Numerical Methods for Civil Engineering

– Notes of the Course–

Riccardo Sacco

January 14, 2013

To my Family and Friends

Preface

This text contains the notes of the course "Numerical Methods in Civil Engineering", which I have been helding (in Italian) over the last eight years for the MSc in Civil Engineering at Politecnico di Milano. The material is organized into 6 parts:

- Part I: Foundations (Chapts. 1, 2 and 3)
- Part II: Elliptic Problems (Chapts. 4 and 5)
- Part III: The GFEM for Elliptic Problems in 1D and 2D (Chapts. 6 and 7)
- Part IV: The GFEM for Linear Elasticity (Chapts. 8 and 9)
- Part V: Examination Problems with Solution (Chapt. 10)
- Part VI: Appendices (A, B and C)

Specifically:

- **Chapts.** 1, 2 and 3 provide an introduction to Numerical Mathematics, Numerical Linear Algebra and Approximation Theory.
- **Chapts.** 4 and 5 illustrate the weak formulation of a boundary value problem and its numerical approximation using the Galerkin Finite Element Method (GFEM).
- **Chapts.** 6 and 7 address the numerical study of 1D model problems and the implementation in 2D of the finite element model for an advection-diffusion-reaction boundary value problem.
- **Chapts.** 8 and 9 deal with the weak formulation and GFE approximation of the Navier-Lamè equations for linear isotropic elasticity in compressible and incompressible regimes.
- **Chapt.** 10 illustrates the complete solution of exercises recently proposed in class exams at the end of the Course.
- **Appendices** A, B and C give a short review of the essentials in Linear Algebra, Functional Analysis and Differential Calculus that are extensively used throughout the text.

All computations have been performed using `Matlab` on a standard laptop running under `Linux` OS. Methodologies and algorithms have been implemented in 1D and 2D codes by Marco Restelli. The codes can be downloaded at the official web page of the Course: `http://www1.mate.polimi.it/CN/MNIC/` to which the reader is kindly referred for any further information.

Milano, *Riccardo Sacco*
January 2013

Acknowledgements

It is a great pleasure for me to acknowledge here the fundamental contribution given by Marco Restelli and Luca Dedè for their computer lab class teaching and for developing the numerical software and tutorial exercises which constitute the supporting backbone of the theoretical part of the course. I also wish to gratefully thank Carlo de Falco for his teaching assistance in computer lab classes during the year 2010, and Chiara Lelli for her assistance during the year 2012, where for the first time the teaching language switched from Italian to English. Last, but certainly not least, my thanks go also to Paola Causin, for fruitful discussions and ten years of common hard work in the development, analysis and implementation of finite element methodologies in Continuum Mechanics, and to Prof. Maurizio Verri, for his fantastically critical proof-reading of the first draft of these notes.

Contents

I Foundations **1**

1 Numerical Mathematics **5**
1.1 The continuous model 5
1.2 The numerical model 9
1.3 The chain of errors 11
1.4 Errors and error analysis 12
1.5 Floating-point numbers 13

2 Numerical Linear Algebra **17**
2.1 Linear algebraic systems 17
2.2 Direct methods for linear systems 19
2.3 Stability analysis 24

3 Approximation Theory **27**
3.1 Interpolation 27
3.1.1 Basis functions 29
3.1.2 Finite element interpolation and error analysis 30
3.2 Quadrature 32

II Elliptic Problems **35**

4 Weak Formulations **39**
4.1 Elliptic boundary value problems 39
4.2 Weak solution of a BVP: the 1D case 41
4.3 Weak solution of a BVP: the 2D case 46
4.3.1 Non-homogeneous Dirichlet problem in 2D 48
4.3.2 Non-homogeneous Neumann problem in 2D 49

4.3.3 Mixed problem in 2D 50
4.4 Well-posedness analysis: the Lax-Milgram Lemma 51

5 The GFEM **55**
5.1 The Galerkin method 55
5.2 The Galerkin Finite Element Method 60
5.2.1 Error analysis 60
5.3 Experimental convergence study of the GFEM 63
5.3.1 BVP with smooth solution 63
5.3.2 BVP with a non-smooth solution 64

