velocity space $\boldsymbol{V}_h$	Pressure space M_h	Inf-Sup stable	Comment
$\mathbb{P}^1$	$\mathbb{P}^1$	No	
$\mathbb{P}^1$	$\mathbb{P}^0$	No	“Locking effect”
$\mathbb{P}^{k+1}$	$\mathbb{P}^k$	Yes	$k \geq 1$, “Taylor–Hood”

9.3 Exercises

A Definitions

A.1 Mapping

Definition A.1 (Mapping). Let E and F be two sets, a mapping

$$\begin{array}{rccc} f: & E & \to & F \\ & x & \mapsto & f(x) \end{array}$$

is a relation which, to any element $x \in E$, associates an element $y = f(x) \in F$.

Definition A.2 (Linear mapping). Let E and F be two $\mathbb{K}$-vector spaces, the mapping $f : E \to F$ is linear if:

1. $\forall\, x, y \in E,\ f(x+y) = f(x) + f(y)$
2. $\forall\, \lambda \in \mathbb{K}, y \in E,\ f(\lambda x) = \lambda f(x)$

A.2 Spaces

Definition A.3 (Vector space (on the left)). Let $(\mathbb{K}, +, \times)$ be defined such that $(\mathbb{K}, +)$ is an Abelian additive group and $(\mathbb{K}, \times)$ is an Abelian multiplicative group, $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

$(\mathbb{K}, +)$	$(\mathbb{K}, \times)$
"+" commutative and associative $0_{\mathbb{K}}$ neutral for '+" "+" admits an opposite	"×" commutative and associative $\mathbb{1}_{\mathbb{K}}$ neutral for '×" "×" admits an inverse
"×" is distributive with respect to "+"	

$(E, +, \cdot)$ is a vector space on $(\mathbb{K}, \times)$ if:

1. $(E, +)$ is an additive Abelian group (same properties as $(\mathbb{K}, +)$).
2. The operation $\cdot\ : \mathbb{K} \times E \to E$ satisfies:

distributive w.r.t "$+_E$" on the left	$\lambda \cdot (u + v) = \lambda \cdot u + \lambda \cdot v$
distributive w.r.t "$+_{\mathbb{K}}$" on the right	$(\lambda + \mu) \cdot u = \lambda \cdot u + \mu \cdot u$
associative w.r.t "×"	$(\lambda \times \mu) \cdot u = \lambda \cdot (\mu \cdot u)$
$\mathbb{1}_{\mathbb{K}}$ neutral element on the left	$\mathbb{1}_{\mathbb{K}} \cdot u = u$

In short the vector space structure allows writing any $\boldsymbol{u} \in E$ as linear combinations of elements $\{\boldsymbol{v}_i\}$ of E called *vectors* with elements $\{\lambda_i\}$ of $\mathbb{K}$ called *scalars* as coefficients,

$$\boldsymbol{u} = \sum_i \lambda_i \boldsymbol{v}_i$$

and both the multiplications for vectors and scalars are distributive with respect to the additions. In this document we only consider real vector spaces, $\mathbb{K} = \mathbb{R}$.

Definition A.4 (Norm). Let E be a $\mathbb{K}$-vector space, the application

$$\|\cdot\|\ :\ E \to \mathbb{R}^+$$

is a norm if the following properties are satisfied:

1. Separation: $\forall\, x \in E,\ (\ \|x\|_E = 0\ ) \Rightarrow (\ x = 0_E\ )$
2. Homogeneity: $\forall\, \lambda \in \mathbb{K},\ \forall\, x \in E,\ \|\lambda x\|_E = |\lambda|\ \ \|x\|_E$

3. Subadditivity: $\forall\, x, y \in E$, $\|x+y\|_E \leq \|x\|_E + \|y\|_E$

Note A.5. The third property is usually called *triangle inequality.*

Definition A.6 (Equivalent norms)**.** Let E be a $\mathbb{K}$-vector space, norm $\|\cdot\|_{EE}$ is said equivalent to $\|\cdot\|_E$ if there exist $C_1, C_2 > 0$ such that:

$$C_1 \|u\|_E \leq \|u\|_{EE} \leq C_2 \ \|u\|_E \quad , \forall\, u \in E$$

Definition A.7 (Seminorm)**.** Let E be a $\mathbb{K}$-vector space, the application

$$\|\cdot\| \ : \ E \to \mathbb{R}^+$$

is a seminorm if it satisfies properties (A.4).2 and (A.4).3.

