Numerical Solution of Partial Differential Equations Using Finite Element Methods

Aurélien Larcher

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, then updated for the course "Numerical Solution of Partial Differential Equations Using Element Methods" (TMA4220) at NTNU, during Fall 2018. It is in not 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 knowledge of Numerical Analysis.

Content

The course goes through the basic theory of the Finite Element Method during the first six lectures while the last three lectures are 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 Finite Element approximation spaces.
5. Polynomial approximation and error analysis.
6. Time dependent problems.
7. Mesh generation and adaptive control.
8. Stabilized finite element methods.
9. Mixed problems.

The intent is to introduce the practicals aspects of the methods without hiding the mathematical issues but without necessarily exposing the details of the proof. 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 four 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*: Is the solution to the approximate problem stable with respect to data and "well-behaved"?

4. *Maximum principle, Physical properties*: Does the discrete solution reproduce features of the physical solution, like satisfying physical bounds or energy/entropy inequalities?

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 is a direction towards deriving discrete counterparts (usually in terms of inequalities) and ensuring that numerical algorithms possess good properties.

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 addressing some technical details is left to more serious and comprehensive works referenced in the bibliography.

Literature

At KTH 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 10.1 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 Analysis. On the other hand, *Mathematical Tools for the Study of the Incompressible Navier–Stokes Equations and Related Models* [1], while retaining all the difficulties of analysis, offers a really didactic approach of PDEs for fluid problems in a clear and rigorous manner.

Chapter 1

Weak formulation of elliptic Partial Differential Equations

1.1 Historical perspective

The physics of phenomena encountered in engineering applications is often modelled under the form of a boundary and initial value problems. They consist of relations describing the evolution of physical quantities involving partial derivatives of physical quantities with respect to space and time, such relations are called Partial Differential Equations (PDEs). Problems involving only variations in space on a domain $\Omega \subset \mathbb{R}^d$ are called Boundary Value Problems (1.1) as they involve the description of the considered physical quantities at the frontier of the physical domain $\partial\Omega$.

$$\left|\begin{array}{l} \text{Find } u : \mathbb{R}^d \to \mathbb{R}^n \text{ such that:} \\ \mathrm{A}u(\boldsymbol{x}) = f(\boldsymbol{x}) \quad , \forall\, \boldsymbol{x} \in \Omega \\ + \text{ Boundary conditions on } \partial\Omega \end{array}\right. \tag{1.1}$$

with A a *differential operator*, *i.e.* involving partial derivatives of u, like the first derivatives with respect to each axis

$$\frac{\partial}{\partial x_i}$$

for $i = 1, \cdots, d$, or more generally

$$\frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_i} x_i \cdots \partial^{\alpha_d} x_d}$$

given the multi-index $\alpha = (\alpha_1, \cdots, \alpha_i, \cdots, \alpha_d)$, and $|\alpha|$ the module of the multi-index is the order of the derivative.

Equations describing the evolution in time of physical quantities required the definition of an initial condition in time, and are therefore called Initial Value Problems. They consist of the coupling of an Ordinary Differential Equation

(ODE) in time, a Cauchy Problem (1.2), with a boundary value problem in space.

$$\left|\begin{array}{l} \text{Find } y : \mathbb{R} \to \mathbb{R}^n \text{ such that:} \\ y'(t) = F(t, y(t)),\ \forall\, t \in [0, T) \\ + \text{ Intial condition at } t = 0\text{: } y(t = 0) = y_0 \end{array}\right. \tag{1.2}$$

with $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.

The term *Finite Element Method* denotes a family of approaches developed to compute an approximate solution to boundary and initial value problems.

Example 1.1.1 (Partial Differential Equations)**.** A few usual mathematical models are listed below.

- Transfer of heat/mass by conduction/diffusion, "Fourier's Law":

  $$-\kappa \Delta T = f$$

  with κ constant conductivity/diffusivity, and for example T temperature.

- Unsteady heat equation:

  $$\frac{\partial T}{\partial t} - \kappa \Delta T = f$$

  with κ thermal diffusivity, and T temperature.

- Transport of a passive scalar field:

  $$\frac{\partial c}{\partial t} + \beta \cdot \nabla c = f$$

  with β advective vector field, and for example c a concentration.

- Burgers equation in one dimension (Forsyth, 1906, and Burgers, 1948):

  $$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0$$

  with ν a viscosity, possibly zero in the inviscid case.

- Wave equation in one dimension (D'Alembert, 1747):

  $$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0$$

  with c a celerity.

- Euler equations, inviscid and incompressible case (Euler, 1757):

  $$\begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \boldsymbol{\nabla})\boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{f} \\ \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0 \end{cases}$$

  with ν a viscosity, possibly zero in the inviscid case.

- Navier–Stokes equations, homogeneous, incompressible and isothermal case (Navier, 1822, then 19th century):

$$\begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{u} - \boldsymbol{\nabla}\cdot\Big(\tau(\boldsymbol{u},p)\Big) = \boldsymbol{f} \\ \boldsymbol{\nabla}\cdot\boldsymbol{u} = 0 \end{cases}$$

  with $\tau(\boldsymbol{u},p) = \nu\boldsymbol{\nabla}\boldsymbol{u} - p\mathbb{I}$ for a Newtonian fluid.

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. The main motivation is the existence of discontinuous solutions produced by classes of PDEs or irregular data, for which the classical definition of a derivative is not suitable. The concept of *weak derivative* introduced by L. Schwartz in his Theory of Distributions, was used to seek solutions to Partial Differential Equations like general elliptic and parabolic problems (J.-L. Lions) or Navier–Stokes (J. Leray). The idea is to replace the pointwise approach of the classical (strong) derivative

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

which requires regularity (derivability, continuity) at any point in space and time by the derivation in the distributional sense, *i.e.* by considering the action of linear forms on smooth functions, such as

$$T : \varphi \mapsto \int_\Omega f\varphi$$

with φ a sufficiently regular function such that the integral is defined.

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:

$$-\boldsymbol{\Delta}u(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \forall\, \boldsymbol{x}\in\Omega \tag{1.3a}$$

$$u(\boldsymbol{x}) = 0, \qquad \forall\, \boldsymbol{x}\in\partial\Omega \tag{1.3b}$$

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

$$\boldsymbol{\Delta} = \sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2} \tag{1.4}$$

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

Definition 1.2.1 (Classical solution)**.** A classical solution (or strong solution) of Problem (1.3) is a function $u \in \mathrm{C}^2(\Omega)$ satisfying relations (1.3a) and (1.3b).

Problem (1.3) 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 (1.3) and let us test Equation (1.3a) 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:

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

For simplicity, we recall the one-dimensional case:

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

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 function space such that the previous relation makes sense.

A weak formulation of Problem (1.3) consists in solving:

$$\left|\begin{array}{l} \text{Find } u \in H, \text{ given } f \in V', \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u\cdot\boldsymbol{\nabla} v\,\mathrm{d}\boldsymbol{x} = \int_\Omega fv\,\mathrm{d}\boldsymbol{x} \quad ,\ \forall\, v \in V \end{array}\right. \tag{1.5}$$

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

1.2.2 Formal passage from weak solution to classical solution

Provided that the weak solution to Problem (1.5) 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	$\boldsymbol{\nabla} u \cdot \mathbf{n} = 0$	"natural" boundary condition

Essential boundary conditions are embedded in the function 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 (1.5) 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 the 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 function 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 spaces

We will now establish that any weak solution "lives" in $\mathrm{H}^1_0(\Omega)$ and that the natural space for test functions is the same space.

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, we 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 function 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 1.3.1 (Trace Theorem)**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$ with piecewise C^1 boundary, then there exists a linear application $\gamma : \mathrm{H}^1(\Omega) \to \mathrm{L}^2(\partial\Omega)$ continuous on $\mathrm{H}^1(\Omega)$ such that $\gamma(u) = 0 \Rightarrow u \in \mathrm{Ker}(\gamma)$.*

The regularity of the solution itself depends on the nature of the differential operators involved in the problem (*e.g.* up to which order should be derivatives controlled?), but also on the data of the problem: regularity of the domain and right-hand side.

The weak formulation of Problem (1.3) reads then:

$$\left| \begin{array}{l} \text{Find } u \in \mathrm{H}^1_0(\Omega), \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x} \quad , \forall\, v \in \mathrm{H}^1_0(\Omega) \end{array} \right. \tag{1.7}$$

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:

$$\left| \begin{array}{l} \text{Find } u \in V, \text{ such that:} \\ a(u, v) = L(v) \quad , \forall\, v \in V \end{array} \right. \tag{1.8}$$

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

Proposition 1.4.1 (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*

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

For example, in the previous section for Problem (1.7), 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,

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

The continuity of these two forms comes directly from that they are respectively the inner-product in $\mathrm{H}^1_0(\Omega)$, and the L^2 inner-product with $f \in \mathrm{L}^2(\Omega)$: the Cauchy–Schwarz inequality gives directly a continuous control of the image of the forms by the norms of its arguments.

Definition 1.4.2 (Topological dual space)**.** The topological dual space V' of a normed vector space V is the vector space of continuous linear forms on V equipped with the norm:

$$\|f\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|f(x)|}{\|x\|_V}$$

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

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

This property is also know as V–ellipticity.

1.5 Well-posedness

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

Theorem 1.5.1 (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 L is continuous for $\|\cdot\|_a$. If the bilinear form $a(\,\cdot\,,\,\cdot\,)$ is not symmetric then Theorem 1.5.1 (Riesz–Fréchet) does not apply.

Theorem 1.5.2 (Lax–Milgram)**.** *Let H be a Hilbert space. Provided that $a(\,\cdot\,,\,\cdot\,)$ is a coercive continuous bilinear form on $H \times H$ and $L(\,\cdot\,)$ is a continuous linear form on H, Problem* (1.8) *admits a unique solution $u \in H$.*

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

Exercises for this section cover some preliminary notions introduced for the weak formulation of PDEs.

Exercise 1.6.1 (Based on Exercise 4 from [8])**.**
Answer the following questions.

(a) Discuss whether the set $S = \left\{ v \in \mathrm{C}_c^\infty((0,1)) : v(\frac{1}{2}) = 1 \right\}$ is a vector space.
(b) For $V = \mathrm{H}_0^1((0,1))$, show that $L(v) = \int_0^1 xv \,\mathrm{d}x$ defines a linear functional. Recall the definition of the topological dual V' and show that L is continuous for $x \in V$.
(c) For $V \equiv \mathbb{R}$ discuss whether $(\, u \,,\, v \,)_V = |u||v|$ is an inner-product in V.
(d) Does $|u|_{\mathrm{H}^1(\Omega)} = \|\nabla u\|_{\mathrm{L}^2(\Omega)}$, $\Omega \in \mathbb{R}^2$ define a norm in $\mathrm{H}^1(\Omega)$? Explain why.
(e) Assess whether the function $f(x) = x^{3/4}$ an element of the following spaces: $\mathrm{L}^2((0,1))$, $\mathrm{H}^1((0,1))$, $\mathrm{H}^2((0,1))$.
(f) For $v = e^{-10x}$ and $\Omega = (0,1)$, is the relation $|u|_{\mathrm{H}^1(\Omega)} = |u|_{\mathrm{H}^2(\Omega)}$ satisfied?

Exercise 1.6.2.
Let us consider the following problem posed on the domain $\Omega = (0,1)$, with κ a real coefficient, and $f \in \mathrm{L}^2(\Omega)$:

$$\left| \begin{array}{l} \text{Find } u \in \mathrm{H}_0^1(\Omega) \text{ such that:} \\ \displaystyle\int_\Omega \kappa \frac{\partial u}{\partial x} v \,\mathrm{d}x + \int_\Omega \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} \,\mathrm{d}x + \int_\Omega uv \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x, \qquad v \in \mathrm{H}_0^1(\Omega) \end{array} \right. \tag{1.9}$$

(a) Formulate the strong problem corresponding to weak formulation (1.9).
(b) Discuss the existence and uniqueness of the solution to Problem (1.9).

Exercise 1.6.3.
Let us consider the biharmonic equation posed on the domain $\Omega = (0,1)$:

$$\boldsymbol{\Delta}^2 u(x) = f(x), \qquad \forall\, x \in \Omega \tag{1.10a}$$

with $f \in \mathrm{L}^2(\Omega)$, and satisfying the boundary condition on $\partial\Omega$

$$u(x) = u'(x) = 0, \qquad \forall\, x \in \partial\Omega \tag{1.10b}$$

(a) For $f \equiv 1$ give a solution to Problem (1.10).
(b) Derive a weak formulation (WF) of Problem (1.10).
(c) Specify the solution space and the test space.
(d) Show that there exists a unique solution u to (WF) belonging to the chosen solution space.

Exercise 1.6.4.
Let us consider the Helmholtz equation posed on the domain $\Omega = (0,1)$, given κ a real coefficient:

$$-u''(x) + \kappa u(x) = f(x), \qquad \forall\, x \in \Omega \tag{1.11a}$$

with $f \in \mathrm{L}^2(\Omega)$,

$$u(x) = 0, \qquad \forall\, x \in \partial\Omega \tag{1.11b}$$

(a) Derive a weak formulation (WF) of Problem (1.11).
(b) Specify the solution and test spaces.
(c) What is the nature of the bilinear form for $\kappa = 1$?
(d) Prove that the problem is well-posed for $\kappa = 0$ and $\kappa > 0$.
(e) Comment on the difficulty posed by the case $\kappa < 0$.
(f) The boundary condition is now given by:

$$u(0) - u'(0) = 0, \quad u'(1) = -1 \tag{1.12}$$

Derive a weak formulation for the Problem (4.2a)–(1.12) and show that it admits a unique solution. To prove the coercivity the following relation can be used:

$$v(1) = v(x) + \int_x^1 v'(t)\,\mathrm{d}t$$

Chapter 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.

Question: Can we construct an approximation to Problem (1.3) which is also well-posed?

2.1 Approximate problem

In the previous section we showed how a classical PDE problem such as Problem (1.3) can be reformulated as a weak problem. The abstract problem for this class of PDE reads then:

$$\left| \begin{array}{l} \text{Find } u \in V, \text{ such that:} \\ \\ a(u,v) = L(v) \quad , \forall\, v \in V \end{array} \right. \tag{2.1}$$

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: we want now to construct an approximate solution u_n to the Problem (2.1) then prove that the solution to the obtained approximate problem exists and is unique.

2.2 Ritz method for symmetric bilinear forms

2.2.1 Variational formulation and minimization problem

The idea behind the Ritz method is to replace the solution space V (which is infinite dimensional) by a finite dimensional subspace $V_n \subset V$, $\dim(V_n) = n$.

Problem (2.2) is the approximate weak problem by the Ritz method:

$$\left|\begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ a(u_n, v_n) = L(v_n) \quad ,\forall\ v_n \in V_n \end{array}\right. \tag{2.2}$$

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 (2.3) is the equivalent approximate variational problem under minimization form:

$$\left|\begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ J(u_n) \leq J(v_n) \quad ,\forall\ v_n \in V_n \\ \text{with } J(v_n) = \dfrac{1}{2}a(v_n, v_n) - L(v_n) \end{array}\right. \tag{2.3}$$

Proposition 2.2.1 (Equivalence of weak and variational formulations)**.** *Problem 2.2 and 2.3 are equivalent.*

Before moving to the well-posedness of the approximate variational problem some definitions are introduced to characterize the solution of minimization problems, then the equivalence of formulations for the Poisson problem with homogeneous Dirichlet boundary conditions in one dimension of space is given as example.

Definition 2.2.2 (Directional derivative)**.** Let V be a Hilbert space, for any $u \in V$ the relation:

$$J'(u; w) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(J(u + \varepsilon w) - J(u)\big) \tag{2.4}$$

defines $J'(\cdot;\cdot) : V \times V \to \mathbb{R}$ derivative of the functional J at u in the direction w.

Definition 2.2.3 (Fréchet derivative)**.** Let V be a Hilbert space, J is Fréchet-derivable at u if:

$$J(u + v) = J(u) + L_u(v) + \varepsilon(v)\|v\|_V \tag{2.5}$$

with L_u a continuous linear form on V and $\varepsilon(v) \to 0$ as $v \to 0$.

Proposition 2.2.4 (Optimality conditions)**.** *Let V be a Hilbert space and J a twice Fréchet-derivable functional, $u_0 \in V$ is solution to*

$$\inf_{v \in V} J(v) \tag{2.6}$$

if the following conditions are satisfied:

1. $J'(u_0) = 0$ *(Euler condition).*

2. $(J''(u)w\ ,\ w) \geq 0$ *(Legendre condition).*

Both conditions can be interpreted in terms of the simpler case of real functions: the first one requires that the first derivative cancels so that u_0 is an extremum, while the second condition is a convexity argument. Moreover, a sufficient condition is given by $(J''(\tilde{u})w\ ,\ w) \geq 0$ for any $\tilde{u}$ in a neighbourhood of u_0 (strong Legendre condition). The coercivity of the bilinear form $a(\cdot,\cdot)$ is an even stronger condition equivalent to: $\exists \alpha > 0$ such that $(J''(\tilde{u})w\ ,\ w) \geq \alpha(w,w)$.

Example 2.2.5. Equivalence of weak and variational formulations for the Dirichlet problem posed on $\Omega = (0,1)$. Let us derive the expression of $J'(u;w)$ defined by (2.4) given $\varepsilon > 0$ and $w \in V$.

First let us verify that if u solves the minimization problem then it solves the corresponding weak problem.

$$\begin{aligned} J(u+\varepsilon w) &= \frac{1}{2}\int_\Omega \left[(u+\varepsilon w)'\right]^2 \mathrm{d}\boldsymbol{x} - \int_\Omega f(u+\varepsilon w)\,\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int_\Omega \left[(u')^2 + 2\varepsilon u'w' + \varepsilon^2 (v')^2\right] \mathrm{d}\boldsymbol{x} - \int_\Omega fu\,\mathrm{d}\boldsymbol{x} - \varepsilon\int_\Omega fw\,\mathrm{d}\boldsymbol{x} \\ &= J(u) + \varepsilon\left[\int_\Omega u'w'\,\mathrm{d}\boldsymbol{x} - \int_\Omega fw\,\mathrm{d}\boldsymbol{x}\right] + \frac{1}{2}\,\varepsilon^2 \int_\Omega (w')^2\,\mathrm{d}\boldsymbol{x} \end{aligned}$$

Writing the derivative gives,

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\big(J(u+\varepsilon w) - J(u)\big) = \lim_{\varepsilon\to 0}\left[\int_\Omega u'w'\,\mathrm{d}\boldsymbol{x} - \int_\Omega fw\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\,\varepsilon|w|_{\mathrm{H}^1_0}\right]$$

so that the Euler condition holds if for any $w \in V = \mathrm{H}^1_0(\Omega)$

$$J'(u;w) = \int_\Omega u'w'\,\mathrm{d}\boldsymbol{x} - \int_\Omega fw\,\mathrm{d}\boldsymbol{x} = 0$$

In this case the functional J is Fréchet-derivable as L_u is linear.

Secondly, the other way around considering that the weak formulation holds for the test function $\varepsilon w \in V$ then in the relation

$$J(u+\varepsilon w) = J(u) + \varepsilon\left[\int_\Omega u'w'\,\mathrm{d}\boldsymbol{x} - \int_\Omega fw\,\mathrm{d}\boldsymbol{x}\right] + \frac{1}{2}\,\varepsilon^2\int_\Omega (w')^2\,\mathrm{d}\boldsymbol{x}$$

the second term of the right-hand side cancels, and the third term is non-negative, then

$$J(u+\varepsilon w) \geq J(u)$$

so that u is solution to the minimization problem.

The same result holds for the continuous problem in V and the approximation in V_n since only requirement is to work in a Hilbert space. Actually the following result for the Dirichlet problem is due to Stampacchia which characterizes the solution to the weak problem in term of minimization.

Theorem 2.2.6 (Stampacchia)**.** *Let $a(\cdot,\cdot)$ be a bilinear coercive continuous form on* H *a Hilbert space, and K be a convex closed non-empty subset of H. Given $\phi \in H'$, $\exists! u \in K$ such that*

$$a(\, u\, ,\, v-u\,) \geq \langle\, \phi\, ,\, v-u\,\rangle_{H',H}, \qquad \forall\, v \in K$$

and if a is symmetric then

$$u = \underset{v\in K}{\operatorname{argmin}} \left\{ \frac{1}{2} a(\, v\, ,\, v\,) - \langle\, \phi\, ,\, v\,\rangle_{H',H} \right\}$$

The solution can be seen as satisfying a minimization of energy, also called Dirichlet principle.