III The GFEM for Elliptic Problems in 1D and 2D **67**

6 Elliptic problems: theory and finite elements **71**
6.1 Reaction-diffusion model problem 72
6.1.1 Weak formulation 73
6.1.2 Galerkin finite element approximation 74
6.1.3 The linear system and the discrete maximum principle . . 75
6.1.4 Stabilization: the method of lumping of the reaction matrix 80
6.2 Advection-diffusion model problem 84
6.2.1 Weak formulation 86
6.2.2 Galerkin finite element approximation 87
6.2.3 The linear system and the discrete maximum principle . . 88
6.2.4 Stabilization: the method of artificial diffusion 91

7 2D Implementation of the GFEM **99**
7.1 Weak formulation 99
7.2 Geometrical discretization 100
7.3 Polynomial spaces in 2D 101
7.4 The approximation space 102
7.5 GFE approximation 104
7.6 The linear system 104
7.7 The assembly phase 105

IV The GFEM for Linear Elasticity 109

8 Compressible elasticity 113
8.1 Essentials of solid mechanics . . . 113
8.1.1 The relation $\sigma - \varepsilon$ . . . 115
8.1.2 Linear isotropic elasticity . . . 117
8.2 The Navier-Lamè Model . . . 118
8.3 The weak formulation . . . 120
8.4 Existence and uniqueness of the weak solution . . . 122
8.5 Two-dimensional models in elasticity . . . 125
8.5.1 Plane stress . . . 126
8.5.2 Plane strain . . . 127
8.6 The GFE approximation in the 2D case . . . 128
8.6.1 The Galerkin FE problem . . . 129
8.6.2 Local approximations and matrices . . . 129
8.6.3 Convergence analysis . . . 133
8.7 Numerical examples . . . 134
8.7.1 Example 1: patch test (constant stress) . . . 134
8.7.2 Example 2: experimental convergence analysis . . . 136

9 Incompressible elasticity 141
9.1 The incompressible regime . . . 141
9.2 Volumetric locking: examples . . . 143
9.3 Two-field model for linear elasticity . . . 146
9.4 Weak formulation of the two-field model . . . 148
9.5 Matrix block form of the Herrmann system . . . 149
9.6 Well-posedness analysis of the two-field model . . . 150
9.7 Energy formulation of incompressible elasticity . . . 154
9.8 GFE approximation of the two-field model . . . 156
9.9 Unique solvability and error analysis . . . 158
9.10 The discrete inf-sup condition . . . 161
9.11 Finite elements for incompressible elasticity . . . 162
9.11.1 Discontinuous pressures . . . 162
9.11.2 Continuous pressures . . . 165
9.12 Locking and pressure spurious modes . . . 168
9.12.1 Discontinuous pressure FE space . . . 169
9.12.2 Continuous pressure FE space . . . 172
9.13 Convergence analysis . . . 174

V Examination Problems with Solution 179

10 Solved problems 183
10.1 Examination of July 09, 2012 . . . 184
10.1.1 Solution of Exercise 1 . . . 186
10.1.2 Solution of Exercise 2 . . . 194
10.1.3 Solution of Exercise 3 . . . 197

VI Appendices 207

A Linear Algebra 211
A.1 Vector spaces . . . 211
A.2 Vector and matrix norms . . . 212
A.3 Matrices . . . 215

B Functional Analysis 219
B.1 Metric spaces . . . 219
B.2 Complete metric spaces . . . 222
B.3 Normed spaces . . . 226
B.4 Banach spaces . . . 226
B.5 Hilbert spaces . . . 228
B.6 Sobolev spaces in 1D . . . 231
B.7 Sobolev spaces in $\mathbb{R}^d$, $d \geq 2$ . . . 236

C Differential Calculus 241
C.1 Differential operators, useful formulas and properties . . . 241
C.1.1 First-order operators . . . 241
C.1.2 Second-order operators . . . 242
C.1.3 Green's formula . . . 242
C.2 Elliptic operators . . . 243
C.2.1 Maximum principle for 2nd order elliptic operators . . . 243
C.3 Tensors . . . 244
C.3.1 Operations/operators on tensors . . . 245

Part I

Foundations

This part contains a short (and obviously incomplete) summary of basic elements of Numerical Mathematics, including Numerical Linear Algebra and Approximation Theory, which are extensively used throughout the remainder of the text.