Definition A.8 (Scalar product)**.** Let E be a $\mathbb{R}$-vector space, the bilinear mapping

$$(\ \cdot\ ,\ \cdot\ ) \ : \ E \times E \to \mathbb{R}$$

is a scalar (or inner) product of E if it satisfies the following three properties:

1. Symmetry: $\forall\, x, y \in E$, $(\ x\ ,\ y\ ) = (\ y\ ,\ x\ )$
2. Positivity: $\forall\, x \in E$, $(\ x\ ,\ x\ ) \geq 0$
3. Definiteness: $(\ (\ x\ ,\ x\ ) = 0\ ) \Rightarrow (\ x = 0\ )$

B Duality in finite dimension

Definition B.1 (Dual space)**.** Let E be a finite dimensional real vector space, its dual $E^\star$ is the space of linear forms on E, denoted by $\mathcal{L}(E;\mathbb{R})$.

Definition B.2 (Dual basis)**.** Let E be a finite dimensional real vector space, $\dim(E) = N$ and $\mathcal{B} = (e_1, \cdots, e_N)$ a basis of E. Let us denote, for any $i, j \in [\![1, N]\!]$, by:

$$\begin{array}{cccc} e_i^\star : & E & \to & \mathbb{R} \\ & e_j & \mapsto & \delta_{ij} \end{array}$$

the i-th coordinate. The dual family of $\mathcal{B}$, $\mathcal{B}^\star = (e_1^\star, \cdots, e_N^\star)$ is a basis of $E^\star$.

Thus we can write any element $u \in E$ as:

$$u = \sum_{i=1}^{N} e_i^\star(u) e_i$$

Proving that $\mathcal{B}^\star$ is a basis of $E^\star$ requires that $\{e_i\}$ generates $E^\star$ and that its elements are linearly independent. The corollary of the first condition is that $\dim(\mathcal{B}^\star) = N$.

C Functional spaces

C.1 Banach and Hilbert spaces

Definition C.1 (Cauchy criterion)**.** Let $(E, \|\cdot\|_E)$ be a normed vector space and $(v^n)_{n\in\mathbb{N}}$ be a sequence of element of E which satisfies:

$$\forall\, \epsilon > 0,\ \exists N \text{ such that } \forall\, p, q \geq N,\ \|v^p - v^q\|_E \leq \epsilon$$

then $(v^n)_{n\in\mathbb{N}}$ is a Cauchy sequence in E.

Definition C.2 (Banach space)**.** A Banach space $(E, \|\cdot\|_E)$ is a normed vector space with is complete with respect to the norm $\|\cdot\|_E$, *i.e.* Cauchy sequences converge in E.

Definition C.3 (Hilbert space)**.** Let E be a $\mathbb{K}$-vector space and $(\,\cdot\,,\,\cdot\,)$ be a sesquilinear form on the left (or bilinear form if $\mathbb{K} = \mathbb{R}$),

$$\begin{aligned}
(\, x_1 + x_2 \,,\, y \,) &= (\, x_1 \,,\, y \,) + (\, x_2 \,,\, y \,) \\
(\, x \,,\, y_1 + y_2 \,) &= (\, x \,,\, y_1 \,) + (\, x \,,\, y_2 \,) \\
(\, \lambda x \,,\, y \,) &= \lambda(\, x \,,\, y \,) \\
(\, x \,,\, \lambda y \,) &= \bar{\lambda}(\, x \,,\, y \,)
\end{aligned}$$

which is also positive definite on E,

$$\forall\, x \neq 0_E,\ (\,\cdot\,,\,\cdot\,) > 0$$

then $(E, (\,\cdot\,,\,\cdot\,))$ is a pre-Hilbertian space. Moreover, if E is complete with respect to the norm defined by $(\,\cdot\,,\,\cdot\,)$, it is a Hilbert space.