2.2.2 Well-posedness

Theorem 2.2.7 (Well-posedness)**.** *Let V be a Hilbert space and V_n a finite dimensional subspace of V, $\dim(V_n) = n$, Problem* (2.2) *admits a unique solution u_n.*

Proof. Given that the weak formulation differs only by introducing finite dimensional subspaces the proof could conclude directly with the Lax–Milgram Theorem. Instead we show that there exists a unique solution to the equivalent minimization problem (2.3) by explicitly constructing an approximation $u_n \in V_n$ decomposed uniquely on a basis $(\varphi_1, \cdots, \varphi_n)$ of V_n:

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

In practice this basis is not any basis but the one constructed to define the approximation space V_n: to one chosen approximation space will correspond one carefully constructed basis. 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 minimization 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 vector of algebraic unknowns also called *degrees of freedom*

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

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

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

Proposition 2.2.8 (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.*

As a consequence to Proposition 2.2.8 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 $J'(u_n) = 0$ □

The general setting for Galerkin methods will be to construct approximate solutions of the form:

$$u_n = \sum_{j=1}^{n} u_j \varphi_j \tag{2.8}$$

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. Since V_n is finite dimensional, there exist a unique decomposition (2.8) on the basis. This basis can be chosen in a way that seems natural so that in practice we will construct one basis for a given type of space V_n and which will define the approximation properties (the basis itself is not unique but we need to choose one that possesses good properties, in a similar fashion that it is more suitable to work with the canonical basis in Euclidean spaces).

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 $n \to \infty$? The ingredients are similar to the Lax principle: stability and consistency implies convergence.

Lemma 2.2.9 (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* (2.1) *and the solution to approximate Problem* (2.2)*. 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:

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

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 2.2.10 (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* (2.1) *and the solution to approximate Problem* (2.2) *, 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 (2.2.9) 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 (2.2.10) gives a control on the discretization error $e_n = u - u_n$ which is *quasi-optimal* in the V-norm (*i.e.* bound multiplied by a constant).

Lemma 2.2.11 (Stability)**.** *Any solution $u_n \in V_n$ to Problem* (2.2) *satisfies:*

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

Proof. Direct using the coercivity and the dual norm. First, the inequality

$$\alpha \|u_n\|_V^2 \leq a(u_n, u_n) = L(u_n)$$

holds for some $\alpha > 0$, and secondly by definition of the dual norm,

$$\|L\|_{V'} = \sup_{v \in V, v \neq 0} \frac{L(v)}{\|v\|_V}$$

so that

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

□

2.2.4 Method

Algorithm 2.2.12 (Ritz 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 minimization problem*

2.3 Galerkin method

2.3.1 Formulation

We use a similar approach as for the Ritz 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 (2.9) is the approximate weak problem by the Galerkin method:

$$\left| \begin{array}{l} \text{Find } u_n \in V_n,\ V_n \subset V, \text{ such that:} \\ a(u_n, v_n) = L(v_n) \quad ,\forall\ v_n \in V_n \end{array} \right. \tag{2.9}$$

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. Consequently, whenever needed we will refer to the following proposition:

Proposition 2.3.1 (Galerkin orthogonality)**.** *Let $u \in V$, $u_n \in V_n$ respectively the solution to Problem* (2.1) *and the solution to approximate Problem* (2.9)*, then:*

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

Proof. Direct consequence of the consistency of the method. The approximate problem

$$a(\ u_n, v_n\ ) = L(v_n) \quad , \forall\ v_n \in V_n \tag{2.10}$$

is obtained by replacing V by a finite dimensional space V_n but the relation

$$a(\ u, v\ ) = L(v) \quad , \forall\ v \in V \tag{2.11}$$

from the weak formulation is otherwise the same: $a(\cdot,\cdot)$ and $L(\cdot)$ are unchanged. Since $V_n \subset V$, it is possible to choose $v \in V_n$ in Equation (2.11), then

$$a(\ u_n, v_n\ ) = a(\ u, v_n\ ) \quad , \forall\ v_n \in V_n \tag{2.12}$$

which gives directly the orthogonality property. □

Lemma 2.3.2 (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* (2.1) *and the solution to approximate Problem* (2.9)*, 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 Method

Algorithm 2.3.3 (Galerkin 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* $\mathrm{A}\mathbf{u} = \mathbf{b}$

2.4 Boundary conditions

2.5 Exercises

Exercise 2.5.1.
Given an abstract weak problem posed in a Hilbert space V:

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

and a minimization problem

$$\left|\begin{array}{l} \text{Find } u \in V,\ V, \text{ such that:} \\ J(u) \leq J(v) \quad ,\forall\, v \in V \\ \text{with } J(v) = \dfrac{1}{2}a(v,v) - L(v) \end{array}\right.$$

(a) Show the equivalence of the formulations when a is bilinear symmetric positive definite and L linear.
(b) Show that if $V = \mathbb{R}^n$ the minimization problem can be recast into a strictly convex quadratic form $J(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathrm{A}\mathbf{u} - \mathbf{u}^T\mathbf{b}$ and the unique solution satisfies $\mathrm{A}\mathbf{u} = \mathbf{b}$.

Exercise 2.5.2.
Let us consider the Poisson problem posed on the domain $\Omega = (0,1)$:

$$-u''(x) = f(x), \qquad \forall\, x \in \Omega \tag{2.13a}$$

with $f \in \mathrm{L}^2(\Omega)$, and satisfying the boundary condition on $\partial\Omega$

$$u(x) = 0, \qquad \forall\, x \in \partial\Omega \tag{2.13b}$$

(a) For $f \equiv 1$ give a solution to Problem (2.13).
(b) Find the weak formulation (WF) of Problem (2.13) and specify the function spaces.
(c) Is this problem well-posed?
(d) Justify that it is possible to reformulate this problem into a minimization problem?
(e) Derive the minimization functional $J(u)$.
(f) Let $w_1 = a_1 \sin(\pi x)$. Find the value of the amplitude a_1 minimizes $J(w_1)$. How does a_1 compare with the maximum of the exact solution u?
(g) Show that $J(w_1) > J(u)$ and interpret.
(h) Let $\phi_i = \sin\big((2i-1)\pi x\big)$, $i \in \mathbb{N}$. Verify that these function are infinitely differentiable and satisfy $\phi_i(0) = \phi_i(1) = 0$. Compute coefficients

$$a_{ij} = \int_0^1 \phi_i'(x)\phi_j'(x)\,\mathrm{d}x\,, \qquad b_i = \int_0^1 f(x)\phi_i(x)\,\mathrm{d}x$$

(i) Given a finite dimensional space $V_n = \mathrm{span}\{\phi_i\}_{1\leq i\leq n}$, express the linear system obtained by the Galerkin method and give the solution.

Chapter 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 now to define the discrete space V_h and construct a basis $(\varphi_1, \cdots, \varphi_{N_{V_h}})$ of V_h, $N_{V_h} = \dim(V_h)$, on which the discrete solution is decomposed as

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

with $\{u_j\}$ 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*.

Previously no assumption was made on the finite dimensional space V_n aside from that $V_n \subset V$. The change of notation to V_h is to reflect that the discrete space V_h will be characterized more carefully as an *approximation space* by constructing the shape functions and by defining the degrees of freedom.

To construct the Finite Element space V_h, three ingredients are introduced:

1. An admissible mesh $\mathcal{T}_h$ generated by a tessellation of domain Ω.

2. A reference Finite Element $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ to construct a basis of V_h and define the meaning of u_j.

3. A mapping that generates a Finite Element $(K, \mathcal{P}, \Sigma)$ for any cell in the mesh from the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$.

As a preliminary step the approximation of the Poisson problem in one dimension by linear Lagrange Finite Elements is described to give an overview of the methodology without hitting the technical difficulties. Concepts and notations for the discretization of the physical domain are then introduced. Provided that all the requirements are identified, a framework for all the Finite Element methods is introduced by stating the definition of a Finite Element. Finally, the generation of the Finite Element space from a reference Finite Element will be described. Some examples of Finite Element spaces are listed at the end of the chapter.

3.1 A preliminary example in one dimension of space

For the sake of completeness, steps performed to derive a Galerkin method for the Poisson Problem 1.3 on $\Omega = (0, 1)$ are sketched below to recapitulate the methodology.

3.1.1 Weak formulation

A solution is sought in the distributional sense by testing the equation against smooth functions,

$$-\int_\Omega u''(x)v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x\,, \qquad \forall\, v \in \mathrm{C}_c^\infty(\Omega)$$

then reporting derivatives on the test functions using the integration by part

$$-\int_\Omega u''(x)v(x)\,\mathrm{d}x = -\underbrace{[u'(x)v(x)]_0^1}_{=0} + \int_\Omega u'(x)v'(x)\,\mathrm{d}x$$

the weak formulation consists of finding $u \in V$ such that

$$\int_\Omega u'(x)v'(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x\,, \qquad \forall\, v \in V$$

given $f \in \mathrm{L}^2(\Omega)$. The choice of solution space and test space is guided by the equation and the data: in this case $V = \mathrm{H}_0^1(\Omega)$ since v and v' should be controlled in $\mathrm{L}^2(\Omega)$, and homogeneous Dirichlet boundary conditions are imposed.

3.1.2 Galerkin method

The approximate problem by a Galerkin method consists of seeking a discrete solution u_h in a finite dimensional space $V_h \in V$, such that

$$\int_\Omega {u_h}'(x) v'(x)\,\mathrm{d}x = \int_\Omega f(x) v(x)\,\mathrm{d}x\,, \qquad \forall\, v \in V_h$$

and given a basis $\{\varphi_j\}$ of V_h any function $w \in V_h$ can be written as

$$w_h = \sum_{j=1}^{N_{V_h}} \underbrace{w_j}_{\in \mathbb{R}}\ \varphi_j$$

with $N_{V_h} = \dim V_h$, and moreover

$$w_h' = \sum_{j=1}^{N_{V_h}} \underbrace{w_j}_{\in \mathbb{R}}\ \varphi_j'$$

by simple application of the derivation on the linear combination. Bounds of the sum will be omitted to simplify the notation,

$$\sum_{j=1}^{N_{V_h}} \sim \sum_j$$

when there is no possible confusion.

Inserting the Galerkin decomposition in the weak formulation and using the commutativity of the derivation with the linear combinations,

$$\int_\Omega \Big(\sum_j u_j\ \varphi_j'(x)\Big)\Big(\sum_i v_i\ \varphi_i'(x)\Big)\,\mathrm{d}x = \int_\Omega f(x)\Big(\sum_i v_i\ \varphi_i(x)\Big)\,\mathrm{d}x$$

and the integration can also commute with the linear combinations,

$$\sum_j \sum_i u_j v_i \int_\Omega \varphi_j'(x)\varphi_i'(x)\,\mathrm{d}x = \sum_i v_i \int_\Omega f(x)\varphi_i(x)\,\mathrm{d}x$$

so that up to some cosmetic reordering, for any $v \in V$

$$\sum_i \sum_j v_i \int_\Omega \varphi_i'(x)\varphi_j'(x)\,\mathrm{d}x u_j = \sum_i v_i \int_\Omega f(x)\varphi_i(x)\,\mathrm{d}x$$

the relation to the linear system of algebraic equations becomes evident,

$$\mathbf{v}^T \mathrm{A}\ \mathbf{u} = \mathbf{v}^T \mathbf{b}$$

for any $\mathbf{v} = [v_i] \in \mathbb{R}^{N_{V_h}}$, with

$$\mathrm{A} = \left[\int_\Omega \varphi_j'(x)\varphi_i'(x)\,\mathrm{d}x\right]_{ij}, \quad \mathbf{b} = \left[\int_\Omega f(x)\varphi_i(x)\,\mathrm{d}x\right]_i$$

and the discrete solution is represented by the solution vector $\mathbf{u} = [u_j] \in \mathbb{R}^{N_{V_h}}$. Computing contributions A_{ij} and $\mathbf{b}_i$ is possible as soon as the basis $\{\varphi\}_j$ is constructed explicitly.

Remark 3.1.1. The choice of indices i and j in the previous expressions follows the usual convention for row and column indices. The matrix A represents a linear application from the solution space H to the trial space V: therefore the solution space is the column space, while the trial space is the row space. In the case of Galerkin approximations where $\mathrm{H} = V$ – and even more when the bilinear form is symmetric – it is tempting to choose the indices arbitrarily, but for the sake of consistency following the convention is recommended.

3.1.3 Construction of the discrete space

A discretization of the computational domain $\bar{\Omega} = [0, 1]$ is constructed by partitioning the interval into disjoints subintervals $K_i = [x_i, x_{i+1}]$, $1 \leq i \leq N_K$.

$$\bar{\Omega} = \bigcup_{i=1}^{N_K} K_i \,, \qquad \mathring{K}_i \cap \mathring{K}_j = \varnothing$$

The family of cells $\{K_i\}$ defines a *mesh* noted $\mathcal{T}_h$. A partition of $[0, 1]$ into N_K subintervals $[x_i, x_{i+1}]$ corresponds to a family of $N_K + 1$ points $\{x_i\}$ which are the *vertices* of the mesh $\mathcal{T}_h$.

In this example the approximation space is constructed with piecewise linear Lagrange polynomials. The chosen discrete space is

$$V_h = \{v \in \mathrm{C}^0(\bar{\Omega}) \cap \mathrm{H}^1_0(\Omega) : \forall\, K \in \mathcal{T}_h, v|_K \in \mathbb{P}_1(K)\} \tag{3.1}$$

which consists of functions continuous over $\bar{\Omega}$ that are linear on each cell K, and $V_h \subset V$ so that the approximation is H^1_0–conformal. A continuous piecewise linear function $v_h \in V_h$ is the linear interpolate of $v \in V$ on $\mathcal{T}_h$ if $v_h = \mathcal{I}_{V_h}$ with the interpolation operator

$$\begin{array}{rcccl} \mathcal{I}_{V_h} & : & V & \to & V_h \\ & & v & \mapsto & \displaystyle\sum_{i=1}^{N_{V_h}} v(\xi_i)\varphi_i \end{array} \tag{3.2}$$

such that $N_{V_h} = \dim V_h$ and $\{\xi_i\}$ is a family of N_{V_h} distinct nodes. In this particular case of a linear interpolation $N_{V_h} = N_K + 1$ and the family $\{\xi_i\}$ is identified with the vertices $\{x_i\}$.

The construction of the basis $\{\varphi_j\}$ of V_h should produce a linear system that can be solved easily so the matrix should be as sparse as possible. Given the expression of the contributions A_{ij} the requirement is that functions φ_j overlap as little as possible with each other: the support of φ_j should be reduced so that contributions are non-zero only for neighbouring φ_i functions; this choice is consistent with the locality of differential operators.

For any x_i the shape function is defined as

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

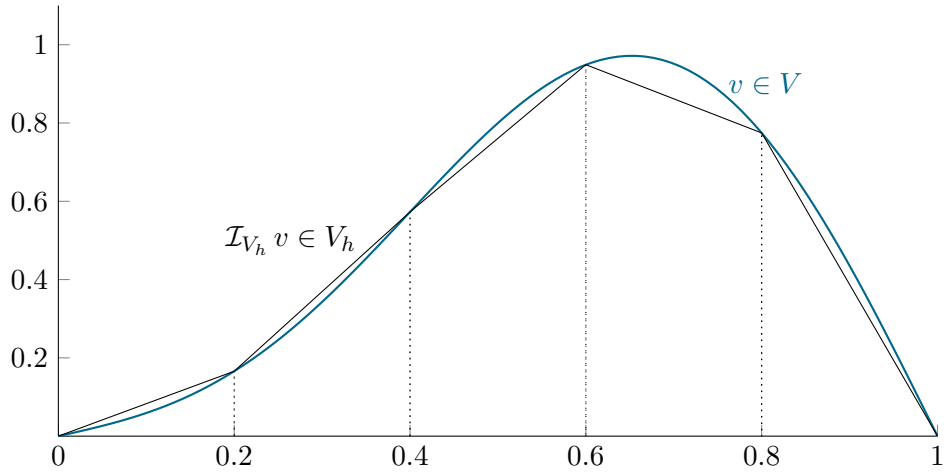


Figure 3.1: The function $v(x) = 0.8\sin(2\pi x) - 0.32\sin(4\pi x)$ and its linear interpolate $\mathcal{I}_{V_h}\, v$ with 6 equidistributed nodes.

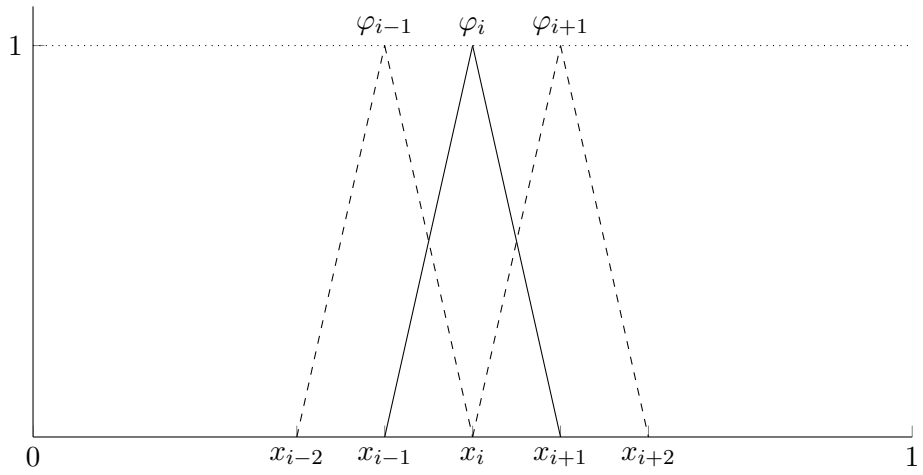


Figure 3.2: Shape functions for Lagrange $\mathbb{P}_1$ on a unit interval discretized into subintervals $K_i = [x_i, x_{i+1}]$.

so that the support of constructed functions φ_i overlaps only with φ_{i-1} and φ_{i+1}, and the expression of all the functions can be obtained from one another by an affine transformation; in the case of a uniform grid, $|x_{i+1} - x_i| = h$, shape functions are obtained simply by translation. Therefore it would be convenient to define shape functions on a reference interval, then apply an affine transformation to generate any φ_i. Shape functions are considered on a *reference cell* $\hat{K}$ which is the unit interval, then transported to any subinterval K_i using the geometric transformation from $\hat{K}$ to K.

In the *reference cell* $\hat{K} = [0, 1]$ shape functions depicted in Figure 3.1.3 have expressions

$$\begin{cases} \hat{\varphi}_0(\hat{x}) &= 1 - \hat{x} \\ \hat{\varphi}_1(\hat{x}) &= \hat{x} \end{cases}$$

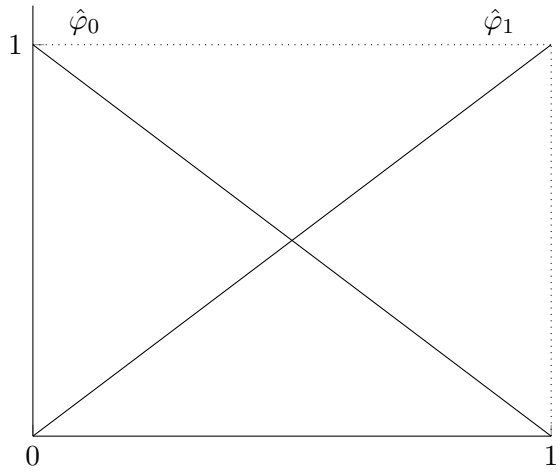


Figure 3.3: Shape functions for Lagrange $\mathbb{P}_1$ on the interval $\hat{K} = [0, 1]$.

and the affine mapping T_K mapping $\hat{K}$ to K in one dimension is given by $T_K : \hat{x} \mapsto b + a\hat{x}$, for some $a, b \in \mathbb{R}$. The affine mapping to any subinterval K_i is given by the change of coordinates

$$T_K : \hat{x} \in \hat{K} \mapsto x = x_i + (x_{i+1} - x_i)\hat{x} \in K_i$$

which is consistent with the definition of the global shape functions and reference shape functions, since

$$\begin{cases} \hat{\varphi}_0(\hat{x}) &= 1 - \hat{x} \\ \varphi_i(x) &= \dfrac{x_{i+1} - x}{x_{i+1} - x_i} = 1 - \dfrac{x - x_i}{x_{i+1} - x_i} \end{cases} \qquad \begin{cases} \hat{\varphi}_1(\hat{x}) &= \hat{x} \\ \varphi_{i+1}(x) &= \dfrac{x - x_i}{x_{i+1} - x_i} \end{cases}$$

so that $\hat{x} = (x - x_i)(x_{i+1} - x_i)^{-1}$ as expected.