Chapter 1

Essentials of Numerical Mathematics

Abstract

In this chapter, we review the basic foundations of Numerical Mathematics that will be used thoroughly in the remainder of the text.

In particular, several elements of Linear Algebra and Numerical Analysis will be addressed, including solution methods for linear systems and function approximation using polynomials.

1.1 The continuous model

In this section, we introduce a general approach to the study of a mathematical representation of a given physical problem. Such a representation is referred to, from now on, as *the continuous model* or, *the continuous problem*, and takes the following form:

$$(\mathscr{P}) \qquad \begin{cases} \text{given } d \in \mathscr{D}, & \text{find } x \in V \text{ such that} \\ F(x,d) = 0 \end{cases}$$

where: d are the data, $\mathscr{D}$ is the set of admissible data (i.e., the data for which $(\mathscr{P})$ admits a solution), x is the solution, to be sought in a space V (see Def. A.1.1), while F is the functional relation between d and x.

Example 1.1.1. Compute $I_f := \int_a^b f(t)\,dt$ with $f \in C^0([a,b])$:

$$d = \{f, a, b\}, \qquad x = I_f$$
$$F(x,d) = \int_a^b f(t)\,dt - x, \qquad V = \mathbb{R}.$$

Example 1.1.2. Solve the linear algebraic system $A\mathbf{x} = \mathbf{b}$:

$$d = \left\{a_{ij}, b_i\right\}, i,j = 1,\ldots,n, \qquad x = \mathbf{x}$$
$$F(x,d) = \mathbf{b} - A\mathbf{x}, \qquad V = \mathbb{R}^n.$$

Example 1.1.3. Solve the Cauchy problem:

$$\begin{cases} y'(t) = f(t, y(t)) & t \in (t_0, t_0+T) \\ y(t_0) = y_0 \end{cases}$$

where:

$$d = \{f, t_0, T, y_0\}, \qquad x = y(t)$$
$$F(x,d) = \begin{cases} f - y' = 0 & t \in (t_0, t_0+T) \\ y_0 - y(t_0) = 0 & t = t_0 \end{cases}, \qquad V = C^1(t_0, t_0+T).$$

Definition 1.1.4 (Continuous dependence on data)**.** *Problem ($\mathscr{P}$) is said to be well-posed (or stable) if it admits a unique solution x continuously depending on the data. Should ($\mathscr{P}$) not enjoy such a property, it is said to be not well-posed (or unstable).*

Example 1.1.5 (Unstable problem). Let a be a real parameter. Let $n = n(a)$ denote the number of real roots of the polynomial $p(t) = t^4 - t^2(2a-1) + a(a-1)$. It is easy to see that:

$$n(a) = \begin{cases} 0 & a < 0 \\ 2 & 0 \le a < 1 \\ 4 & a \ge 1. \end{cases}$$

Therefore, the problem of determining $n(a)$ is not well-posed because n is not continuous at $a = 0$ and $a = 1$.

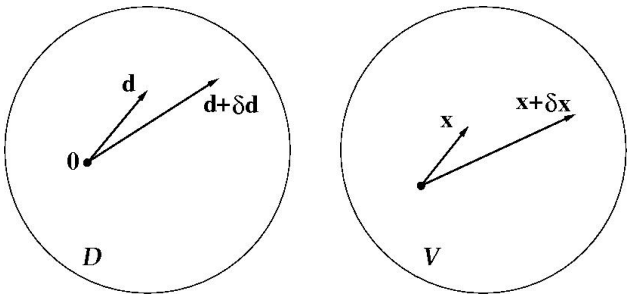


Figure 1.1: Continuous problem with perturbation on data and solution.