Definition C.4 (Hilbertian norm)**.**

$$\frac{1}{2}\left(\|x\|_E^2 + \|y\|_E^2\right) = \left\|\frac{x+y}{2}\right\|_E^2 + \left\|\frac{x-y}{2}\right\|_E^2$$

Remark C.5. This is basically the parallelogram identity. This inequality is useful to check that a norm is generated from a scalar product.

Theorem C.6 (Projection on a convex subset)**.** *Let* H *be a Hilbert space and* $K \subset \mathrm{H}$ *be a convex closed non-empty subset,* $\forall\, x \in \mathrm{H}$ *there exists a unique* $x_0 \in K$ *such that*

$$\|x - x_0\|_{\mathrm{H}} = \inf_{y \in K} \|x - y\|_{\mathrm{H}}$$

with x_0 *the projection of* x *onto* K *and we denote it by* $x_0 = \mathrm{P}_K\, x$

C.2 Spaces of continuous functions

$$\mathrm{C}^k(\Omega) = \left\{u \in \mathrm{C}^0(\Omega) : u' \in \mathrm{C}^{k-1}(\Omega)\right\}$$

$$\mathrm{C}_c^\infty(\Omega) = \{u \in \mathrm{C}^\infty(\Omega) \text{ with compact support in } \Omega\}$$

C.3 Lebesgue spaces

$$\mathrm{L}^p(\Omega) = \left\{u : \int_\Omega |u(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x} < \infty\right\}$$

Remark C.7. Lebesgue spaces L^p, $1 \geq p \geq \infty$ are Banach spaces for the norm

$$\|\cdot\|_{\mathrm{L}^p(\Omega)} = \left(\int_\Omega |u(\boldsymbol{x})|^p\right)^{1/p}$$

and L^2 is a Hilbert space endowed with the scalar product

$$(\, u \,,\, v \,)_{\mathrm{L}^2(\Omega)} = \int_\Omega u\, v \,\mathrm{d}\boldsymbol{x} \tag{31}$$

C.4 Hilbert–Sobolev spaces

$$\mathrm{H}^s(\Omega) = \big\{u \in \mathrm{L}^2(\Omega) \,:\, \mathbf{D}^\alpha u \in \mathrm{L}^2(\Omega)\,, 1 \leq \alpha \leq s\big\}$$

C.5 Sobolev spaces

$$\mathrm{W}^{m,p}(\Omega) = \{u \in \mathrm{L}^p(\Omega) \,:\, \mathbf{D}^\alpha u \in \mathrm{L}^p(\Omega)\,, 1 \leq \alpha \leq m\}$$

Remark C.8. H^s spaces are $\mathrm{W}^{s,2}$ spaces.

D Inequalities

Lemma D.1 (Cauchy–Schwarz)**.** *Let E be a $\mathbb{K}$-vector space, any positive sesquilinear form $(\,\cdot\,,\,\cdot\,)$ on E satisfies the inequality:*

$$(\,u\,,\,v\,) \leq \|u\|_E\,\|v\|_E$$

Remark D.2. In particular any scalar product satisfies the Cauchy–Schwarz inequality. For example:

$$(\,u\,,\,v\,)_{\mathrm{L}^2(\Omega)} = \int_\Omega u\,v\,\mathrm{d}\boldsymbol{x} \leq \|u\|_{\mathrm{L}^2(\Omega)}\,\|v\|_{\mathrm{L}^2(\Omega)}$$

Lemma D.3 (Young)**.** *Let $a, b > 0$ be two real numbers:*

$$ab \leq \frac{1}{p}\Big(\frac{a}{\gamma}\Big)^p + \frac{1}{q}\Big(b\gamma\Big)^q$$

with $\dfrac{1}{q} + \dfrac{1}{p} = 1$ *and* $\gamma > 0$.