Remark 3.1.2 (Link to higher dimensions of space)**.** The affine mapping in $\mathbb{R}^d$ is given a relation of the type $T_K : \hat{\boldsymbol{x}} \mapsto \mathbf{b}_K + \mathrm{B}_K\hat{\boldsymbol{x}}$, with $\mathbf{b}_K \in \mathbb{R}^d$ and $\mathrm{B}_K \in \mathbb{R}^{d\times d}$. The matrix B_K is the Jacobian of T_K, noted J_{T_K}. In one dimension of space J_{T_K} contains only one entry $\partial T_K/\partial\hat{x} = (x_{i+1} - x_i) = |K|$, and its inverse $\mathrm{J}_{T_K}^{-1}$ is obviously $(x_{i+1} - x_i)^{-1} = |K|^{-1}$.

3.1.4 Transport of Finite Element contributions

In the case of linear Lagrange elements the transport of Finite Element contributions corresponds to the change of coordinates T_K. The change of variable $x = T_K(\hat{x})$ is sketched for the mass matrix and the stiffness matrix, given a cell $K = [x_{i+1} - x_i]$.

Firstly, let us consider contributions to the mass matrix M,

$$\mathrm{M}_{ij} = \int_K \varphi_j(x)\varphi_i(x)\,\mathrm{d}x$$

which can be rewritten, given that $x = T_K(\hat{x}$ and $\mathrm{d}x = (x_{i+1} - x_i))\,\mathrm{d}\hat{x}$,

$$\mathrm{M}_{ij} = \int_{\hat{K}} \varphi_j(T_K(\hat{x}))\varphi_i(T_K(\hat{x}))(x_{i+1} - x_i)\,\mathrm{d}\hat{x}$$

Since $\varphi_i \circ T_K = \hat{\varphi}_i$ and $(x_{i+1} - x_i) = |K|$, then

$$\mathrm{M}_{ij} = |K| \int_{\hat{K}} \hat{\varphi}_j(\hat{x})\hat{\varphi}_i(\hat{x})\,\mathrm{d}\hat{x}$$

so that contributions for $K \in \mathcal{T}_h$ to the mass matrix can obtained by a scaling of the contribution on $\hat{K}$.

Secondly, let us consider contributions to the stiffness matrix K,

$$\mathrm{K}_{ij} = \int_K \partial_x\varphi_j(x)\ \partial_x\varphi_i(x)\,\mathrm{d}x$$

which can be rewritten using the chain rule for the derivation of $\varphi_i = \hat{\varphi}_i \circ T_K^{-1}$, formally $\partial_x\varphi_i = \partial_{\hat{x}}\hat{\varphi}_i\ \partial_x T_K^{-1}$ with $\partial_x T_K^{-1} = |K|^{-1}$

$$\mathrm{K}_{ij} = \int_{\hat{K}} \frac{1}{|K|}\partial_{\hat{x}}\hat{\varphi}_j(\hat{x})\ \frac{1}{|K|}\partial_{\hat{x}}\hat{\varphi}_i(\hat{x})|K|\,\mathrm{d}\hat{x}$$

therefore,

$$\mathrm{K}_{ij} = \frac{1}{|K|}\int_{\hat{K}} \hat{\varphi}_j(\hat{x})\hat{\varphi}_i(\hat{x})\,\mathrm{d}\hat{x}$$

so that contributions for $K \in \mathcal{T}_h$ to the stiffness matrix can also obtained by a scaling of the contribution on $\hat{K}$.

3.1.5 Generalization of the methodology

The example of linear Lagrange in one dimension provides a good overview of the methodology for constructing discrete spaces using Finite Elements. The steps can be summarized as:

1. Choose a function space suggested by the weak formulation.
2. Discretize the computational domain into a mesh.
3. Choose a polynomial approximation space.
4. Define an interpolation operator.
5. Provide a definition of a reference element.
6. Construct a mapping to generate the discrete space.

The generalization of this approach can take different directions:

- Extending to multiple dimensions in space.
- Increasing the polynomial order.
- Using a different approximation space.
- Controlling the solution in other means than pointwise values.

3.2 Admissible mesh

Definition 3.2.1 (Mesh)**.** Let Ω be polygonal ($d = 2$) or polyhedral ($d = 3$) subset of $\mathbb{R}^d$, we define $\mathcal{M}$ (a triangulation $\mathcal{T}_h$ in the simplicial case) as a finite family $\{K_i\}$ of disjoint convex non-empty subsets of Ω named cells. Moreover $\mathcal{V}(\mathcal{T}_h) = \{\mathrm{v}_i\}$ denotes the set a vertices of $\mathcal{M}$, $\mathcal{E}(\mathcal{M}) = \{\sigma_{KL} = K \cap L\}$ denotes the set of facets, which are edges ($d = 2$) or faces ($d = 3$).

Definition 3.2.2 (Mesh size)**.**

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

with $\mathrm{diam}(K)$ with diameter of the cell, *i.e.* the maximum distance between two points of K.

Definition 3.2.3 (Geometrically conforming mesh)**.** A mesh is said geometrically conforming if two neighbouring cells share either exactly one vertex, exactly one edge in the case $d = 2$, or in the case $d = 3$ exactly one face.

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, ...

3.3 Definition of a Finite Element

In the frame of the Galerkin method, given a function space V, a discrete space $V_h = \mathrm{span}\{\phi_j\} \subset V$ was introduced, this section describes how to build such space by constructing an abstraction called *Finite Element*.

Definition 3.3.1 (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 of functions $p : K \to \mathbb{R}$, which is the space of shape functions; $\dim(\mathcal{P}) = N_{\mathcal{P}}$.
- Σ is a set $\{\sigma\}_j$ of linear forms,

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

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

Practically, the definition requires first to consider 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 $N_{\mathcal{P}}$ geometrical nodes $\{\xi_i\}$, $\sigma_i(\varphi_j) = \varphi_j(\xi_i)$). The local shape functions $\{\varphi_i\}$ are then constructed so as to ensure unisolvence. The dimension of $\mathcal{P}$ is usually called *dimension of the Finite Element*.

Proposition 3.3.2 (Determination of the local shape functions)**.** *Let $\{\sigma_i\}_{1\leq i\leq N_{\mathcal{P}}}$ be the set of local degrees of freedoms, the local shape functions are defined as $\{\varphi_i\}_{1\leq i\leq N_{\mathcal{P}}}$ a basis of $\mathcal{P}$ such that,*

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

Definition 3.3.3 (Unisolvence)**.** A Finite Element is said unisolvent if for any vector $(\alpha_1, \cdots, \alpha_{N_{\mathcal{P}}}) \in \mathbb{R}^N_{\mathcal{P}}$ there exists a unique representant $p \in \mathcal{P}$ such that $\sigma_i(p) = \alpha_i$, $\forall\, i \in [\![1, N_{\mathcal{P}}]\!]$.

The unisolvence property of a Finite Element is equivalent to construct Σ as dual basis of $\mathcal{P}$, thus any function $p \in \mathcal{P}$ can be expressed 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 $(p_1, \cdots, p_{N_{\mathcal{P}}})$ uniquely represents a function of $\mathcal{P}$. Defining Σ as dual basis of $\mathcal{P}$ is equivalent to:

$$\dim(\mathcal{P}) = \mathrm{card}(\Sigma) = N_{\mathcal{P}} \tag{3.3a}$$

$$\forall\, p \in \mathcal{P},\ (\sigma_i(p) = 0, 1 \leq i \leq N) \Rightarrow (p = 0) \tag{3.3b}$$

in which Property (3.3a) ensures that Σ generates $\mathcal{L}(\mathcal{P}, \mathbb{R})$ and Property (3.3b) 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.

As a first step, the Galerkin decomposition of functions $u_h \in V_h$ was introduced given a basis of V_h and degrees of freedom u_j, but without giving a proper definition of the latter; we just assumed that they are real coefficients. In a second step, the Lagrange interpolation operator of Definition 3.2 was introduced to interpolate functions in V as continuous piecewise linear functions in V_h. Degrees of freedom u_j were then defined as $u(\xi_j)$ the value of the function u at the Lagrange node ξ_j. Now that a definition a Finite Element was stated, in a similar fashion a global interpolation operator can be defined for general Finite Elements, *i.e.* with degrees of freedom defined by linear forms σ_j.

Definition 3.3.4 (Global interpolation operator)**.**

$$\begin{aligned} \mathcal{I}_{V_h} : \quad V &\to V_h \\ v &\mapsto \sum_{j=1}^{N_{V_h}} \sigma_j(v)\, \varphi_j \end{aligned}$$

Given that the Finite Element is defined on a cell K, the corresponding local interpolation operator can be introduced.

Definition 3.3.5 (Local interpolation operator – [4] page 20)**.**

$$\begin{aligned} \mathcal{I}_{K,V} : \quad V(K) &\to \mathcal{P} \\ v &\mapsto \sum_{j=1}^{N_{\mathcal{P}}} \sigma_j(v)\, \varphi_j \end{aligned}$$

Remark 3.3.6. The notation using the dual basis can be confusing but with the relation $\sigma_i(p) = p(\xi_i)$ in the nodal Finite Element case it is easier to understand that the set Σ of linear forms defines how the interpolated function $\mathcal{I}_{K,V}\, u$ 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 $u_i = u(\xi_i)$ at geometrical nodes $\{\xi_i\}$, 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)
$$\sigma_i(v) = \int_\xi v \cdot \mathbf{n}_\xi \, \mathrm{d}s$$
- 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)
$$\sigma_i(v) = \int_\xi v \cdot \boldsymbol{\tau}_\xi \, \mathrm{d}s$$

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.

Remark 3.3.7. The Finite Element approximation is said H-conformal if $V_h \subset \mathrm{H}$ and is said non-conformal is $V_h \not\subset \mathrm{H}$. 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.

3.4 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 dimensions the unit triangle with vertices $\{(0, 0), (1, 0), (0, 1)\}$, and in three dimensions the unit tetrahedron with vertices $\{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)\}$.

Using this approach any Finite Element $(K, \mathcal{P}, \Sigma)$ on the mesh can be generated from $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ provided that a mapping can be constructed such that $(K, \mathcal{P}, \Sigma)$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ possess equivalent properties. In particular, an important property is the unisolvence: if a Finite Element is equivalent to another Finite Element which is unisolvent, then it is also unisolvent.

Definition 3.4.1 (Affine-equivalent Finite Elements)**.** Two Finite Elements $(K, \mathcal{P}, \Sigma)$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ are said *affine-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 the *affine mapping* T_K but this is not suitable in general. An auxiliary mapping is needed to transfer the approximation space on $\hat{K}$ to the approximation space on K. The following definition extends the affine-equivalence to a general equivalence property between Finite Elements.

Definition 3.4.2 (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 bicontinuous bijection Φ_K from $V(K)$ onto $V(\hat{K})$ such that $(K, \mathcal{P}, \Sigma)$ is generated from $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$:

$$\begin{aligned} K &= T_K(\hat{K}) \\ \mathcal{P} &= \{\Phi_K^{-1}(\hat{p}), \forall\, \hat{p} \in \hat{\mathcal{P}}\} \\ \Sigma &= \{\sigma_{K,j} : \sigma_{K,j}(\hat{p}) = \hat{\sigma}_{K,j}(\Phi_K^{-1}(\hat{p})), \forall\, \hat{p} \in \hat{\mathcal{P}}\} \end{aligned}$$

Given that any Finite Element $(K, \mathcal{P}, \Sigma)$ can be generated from a reference Finite Element $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$, then global interpolation properties for the entire space V_h (spanned by collecting all the shape functions) can be inferred from local interpolation properties on $\hat{K}$. More precisely, without going into the details, the following result is specified in [4]:

$$\begin{array}{ccc} V(K) & \xrightarrow{\Phi_K} & V(\hat{K}) \\ \Big\downarrow{\scriptstyle \mathcal{I}_{K,V}} & & \Big\downarrow{\scriptstyle \mathcal{I}_{\hat{K},V}} \\ \mathcal{P} & \xrightarrow{\Phi_K} & \hat{\mathcal{P}} \end{array}$$

which means that the interpolation operators and the transport of the elements commute: $\mathcal{I}_{\hat{K},V} \circ \Phi_K = \Phi_K \circ \mathcal{I}_{K,V}$.

Remark 3.4.3. Note that in the literature the mapping between spaces is defined from $V(K)$ to $V(\hat{K})$, while the geometric mapping between cells is defined from $\hat{K}$ to K. The affine-equivalence consists in the case Φ_K coinciding with T_K^{-1}. For Lagrange we defined $\Phi_K : \mathrm{C}^0(K) \to \mathrm{C}^0(\hat{K})$ given by $v \mapsto v \circ T_K$. In that case $\mathcal{P} = \operatorname{span}\left\{\hat{\varphi} \circ T_K^{-1}\right\}$ and the degrees of freedom are defined by $\Sigma = \left\{\sigma_i : \sigma_i(v) = \hat{\sigma}_i(\Phi_K(v)) = \Phi_K(v)(\hat{\xi}_i) = v \circ T_K(\hat{\xi}_i)\right\}$. If $\xi_i = T_K(\hat{\xi}_i)$ then the definition is consistent.

3.5 Method

Algorithm 3.5.1 (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 generate* $(K, \mathcal{P}, \Sigma)$.
4. *Construct a basis for* V_h *by collecting all the shape functions of Finite Elements* $\{(K, \mathcal{P}, \Sigma)\}_{K \in \mathcal{T}_h}$ *sharing the same degree of freedom.*

3.6 Exercises

Exercise 3.6.1.
Let us consider the Poisson problem posed on the domain $\Omega = (0,1)$:

$$-u''(x) = f(x), \qquad \forall\, x \in \Omega \tag{3.4a}$$

with $f \in \mathrm{L}^2(\Omega)$, and satisfying the boundary condition on $\partial\Omega$

$$u(x) = 0, \qquad \forall\, x \in \partial\Omega \tag{3.4b}$$

The domain $\bar{\Omega}$ is discretized into a family of subintervals $[x_i, x_{i+1}]$, $i = 0, \cdots, N$, and Problem 3.4 is approximated by a linear Lagrange finite element method. The approximation space is the space of continuous piecewise linear functions $V_h = \{\varphi_i\}_{0 \le i \le N}$ with

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x_{i-1} \le x \le x_i,\ i \neq 0 \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x_i \le x \le x_{i+1},\ i \neq N \\ 0 \quad \textit{otherwise} \end{cases} \tag{3.5}$$

(a) Find the weak formulation of Problem 3.4.
(b) Prove that $a(u - u_h, \varphi_i) = 0$, for $i = 1, \cdots, N-1$.
(c) Prove that for any $v \in \mathrm{H}^1([0,1]) \cap \mathrm{C}^0([0,1])$, $i = 1, \cdots, N-1$:

$$a(v, \varphi_i) = \frac{1}{h}\big[-v(x_{i-1}) + 2v(x_i) - v(x_{i+1})\big] \tag{3.6}$$

Let us consider $f(x) = x^4$:

(a) Find the expression of the solution to Problem 3.4.
(b) Give the expression of the linear system obtained by the suggested method on a uniform grid, *i.e.* $x_i = ih$, $i = 0, \cdots, N$.
(c) Implement a program computing the discrete solution u_h using the suggested method.
(d) Plot the discrete solution u_h, the exact solution u, and the error $|u - u_h|$ with $N = 8, 16, 32$.
(e) Implement a function computing the L^2 error-norm $\|u - u_h\|_{\mathrm{L}^2(\Omega)}$ and plot the value for different values of N.
(f) Modify the program to handle non-uniform grids, given a list of node coordinates $\{x_i\}_{0 \le i \le N}$.
(g) Based on the error $|u - u_h|$ suggest a distribution of the nodes $\{x_i\}_{0 \le i \le N}$, repeat the same study, then compare the error values.

Chapter 4

Simplicial Lagrange Finite Elements

4.1 Definitions

Vector spaces of polynomials are used as approximation spaces to construct Finite Elements.

Definition 4.1.1 (Space of polynomials with real coefficients)**.** Let $K \subset \mathbb{R}^d$, $k \in \mathbb{N}$, $\mathbb{P}_k(K)$ is the vector space of polynomials with real coefficients of degree k on K, and the canonical basis is given by the family $\left\{x_1^{\alpha_1} \cdots x_i^{\alpha_i} \cdots x_d^{\alpha_d} : |\alpha| = k\right\}$, with x_i the i-th coordinate of $\boldsymbol{x} \in \mathbb{R}^d$.

The dimension of the space is given by

$$\dim(\mathbb{P}_k(\mathbb{R}^d)) = \frac{1}{d!}\prod_{i=1}^{d}(k+i)$$

so that in particular $\dim(\mathbb{P}_k(\mathbb{R}^1)) = k+1$, which means that such polynomials will be uniquely defined by its values at $k+1$ nodes.

Example 4.1.2. Linear and quadratic polynomials in different dimensions of space are listed below.

$$\begin{aligned}
\mathbb{P}_1(\mathbb{R}^1) &= \operatorname{span}\{1, x\} \\
\mathbb{P}_1(\mathbb{R}^2) &= \operatorname{span}\{1, x, y\} \\
\mathbb{P}_1(\mathbb{R}^3) &= \operatorname{span}\{1, x, y, z\} \\
\mathbb{P}_2(\mathbb{R}^1) &= \operatorname{span}\left\{1, x, x^2\right\} \\
\mathbb{P}_2(\mathbb{R}^2) &= \operatorname{span}\left\{1, x, y, xy, x^2, y^2\right\} \\
\mathbb{P}_2(\mathbb{R}^3) &= \operatorname{span}\left\{1, x, y, z, xy, yz, xz, x^2, y^2, z^2\right\}
\end{aligned}$$

Simplicial Lagrange Finite Elements are considered, for which the approximation space $\mathcal{P}$ will be polynomials on K, a d-simplex. Simplices are a generalization of triangles to d dimensions, which consists of intervals ($d = 1$), triangles ($d = 2$), or tetrahedra ($d = 3$).

Definition 4.1.3 (Simplex)**.** Let $\{\mathrm{v}_i\}_{0\leq i\leq d}$ be a family of $d+1$ points of $\mathbb{R}^d$ that do not belong to the same hyperplane, the associated d-simplex K is the convex hull of these points. Points $\{\mathrm{v}_i\}$ are called *vertices* of the simplex, and pairs $\epsilon_{ij} = (\mathrm{v}_i, \mathrm{v}_j)$, $i \neq j$, consist of the eddges. The diameter of the simplex is the maximum Euclidean distance between two vertices,

$$\mathrm{diam}(K) = \max_{0\leq i,j\leq d} \|\mathrm{v}_i - \mathrm{v}_j\|$$

The convex hull is the minimum convex subset of $\mathbb{R}^d$ enclosing the points. The condition that all points are not in the same hyperplane means that the convex hull does not degenerate into a lower-dimensional entity; for instance in two dimensions, a triangle degenerates to a segment when points are aligned, and in three dimensions, a tetrahedron degenerates if all points are in the same plane. Therefore a degenerate simplex has a zero d-dimensional Lebesgue measure.

Consider the matrix M_d of $\mathbb{R}^{(d+1)\times(d+1)}$ consisting of column vectors with coordinates of vertices $\{\mathrm{v}_i\}$ of K, and completed by a unit row.

$$M_1 = \begin{bmatrix} x_0 & x_1 \\ 1 & 1 \end{bmatrix} \qquad M_2 = \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix} \qquad M_3 = \begin{bmatrix} x_0 & x_1 & x_2 & x_3 \\ y_0 & y_1 & y_2 & y_3 \\ z_0 & z_1 & z_2 & z_3 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

The determinant of M_d gives the signed d-dimensional measure (volume) of the corresponding simplex K.

$$\det(M_d) = \pm d!|K|$$

In particular, if all the points are contained in the same hyperplane then any vertex is the linear combinations of others so that the determinant is zero. Given that the sign of the determinant depends on permutations of the matrix M_d, in practice the meaning of the sign is the *orientation* of the simplex which depends on how vertices are numbered. If one substract the first column from all the other columns then the link to the Jacobian matrix of the affine mapping T_K introduced ealier is direct: the determinant of J_{T_K} is equal to $d!|K|$ and a negative determinant means that the simplex K is inverted with respect to $\hat{K}$.

4.2 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 degree 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 $\left\{1, x, x^2, \cdots, x^k\right\}$. 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 4.2.1 (Lagrange polynomials – [4] page 21, [6] page 76)**.** The Lagrange polynomial of degree k associated with node ξ_m reads

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

and

$$\sum_{i=0}^{k}\mathcal{L}_i^k(x) = 1$$

Proposition 4.2.2 (Nodal basis)**.**

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

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

Theorem 4.2.3 (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 $\forall\, x \in [a,b]$:*

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

4.3 Construction of the Finite Element space

4.3.1 A nodal element

Let us take $\{\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_N\}$ a family of points of K such that $\sigma_i(p) = p(\boldsymbol{\xi}_i)$, $1 \leq i \leq N$:

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

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

$$p(\boldsymbol{\xi}_j) = \sum_{i=1}^{N}\sigma_i(p)\,\underbrace{\varphi_i(\boldsymbol{\xi}_i)}_{\delta_{ij}} \quad , \, 1 \leq i,j \leq N$$

which reduces to:

$$p(\boldsymbol{\xi}_j) = \sigma_i(p)$$

Remark 4.3.1 (Support of shape functions)**.** The polynomial basis being defined such that $\varphi_j(\boldsymbol{\xi}_i) = \delta_{ij}$ then any shape function φ_i has support on the union of cells containing the node $\boldsymbol{\xi}_i$.