To characterize mathematically the concept of stability of $(\mathscr{P})$, we consider a *perturbation* of this latter (cf. Fig. 1.1):

$$(\mathscr{P}_\delta) \qquad \begin{cases} \text{given } (d+\delta d) \in \mathscr{D}, & \text{find } (x+\delta x) \in V \text{ such that} \\ F(x+\delta x, d+\delta d) = 0 \end{cases}$$

where δd is a perturbation of the data d such that $(\mathscr{P}_\delta)$ is still well posed, and δx is the corresponding perturbation of the solution.

Definition 1.1.6 (Stability of $(\mathscr{P})$)**.** *We say that $(\mathscr{P})$ is stable if*

$$\begin{aligned} &\exists \eta_0 = \eta_0(d), \quad \exists K_0 = K_0(d) \quad \textit{such that} \\ &\|\delta d\|_{\mathscr{D}} \le \eta_0 \qquad \Rightarrow \qquad \|\delta x\|_V \le K_0 \|\delta d\|_{\mathscr{D}}, \end{aligned} \tag{1.1}$$

where $\|\cdot\|_{\mathscr{D}}$ and $\|\cdot\|_V$ are appropriate norms for $\mathscr{D}$ and V, respectively (see Def. A.2.1).

The above definition tells us that, provided a limiting threshold η_0 is assumed for the admissible perturbations, the corresponding perturbation δx on the solution can be controlled by η_0, modulo an amplification constant K_0. Notice that *both* η_0 and K_0 depend, in general, on the reference data d. In this sense, the stability emanating from Def. 1.1.6 is "local". To get a "global" estimate of the stability of $(\mathscr{P})$, it is clear that we need to "explore" the *whole* admissible set $\mathscr{D}$. This leads to the next definition.

Definition 1.1.7 (Condition number of $(\mathscr{P})$)**.** *The relative condition number associated with problem $(\mathscr{P})$ is*

$$K(d) = \sup_{\delta d \neq 0} \frac{\|\delta x\|_V / \|x\|_V}{\|\delta d\|_{\mathscr{D}} / \|d\|_{\mathscr{D}}} \qquad (d+\delta d) \in \mathscr{D}. \tag{1.2}$$

If $x = 0$ or $d = 0$, the relative condition number is replaced by the absolute condition number associated with problem ($\mathscr{P}$)

$$K_{abs}(d) = \sup_{\delta d \neq 0} \frac{\|\delta x\|_V}{\|\delta d\|_{\mathscr{D}}} \qquad (d + \delta d) \in \mathscr{D}. \tag{1.3}$$

Roughly speaking, we say that ($\mathscr{P}$) is "well-conditioned" whenever $K(d)$ is "small", while ($\mathscr{P}$) is "ill-conditioned" whenever $K(d)$ is "large".

Example 1.1.8 (Condition number of a matrix A). In this example, we compute an estimate of the relative condition number in the case of Ex. 1.1.2. We use the p-norm (A.1) for $\mathbf{x}$, $\delta\mathbf{x}$, $\mathbf{b}$ and $\delta\mathbf{b}$, and the induced p-norm (A.3) for A and δA. Moreover, we assume that $\delta A = 0$, so that only the right-hand side $\mathbf{b}$ is subject to a perturbation $\delta\mathbf{b}$. Then, relation (1.2) yields

$$\begin{aligned} K(d) &= \sup_{\delta\mathbf{b}\neq\mathbf{0}} \frac{\|\delta\mathbf{x}\|_p/\|\mathbf{x}\|_p}{\|\delta\mathbf{b}\|_p/\|\mathbf{b}\|_p} = \sup_{\delta\mathbf{b}\neq\mathbf{0}} \frac{\|A^{-1}\delta\mathbf{b}\|_p}{\|\mathbf{x}\|_p} \frac{\|A\mathbf{x}\|_p}{\|\delta\mathbf{b}\|_p} \\ &\leq \frac{\|A^{-1}\|_p\|\delta\mathbf{b}\|_p}{\|\mathbf{x}\|_p} \frac{\|A\|_p\|\mathbf{x}\|_p}{\|\delta\mathbf{b}\|_p} \equiv K_p(A) \end{aligned}$$