Remark D.4. In particular, the following inequality is commonly used for energy estimates:

$$ab \leq \frac{1}{2}\Big(\frac{a}{\gamma}\Big)^2 + \frac{1}{2}(b\gamma)^2$$

Lemma D.5 (Generalized Hölder)**.** *Let $u \in \mathrm{L}^p(\Omega)$, $v \in \mathrm{L}^q(\Omega)$, with $1 \leq p < \infty$, then:*

$$\|u\,v\|_{\mathrm{L}^r(\Omega)} \leq \|u\|_{\mathrm{L}^q(\Omega)}\,\|v\|_{\mathrm{L}^q(\Omega)}$$

with

$$\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$$

Lemma D.6 (Minkowski)**.**

$$\|u + v\|_{\mathrm{L}^p(\Omega)} \leq \|u\|_{\mathrm{L}^p(\Omega)} + \|v\|_{\mathrm{L}^p(\Omega)}$$

Remark D.7. The previous result is basically the triangle inequality for the L^p–norm.

Lemma D.8 (Poincaré)**.** *Let Ω be an open bounded subset, for any $1 \leq p < \infty$ there exists a constant real number $c_P > 0$ such that $\forall\, u \in \mathrm{W}_0^{1,p}(\Omega)$:*

$$c_P\,\|u\|_{\mathrm{L}^p(\Omega)} \leq \|\boldsymbol{\nabla} u\|_{\mathrm{L}^p(\Omega)}$$

Remark D.9. As a Corollary usefull for the Poisson problem that we address, we get that $\|\boldsymbol{\nabla} u\|_{\mathrm{L}^2(\Omega)}$ defines an equivalent norm to $\|u\|_1$ on $\mathrm{H}_0^1(\Omega)$.

Lemma D.10 (Clarkson)**.** *Let $1 < p < \infty$, and u, v be two functions of $\mathrm{L}^p(\Omega)$, then:*

1. *for $p \geq 2$*

$$\left\|\frac{u+v}{2}\right\|_{\mathrm{L}^p(\Omega)}^2 + \left\|\frac{u-v}{2}\right\|_{\mathrm{L}^p(\Omega)}^2 \leq \frac{1}{2}\big(\|u\|_{\mathrm{L}^p(\Omega)}^2 + \|v\|_{\mathrm{L}^p(\Omega)}^2\big)$$

2. *for $p < 2$*

$$\left\|\frac{u+v}{2}\right\|_{\mathrm{L}^{p'}(\Omega)}^2 + \left\|\frac{u-v}{2}\right\|_{\mathrm{L}^{p'}(\Omega)}^2 \leq \left(\frac{1}{2}\|u\|_{\mathrm{L}^p(\Omega)}^2 + \frac{1}{2}\|v\|_{\mathrm{L}^p(\Omega)}^2\right)^{1/(p-1)}$$

Remark D.11. These inequalities are basically parallelogram inequalities generalized to L^p spaces.

References

[1] F. Boyer and P. Fabrie. *Mathematical Tools for the Study of the Incompressible Navier–Stokes Equations and Related Models*, volume 183 of *Springer Series: Applied Mathematical Sciences*. Springer, 2013.

[2] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Springer Series: Texts in Applied Mathematics*. Springer, 2008.

[3] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

[4] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Springer Series: Applied Mathematical Sciences*. Springer, 2004.

[5] R. Herbin. Analyse Numérique des EDPs. Notes de cours de Master 2, 2006.

[6] P. Hansbo K. Eriksson, D. Estep and C. Johnson. *Computational Differential Equations*. Press Syndicate of the University of Cambridge, 1996.

[7] J.-C. Latché. Méthodes d'Éléments Finis pour quelques problèmes elliptiques linéaires issus de la mécanique des fluides. Notes de cours de Master 2, 2002.