4.3.2 Reference Finite Element

To make the connection between the abstract Definition 3.3.1 and a simple concrete example, the reference element for Lagrange $\mathbb{P}_1$ in one dimension is given by the following triple $(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$.

Definition 4.3.2 ($(\hat{K}, \hat{\mathcal{P}}, \hat{\Sigma})$ for Lagrange $\mathbb{P}_1$ 1D)**.** The Finite Element space Lagrange $\mathbb{P}_1$ 1D approximating

$$V_h = \left\{ v \in \mathrm{C}^0(\bar{\Omega}) \cap \mathrm{H}^1(\Omega) : v\big|_K \in \mathbb{P}_1(K), \forall\, K \in \mathcal{T}_h \right\}$$

is given by

- $\hat{K}$ is the unit interval $[0, 1]$.
- $\hat{\mathcal{P}}$ is the space of linear polynomials $\mathbb{P}_1([0, 1])$ with the basis $(\hat{\varphi}_0, \hat{\varphi}_1)$ by Definition 4.2.1 of $\mathcal{L}_i^1$,

$$\hat{\varphi}_0(\hat{x}) = \mathcal{L}_0^1(\hat{x}) = 1 - \hat{x}, \qquad \hat{\varphi}_1(\hat{x}) = \mathcal{L}_1^1(\hat{x}) = \hat{x}$$

- $\hat{\Sigma}$ is the set of linear forms evaluating the function at Lagrange nodes $\xi_0 = 0$ and $\xi_1 = 1$, $\{\hat{\sigma}_0 : v \mapsto v(\xi_0), \hat{\sigma}_1 : v \mapsto v(\xi_1)\}$.

Consequently the local interpolation operator is

$$\begin{array}{rccl} \mathcal{I}_{K,V} & : & V(K) & \to & V_h(K) \\ & & v & \mapsto & v(\xi_0)\varphi_0 + v(\xi_1)\varphi_1 \end{array}$$

Proposition 4.3.3. $(K, \mathcal{P}, \Sigma)$ *by Definition 4.3.2 is a unisolvent* H^1*-conformal Finite Element.*

Proof. During the lecture we proved that:

- $V_h \subset \mathrm{H}^1(\Omega)$, since piecewise linear and piecewise constant functions belong to $\mathrm{L}^2(\Omega)$.
- (φ_i) is a basis of $\mathbb{P}_1(K)$ since shape functions are linearly independent as they satisfy $\varphi_i(\xi_j) = \delta_{ij}$, and they generate the space V_h as any piecewise linear function coincides with its interpolate.
- (σ_i) is a dual basis of $\mathbb{P}_1(K)$ since $\sigma_i(\varphi_j) = \delta_{ij}$.

□

4.3.3 Lagrange $\mathbb{P}_k$ elements

The extension to Lagrange $\mathbb{P}_k$ is natural as it boils down to construct a basis of $\mathbb{P}_k$ using $(\mathcal{L}_i^k)_{0 \le i \le k}$ and choose the degrees of freedom at the corresponding Lagrange nodes $\{\xi_i\}_{0 \le i \le k}$. Given that the unisolvence for Lagrange $\mathbb{P}_1$ is a direct consequence of the construction from a nodal basis, the same argument applies for higher polynomial order as long as degrees of freedom are located at Lagrange nodes. As illustration Table 4.3.3 depicts the shape functions for $k = 1, 2, 3$ in one dimension with equidistributed nodes (but other distributions are possible and also other polynomials can be used to build nodal elements).

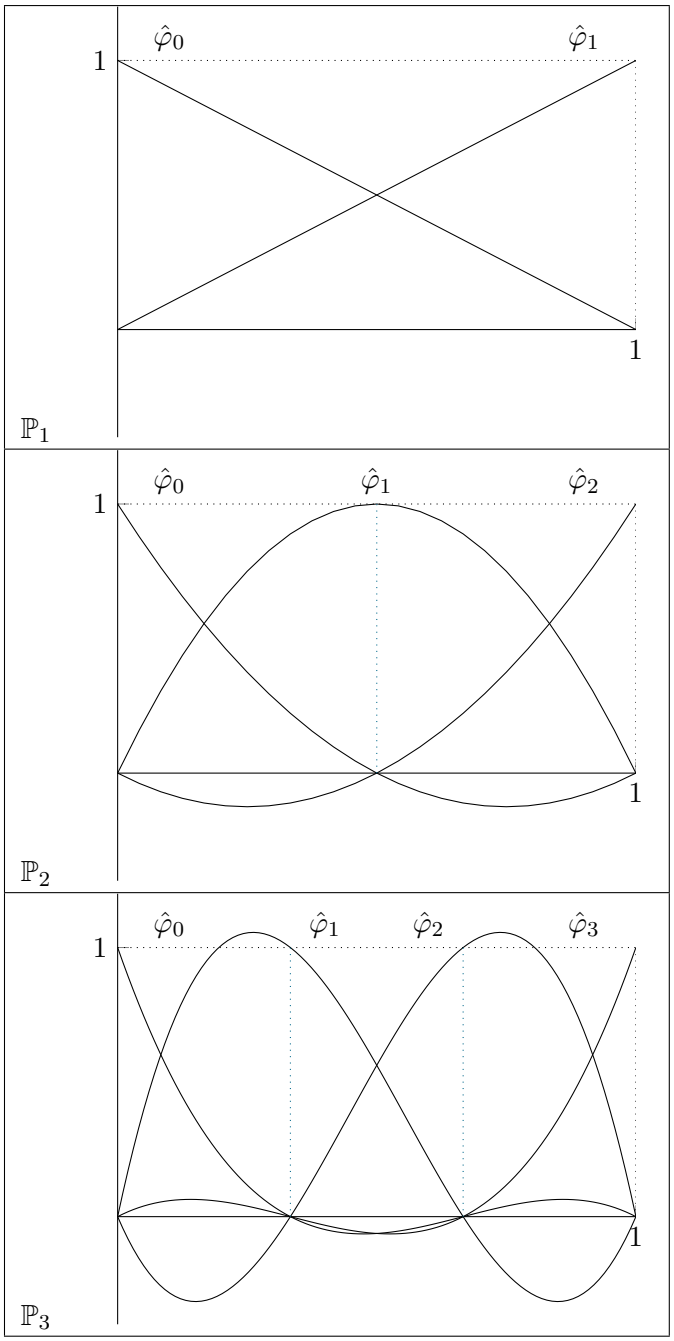


Table 4.1: Shape functions for Lagrange $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3$ on the interval $\hat{K} = [0, 1]$.

4.4 Extension to multiple dimensions

4.4.1 Barycentric coordinates

Lagrange polynomials (4.2.1) construct directly one-dimensional shape functions, while in higher dimensions they can be reformulated in terms of barycen-

tric coordinates. High-order Lagrange basis can also be expressed as polynomials of barycentric coordinates.

Definition 4.4.1 (Barycentric coordinates)**.** Let us consider K a d-simplex with vertices $\{\mathrm{v}_i\}_{0\leq i\leq d}$, any point $\boldsymbol{x}\in K$ satisfies

$$\boldsymbol{x}=\sum_i \lambda_i(\boldsymbol{x})\mathrm{v}_i$$

where barycentric coordinates are obtained by relation

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

with $\mathbf{n}_i$ the unit outward normal to the facet opposite to v_i, and v_f a vertex belonging to this facet.

The geometric interpretation of barycentric coordinates is given by

$$\lambda_i(\boldsymbol{x})=\frac{\mathrm{meas}(K_i)}{|K|}$$

with $\mathrm{meas}(K_i)$ the signed measure of $K_i(\boldsymbol{x})$ the d-simplex constructed with point $\boldsymbol{x}$ and the facet opposite to v_i. In particular, points located within K_i have non-negative λ_i, and the point $\boldsymbol{x}_g$ satisfying equal weight $\lambda_i(\boldsymbol{x}_g)=(d+1)^{-1}$ is the isobarycentre. In practice this property can be used to check if a point is inside a simplex: if the signed measure of one λ_i is negative or if $\sum_i|K_i(\boldsymbol{x})|>|K|$ then the point is outside of the simplex.

Example 4.4.2 (Lagrange $\mathbb{P}_1$ 1D)**.** In one dimension of space, barycentric coordinates on $K=[x_0,x_1]$ are

$$\begin{cases}\lambda_0(x) &= 1-\dfrac{x-x_0}{x_1-x_0}=\dfrac{x_1-x}{x_1-x_0}\\ \lambda_1(x) &= 1-\dfrac{x-x_1}{x_0-x_1}=\dfrac{x-x_0}{x_1-x_0}\end{cases}$$

which is exactly the same expression as linear shape functions φ_0 and φ_1.

Example 4.4.3 (Lagrange $\mathbb{P}_1$ 2D)**.** In two dimensions of space, barycentric coordinates on the unit triangle $\hat{K}$ depicted Figure 4.4.3 are

$$\begin{cases}\lambda_0(\boldsymbol{x}) &= 1-\dfrac{(\hat{x},\hat{y})\cdot(1,1)}{(1,0)\cdot(1,1)} &= 1-\hat{x}-\hat{y}\\ \lambda_1(\boldsymbol{x}) &= 1-\dfrac{(\hat{x}-1,\hat{y})\cdot(-1,0)}{(-1,0)\cdot(-1,0)} &= \hat{x}\\ \lambda_2(\boldsymbol{x}) &= 1-\dfrac{(\hat{x},\hat{y}-1)\cdot(0,-1)}{(0,-1)\cdot(0,-1)} &= \hat{y}\end{cases}$$

which is the linear Lagrange basis on $\hat{K}$ (since the normal is at the numerator and the denominator, there was no need to normalize the vector).

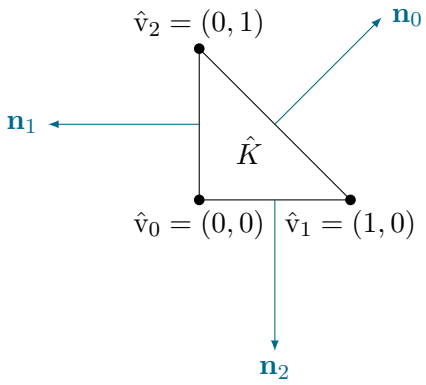


Figure 4.1: Unit triangle with outward facet normals

We can verify easily that shape functions

$$\begin{cases} \hat{\varphi}_0(\hat{\boldsymbol{x}}) &= 1 - \hat{x} - \hat{y} \\ \hat{\varphi}_1(\hat{\boldsymbol{x}}) &= \hat{x} \\ \hat{\varphi}_2(\hat{\boldsymbol{x}}) &= \hat{y} \end{cases}$$

form a nodal basis and that $\hat{\varphi}_0(\hat{\boldsymbol{x}}) + \hat{\varphi}_1(\hat{\boldsymbol{x}}) + \hat{\varphi}_2(\hat{\boldsymbol{x}}) = 1$ for any $\hat{\boldsymbol{x}}$.

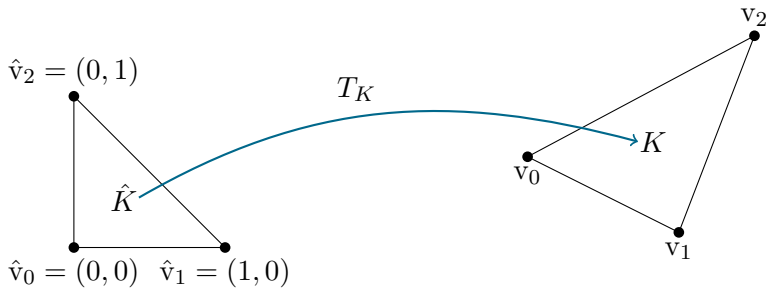
Lagrange elements of polynomial degree $k = 1, 2, 3$ can be expressed using barycentric coordinates in higher dimensions, the shape functions are given by:

$$\begin{aligned}
k = 1, \quad & \lambda_i && , \ 0 \le i \le d \\
k = 2, \quad & \lambda_i \, (2\lambda_i - 1) && , \ 0 \le i \le d \\
& 4 \, \lambda_i \, \lambda_j && , \ 0 \le i < j \le d \\
k = 3, \quad & \frac{1}{2} \, \lambda_i \, (3\lambda_i - 1) \, (3\lambda_i - 2) && , \ 0 \le i \le d \\
& \frac{9}{2} \, \lambda_i \, (3\lambda_i - 1) \, \lambda_j && , \ 0 \le i, j \le d, i \ne j \\
& 27 \, \lambda_i \, \lambda_j \, \lambda_k && , \ 0 \le i < j < k \le d
\end{aligned}$$

4.4.2 Affine transformation

In Chapter 3 the one-dimensional affine mapping between the unit interval and any subinterval $K = [x_i, x_{i+1}]$ of a one-dimensional mesh $\mathcal{T}_h$ was given by Equation (3.1.3); the link to higher dimensions in space was then briefly discussed. The following example describes how the affine mapping $T_K : \hat{K} \to K$ is defined for a triangle in $\mathbb{R}^2$.

The shape of the reference triangle $\hat{K}$ is defined by vectors $\hat{\mathrm{v}}_1 - \hat{\mathrm{v}}_0$ and $\hat{\mathrm{v}}_2 - \hat{\mathrm{v}}_0$, and in the same fashion the shape of any triangle K is defined by vectors $\mathrm{v}_1 - \mathrm{v}_0$



and $\mathrm{v}_2 - \mathrm{v}_0$. The affine mapping is a simple change of coordinates but the detail is given below for the sake of completeness.

$$\left\{ \begin{array}{rcccl} \mathrm{v}_1 & = & T_K(\hat{\mathrm{v}}_1) & = & \mathrm{v}_0 + (\mathrm{v}_1 - \mathrm{v}_0) \\ \mathrm{v}_2 & = & T_K(\hat{\mathrm{v}}_2) & = & \mathrm{v}_0 + (\mathrm{v}_2 - \mathrm{v}_0) \end{array} \right.$$

and any point $\boldsymbol{x}$ of K can be expressed in terms of the relation

$$\boldsymbol{x} = \mathrm{v}_0 + \lambda(\mathrm{v}_1 - \mathrm{v}_0) + \mu(\mathrm{v}_2 - \mathrm{v}_0)$$

with given λ and μ. The reference triangle is defined by the canonical basis of $\mathbb{R}^2$ as $(\hat{\mathrm{v}}_1 - \hat{\mathrm{v}}_0, \hat{\mathrm{v}}_2 - \hat{\mathrm{v}}_0) = (\boldsymbol{e}_x, \boldsymbol{e}_y)$ so that the affine mapping

$$\boldsymbol{x} = \mathrm{v}_0 + \mathrm{B}_K \hat{\boldsymbol{x}}$$

satisfies $T_K(\boldsymbol{e}_x) = (\mathrm{v}_1 - \mathrm{v}_0)$ and $T_K(\boldsymbol{e}_y) = (\mathrm{v}_2 - \mathrm{v}_0)$. The matrix B_K is then the matrix of the corresponding change of basis composed of column vectors $\mathrm{v}_j - \mathrm{v}_0$, thus

$$\boldsymbol{x} = \mathrm{v}_0 + \begin{bmatrix} \mathrm{v}_{1,x} - \mathrm{v}_{0,x} & \mathrm{v}_{2,x} - \mathrm{v}_{0,x} \\ \mathrm{v}_{1,y} - \mathrm{v}_{0,y} & \mathrm{v}_{2,y} - \mathrm{v}_{0,y} \end{bmatrix} \hat{\boldsymbol{x}}$$

Definition 4.4.4 (Affine mapping from reference simplex in $\mathbb{R}^d$)**.** The generalization of the affine mapping in $\mathbb{R}^d$ from the reference simplex $\hat{K} = \{\hat{\mathrm{v}}_i\}_{0 \le i \le d}$ to $K = \{\mathrm{v}_i\}_{0 \le i \le d}$ is given by $\boldsymbol{x} = \mathrm{v}_0 + \mathrm{J}_{T_K} \hat{\boldsymbol{x}}$, with

$$\mathrm{J}_{T_K} = \left[\frac{\partial T_K^i}{\partial x_j} \right]_{ij}$$

given by column vectors $(\mathrm{v}_j - \mathrm{v}_0)$.

While the change of coordinates for the mass matrix does not pose any difficulty, the case of the stiffness matrix requires some precisions. The derivation of the composition of two functions reads

$$\boldsymbol{\nabla} \varphi = \boldsymbol{\nabla}(\hat{\varphi} \circ T_K^{-1}) = (\boldsymbol{\nabla} \hat{\varphi} \circ T_K^{-1}) \cdot \mathrm{J}_{T_K^{-1}}$$

which can be also written formally component by component

$$\frac{\partial \varphi}{\partial x_i} = \sum_j \frac{\partial \varphi}{\partial \hat{x}_j} \frac{\partial \hat{x}_j}{\partial x_i} = \sum_j \frac{\partial \varphi}{\partial \hat{x}_j} \left[\mathrm{J}_{T_K^{-1}} \right]_{ji}$$

and can be interpreted as the decomposition of variation $\mathrm{d}\boldsymbol{x}$ along each axis in terms of $\mathrm{d}\hat{\boldsymbol{x}}$. Moreover the Jacobian matrix of the inverse mapping is the inverse of the Jacobian matrix

$$\mathrm{J}_{T_K^{-1}} = (\mathrm{J}_{T_K} \circ T_K^{-1})^{-1}$$

so that

$$\boldsymbol{\nabla}\varphi = \left[(\mathrm{J}_{T_K} \circ T_K^{-1})^{-1}\right]^T (\boldsymbol{\nabla}\hat{\varphi} \circ T_K^{-1})$$

and since the Jacobian matrix is constant on each cell K, it can be simplified as

$$\boldsymbol{\nabla}\varphi = \left[\mathrm{J}_{T_K}^{-1}\right]^T (\boldsymbol{\nabla}\hat{\varphi} \circ T_K^{-1})$$

4.5 Local equation for Lagrange $\mathbb{P}_1$ in one dimension

The approximation of Problem (1.7) by Lagrange $\mathbb{P}_1$ elements on domain $\Omega = (0,1)$ reads:

$$\left|\begin{array}{l} \text{Find } u \in V_h, \text{ given } f \in \mathrm{L}^2(\Omega), \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla}u \cdot \boldsymbol{\nabla}v \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x \quad , \forall\, v \in V_h \end{array}\right. \tag{4.1a}$$

with the approximation space V_h chosen as:

$$V_h = \left\{v \in \mathrm{C}^0(\bar{\Omega}) \cap \mathrm{H}_0^1(\Omega) : v\big|_K \in \mathbb{P}_1(K), \forall\, K \in \mathcal{T}_h\right\} \tag{4.1b}$$

The interval $\bar{\Omega} = [0,1]$ is discretized by partitioning into disjoints subintervals $[x_n, x_{n+1}]$, $1 \le n \le N_K$ of length $h = 1/N_K$. Steps to obtain a weak formulation and deriving a discrete problem were detailed in Section 3.1.

Expressing the local equation for any subinterval $K = [x_n, x_{n+1}]$ consists of assembling a matrix corresponding to contributions

$$\mathrm{A}_{ij} = \int_K \partial_x\varphi_j(x)\ \partial_x\varphi_i(x)\,\mathrm{d}x$$

for shape functions φ_j and φ_i which have support on K. Given that the dimension of the Lagrange $\mathbb{P}_1$ element in one dimension is two, with two shape functions φ_n and φ_{n+1}, the local matrix is of dimension 2×2. The derivative of φ_n and φ_{n+1} is constant on K and of opposite signs: $\varphi_n(x) = -1$ and $\varphi_{n+1}(x) = +1$.

$$\mathrm{A}_K = \frac{1}{h}\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}$$

with row and column indices of the local matrix mapping to row and column indices $(n, n+1)$ of the global matrix. Therefore assembling the local equation into the global matrix consists of adding entries of A_K to the submatrix with row and column indices $(n, n+1)$. Since each node x_n has two adjacent subintervals $[x_{n-1}, x_n]$ and $[x_n, x_{n+1}]$, inner nodes (which are not on the boundary) will see

two contributions $+1$ on the diagonal, one contribution -1 for columns $n-1$, and one contribution -1 for columns $n+1$, scaled by factor $1/h$.

$$\mathrm{A}_i = \frac{1}{h}\begin{bmatrix} 0 & \cdots & 0 & -1 & \underbrace{+2}_{a_{ii}} & -1 & 0 & \cdots & 0 \end{bmatrix}$$

If the partition of the interval is not uniform then the assembly of the local equation should be modified,

$$\mathrm{A}_K = \frac{1}{|K|}\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}$$

with $|K| = |x_{n+1} - x_n|$.