where

$$K_p(A) := \|A\|_p\|A^{-1}\|_p \tag{1.4}$$

is the p-condition number of matrix A. It is easy to see that

$$K_p(A) \geq 1$$

and that $K_p(I) = 1$, I being the identity matrix of order n. By definition, $K_p(A) = +\infty$ if A is singular. In the important case where A is symmetric and positive definite, we have

$$\begin{aligned} K_2(A) &= \|A\|_2\|A^{-1}\|_2 = \sqrt{\rho(AA^T)}\sqrt{\rho(A^{-1}A^{-T})} \\ &= \rho(A)\rho(A^{-1}) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \end{aligned}$$

where $\lambda_{max}(A)$ and $\lambda_{min}(A)$ are the extreme eigenvalues of A and $\rho(A)$ is the spectral radius of A (see (A.4)).

1.2 The numerical model

Solving explicitly the continuous problem $(\mathscr{P})$ is in general a difficult, if not even impossible, task. For this reason, we introduce *a family of numerical problems depending on the discretization parameter* $h > 0$:

$$(\mathscr{P})_h \qquad \begin{cases} \text{given } d_h \in \mathscr{D}_h, \qquad \text{find } x_h \in V_h \text{ such that} \\ F(x_h, d_h) = 0 \end{cases}$$

where d_h and x_h are the *approximate* data and solution, to be sought in suitable subspaces $\mathscr{D}_h \subseteq \mathscr{D}$ and $V_h \subseteq V$. The main difference between $(\mathscr{P})_h$ and its continuous counterpart $(\mathscr{P})$ is that the former is constructed in such a way that x_h is a *computable* quantity, unlike x. For this reason, we refer to $(\mathscr{P})_h$ as the *numerical method*. Our request is that

$$\lim_{h\to 0} x_h = x, \tag{1.5}$$

that is, we want the numerical solution *to converge* to the solution of the continuous problem, as the discretization parameter becomes arbitrarily small. In order convergence to occur, we necessarily need that:

1. the approximate data to be convergent, i.e.

$$\lim_{h\to 0} d_h = d;$$

2. $(\mathscr{P})_h$ to be *consistent*, i.e.

$$\lim_{h\to 0} \underbrace{(F_h(x,d) - F(x,d))}_{:=\mathscr{R}_h(x,d)} = 0 \tag{1.6}$$

   $\mathscr{R}_h(x,d)$ being the *residual* of problem $(\mathscr{P})_h$, obtained by forcing into the numerical problem the solution and data of the continuous problem.

Example 1.2.1 (Numerical quadrature). For $f \in C^0([a,b])$, let $x = I_f$, $x_h = I_{f,h} \simeq I_f$ and

$$F_h(x_h, d_h) = I_{f,h} - x_h := \left(\sum_{k=1}^{N_h} h f(\bar{t}_k) \right) - x_h$$

where $N_h \geq 1$ is the number of subintervals in which $[a,b]$ is divided, each of width $h = (b-a)/N_h$, while $\bar{t}_k := (t_{k-1} + t_k)/2$, $k = 1, \ldots, N_h$, is the midpoint of each

subinterval. Notice that $h \to 0 \Leftrightarrow N_h \to \infty$, but the product hN_h remains constant and equal to $b-a$. The approximate formula $I_{f,h}$ to compute I_f is known as the composite midpoint quadrature rule and its geometrical representation is given in Fig. 1.2.

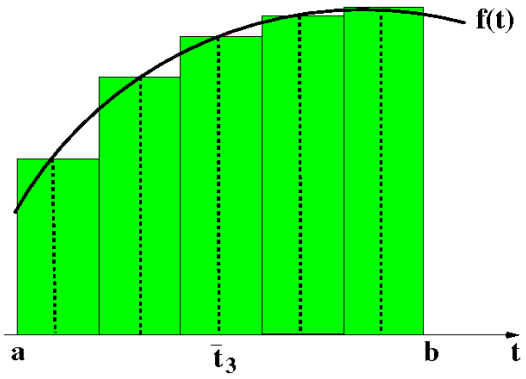


Figure 1.2: The composite midpoint quadrature rule. The shaded scaloid is the approximation of I_f.

Let us check that $(\mathscr{P})_h$ is consistent. Assuming $f \in C^2([a,b])$, it can be shown that

$$\begin{aligned} \mathscr{R}_h(x,d) &= \left(\sum_{k=1}^{N_h} hf(\bar{t}_k)\right) - x - \left(\int_a^b f(t)\,dt - x\right) \\ &= \left(\sum_{k=1}^{N_h} hf(\bar{t}_k)\right) - \int_a^b f(t)\,dt = \frac{b-a}{24}h^2 f''(\xi) \end{aligned}$$

where $\xi \in (a,b)$, so that (1.6) is satisfied. As for convergence, we notice that the midpoint quadrature rule $I_{f,h}$ coincides with the average Riemann sum, so that, by definition of integral of a function between a and b we have

$$\lim_{h\to 0} x_h = I_f = x \qquad \forall f \in C^0([a,b])$$

and thus x_h converges to x.

The following general result is the milestone of Numerical Analysis.

Theorem 1.2.2 (Equivalence theorem (Lax-Richtmyer))**.** *The numerical method $(\mathscr{P})_h$ is convergent iff it is consistent and stable. Consistency is expressed by* (1.6)*, while stability is expressed by* (1.1) *provided to replace d, x, $\mathscr{D}$ and V with d_h, x_h, $\mathscr{D}_h$ and V_h.*

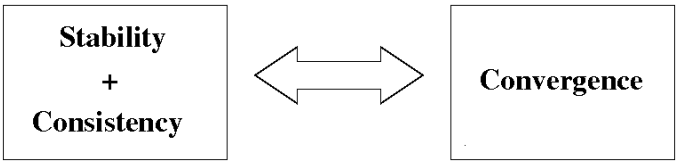


Figure 1.3: The Lax-Richtmyer paradigm.

1.3 The chain of errors

So far, we have introduced the notion of error as that of being, in general, the difference between the solution of the continuous problem, which we call from now on, the *exact solution*, and the solution of the numerical problem, which we call the *approximate, or, numerical solution*. As a matter of fact, our definition of error is not completely precise, because we are neglecting, at least, two other important sources of error, namely, the so-called *modeling error* and *rounding error*.

The modeling error is associated with a possible inaccuracy of the mathematical model describing the real physical application at hand, and/or a possible inaccuracy in the data entering the model formulation (due, for instance, to measurement machine tolerances or statistical fluctuations of the phenomena under investigation). As such, the modeling error is somewhat outside our possibility of control and, consequently, reduction.

The rounding error is, instead, inherently related to the computational process that is implemented in a computer algorithm, i.e., a sequence of deterministic machine operations that are run using a software environment (Matlab) on a PC characterized by a specific hardware (Intel Pentium IV processor) and a specific OS (Linux, Windows 7). Depending on the machine arithmetic being used, the rounding error can be monitored and accurately estimated, but not completely eliminated.

In mathematical terms, we have:

- $e_m := x_{ph} - x$: modeling error. It is the difference between the solution x_{ph} of the real physical problem and the solution x of the mathematical model;
- $e_h := x - x_h$: numerical error. It is the difference between the solution x of the mathematical model and the solution x_h of the numerical problem;
- $e_r := \widehat{x}_h - x_h$: rounding error. It is the difference between the solution x_h of the numerical model and the solution $\widehat{x}_h$ that is *actually* produced by the

computational process.

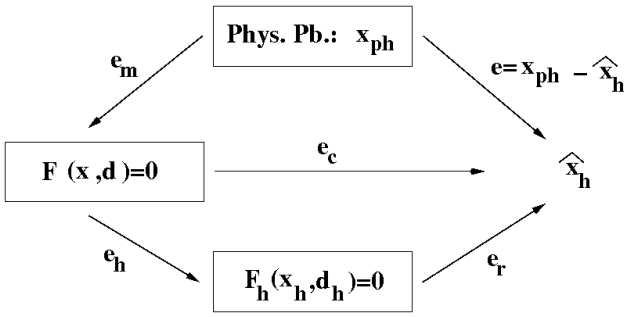


Figure 1.4: Sources of error in a computational process.