4.6 Exercises

Exercise 4.6.1.
Let us consider the Helmholtz problem posed on the domain $\Omega = (0, 1)$, given κ a real coefficient:

$$-u''(x) + \kappa u(x) = f(x), \qquad \forall\, x \in \Omega \tag{4.2a}$$

with $f \in \mathrm{L}^2(\Omega)$,

$$u(x) = 0, \qquad \forall\, x \in \partial\Omega \tag{4.2b}$$

Let us consider the analytic solution for $\kappa = 1$ and $f(x) = \sin(\pi x)$ given by $u(x) = \sin(\pi x)/(1 + \pi^2)$.

(a) Derive a weak formulation of Problem (4.2).
(b) Show that the problem solved by a Galerkin method using the discrete space $V_h = \operatorname{span}\{\varphi_i\}$, $1 \leq i \leq N$, can be written under the form of a linear system
$$(\mathrm{A} + \kappa \mathrm{M})\mathbf{u} = \mathbf{b}$$
(c) Express the linear system when the problem is approximated with $V_h = X_h^1$ the space of linear Lagrange finite element on a uniform grid.
(d) Implement a program to solve the problem with $V_h = X_h^1$ and compare the discrete solution with the suggested analytic solution.
(e) Express the linear system when the problem is approximated with $V_h = X_h^2$ the space of quadratic Lagrange finite element on a uniform grid.
(f) Modify the linear system for a general non-uniform grid, and for testing use vertices:
$$\mathcal{V} = \{0.0, 0.1, 0.25, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.8, 0.9, 1.0\}$$
(g) Write a function computing integrals by the quadrature formula:
$$\int_0^1 g(x)\,\mathrm{d}x \approx \frac{1}{2}\big[g(c_1) + g(c_2)\big], \quad c_1 = \frac{1}{2} + \frac{\sqrt{3}}{6}, \quad c_2 = \frac{1}{2} - \frac{\sqrt{3}}{6}$$
to approximate contributions for each element to the load vector $\mathbf{b}$.
(h) Implement a program to solve the problem with $V_h = X_h^2$ and compare the discrete solution with the suggested analytic solution.
(i) Modify the program to implement the boundary condition $u(0) = 1$, $u'(x) = 2$. Which changes are required?

Chapter 5

Error analysis

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

— on the one hand, in the analysis of Ritz and Galerkin methods, consistency estimates like Céa's Lemma give a control on the approximation 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 (4.2.3) gives a control on the interpolation error $e_\pi = u - \pi_k\, u$, *i.e.* the difference between any function and its interpolate by Lagrange polynomials of order k.

For consistency with the notation introduced in Chapter 3 for the Lagrange interpolation operator in the Finite Element setting, ${\mathcal{I}_h}^k$ will stand for the Lagrange $\mathbb{P}_k$ interpolation operator in this section.

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

5.1 Preliminary discussion on the Poisson problem

In the previous chapters, the weak problem derived for the Poisson problem with homogeneous Dirichlet boundary conditions and $f \in \mathrm{L}^2(\Omega)$,

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

was approximated by a Galerkin problem,

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

by simply looking for a finite dimensional solution.

In a first stage, we verified that the weak problem is well-posed for the chosen solution and test spaces, and that the exact solution u and discrete solution u_h satisfy both a relation of the type

$$|u|_{\mathrm{H}^1(\Omega)} \leq C\|f\|_{\mathrm{L}^2(\Omega)}$$

with $C > 0$, which we denoted as *a priori* estimate.

In a second stage, without making any further assumption on the approximation space V_h, a general result was introduced showing that the distance between the exact solution and the approximate solution $\|u - u_h\|_V$ is bounded by the distance of u to the best approximation of u in V_h; this is Céa's Lemma. For the sake of completeness, the estimate corresponding to Céa's Lemma is given in the case of the Poisson problem with homogeneous Dirichlet boundary conditions.

$$E = |u - u_h|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega |\boldsymbol{\nabla}(u - u_h)|^2 \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{\nabla}(u - u_h) \cdot \boldsymbol{\nabla}(u - u_h) \, \mathrm{d}\boldsymbol{x}$$

For any $v_h \in V_h$

$$E = \int_\Omega \boldsymbol{\nabla}(u - u_h) \cdot \boldsymbol{\nabla}(u - v_h) \, \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{\nabla}(u - u_h) \cdot \boldsymbol{\nabla}(v_h - u_h) \, \mathrm{d}\boldsymbol{x}$$

and the second term cancels by virtue of consistency (Galerkin orthogonality), so that Cauchy–Schwarz gives

$$E = |u - u_h|^2_{\mathrm{H}^1(\Omega)} \leq \left(\int_\Omega |\boldsymbol{\nabla}(u - u_h)|^2 \, \mathrm{d}\boldsymbol{x} \right)^{1/2} \left(\int_\Omega |\boldsymbol{\nabla}(u - v_h)|^2 \, \mathrm{d}\boldsymbol{x} \right)^{1/2}$$

thus

$$|u - u_h|_{\mathrm{H}^1(\Omega)} \leq |u - v_h|_{\mathrm{H}^1(\Omega)}$$

As introduced, the relation

$$|e_h|_{\mathrm{H}^1(\Omega)} \leq |u - v_h|_{\mathrm{H}^1(\Omega)} \tag{5.1}$$

means that the *approximation error* $e_h = u - u_h$ is controlled as soon as we are able to bound the distance between u and its best representant in V_h.

In a third stage, we introduced Finite Element approximation spaces based on Lagrange polynomials, Definition 4.2.1. Their local interpolation operator given here in one dimension,

$$\mathcal{I}_K^k \, v = \sum_{j=0}^{k} v(\xi_j) \mathcal{L}_j^k$$

with Lagrange nodes $\{\xi_j\}$, satisfies Inequality (4.2.3) controlling the *interpolation error* $e_\imath = v - \mathcal{I}_K^k \, v$ pointwise on K.

Inequality (5.1) can be written

$$|e_h|_{\mathrm{H}^1(\Omega)} \leq |e_\imath|_{\mathrm{H}^1(\Omega)} \tag{5.2}$$

which means that the *approximation error* is bounded by the *interpolation error*. From pointwise polynomial interpolation estimates we need to derive inequalities in Sobolev and Lebesgue norms, so that the right-hand side of Equation (5.2) can be bounded by a function ϵ of the mesh size $h_\mathcal{T}$, such that $\epsilon(h_\mathcal{T}) \to 0$ as $h_\mathcal{T} \to 0$; convergence of the approximation is then ensured.

Inequalities of the form

$$\|e_h\| \leq C\epsilon(h_\mathcal{T})$$

for some constant C depending on the domain and the exact solution, are called *a priori error estimates*. Usually $\epsilon(h_\mathcal{T}) \sim O(h_\mathcal{T}^r)$, with $r > 0$ the convergence rate of the method, which represents the expected *accuracy*. In this case sequences of approximate solution $(u_h)_{h_\mathcal{T}}$ converge to the exact solution u, as fast as the function ϵ allows, when $h_\mathcal{T}$ tends to zero.

The goal of this section is to derive such error estimates after proving required interpolation inequalities. In particular the space of continuous linear Lagrange elements

$$V_h = \{v \in \mathrm{C}^0(\bar{\Omega}) \cap \mathrm{H}_0^1(\Omega) : \forall\, K \in \mathcal{T}_h, v\big|_K \in \mathbb{P}_1(K)\} \tag{5.3}$$

will be considered, then more general results will be stated without proof.

5.2 Stability of the Lagrange interpolation operator

Before moving to interpolation inequalities, the *stability* of the interpolation operator should be proved: such estimate shows that the interpolate of any function in V is also in V. The exposé is restricted to the one-dimensional case and detailed so that anybody without prior experience in estimates should be able to follow the procedure.

Interpolation properties need to be reformulated in terms of estimates in L^2 norms. First they are expressed elementwise, then on the entire domain by collecting contributions over the mesh. To give a better idea, let us introduce below an ingredient used for subsequent estimates, considering a function $v \in \mathrm{H}^1(I)$ with I an interval, and two points $\xi, x \in I$.

$$\begin{aligned} |v(x) - v(\xi)| &= \left| \int_\xi^x v'(s)\,\mathrm{d}s \right| \\ &\leq \int_\xi^x \left|v'(s)\right| \,\mathrm{d}s \\ &\leq |x-\xi|^{1/2} \left(\int_\xi^x \left|v'(s)\right|^2 \,\mathrm{d}s \right)^{1/2} \end{aligned}$$

using Cauchy–Schwarz with $|v'(s)|$ and $\mathbb{1}_{[\xi,x]}$, the indicator function on $[\xi, x]$. Since we assumed that $v \in \mathrm{H}^1(I)$, then the right-hand side of

$$|v(x) - v(\xi)| \leq |x-\xi|^{1/2} |v|_{\mathrm{H}^1([\xi,x])} \tag{5.4}$$

is bounded.

Now let us consider that ξ realizes the minimum of $|v|$ on I, then

$$|v(x)| \leq |x-\xi|^{1/2}|v|_{\mathrm{H}^1([\xi,x])} + |v(\xi)|\,, \quad \xi = \operatorname*{argmin}_{s\in I} |v(s)|$$

using the second triangle inequality.

Immediately if $|v(\xi)| = 0$ the estimate gives for any $x \in I$

$$|v(x)| \leq |I|^{1/2}|v|_{\mathrm{H}^1(I)} \tag{5.5}$$

but otherwise we can bound $|v(\xi)|$ using

$$|v(\xi)| = |x-\xi|^{-1}\int_\xi^x |v(s)|\ \mathrm{d}s \leq |x-\xi|^{-1/2}\|v\|_{\mathrm{L}^2([\xi,x])}$$

so that

$$|v(x)| \leq |x-\xi|^{1/2}|v|_{\mathrm{H}^1([\xi,x])} + |x-\xi|^{-1/2}\|v\|_{\mathrm{L}^2([\xi,x])}$$

thus for any $x \in I$

$$|v(x)| \leq |I|^{1/2}|v|_{\mathrm{H}^1(I)} + |I|^{-1/2}\|v\|_{\mathrm{L}^2(I)} \tag{5.6}$$

in other words

$$\|v\|_{\mathrm{L}^\infty(I)} \leq |I|^{1/2}|v|_{\mathrm{H}^1(I)} + |I|^{-1/2}\|v\|_{\mathrm{L}^2(I)} \tag{5.7}$$

Proposition 5.2.1 (H^1-stability of the Lagrange $\mathbb{P}_1$ interpolation operator)**.** *There exists a constant $C > 0$ such that,*

$$\|\mathcal{I}_h^1 v\|_{\mathrm{H}^1(\Omega)} \leq C\ \|v\|_{\mathrm{H}^1(\Omega)} \tag{5.8}$$

Proof. Since $\|\cdot\|^2_{\mathrm{H}^1} = \|\cdot\|^2_{\mathrm{L}^2} + |\cdot|^2_{\mathrm{H}^1}$ then the L^2 norm of the interpolate and its derivative should be controlled. Since controlling the derivative of a function in L^2 gives a control on the function in L^2 (Poincaré Inequality), it is natural to start looking for an estimate of $|\cdot|_{\mathrm{H}^1}$, then move to $\|\cdot\|_{\mathrm{L}^2}$.

(i) Estimate in $|\cdot|_{\mathrm{H}^1}$:
Let us consider the restriction of the interpolation operator to any $K = [x_i, x_{i+1}] \in \mathcal{T}_h$

$$\big(\mathcal{I}_h^1 v\big)'\Big|_K = \big(\mathcal{I}_K^1 v\big)' = h_K^{-1}\big(v(x_{i+1}) - v(x_i)\big)$$

which is constant over K. The elementwise H^1 semi-norm of the interpolate can be bounded using Inequality (5.4)

$$\begin{aligned}
|\mathcal{I}_K^1 v|_{\mathrm{H}^1(K)} &= \left(\int_K h_K^{-2}\,|v(x_{i+1}) - v(x_i)|^2\ \mathrm{d}\boldsymbol{x}\right)^{1/2} \\
&\leq h_K^{-1/2}\,|v(x_{i+1}) - v(x_i)| \\
&\leq h_K^{-1/2} h_K^{1/2}|v|_{\mathrm{H}^1(K)} \\
&\leq |v|_{\mathrm{H}^1(K)}
\end{aligned}$$

Summing over K and using the definition of $h_{\mathcal{T}}$,

$$|\mathcal{I}_h^1 v|_{\mathrm{H}^1(\Omega)} \leq |v|_{\mathrm{H}^1(\Omega)}$$

(ii) Estimate in $\|\cdot\|_{\mathrm{L}^2}$:
In this case we do not need to prove an elementwise estimate first as we already introduced a control of pointwise values in terms of the H^1 seminorm and the L^2 norm: it boils down to estimate the maximum attained by the linear interpolate.

$$\begin{aligned}\|\mathcal{I}_h^1 v\|_{\mathrm{L}^2(\Omega)} &= \left(\int_\Omega \left|\mathcal{I}_h^1 v\right|^2 \,\mathrm{d}\boldsymbol{x}\right)^{1/2} \\ &\leq \left(\int_\Omega \|\mathcal{I}_h^1 v\|^2_{\mathrm{L}^\infty(\Omega)} \,\mathrm{d}\boldsymbol{x}\right)^{1/2} \\ &\leq |\Omega|^{1/2}\|\mathcal{I}_h^1 v\|_{\mathrm{L}^\infty(\Omega)} \\ &\leq |\Omega|^{1/2}\|v\|_{\mathrm{L}^\infty(\Omega)}\end{aligned}$$

the last line is given by the L^∞ stability of the linear interpolation: the function and its linear interpolate coincide at Lagrange nodes so that $\|\mathcal{I}_h^1 v\|_{\mathrm{L}^\infty(\Omega)} \leq \|v\|_{\mathrm{L}^\infty(\Omega)}$. Therefore, using Inequality (5.7)

$$\|\mathcal{I}_h^1 v\|_{\mathrm{L}^2(\Omega)} \leq |\Omega|^{1/2}|v|_{\mathrm{H}^1(\Omega)} + |\Omega|^{-1/2}\|v\|_{\mathrm{L}^2(\Omega)}$$

we can conclude using the Poincaré Inequality with constant c_p,

$$\|\mathcal{I}_h^1 v\|_{\mathrm{L}^2(\Omega)} \leq (|\Omega| + c_p^{-1})|v|_{\mathrm{H}^1(\Omega)}$$

□

Inequality (5.8) is also called *uniform continuity* in zero of the interpolation operator since C should not depend on $\mathcal{T}_h$; but possibly depends on Ω.

5.3 *A priori* error estimate with Lagrange $\mathbb{P}_1$

Proposition 5.3.1 (Interpolation Inequalities for Lagrange $\mathbb{P}_1$)**.** *There exists two positive constants C_1 and C_0 such that for any $v \in \mathrm{H}^2(\Omega)$*

$$|v - \mathcal{I}_h^1 v|_{\mathrm{H}^1(\Omega)} \leq C_1 h_{\mathcal{T}}\, |v|_{\mathrm{H}^2(\Omega)} \tag{5.9}$$

$$\|v - \mathcal{I}_h^1 v\|_{\mathrm{L}^2(\Omega)} \leq C_0 h_{\mathcal{T}}^2 |v|_{\mathrm{H}^2(\Omega)} \tag{5.10}$$

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

Proof. The proof is sketched in one dimension, based on a decomposition of the error per element and on the Mean-Value Theorem (also known as Rolle Theorem). The global interpolation error is then recovered by summing over the cells, given that

$$\|e_\imath\|^2_{\mathrm{L}^2(\Omega)} = \sum_{K\in\mathcal{T}_h}\int_K |e_\imath(x)|^2 \,\mathrm{d}\boldsymbol{x} = \sum_{K\in\mathcal{T}_h}\|e_\imath\|^2_{\mathrm{L}^2(K)}$$

This makes sense since the polynomial interpolation estimate is defined pointwise, it is then a local property. In the same spirit as the stability of the

interpolation operator of Proposition 5.2.1 which we proved first elementwise, the estimate is derived in H^1 semi-norm then in L^2 norm.

The main technical ingredient is given by Inequality (5.4) for a given function $w \in \mathrm{H}^1(K)$,

$$|w(x) - w(\xi)| \leq |x - \xi|^{1/2} |w|_{\mathrm{H}^1([\xi,x])} \tag{5.11}$$

with $\xi, x \in K$. If we choose ξ such that $w(\xi)$ cancels and if we integrate the square of the expression over any $K \in \mathcal{T}_h$, then we get a control of w in $\mathrm{L}^2(K)$,

$$\|w\|_{\mathrm{L}^2(K)} \leq |x - \xi|^{1/2} \left(\int_K |w|^2_{\mathrm{H}^1([\xi,x])} \, \mathrm{d}\boldsymbol{x} \right)^{1/2} \leq h_K |w|_{\mathrm{H}^1(K)} \tag{5.12}$$

(i) Estimate in $|\cdot|_{\mathrm{H}^1}$:
The elementwise estimate on any $K = [x_i, x_{i+1}]$ is obtained by taking $w = (v - \mathcal{I}_h^1 v)'$ in Inequality (5.11). Given that x_i and x_{i+1} are Lagrange nodes, $\big(v - \mathcal{I}_h^1 v\big)(x_i) = 0$ and $\big(v - \mathcal{I}_h^1 v\big)(x_{i+1}) = 0$, then by virtue of the Mean-Value Theorem, there exists $\xi \in [x_i, x_{i+1}]$ such that the derivative cancels, $(v - \mathcal{I}_h^1 v)'(\xi) = 0$. Moreover $|(v - \mathcal{I}_h^1 v)'|_{\mathrm{H}^1(K)} = |v'|_{\mathrm{H}^1(K)}$ since $(v - \mathcal{I}_h^1 v)'' = v'' - (\mathcal{I}_h^1 v)''$ by linearity and $(\mathcal{I}_h^1 v)''$ is identically zero, so we get directly

$$\|(v - \mathcal{I}_h^1 v)'\|_{\mathrm{L}^2(K)} \leq h_K |v'|_{\mathrm{H}^1(K)}$$

in the same fashion as Inequality (5.12), which can be rewritten under the expected form

$$|v - \mathcal{I}_h^1 v|_{\mathrm{H}^1(K)} \leq h_K |v|_{\mathrm{H}^2(K)} \tag{5.13}$$

(ii) Estimate in $\|\cdot\|_{\mathrm{L}^2}$:
The elementwise estimate on any $K = [x_i, x_{i+1}]$ is obtained by taking $w = (v - \mathcal{I}_h^1 v)$ in Inequality (5.11). Given that $\big(v - \mathcal{I}_h^1 v\big)(x_i) = 0$ and $\big(v - \mathcal{I}_h^1 v\big)(x_{i+1}) = 0$, a similar argument as for the H^1 semi-norm can be used with $\xi = x_i$. Inequality (5.12) with $w = (v - \mathcal{I}_h^1 v)$ reads

$$\|v - \mathcal{I}_h^1 v\|_{\mathrm{L}^2(K)} \leq h_K |v - \mathcal{I}_h^1 v|_{\mathrm{H}^1(K)}$$

so that using Inequality (5.13)

$$\|v - \mathcal{I}_h^1 v\|_{\mathrm{L}^2(K)} \leq h_K^2 |v|_{\mathrm{H}^2(K)}$$

In both cases global estimates are obtained by summing over $K \in \mathcal{T}_h$ and factoring $h_{\mathcal{T}} = \max_{K \in \mathcal{T}_h}(h_K)$. □

Remark 5.3.2 (Convergence order in H^1 norm)**.** Using the definition of the norm

$$\|v - \mathcal{I}_h^1 v\|^2_{\mathrm{H}^1(\Omega)} = \|v - \mathcal{I}_h^1 v\|^2_{\mathrm{L}^2(\Omega)} + |v - \mathcal{I}_h^1 v|^2_{\mathrm{H}^1(\Omega)}$$

we get

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

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

Thus we verify that the approximation is first order in H^1 norm if $|v|_{\mathrm{H}^2(\Omega)}$ is bounded.

A more general result Proposition 5.3.3 can be proved for Lagrange $\mathbb{P}_k$ Finite Elements, which we verify is equivalent to Proposition 5.3.1 for $s = 1$, $k = 1$.