Letting $e_c := e_h + e_r$ the *computational error*, we define the *global error* as

$$e = x_{ph} - \widehat{x}_h = \underbrace{x_{ph} - x}_{e_m} + \underbrace{x - \widehat{x}_h}_{e_c} = x_{ph} - x + (\underbrace{x - x_h}_{e_h} + \underbrace{x_h - \widehat{x}_h}_{e_r}). \tag{1.7}$$

The whole chain of errors is pictorially represented in Fig. 1.4. It is important to notice that the *only* quantity that is available to the user is $\widehat{x}_h$; all the other solutions of the various intermediate steps of the process are virtually unaccessible. In particular, throughout the remainder of this text, we shall always neglect the modeling error in the analysis of the numerical methods of processes. Moreover, for ease of notation, we shall write x_h instead of $\widehat{x}_h$, unless otherwise specified, keeping the presence of the rounding error always understood.

1.4 Errors and error analysis

Definition 1.4.1 (Absolute and relative errors)**.** *For an appropriate norm $\|\cdot\|$, we have*

$$\begin{aligned} E_{abs}(x_h) &:= \|x - x_h\|, && \textit{absolute error} \\ E_{rel}(x_h) &:= \frac{\|x - x_h\|}{\|x\|}, && \textit{relative error}, \qquad x \neq 0. \end{aligned} \tag{1.8}$$

Definition 1.4.2 (Order of convergence)**.** *We say that $\widehat{x}_h$ converges to x with order $p > 0$ with respect to the discretization parameter h, if $\exists$ a positive constant C independent of h, such that*

$$e_c \leq Ch^p. \tag{1.9}$$

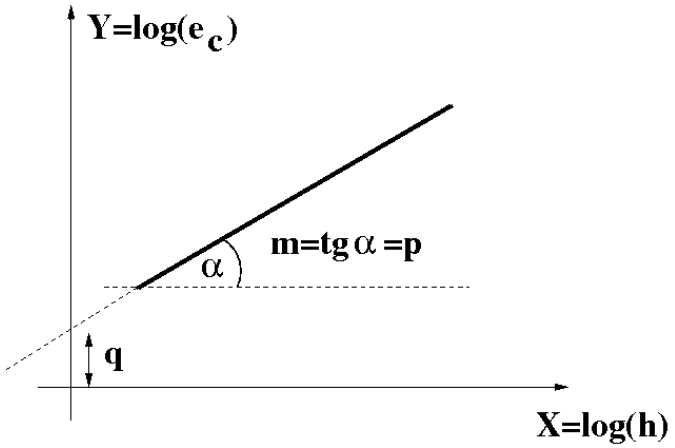


Figure 1.5: Error in log-log scale.

Remark 1.4.3. Using logarithmic scale, we have

$$\underbrace{\log(e_c)}_{Y} \simeq \underbrace{\log(C)}_{q} + \underbrace{p}_{m}\underbrace{\log(x)}_{X}$$

that is, the log-plot of the error is a straight line in the X-Y plane, whose slope m gives us immediately the order of convergence of the method. In the case of Ex. 1.2.1, $p = 2$ and $C = \dfrac{b-a}{24} \max\limits_{t\in[a,b]} |f''(t)|$.

1.5 Machine representation of numbers and rounding error

When working on a computer machine, any real number x is *represented* in the hardware by another real number, called *floating-point number* and denoted by $fl(x)$, equal to

$$fl(x) = (-1)^s \cdot (0.\underbrace{a_1 a_2 \dots a_t}_{m})\beta^e \tag{1.10}$$

where:

- β is the machine base;
- t is the number of significant digits;

- m is the mantissa;
- e is the exponent;
- s is the "sign" bit, with $s = 0$ if $x \geq 0$ and $s = 1$ if $x < 0$.