Proposition 5.3.3 (Interpolation Inequality for Lagrange $\mathbb{P}_k$, [4])**.** *Given $0 \leq s \leq k$, there exists a positive constant C such that for any $v \in \mathrm{H}^{s+1}(\Omega)$,*

$$\|v - \mathcal{I}_h^1 v\|_{\mathrm{L}^2(\Omega)} + h_{\mathcal{T}}|v - \mathcal{I}_h^1 v|_{\mathrm{H}^1(\Omega)} \leq C h_{\mathcal{T}}^{s+1}|v|_{\mathrm{H}^{s+1}(\Omega)}$$

One important remark is that the convergence order depends on the regularity of the solution since the order is $s + 1$ for a solution in $\mathrm{H}^{s+1}(\Omega)$. Unfortunately the Lagrange $\mathbb{P}_1$ Finite Element is only H^1–conformal, so *a priori* it cannot represent functions of H^2 accurately. Given that the solution space is H^1 the interpolation inequality of Proposition 5.3.3 only applies with $s = 0$ so that the convergence order is only one in L^2 norm in the general case, and we have only a weak convergence result in H^1. In the next section we show that the convergence rate for the Poisson problem approximated with Lagrange $\mathbb{P}_1$ is actually second order in L^2 norm and first order in H^1 semi-norm: this is one order more than expected from interpolation inequalities.

5.4 Superconvergence

The following result shows the the convergence properties of the method is not only limited by interpolation inequalities. Indeed, using a result by Aubin and Nitsche, we show that even if the approximation is not H^2-conformal, we can improve the error estimate by one order: the convergence in L^2 becomes then second order in $h_{\mathcal{T}}$. The idea behind this result is that if u is the weak solution to the Poisson equation then it is not only in $\mathrm{H}_0^1(\Omega)$ since the differential operator involves second order derivatives: we say that u is regularized due to the ellipticity of the operator.

Theorem 5.4.1 (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* (1.3) *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 $u \in \mathrm{H}_0^1(\Omega)$ is solution to the Poisson problem, then by elliptic regularity and density of $\mathrm{H}^2(\Omega)$ in $\mathrm{H}^1(\Omega)$), $u \in \mathrm{H}^2(\Omega)$, thus $\exists\, C_u > 0$ such that:

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

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

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq C_u\, h_{\mathcal{T}}\, \|f\|_{\mathrm{L}^2(\Omega)} \tag{5.14}$$

Let us introduce the following auxiliary problem:

$$-\boldsymbol{\Delta}\varphi(\boldsymbol{x}) = e_h(\boldsymbol{x}) \quad , \boldsymbol{x} \in \Omega \tag{5.15a}$$

$$\varphi(\boldsymbol{x}) = 0 \quad , \boldsymbol{x} \in \partial\Omega \tag{5.15b}$$

obtained by a duality argument, formally

$$-(\ \boldsymbol{\Delta} v\ ,\ w\ ) = (\ \nabla v\ ,\ \nabla w\ ) = -(\ v\ ,\ \boldsymbol{\Delta} w\ )$$

by integration by parts; since the adjoint operator of the Laplace operator is the Laplace operator itself, it is said *self-adjoint*. The motivation of introducing this equation is to control the *approximation error* e_h in L^2:

$$(\ e_h\ ,\ e_h\ ) = (\ e_h\ ,\ -\boldsymbol{\Delta}\phi\ ) = (\ \nabla e_h\ ,\ \nabla\phi\ ) = (\ -\boldsymbol{\Delta} e_h\ ,\ \phi\ )$$

for $\phi \in \mathrm{H}_0^1(\Omega)$ called *dual solution* satisfying $-\boldsymbol{\Delta}\phi = e_h$.

Similarly to the Poisson equation, the weak formulation of Problem 5.15 reads:

$$\left|\begin{array}{l} \text{Find } \varphi \in \mathrm{H}_0^1(\Omega), \text{ given } e_h \in \mathrm{L}^2(\Omega), \text{ such that:} \\ \displaystyle\int_\Omega \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\phi \,\mathrm{d}\boldsymbol{x} = \int_\Omega e_h \phi \,\mathrm{d}\boldsymbol{x} \quad ,\ \forall\ \phi \in \mathrm{H}_0^1(\Omega) \end{array}\right. \tag{5.16}$$

Since e_h is bounded in $\mathrm{L}^2(\Omega)$ then the same regularity result holds for the auxiliary Problem (5.15), $\exists C_\varphi > 0$ such that:

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

so we have from the interpolation inequality for φ

$$\|\varphi - \varphi_h\|_{\mathrm{H}^1(\Omega)} \leq C_\varphi\ h_{\mathcal{T}}\ \|e_h\|_{\mathrm{L}^2(\Omega)} \tag{5.17}$$

Let us try to bound the L^2 norm of the approximation error by noticing that its amounts to take $\phi = e_h$ in (5.16):

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

If we consider the approximate of Problem (5.16) by Galerkin 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 subtract 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 approximation errors appear since we control them by Equation (5.14) and (5.17):

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

Replacing by the bounds from the interpolation inequalities we get:

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_u\ C_\varphi\ h_{\mathcal{T}}^2\ \|f\|_{\mathrm{L}^2(\Omega)}$$

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

The conclusion of this result is that the observed convergence order may be different than the order suggested by the interpolation inequality. As seen in this example it may be improved if the differential operator has a regularizing effect, but can also be influenced by other factors like the regularity of the boundary or the discretization of the computational domain.

5.5 Exercises

Chapter 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 2.2.11.

6.1 Time marching schemes

In this section, problems describing the evolution in time of an unknown u will be considered. Since the unknown depends on the space coordinates and on the time, such evolution problem will be posed on a domain which is the open cylinder $\mathrm{Q} = \Omega \times (0, T)$.

The Initial and Boundary Value problem for a Partial Differential Equation will therefore take a form such as

$$\left| \begin{array}{ll} \text{Find } u(\boldsymbol{x}, t) \text{ satisfying:} & \\ & \\ \partial_t u(\boldsymbol{x}, t) + \mathrm{A}\ u(\boldsymbol{x}, t) = f(\boldsymbol{x}, t) & , \forall\ (\boldsymbol{x}, t) \in \mathrm{Q} \\ u(\boldsymbol{x}, t) = g(\boldsymbol{x}, t) & , \forall\ \boldsymbol{x} \in \partial\Omega, t \in (0, T) \\ u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) & , \forall\ \boldsymbol{x} \in \partial\Omega \end{array} \right.$$

in the case of a Dirichlet problem which is first-order in time, and with A a differential operator in space. The differential operator and the right-hand side may depend on u, in which case the problem becomes non-linear.

The equation can be recast under the form of a Cauchy problem

$$\dot{u}(\boldsymbol{x}, t) = F(\boldsymbol{x}, t; u) \tag{6.1}$$

so that an evolution problem can be seen as the coupling between a partial differential equation in space, and an ordinary differential equation in time.

This suggests that two discretizations may be considered: a Finite Element discretization in space which was discussed for elliptic problems in Chapter 2 and Chapter 3, and a time discretization. Similarly to the one-dimensional spatial case, the time discretization consists of solving the problem on a partition of $(0, T)$. Let us define a family $\{t^n\}_{0 \leq n \leq N}$ of $N + 1$ discrete times, with $t^0 < \cdots < t^N$, and integrate the equation on each subinterval $[t^{n-1}, t^n]$, $1 \leq n \leq N$,

characterized by the time-step $\delta t^n = t^n - t^{n-1}$. Marching in time consists of solving a succession of problems at discrete times t^n, $1 \leq n \leq N$ given solutions at previous discrete times.

For example, in the case of a first order approximation in time, the discrete time-derivative is

$$\partial_{t,n}\, u = \frac{u^n - u^{n-1}}{\delta t^n} \tag{6.2}$$

for $n = 1, \ldots, N$, so that Relation (6.1) reads

$$u^n = u^{n-1} + \delta t^n F(\boldsymbol{x}, t; u) \tag{6.3}$$

and the treatment of term $F(\boldsymbol{x}, t; u)$ is left to be determined as it can be expressed as a function of u^n but also of solutions u^{n-k}, $k = 1, \ldots, n$ at previous time-steps.

The choice of the time-derivative and the way $F(\boldsymbol{x}, t; u)$ is expressed will define the type of numerical scheme. For example, in Relation (6.3) $F(\boldsymbol{x}, t; u)$ can be evaluated at time t^n,

$$u^n = u^{n-1} + \delta t^n F^n(\boldsymbol{x}, t; u) \tag{6.4}$$

or at time t^{n-1},

$$u^n = u^{n-1} + \delta t^n F^{n-1}(\boldsymbol{x}, t; u) \tag{6.5}$$

which correspond respectively to Backward Euler and Forward Euler schemes. The former is an *implicit scheme* as the term $F(\boldsymbol{x}, t; u)$ depends on the unknown u^n, while the latter is an *explicit scheme* as the term $F(\boldsymbol{x}, t; u)$ is expressed in terms of u^{n-1} which is known. In a more general fashion, the theta-scheme reads,

$$u^n = u^{n-1} + \delta t^n \big[\theta F^n(\boldsymbol{x}, t; u) + (1-\theta) F^{n-1}(\boldsymbol{x}, t; u)\big] \tag{6.6}$$

with parameter $\theta \in [0, 1]$, so that Backward Euler is recovered for $\theta = 1$, Forward Euler is recovered for $\theta = 0$, and Crank–Nicolson corresponds to the choice $\theta = 1/2$.

Stability and accuracy properties of the numerical scheme will depend on which terms are chosen as implicit or explicit: without going into the details and as a general rule implicit schemes tend to be more stable while explicit schemes will be limited by a condition on the time-step. You can refer Von Neumann stability analysis for ordinary differential equations, and the Courant–Friedrichs–Levi (CFL) condition.

Regardless of the numerical scheme, properties of solutions will also depend on the regularity of the initial condition and the nature of the differential operator. Parabolic equations involving an elliptic operator will enjoy a smoothing property, given that energy dissipation is induce by diffusion-type operators, while hyperbolic equations may give rise to discontinuities and see the propagation of shocks.

6.2 *A priori* stability estimate

In a similar fashion as elliptic problems, *a priori* estimates can be derived by a careful choice of test function.

6.2.1 Heat equation

In the continuity of the Poisson problem the following unsteady problem is considered

$$\left| \begin{array}{l} \partial_t u(\boldsymbol{x},t) - \Delta u(\boldsymbol{x},t) = f(\boldsymbol{x},t) \\ u(\boldsymbol{x},t) = 0 \\ u(\boldsymbol{x},0) = u_0(\boldsymbol{x}) \end{array} \right.$$

which corresponds to the case $\mathrm{A}\ u = -\Delta\ u$.

$$(\ \mathrm{A}\ u\ ,\ v\ )_= - \int_\Omega \Delta\ uv \,\mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x}$$

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:

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

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

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa |u|^2_{\mathrm{H}^1(\Omega)} = 0 \tag{6.7}$$

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 $[0, T]$:

$$\frac{1}{2}\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa \int_0^T |u|^2_{\mathrm{H}^1(\Omega)} \,\mathrm{d}t = 0 \tag{6.8}$$

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

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa |u|^2_{\mathrm{H}^1(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)} \|u\|_{\mathrm{L}^2(\Omega)} \tag{6.9}$$

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.1.8) and then the Young inequality (Lemma D.1.3), we can bound the right-hand side by the data and the H^1 seminorm,

$$\|f\|_{\mathrm{L}^2(\Omega)}\|u\|_{\mathrm{L}^2(\Omega)} \leq \frac{1}{2\gamma^2 c_P^2}\|f\|^2_{\mathrm{L}^2(\Omega)} + \frac{\gamma^2}{2}|u|^2_{\mathrm{H}^1(\Omega)} \tag{6.10}$$

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 subtract the second term of (6.10) to the left-hand side of the estimate, given that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{\mathrm{L}^2(\Omega)} + \frac{2\kappa - \gamma^2}{2}|u|^2_{\mathrm{H}^1(\Omega)} \leq \frac{1}{2\gamma^2 c_P^2}\|f\|^2_{\mathrm{L}^2(\Omega)} \tag{6.11}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2_{\mathrm{L}^2(\Omega)} + \kappa|u|^2_{\mathrm{H}^1(\Omega)} \leq C(c_P)\|f\|^2_{\mathrm{L}^2(\Omega)} \tag{6.12}$$

This inequality yields a control of the L^2 norm and H^1 seminorm of the solution at any time t of the time interval $[0, T]$. Similarly to Equation (6.8), if we integrate over the time, we get

$$\|u(T)\|^2_{\mathrm{L}^2(\Omega)} - \|u(0)\|^2_{\mathrm{L}^2(\Omega)} + \kappa\int_0^T |u|^2_{\mathrm{H}^1(\Omega)}\,\mathrm{d}t \leq C(c_P)\int_0^T \|f\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}t$$

which, by defining,

$$\|v\|_{\mathrm{L}^r(0,T;\mathrm{L}^p(\Omega))} = \left(\int_0^T \|v\|^r_{\mathrm{L}^p(\Omega)}\,\mathrm{d}t\right)^{1/r} \tag{6.13}$$

can be rewritten as

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

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.

H^2

Take $v = -t\Delta u$.

$$-\int_\Omega \big(\partial_t u - \Delta u\big)\, t\, \Delta u \,\mathrm{d}\boldsymbol{x}$$

$$\int_\Omega t\boldsymbol{\nabla}(\partial_t u)\cdot\boldsymbol{\nabla} u \,\mathrm{d}\boldsymbol{x} + \int_\Omega t|\Delta u|^2 \,\mathrm{d}\boldsymbol{x}$$

$$t\int_\Omega \partial_t(\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} u \,\mathrm{d}\boldsymbol{x} + t\int_\Omega |\Delta u|^2 \,\mathrm{d}\boldsymbol{x}$$

$$\frac{t}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u|^2_{\mathrm{H}^1(\Omega)} + |u|^2_{\mathrm{H}^2(\Omega)}$$

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\big(t|u|^2_{\mathrm{H}^1(\Omega)}\big) - \frac{1}{2}|u|^2_{\mathrm{H}^1(\Omega)} + t|u|^2_{\mathrm{H}^2(\Omega)}$$

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

$$+t|u|^2_{\mathrm{H}^2(\Omega)} = +\frac{1}{2}\int_0^T |u|^2_{\mathrm{H}^1(\Omega)} - \frac{T}{2}|u(T)|^2_{\mathrm{H}^1(\Omega)}$$

Using the previous control in H^1 allows us to conclude.

Chapter 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 an approximate solution computed on a given mesh to improve the accuracy?

7.1 *A posteriori* estimates

In Chapter 5 interpolation inequalities were established, such as

$$\begin{aligned} |u - \mathcal{I}_h u|_{\mathrm{H}^1(\Omega)} &\leq C_1 h_{\mathcal{T}} |u|_{\mathrm{H}^2(\Omega)} \\ \|u - \mathcal{I}_h u\|_{\mathrm{L}^2(\Omega)} &\leq C_0 h_{\mathcal{T}}^2 |u|_{\mathrm{H}^2(\Omega)} \end{aligned}$$

in order to bound the right-and side of Céa's inequality

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq \|u - v\|_{\mathrm{H}^1(\Omega)}$$

for $v \in V_h$. As such this provides directly a control on the H^1 norm of the approximation error $e_h = u - u_h$: the obtained estimation is called *a priori* error estimate. This type of result is important to indicate the optimal convergence order of sequence of discrete solutions to the exact solution as the mesh size tends to zero (which means that the dimension of the solution space tends to infinity). Since *a priori* error estimates depend on the exact solution to the problem they does not provide a quality measure of the discrete solution which is *computable*. Instead we would like to derive an error estimate which would hep us determine if the computed discrete solution is accurate. This type of estimation is called *a posteriori* error estimate: it consists of computing *error indicators* depending on the discrete solution u_h and the discretization, providing a quality measure of u_h.

As an introduction, let us consider the exact solution $u \in V$ and its approximation $u_h \in V_h$ and observe that there exists a direct relation between the *approximation error* $e_h = u - u_h \in V$ and the residual of the equation $\mathcal{R}(u_h) \in V'$,

$$(\mathcal{R}(u_h) , v) = L(v) - a(u_h, v)$$

for $v \in V$ so that by consistency of the Galerkin method

$$(\mathcal{R}(u_h) , v) = a(u - u_h, v)$$

and in the case $v \in V_h$ the inner-product cancels.

In case $a(\cdot, \cdot)$ is a coercive continuous bilinear form we can write:

1. Coercivity:

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) \tag{7.1}$$

$$a(u - u_h, u - u_h) = (\mathcal{R}(u_h) , u - u_h)$$

$$\alpha \|e_h\|_V \leq \|\mathcal{R}(u_h)\|_{V'} \tag{7.2}$$

2. Continuity:

$$a(u - u_h, u - u_h) \leq M \|u - u_h\|_V \tag{7.3}$$

$$\|\mathcal{R}(u_h)\|_{V'} = \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u - u_h, v)}{\|v\|_V}$$

$$\|\mathcal{R}(u_h)\|_{V'} \leq M \|e_h\|_V \tag{7.4}$$

Using only the coercivity (7.1) and the continuity (7.3) of the bilinear form, it follows from relations (7.2) and (7.4) that estimating $\mathcal{R}(u_h)$ is equivalent to estimating the *approximation error* in the norm of V. Given that u_h is know the quantity $\mathcal{R}(u_h)$ is computable and can be used to derive *a posteriori* error estimators.

Different strategies can be used for improving the accuracy:

- h-adaptivity: cells with largest error indicators are refined, *i.e.* divided into smaller cells so that the mesh size is decreased: the mesh topology is changed as new cells are inserted.

- p-adaptivity: polynomial order is increased on the cell so that the exponent becomes larger in the error estimate.

- r-adaptivity: points are moved to get smaller cells where error indicators are large: the mesh topology is left unchanged as the transformation is only geometric.

7.2 Residual-based error estimator for Poisson

Let u and u_h be respectively the solutions to Problem (1.7) and the approximate Problem (4.1) by a Lagrange $\mathbb{P}_1$ discretization. Therefore the ideas of the previous section are applied with $V = \mathrm{H}^1_0(\Omega)$ and V_h the space of continuous piecewise linear functions vanishing on $\partial\Omega$.

The objective of this section is to exhibit a control of the H^1 seminorm of the error

$$|e_h|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega \boldsymbol{\nabla} e_h \cdot \boldsymbol{\nabla} e_h \,\mathrm{d}\boldsymbol{x}$$

in terms of the residual of the equation $\mathcal{R}(u_h)$.

By Galerkin orthogonality $a(e_h, v_h) = 0$ for $v_h \in V_h$, in particular testing against $v_h = \mathcal{I}_h\, e_h$ is possible so that

$$|e_h|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega \boldsymbol{\nabla} e_h \cdot \boldsymbol{\nabla}(e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x}$$

The immediate motivation for introducing this test function is to be able to use interpolation inequalities.

Following the ideas of the previous section,

$$|e_h|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla}(e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x}$$

so that by consistency of the Galerkin method

$$|e_h|^2_{\mathrm{H}^1(\Omega)} = \int_\Omega f\ (e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x} \tag{7.5}$$

with the first term being $L(v)$ and the second term $a(u_h, v)$.

Similarly to interpolation inequalities we consider the expression per element $K \in \mathcal{T}_h$. The second term is integrated the by part,

$$\int_K \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla}(e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x} = \int_{\partial K} \boldsymbol{\nabla} u_h \cdot \mathbf{n}\ (e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x} - \int_K \Delta u_h\ (e_h - \mathcal{I}_h\, e_h) \,\mathrm{d}\boldsymbol{x}$$

and combined with the first term to obtain the element residual

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

Summing again over the domain yields

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

with the first term being a volume integral involving the residual of the equation, and the second term being a surface integral involving jump of the normal gradient of u_h across the cell facets.

Using first the Cauchy–Schwarz inequality

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

then the interpolation inequality with constant C_I

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

Consequenlty, we conclude

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

7.3 Dual weighted residual estimate

7.3.1 Adjoint operator

Definition 7.3.1 (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 7.3.2 (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 7.3.3 (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 $\mathrm{L}^2(\Omega)$, then $\mathcal{A}^\star = -\mathrm{D}_x$.

Example 7.3.4 (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*.

7.3.2 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 $\mathcal{A}^\star$:

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

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.* $\alpha = 2$,

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

Consequently,

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

7.4 Method

Definition 7.4.1 (h-adaptivity)**.** Given a tolerance parameter $\epsilon_{tol} > 0$ defining a quality criterion for the computed solution u_h, adapt the mesh such that it satisfies:

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

Algorithm 7.4.2 (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* $(\epsilon_{\mathcal{T}} \geq \epsilon_{tol})$ *:*
     $\rightarrow$ *Generate mesh* $\mathcal{T}_h^{\ell+1}$ *by refining cells with largest values of* ϵ_K.
     Else :
     $\rightarrow$ *Terminate adaptive iterations,* $\ell_{max} = \ell$.

7.5 Exercises

Exercise 7.5.1 (Diffusion–Reaction problem on the unit interval)**.**

Consider the following one-dimensional problem:

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

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:

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq \|h^2 \mathcal{R}(u_h)\|_{\mathrm{L}^2(\Omega)} \|h^{-2}(\eta - \mathcal{I}_h^1\, \eta)\|_{\mathrm{L}^2(\Omega)}$$

   with the discretization error $e_h = u - u_h$, the equation residual $\mathcal{R}(u_h) = f + \partial_x \left(a\, \partial_x\, u_h\right) - c\, u_h$ and the Lagrange $\mathbb{P}_1$ interpolation operator $\mathcal{I}_h^1$. First you should test the dual equation against e_h, then write the expression element-wise to be able to define the residual.

4. Conclude that the *a posteriori* error estimate holds

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

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

Chapter 8

Stabilized methods for advection dominated problems

8.1 An advection–diffusion problem in one dimension

Let us consider the following one-dimensional advection–diffusion problem:

$$-\partial_x \left(\nu(x)\, \partial_x u(x)\right) + \partial_x u(x) = f(x) \quad , \forall\, x \in \Omega = (0,1)$$

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

Chapter 9

Iterative solvers and Multigrid

9.1 Iterative methods

As seen in the previous lecture, direct methods can theoretically compute exact solutions $\boldsymbol{x}\mathbb{R}^N$ to linear systems in the form of:

$$\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$$

with matrix with real coefficients $\mathrm{A} \in M_N(\mathbb{R})$ and given data $\boldsymbol{b} \in \mathbb{R}^N$, in a determined finite number of steps.

As computing the inverse of the matrix is unrealistic, several methods were introduced based on factorizations of the type A = P Q where P and Q have a structure simplifying the resolution of the system: diagonal, banded, triangular.

Methods like LU, Cholevski take advantage of the existence of a decomposition involving triangular matrices while QR for example, involves the construcion of an orthogonal basis. All methods prove to be quite expensive, hard to parallelize due to the sequential nature of the algorithm and prone to error propagation.

Iterative methods have been developed for:

- solving very large linear systems with direct methods is in practice not possible due to the complexity in term of computational operations and data,
- taking advantage of sparse system for which the structure of the matrix can result in dramatic speed-up (this is the case for numerical schemes for PDEs),
- using the fact that some systems like PDEs discretizations are already formulated in an iterative fashion.

In this section, we discuss briefly the computational properties of iterative methods for solving linear systems. Computing the exact solution is not a requirement anymore but instead the algorithm is supposed to converge asymptotically to the exact solution: the algorithm is stopped when the approximate

solution is deemed *close enough* to the exact solution in a sense to be defined. A parameter used as stopping criterion triggers the completion of the algorithm.

The general idea of these methods is to introduce a splitting of the form:

$$\mathrm{A} = \mathrm{G} - \mathrm{H}$$

such the solution $\boldsymbol{x}$ satisfies:

$$\mathrm{G}\boldsymbol{x} = \boldsymbol{b} + \mathrm{H}\boldsymbol{x}$$

Similarly to fixed-point methods we can define a sequence of approximate solutions $\left(\boldsymbol{x}^k\right)$ satisfying relations of the form:

$$\mathrm{G}\hat{\boldsymbol{x}}^{k+1} = \boldsymbol{b} + \mathrm{H}\hat{\boldsymbol{x}}^k$$

with G invertible.

The matrix viewed as a linear mapping in $\mathbb{R}^N$, the counterpart of such approaches is given by the Brouwer Theorem in finite dimension, where a continuous mapping $f : \Omega \to \Omega$ with Ω compact of $\mathbb{R}^N$ admits a fixed-point $\boldsymbol{x}^\star$ satisfying $f(\boldsymbol{x}^\star) = \boldsymbol{x}^\star$ and is contracting.

Methods introduced depend on the iteration defined by the splitting and call for several questions regarding the computational aspects:

1. How can the convergence be ensure?
2. How fast is the convergence?
3. How expensive is each iteration?
4. How does the algorithm behave with respect to numerical error?

The question of the convergence is addressed by proving an estimate on error vectors in terms of iteration error $\hat{\boldsymbol{\epsilon}}^k = \hat{\boldsymbol{x}}^{k+1} - \hat{\boldsymbol{x}}^k$ or global error: $\boldsymbol{\epsilon}^k = \hat{\boldsymbol{x}}^k - \boldsymbol{x}$. The convergence rate α means that $C > 0$, $|\boldsymbol{\epsilon}^{k+1}| \leq C|\boldsymbol{\epsilon}^k|^\alpha$.

For example, substituting $\hat{\boldsymbol{x}}^{k+1} = \mathrm{G}^{-1}\boldsymbol{b} + \mathrm{G}^{-1}\mathrm{H}\hat{\boldsymbol{x}}^k$ in $\hat{\boldsymbol{\epsilon}}^k = \hat{\boldsymbol{x}}^{k+1} - \hat{\boldsymbol{x}}^k$ gives a relation between successive iteration errors:

$$\hat{\boldsymbol{\epsilon}}^k = \mathrm{G}^{-1}\mathrm{H}\,\hat{\boldsymbol{\epsilon}}^{k-1}$$

with $\mathrm{M} = \mathrm{G}^{-1}\mathrm{H}$ the iteration matrix, and recursively $\hat{\boldsymbol{\epsilon}}^k = (\mathrm{G}^{-1}\mathrm{H})^{k+1}\hat{\boldsymbol{\epsilon}}^0$. Convergence is then conditioned to the existence of a contraction factor $K < 1$ such that $\|\hat{\boldsymbol{\epsilon}}^k\|_\infty \leq K\,\|\hat{\boldsymbol{\epsilon}}^{k-1}\|_\infty$ ensuring decrease of the error.

In terms of the matrix M, this translate for the spectral radius $\rho(\mathrm{M})$ as $\rho(\mathrm{M}) < 1$ since in that case $\lim_{k\to\infty} M^k\hat{\boldsymbol{\epsilon}}^0 = 0_{\mathbb{R}^N}$. The smaller the spectral radius, the faster the convergence.

Each method is described briefly and qualitatively with just the necessary ingredients to discuss practical implementations.

9.2 Relaxation methods

Consider the relations for each row $i = 1, \dots, N$:

$$\boldsymbol{x}_i = \frac{1}{a_{ii}}\Big(b_i - \sum_{i \neq j} a_{ij}\boldsymbol{x}_j\Big) \tag{9.1}$$

Let us introduce two methods based on constructing sequences of approximate solutions $\big(\hat{\boldsymbol{x}}^k\big)$, $k \geq 1$ given an initial guess $\hat{\boldsymbol{x}}^0 \in \mathbb{R}^N$ and then associated relaxation methods.

9.2.1 Jacobi, methods of simultaneous displacements

$$\hat{\boldsymbol{x}}_i^{k+1} = \frac{1}{a_{ii}}\Big(b_i - \sum_{i \neq j} a_{ij}\hat{\boldsymbol{x}}_j^k\Big) \tag{9.2}$$

Convergence: the global error $\boldsymbol{\epsilon}^k$ is controlled by

$$\|\boldsymbol{\epsilon}^{k+1}\| \leq \sum_{i \neq j}\Big|\frac{a_{ij}}{a_{ii}}\Big| \, \|\boldsymbol{\epsilon}^k\| \leq K^k \, \|\boldsymbol{\epsilon}^1\|$$

It is then enough if the matrix is strictly diagonally dominant. Expressing the iteration error gives directly that $\mathrm{M} = \mathrm{G}^{-1}\mathrm{H}$ such that $\rho(\mathrm{M}) < 1$.

Algorithm: the splitting is

$$\mathrm{A} = \mathrm{D} - \mathrm{H}$$

with $\mathrm{D} = \mathrm{diag}(A)$, thus

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{D}^{-1}(\boldsymbol{b} + \mathrm{H}\hat{\boldsymbol{x}}^k)$$

Implementation:

1. Parallelization component by component is possible since there is only dependency on $\hat{\boldsymbol{x}}^k$.
2. Memory requirement for storing both $\hat{\boldsymbol{x}}^{k+1}$ and $\hat{\boldsymbol{x}}^k$ at each iteration.

9.2.2 Gauss–Seidel, methods of sucessive displacements

In Jacobi iterations, notice that sequential ordered computation of terms

$$\hat{\boldsymbol{x}}_i^{k+1} = \frac{1}{a_{ii}}\Big(b_i - \sum_{i \neq j} a_{ij}\hat{\boldsymbol{x}}_j^k\Big) \tag{9.3}$$

involves components $\hat{\boldsymbol{x}}_j^k$ which are also computed for $\hat{\boldsymbol{x}}^{k+1}$ if $j < i$.

$$\hat{\boldsymbol{x}}_i^{k+1} = \frac{1}{a_{ii}}\Big(b_i - \sum_{i<j} a_{ij}\hat{\boldsymbol{x}}_j^{k+1} - \sum_{i>j} a_{ij}\hat{\boldsymbol{x}}_j^k\Big) \tag{9.4}$$

Algorithm: the splitting is

$$\mathrm{A} = \mathrm{L} - \mathrm{R}_0$$

with $\mathrm{L} = \mathrm{D} + \mathrm{L}_0$ lower-triangular matrix and R_0 strict upper-triangular matrix, thus

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{D}^{-1}(\boldsymbol{b} - \mathrm{L}_0\hat{\boldsymbol{x}}^{k+1} + \mathrm{R}_0\hat{\boldsymbol{x}}^{k})$$

or

$$\mathrm{L}\ \hat{\boldsymbol{x}}^{k+1} = \mathrm{D}^{-1}(\boldsymbol{b} + \mathrm{R}_0\hat{\boldsymbol{x}}^{k})$$

Recast under the usual form:

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{L}^{-1}(\boldsymbol{b} + \mathrm{R}_0\hat{\boldsymbol{x}}^{k})$$

and the iteration matrix is $\bar{\mathrm{M}} = \mathrm{L}^{-1}\mathrm{R}_0$.

Convergence: the global error $\boldsymbol{\epsilon}^k$ is controlled by

$$\|\boldsymbol{\epsilon}^{k+1}\| \leq \frac{\displaystyle\sum_{i>j}\left|\frac{a_{ij}}{a_{ii}}\right|}{1 - \displaystyle\sum_{i<j}\left|\frac{a_{ij}}{a_{ii}}\right|}\ \|\boldsymbol{\epsilon}^{k}\| \leq \bar{K}^k\ \|\boldsymbol{\epsilon}^{1}\|$$

If the Jacobi contraction factor $K < 1$ then $\bar{K} < 1$. Expressing the iteration error gives directly that $\bar{\mathrm{M}} = \mathrm{L}^{-1}\mathrm{R}_0$ such that $\rho(\bar{\mathrm{M}}) < 1$.

Implementation:

1. Parallelization component by component is not possible easily since there is serialization for each row i due to the dependency on $\hat{\boldsymbol{x}}_j^{k+1}$, $j < i$.

2. Memory requirement is only for storing one vector of $\mathbb{R}^N$ at each iteration.

9.2.3 Relaxation of Jacobi and Gauss-Seidel

Relaxation methods consists of adding a linear combination of the approximate solution at the previous iteration to minimize the spectral radius for convergence, using the relaxation parameter $\gamma \in (0, 1)$.

1. Jacobi Over-Relaxation (JOR):

$$\hat{\boldsymbol{x}}_i^{k+1} = (1-\gamma)\ \hat{\boldsymbol{x}}_i^{k} + \gamma\frac{1}{a_{ii}}\Big(b_i - \sum_{i\neq j} a_{ij}\hat{\boldsymbol{x}}_j^{k}\Big) \tag{9.5}$$

   which reads in matricial form

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{M}_\gamma\hat{\boldsymbol{x}}^{k} + \gamma\mathrm{D}^{-1}\ \boldsymbol{b}$$

   with $\mathrm{M}_\gamma = (1-\gamma)\mathbb{I} + \gamma\mathrm{D}^{-1}\mathrm{H}$

2. Successive Over-Relaxation (SOR):

$$\hat{\boldsymbol{x}}_i^{k+1} = (1-\gamma)\,\hat{\boldsymbol{x}}_i^k + \gamma\frac{1}{a_{ii}}\Big(b_i - \sum_{i<j} a_{ij}\hat{\boldsymbol{x}}_j^{k+1} - \sum_{i>j} a_{ij}\hat{\boldsymbol{x}}_j^k\Big) \tag{9.6}$$

which reads in matricial form

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{M}_\gamma\hat{\boldsymbol{x}}^k + \gamma\mathrm{C}\,\boldsymbol{b}$$

with $\mathrm{M}_\gamma = (1+\gamma\mathrm{D}^{-1}\mathrm{L_0}^{-1})^{-1}\big[(1-\gamma)\mathbb{I}+\gamma\mathrm{D}^{-1}\mathrm{R_0}\big]$ and $\mathrm{C} = (1+\gamma\mathrm{D}^{-1}\mathrm{L_0}^{-1})^{-1}\mathrm{D}^{-1}$

The relaxation parameter γ cannot be known *a priori* and is usually determined by heuristics.

9.2.4 Parallelization of Gauss–Seidel

Overcoming the serialization in Gauss–Seidel is possible if the matrix is sparse. Taking advantage of the fact that components does not all possess connectivities with each other: such dependencies can be built from the sparsity pattern then decoupled graphs identified:

1. Component Dependency-Graph: generate a graph to reorder entries such that dependencies are avoided.

2. Red–Black coloring: special case for two-dimensional problems.

9.3 Krylov-subspace methods

The idea of these methods is that the solution is decomposed on a sequence of orthogonal subspaces.

If A is symmetric definite positive it induces the corresponding scalar product:

$$\langle\, \boldsymbol{x}\,,\, \boldsymbol{y}\,\rangle = (\,\mathrm{A}\boldsymbol{x}\,,\, \boldsymbol{y}\,) = \boldsymbol{y}^T\mathrm{A}\boldsymbol{x}$$

with $(\,\mathrm{A}\cdot\,,\,\cdot\,)$ canonical scalar product in $\mathbb{R}^N$. The vectors $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_N)$ are said A-conjugate if $\boldsymbol{e}_j^T\mathrm{A}\boldsymbol{e}_i = 0$ for $i \neq j$: they are orthogonal for the scalar-product induced by A.

9.3.1 Principle of descent methods: Steepest Gradient

Minimisation of the residual:

$$\boldsymbol{x}^\star = \mathrm{argmin}_{\boldsymbol{x}}\,\mathrm{J}(\boldsymbol{x}) = \frac{1}{2}\langle\, \boldsymbol{x}\,,\, \boldsymbol{x}\,\rangle - \langle\, \boldsymbol{b}\,,\, \boldsymbol{x}\,\rangle$$

Construct a sequence of solutions to approximate minimization problems, given $\hat{\boldsymbol{x}}^k$:

$$\mathrm{J}(\hat{\boldsymbol{x}}^{k+1}) \leq \mathrm{J}(\hat{\boldsymbol{x}}^k)$$

where $\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{e}^{k+1}$, with α_{k+1} a descent factor and $\boldsymbol{e}_{k+1}$ a direction.

For the Steepest Gradient:

1. take the direction given by $-\nabla \mathrm{J}(\hat{\boldsymbol{x}}^k) = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^k$ which is the residual $\boldsymbol{r}_k = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^k$, thus $\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{r}_k$.
2. choose the descent factor α^{k+1} minimizing the functional $\mathrm{J}(\hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{r}_k)$:

$$\alpha_{k+1} = \frac{\boldsymbol{r}_k^T \boldsymbol{b}}{\boldsymbol{r}_k^T \mathrm{A} \boldsymbol{r}_k}$$

The speed of convergence is bounded by $\mathcal{O}\left(1 - \mathcal{C}(\mathrm{A})^{-1}\right)$ with $\mathcal{C}(\mathrm{A})$ the conditioning of A. The gradient direction may not be optimal, Conjugate Gradient methods improve the choice of $\left(\boldsymbol{e}_k\right)$.

9.3.2 Conjugate Gradient

The Conjugate Gradient (CG) is a Krylov-subspace algorithm for symmetric positive definite matrices.

Given $\hat{\boldsymbol{x}}^0$, $\left(\hat{\boldsymbol{x}}^k\right)$ is q sequence of solutions to approximate k-dimensional minimisation problems.

For the Conjugate Gradient:

1. take the direction $\boldsymbol{e}_{k+1}$ such that $\left(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \boldsymbol{e}_{k+1}\right)$ is A-conjugate, thus $\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{e}_{k+1}$.
2. choose the descent factor α^{k+1} minimizing the functional $\mathrm{J}(\hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{r}_k)$, which is defined by

$$\alpha_j = \frac{\boldsymbol{e}_j^T \boldsymbol{b}}{\boldsymbol{e}_j^T \mathrm{A} \boldsymbol{e}_j}$$

and with $\boldsymbol{e}_j^T \boldsymbol{b} \neq 0$ (unless the exact solution is reached).

The construction of $\left(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{k+1}\right)$ is done by orthogonalization of residuals by Gram–Schmidt:

$$\boldsymbol{e}_{k+1} = \boldsymbol{r}_k - \frac{\boldsymbol{e}_k^T \mathrm{A} \boldsymbol{r}_{k-1}}{\boldsymbol{e}_k^T \mathrm{A} \boldsymbol{e}_k} \boldsymbol{e}_k$$

so that $\boldsymbol{r}_{k+1} = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^{k+1} = \boldsymbol{r}^k - \alpha_{k+1}\mathrm{A}\boldsymbol{e}_{k+1}$

After N steps, the A-conjugate basis of $\mathbb{R}^N$ is done and the exact solution is reached:

$$\boldsymbol{x} = \sum_{j=1}^{N} \alpha_j \hat{\boldsymbol{x}}^j$$

For any k, the speed of convergence is bounded by

$$\mathcal{O}\left(\frac{1 - \sqrt{\mathcal{C}(\mathrm{A})}}{1 + \sqrt{\mathcal{C}(\mathrm{A})}}\right)^{2k}$$

in the norm induced by A, with $\mathcal{C}(\mathrm{A})$ the conditioning of A.

The Conjugate Gradient can therefore be seen as a direct methods but in practice:

- the iterative computation of the A-conjugate basis suffers from the same issue of numerical error propagation as the QR factorization leading to a loss of orthogonality,
- the convergence is slow, which makes it unrealistic to compute the exact solution for large systems,

so it is used as an iterative method.

Example algorithm on first steps:

1. Given $\hat{\boldsymbol{x}}^0 = 0$, set $\boldsymbol{r}_0 = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^0$ and $\boldsymbol{e}_1 = \boldsymbol{r}_0$,
2. Take $\hat{\boldsymbol{x}}_1 = \alpha_1 \boldsymbol{e}_1$, then $\alpha_1 \boldsymbol{e}_1^T \mathrm{A} \boldsymbol{e}_1 = \boldsymbol{e}_1^T \boldsymbol{b}$, thus
$$\alpha_1 = \frac{\boldsymbol{r}_0^T \boldsymbol{b}}{\boldsymbol{r}_0^T \mathrm{A} \boldsymbol{r}_0}$$
3. Compute the residual:
$$\boldsymbol{r}_1 = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^1$$
4. Compute the direction:
$$\boldsymbol{e}_2 = \boldsymbol{r}_1 - \frac{\boldsymbol{e}_1^T \mathrm{A} \boldsymbol{r}_0}{\boldsymbol{e}_1^T \mathrm{A} \boldsymbol{e}_1} \boldsymbol{e}_1$$
5. Compute the factor:
$$\alpha_2 = \frac{\boldsymbol{e}_2^T \boldsymbol{b}}{\boldsymbol{e}_2^T \mathrm{A} \boldsymbol{e}_2}$$
6. Update the solution:
$$\hat{\boldsymbol{x}}^2 = \hat{\boldsymbol{x}}^1 + \alpha_2 \boldsymbol{e}_2$$
7. ...

The algorithm iteration reads:

1. Compute the residual:
$$\boldsymbol{r}_k = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}^k$$
2. Compute the direction:
$$\boldsymbol{e}_{k+1} = \boldsymbol{r}_k - \frac{\boldsymbol{e}_k^T \mathrm{A} \boldsymbol{r}_{k-1}}{\boldsymbol{e}_k^T \mathrm{A} \boldsymbol{e}_k} \boldsymbol{e}_k$$
3. Compute the factor:
$$\alpha_{k+1} = \frac{\boldsymbol{e}_{k+1}^T \boldsymbol{b}}{\boldsymbol{e}_{k+1}^T \mathrm{A} \boldsymbol{e}_{k+1}}$$
4. Update the solution:
$$\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1} \boldsymbol{e}_{k+1}$$

which requires two matrix-vector multiplications per loop, $\mathrm{A}\hat{\boldsymbol{x}}^k$ then $\mathrm{A}\boldsymbol{e}_{k+1}$ Using $\boldsymbol{r}_{k+1} = \boldsymbol{r}^k - \alpha_{k+1}\mathrm{A}\boldsymbol{e}_{k+1}$ saves one matrix-vector multiplication.

While the residual norm $\varrho_k = \|\boldsymbol{r}_k\|_2^2$ is big:

1. Compute the projection:
$$\beta_k = \frac{\varrho_k}{\varrho_{k-1}}$$
2. Compute the direction:
$$\boldsymbol{e}_{k+1} = \boldsymbol{r}_k + \beta_k \boldsymbol{e}_k$$
3. Compute the factor:
$$\boldsymbol{w} = \mathrm{A}\boldsymbol{e}_{k+1}; \quad \alpha_{k+1} = \frac{\varrho_k}{\boldsymbol{e}_{k+1}^T \boldsymbol{w}}$$
4. Update the solution:
$$\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1}\boldsymbol{e}_{k+1}$$
5. Update the residual:
$$\boldsymbol{r}^{k+1} = \boldsymbol{r}^k - \alpha_{k+1}\boldsymbol{w}$$

9.3.3 Preconditioners

While seeing the Conjugate Gradient as a pure iterative method relieves from concerns regarding orthogonality loss, the convergence is still slow as soon as the condition number of the matrix is bad.

Preconditioning the system consists in finding a non-singular symmetrix matrix C such that $\tilde{\mathrm{A}} = \mathrm{C}^{-1}\mathrm{A}\mathrm{C}^{-1}$ and the conjugate gradient is applied to

$$\tilde{\mathrm{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$$

with $\tilde{\boldsymbol{x}} = \mathrm{C}^{-1}\boldsymbol{x}$ and $\tilde{\boldsymbol{b}} = \mathrm{C}^{-1}\boldsymbol{b}$.

With:

- $\mathrm{M} = \mathrm{C}^2$
- $\boldsymbol{e}_k = \mathrm{C}^{-1}\tilde{\boldsymbol{e}}_k$
- $\hat{\boldsymbol{x}}_k = \mathrm{C}^{-1}\tilde{\hat{\boldsymbol{x}}}_k$
- $\boldsymbol{z}_k = \mathrm{C}^{-1}\tilde{\boldsymbol{r}}_k$
- $\boldsymbol{r}_k = \mathrm{C}\tilde{\boldsymbol{r}}_k = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}_k$

and M is a symmetric positive definite matrix called the preconditioner.

While the residual norm $\varrho_k = \|\boldsymbol{r}_k\|_2^2$ is big:

1. Solve:
$$\mathrm{M}\boldsymbol{z}_k = \boldsymbol{r}_k$$

2. Compute the projection:

$$\beta_k = \frac{\boldsymbol{z}_k^T \boldsymbol{r}_k}{\boldsymbol{z}_{k-1}^T \boldsymbol{r}_{k-1}}$$

3. Compute the direction:

$$\boldsymbol{e}_{k+1} = \boldsymbol{z}_k + \beta_k \boldsymbol{e}_k$$

4. Compute the factor:

$$\alpha_{k+1} = \frac{\boldsymbol{z}_k^T \boldsymbol{r}_k}{\boldsymbol{e}_{k+1}^T \mathrm{A} \boldsymbol{e}_{k+1}}$$

5. Update the solution:

$$\hat{\boldsymbol{x}}^{k+1} = \hat{\boldsymbol{x}}^k + \alpha_{k+1} \boldsymbol{e}_{k+1}$$

6. Update the residual:

$$\boldsymbol{r}^{k+1} = \boldsymbol{r}^k - \alpha_{k+1} \boldsymbol{w}$$

The linear system $\mathrm{M}\boldsymbol{z}_k = \boldsymbol{r}_k$ should be easy to solve and can lead to fast convergence, typically $\mathcal{O}\left(() \sqrt{N}\right)$. Since

$$\mathrm{M}\boldsymbol{z}_k = \boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}_k$$

Then an iterative relation appears:

$$\hat{\boldsymbol{x}}_{k+1} = \mathrm{M}^{-1}\left(\boldsymbol{b} - \mathrm{A}\hat{\boldsymbol{x}}_k\right)$$

therefore iterative methods like Jacobi, Gauss-Seidel and relaxation methods can be used.

9.4 Power method

This method is used for finding the dominant eigenvalues of a matrix $\mathrm{A} \in M_N(\mathbb{R})$ of N eigenvectors $\left(\boldsymbol{v}_i\right)$ with associated eigenvalues $\left(\lambda_i\right)$ ordered in decreasing module. The eigenvalues are either real or conjugate complex pairs.

Given a random vector $\boldsymbol{x}^0$, construct a sequence of vectors $\left(\hat{\boldsymbol{x}}^k\right)$ such that

$$\hat{\boldsymbol{x}}^{k+1} = \mathrm{A}\hat{\boldsymbol{x}}^k$$

then $\forall k \geq 0$

$$\hat{\boldsymbol{x}}^k = \sum_{i=0}^{N-1} \lambda_i^k \xi_i \boldsymbol{v}_i$$

for some coefficients $\left(\boldsymbol{v}_i\right)$.

Assume that λ_0 is a dominant real eigenvalue and $\xi_0 \neq 0$, then

$$\hat{\boldsymbol{x}}^k = \lambda_0^k\left(\xi_0 \boldsymbol{v}_0 + \boldsymbol{r}_k\right)$$

with the residual $\boldsymbol{r}_k$ defined as

$$\boldsymbol{r}_k = \lambda_0^{-k} \sum_{i=1}^{N-1} \lambda_i^k \xi_i \boldsymbol{v}_i$$

and $\lim_{k\to\infty} \boldsymbol{r}_k = O_{\mathbb{R}^N}$. To the limit $\hat{\boldsymbol{x}}_{k+1} \approx \lambda_0 \hat{\boldsymbol{x}}^k \approx \lambda_0 \xi_0 \boldsymbol{v}0$ almost parallel to the first eigenvector.

- This method is fast to compute the spectral radius for the Jacobi method and relaxation parameters.
- The convergence is geometric and the speed depends on the ratio $|\lambda_1/\lambda_0|$.
- If the matrix is symmetric, the convergence speed can be doubled.
- If λ_0 is very large or very small then taking high powers lead to numerical issues, the algorithm requires a normalization.

9.5 Multigrid methods

TBD

Chapter 10

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 ?

10.1 The Stokes equations

10.1.1 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

$$-\boldsymbol{\nabla}\cdot\sigma(\mathbf{x}) = \varrho(\mathbf{x})\mathbf{g}(\mathbf{x}) \tag{10.1a}$$

with the stress tensor

$$\sigma = \tau - p\,\mathbb{I} \tag{10.1b}$$

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

$$\boldsymbol{\nabla}\cdot\bar{\boldsymbol{u}}(\mathbf{x}) = 0 \tag{10.1c}$$

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 $\partial\Omega_D$

$$\bar{\boldsymbol{u}} = \boldsymbol{u}_D \tag{10.1d}$$

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

$$\sigma\cdot\mathbf{n} = \sigma_N \tag{10.1e}$$

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 (10.1a) and (10.1c) against smooth functions, such that we consider

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

and

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

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

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

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

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

Owing to relation

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

and

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

the weak formulation of Problem (10.1) reads:

Find $(\bar{\boldsymbol{u}},p)\in\boldsymbol{W}\times M$ such that:

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

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.

10.1.2 Abstract weak formulation

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

$$\left|\begin{array}{l} \text{Find } (\bar{\boldsymbol{u}}, p) \in \boldsymbol{W} \times M \text{ such that:} \\ \qquad\qquad a(\bar{\boldsymbol{u}}, \boldsymbol{v}) + b(\boldsymbol{v}, p) = L(\boldsymbol{v}) \quad ,\forall\, \boldsymbol{v} \in \boldsymbol{V} \\ \qquad\qquad b(\bar{\boldsymbol{u}}, q) = 0 \qquad\quad ,\forall\, q \in M \end{array}\right.$$

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

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

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

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

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

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

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 $\sigma \cdot \mathbf{n} = \sigma_N$, 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:

$$\left|\begin{array}{l} \text{Find } (\boldsymbol{u}, p) \in \boldsymbol{V} \times M \text{ such that:} \\ \qquad\qquad a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) = L(\boldsymbol{v}) \qquad\quad ,\forall\, \boldsymbol{v} \in \boldsymbol{V} \\ \qquad\qquad b(\boldsymbol{u}, q) = \langle\, \Psi\, ,\, p\, \rangle_{M',M} \quad ,\forall\, q \in M \end{array}\right. \tag{10.2}$$

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'.

10.1.3 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:

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

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:

Find $(\boldsymbol{u}, p) \in \boldsymbol{V} \times M$ such that:

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) &= L(\boldsymbol{v}) && , \forall\ \boldsymbol{v} \in \boldsymbol{V}_0 \\ b(\boldsymbol{u}, q) &= \langle\ \Psi\ ,\ p\ \rangle_{M',M} && , \forall\ q \in M \end{aligned}$$

Theorem 10.1.1 (Well-posedness of constrained problem)**.** *Let us define the space*

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

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

Find $(\boldsymbol{u}, p) \in \boldsymbol{V}_\Psi \times M$ such that:

$$a(\boldsymbol{u}, \boldsymbol{v}) = L(\boldsymbol{v}) \quad , \forall\ \boldsymbol{v} \in \boldsymbol{V}_0$$

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 10.1.2 (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 10.1.3 (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* (10.1.2)*, 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 (10.1.3) *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,

Theorem 10.1.4 (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 $\mathrm{L}^{p'}_{\int=0}(\Omega)$.*

10.2 The discrete Inf-Sup condition

10.2.1 Results

Let us consider an approximation of Problem 10.2 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 10.2.1 (Well-posedness)**.** *If $\Psi_h \in \mathrm{Im}(\mathbf{B}_h)$ then Problem* (10.2.1) *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 $p_h = \tilde{p}_h + \ker(\mathbf{B}^T{}_h)$ with $\tilde{p}_h \in \ker(\mathbf{B}^T{}_h)^\perp$ unique.*

Theorem 10.2.2 (Convergence – [7] page 21)**.** *Let* $(\boldsymbol{u}, p) \in \boldsymbol{V} \times \tilde{\mathrm{M}}$ *be the solution of Problem* (10.1) *and* $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times \tilde{\mathrm{M}}_h$ *the solution of discrete Problem* (10.2.1) *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.

10.2.2 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"

10.3 Exercises

Appendix A

Definitions

A.1 Mapping

Definition A.1.1 (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.1.2 (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.2.1 (Vector space (on the left))**.** Let $(\mathbb{K}, +, \times)$ be a field *i.e.* defined such that $(\mathbb{K}, +)$ is an Abelian additive group and $(\mathbb{K} \setminus \{0_{\mathbb{K}}\}, \times)$ is an Abelian multiplicative group, $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

$(\mathbb{K}, +)$	$(\mathbb{K} \setminus \{0_{\mathbb{K}}\}, \times)$
"$+$" commutative and associative $0_{\mathbb{K}}$ neutral for '$+$" "$+$" admits an opposite	"$\times$" commutative and associative $\mathbb{1}_{\mathbb{K}}$ neutral for '$\times$" "$\times$" admits an inverse
"$\times$" 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 "$\times$"	$(\lambda \times \mu) \cdot u = \lambda \cdot (\mu \cdot u)$
$\mathbb{1}_{\mathbb{K}}$ 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.2.2 (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.2.3. The third property is usually called *triangle inequality.*

Definition A.2.4 (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.2.5 (Seminorm)**.** Let E be a $\mathbb{K}$-vector space, the application

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

is a seminorm if it satisfies properties (A.2.2).2 and (A.2.2).3.

Definition A.2.6 (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: $(\ (\ x\ ,\ x\ ) = 0\ ) \Leftrightarrow (\ x = 0\ )$

Appendix B

Duality

B.1 In finite dimension

Definition B.1.1 (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.1.2 (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:

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

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 $\dim(\mathcal{B}^\star) = N$.

Appendix C

Function spaces

C.1 Banach and Hilbert spaces

Definition C.1.1 (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.1.2 (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.1.3 (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.1.4 (Hilbertian norm)**.**

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

Remark C.1.5. This is basically the parallelogram identity stating that the sum of the square of the length diagonals is equal to the sum of the square of the length of all the sides. This equality is useful to check that a norm is generated from a scalar product. For Banach spaces like L^p, this becomes the Clarkson inequality.

Theorem C.1.6 (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

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

Remark C.3.1. Lebesgue spaces L^p, $1 \geq p \geq \infty$ are Banach spaces for the norm

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

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

$$(u \, , \, v \,)_{\mathrm{L}^2(\Omega)} = \int_\Omega u \, v \, \mathrm{d}x \tag{C.1}$$

C.4 Hilbert–Sobolev spaces

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

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.5.1. H^s spaces are $\mathrm{W}^{s,2}$ spaces.

Appendix D

Inequalities

D.1 Useful inequalities in normed vector spaces

Lemma D.1.1 (Cauchy–Schwarz)**.** *Let E be a $\mathbb{K}$-vector space, any positive sesquilinear form $(\,\cdot\,,\,\cdot\,)$ on E satisfies the inequality:*

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

Remark D.1.2. In particular any scalar product satisfies the Cauchy–Schwarz inequality. For example:

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

Lemma D.1.3 (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.1.4. 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.1.5 (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.1.6 (Minkowski)**.**

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

Remark D.1.7. The previous result is basically the triangle inequality for the L^p– norm.

Lemma D.1.8 (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.1.9. 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 $\|u\|_{\mathrm{H}^1(\Omega)}$ on $\mathrm{H}_0^1(\Omega)$.

Lemma D.1.10 (Clarkson)**.** *Let $1 < p < \infty$, and u, v be two functions of $\mathrm{L}^p(\Omega)$, then:*

1. *for $p \geq 2$*

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

2. *for $p < 2$*

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

Remark D.1.11. These inequalities are basically parallelogram inequalities generalized to L^p spaces.

Appendix E

Tensor formulæ

This short chapter is a reminder of tensor notations and identities that are useful for usual partial differential equations.

E.1 Operators

A tensor of order p denotes an element of $\mathbb{R}^{d_1 \times \cdots \times d_p}$, with d_i dimension on the i-th axis: a zero-order tensor is a scalar, a first-order tensor is a vector, and a second-order tensor is a matrix. The following operators are defined for second order tensors.

E.1.1 Tensor product

(order $p+q$)

$$(\mathbf{P} \otimes \mathbf{Q})_{ijkl} = \mathbf{P}_{ij}\mathbf{Q}_{kl}\, \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \otimes \mathbf{e}_l \tag{E.1}$$

E.1.2 Dot product (simple contraction)

(order $p+q-2$)

$$(\mathbf{P} \cdot \mathbf{Q}) = \mathbf{Tr}^{(p,p+1)}(\mathbf{P} \otimes \mathbf{Q}) \tag{E.2}$$

with $\mathbf{Tr}^{(p,p+1)}(\cdot)$ the Trace operator with respect to indices p and $p+1$. In index notation, it consists of a summation on indices p and $p+1$ (contraction),

$$(\mathbf{P} \cdot \mathbf{Q}) = \sum_j \mathbf{P}_{ij}\mathbf{Q}_{jk} \tag{E.3}$$

Since summation occurs on a pair of indices, the order of the resulting tensor is reduced by two.

E.1.3 Double-dot product (double contraction)

(order $p+q-4$)

$$(\mathbf{P} : \mathbf{Q}) = \mathbf{Tr}^{(p-1,p+1)}(\mathbf{Tr}^{(p,p+2)}(\mathbf{P} \otimes \mathbf{Q})) \tag{E.4}$$

with two contractions in this case, corresponding in index notation to summations over two pairs of indices,

$$(\mathbf{P} : \mathbf{Q}) = \sum_i \sum_j \mathbf{P}_{ij}\mathbf{Q}_{ij} \tag{E.5}$$

so that the order of the resulting tensor is reduced by four. If $\mathbf{P} = \mathbf{Q}$ this corresponds to the Frobenius norm.

E.1.4 Gradient

The gradient of a scalar field is the first order tensor

$$\boldsymbol{\nabla} f = [\partial_i f]_i \tag{E.6}$$

while the gradient of a vector field is the second order tensor

$$\boldsymbol{\nabla} \boldsymbol{v} = [\partial_j \boldsymbol{v}_i]_{ij} \tag{E.7}$$

such that the derivative is applied to the last index.

E.1.5 Divergence

$$\boldsymbol{\nabla} \cdot (\mathbf{T}) = \nabla \mathbf{T} : \mathbf{G} = \mathbf{Tr}^{(p,p+1)}(\nabla \mathbf{T}) \tag{E.8}$$

with $\mathbf{G}$ the metric tensor, which entries are the scalar product of the chosen basis vectors. If the canonical basis is chosen, it is simply the identity matrix, therefore the divergence is simply the sum of diagonal entries of the gradient.

E.1.6 Curl (Rotational)

$$\nabla \wedge \mathbf{T} = -\nabla \mathbf{T} : \mathbf{H} \tag{E.9}$$

with $\mathbf{H}$ is the orientation tensor, which entries are the mixed product of the chosen basis vectors. The curl operator is also denoted as $\nabla\times$.

E.2 Identities

E.2.1 First order tensors

- Gradient of a vector field:

$$\nabla(f\boldsymbol{v}) = f\nabla\boldsymbol{v} + \boldsymbol{v}\otimes\nabla f \tag{E.10}$$

  which corresponds to the expansion of the derivative of a product, as the index notation shows

$$\partial_j (f\boldsymbol{v}_i) = f\partial_j \boldsymbol{v}_i + \boldsymbol{v}_i \partial_j f \tag{E.11}$$

- Divergence of a vector field:

$$\boldsymbol{\nabla}\cdot(f\boldsymbol{v}) = f\boldsymbol{\nabla}\cdot(\boldsymbol{v}) + \boldsymbol{v}\cdot\nabla f \tag{E.12}$$

which corresponds to the expansion of the derivative of a product, as the index notation shows

$$\partial_i(f\boldsymbol{v}_i) = f\partial_i\,\boldsymbol{v}_i + \boldsymbol{v}_i\partial_i\,f \tag{E.13}$$

- Identity of the advection operator:

$$(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} = \frac{1}{2}\nabla(\boldsymbol{v}\cdot\boldsymbol{v}) + (\nabla\wedge\boldsymbol{v})\wedge\boldsymbol{v} \tag{E.14}$$

- Identity for the Laplace operator:

$$\Delta\boldsymbol{v} = \nabla(\boldsymbol{\nabla}\cdot\boldsymbol{v}) - \nabla\wedge(\nabla\wedge\boldsymbol{v}) \tag{E.15}$$

- Divergence of a vector product:

$$\boldsymbol{\nabla}\cdot(\boldsymbol{u}\wedge\boldsymbol{v}) = \boldsymbol{v}\cdot(\nabla\wedge\boldsymbol{u}) - \boldsymbol{u}\cdot(\nabla\wedge\boldsymbol{v}) \tag{E.16}$$

- Curl of of vector product:

$$\nabla\wedge(\boldsymbol{u}\wedge\boldsymbol{v}) = (\boldsymbol{\nabla}\cdot\boldsymbol{v})\boldsymbol{u} - (\boldsymbol{\nabla}\cdot\boldsymbol{u})\boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{u} - (\boldsymbol{u}\cdot\nabla)\boldsymbol{v} \tag{E.17}$$

E.2.2 Second order tensors

- Dyadic/scalar mixed product:

$$(\boldsymbol{u}\otimes\boldsymbol{v})\cdot\boldsymbol{w} = (\boldsymbol{v}\cdot\boldsymbol{w})\boldsymbol{u} \tag{E.18}$$

- Gradient of a tensor field:

$$\nabla(\mathbf{T}\cdot\boldsymbol{v}) = \mathbf{T}\cdot\nabla\boldsymbol{v} + \boldsymbol{v}\cdot\nabla\mathbf{T}^{\mathrm{T}} \tag{E.19}$$

- Divergence of a tensor field:

$$\boldsymbol{\nabla}\cdot(\boldsymbol{v}\cdot\mathbf{T}) = \boldsymbol{v}\cdot\boldsymbol{\nabla}\cdot(\mathbf{T}) + \mathbf{T}:\nabla\boldsymbol{v} \tag{E.20}$$

which corresponds to the expansion of the derivative of a product, as the index notation shows

$$\partial_j(\boldsymbol{v}_i\mathbf{T}_{ij}) = \boldsymbol{v}_i\partial_j\,\mathbf{T}_{ij} + \mathbf{T}_{ij}\partial_j\,\boldsymbol{v}_i \tag{E.21}$$

- Divergence of a dyadic product:

$$\boldsymbol{\nabla}\cdot(\boldsymbol{u}\otimes\boldsymbol{v}) = (\boldsymbol{\nabla}\cdot\boldsymbol{v})\boldsymbol{u} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{u} \tag{E.22}$$

Bibliography

[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, Université de Provence, 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, Université de Provence, 2002.

[8] A. T. Patera. Finite Element Methods for Elliptic Problems. Lecture Notes, MIT OpenCourseWare, 2003.