Typically, we have $\beta = 2$ and $L \leq e \leq U$, with $L < 0$ and $U > 0$. Using double precision for number representation (which means 8 bytes of memory, corresponding to a machine word of 64 bits for storing $fl(x)$), we have usually $t = 53$, $L = -1021$ and $U = 1024$. From (1.10), it turns out that the range of numbers that can be represented in a computer machine is *finite*, and in particular it can be seen that

$$|fl(x)| \in [\underbrace{\beta^{L-1}}_{x_{min}}, \underbrace{\beta^{U}(1-\beta^{-t})}_{x_{max}}].$$

Matlab coding. The quantities x_{min} and x_{max} are stored in the `Matlab` variables `realmin` and `realmax`.

```
>> realmin, realmax

ans =

  2.2251e-308


ans =

  1.7977e+308
```

Floating-point numbers are distributed in a discrete manner along the real axis. In particular, the power of resolution of a computer machine is characterized by the following quantity.

Definition 1.5.1 (Machine epsilon)**.** *The machine epsilon $\varepsilon_M = \beta^{1-t}$ is the smallest floating point number such that*

$$1 + \varepsilon_M > 1.$$

Matlab coding. The quantity ε_M is stored in the `Matlab` variable `eps`.

```
>> eps

ans =

   2.2204e-16
```

Remark 1.5.2 (Computing `realmax` in `Matlab`). Using the definition of ε_M into the definition of x_{max}, we can write this latter quantity in the following equivalent manner

$$x_{max} = \beta^U(1-\beta^{-t}) = \beta^U(1-\varepsilon_M\beta^{-1}) = \beta^{U-1}(\beta-\varepsilon_M).$$

This expression is implemented in the `Matlab` environment to compute the variable `realmax` without occurring into overflow problems.

In general, $fl(x)$ does not coincide with x, as stated by the following result.

Proposition 1.5.3 (Round-off error)**.** *Let $x \in \mathbb{R}$ be a given number. If $x_{min} \leq |x| \leq x_{max}$, then we have*

$$fl(x) = x(1+\delta) \qquad \textit{with } |\delta| \leq \mathtt{u}, \tag{1.11}$$

where

$$\mathtt{u} = \frac{1}{2}\beta^{1-t} \equiv \frac{1}{2}\varepsilon_M \tag{1.12}$$

is the roundoff unit (or machine precision).

Using (1.11) into the definitions (1.8), we get an estimate of the rounding error e_r introduced in Sect. 1.3

$$E_{rel}(x) = \frac{|x-fl(x)|}{|x|} = |\delta| \leq \mathtt{u}, \qquad E_{abs}(x) = |x-fl(x)| \leq \frac{1}{2}\beta^{e-t}.$$

Chapter 2

Essentials of Numerical Linear Algebra

Abstract

In this chapter, we review the basic foundations of Numerical Linear Algebra that will be used thoroughly in the remainder of the text. Solution methods for linear systems, including direct and iterative methods, will be illustrated.

2.1 Linear algebraic systems

The mathematical problem of the solution of a linear algebraic system consists of finding $\mathbf{x} \in \mathbb{R}^n$ such that

$$A\mathbf{x} = \mathbf{b} \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a given real-valued matrix and $\mathbf{b} \in \mathbb{R}^n$ is a given right-hand side vector.

Theorem 2.1.1. *Problem* (2.1) *admits a unique solution iff A is nonsingular. In such a case, Cramer's rule can be applied to yield*

$$x_i = \frac{\det(A_i)}{\det(A)} \qquad i = 1, \ldots, n. \tag{2.2}$$

Apparently, we are very satisfied with the approach to problem (2.1). Under a reasonable condition (matrix invertibility), there is an *explicit* formula to compute

the solution $\mathbf{x}$. A closer look at the situation shows that this is not completely true. As a matter of fact, the application of formula (2.2) requires to evaluate $(n+1)$ determinants. Assuming to work with a machine capable of performing 10^{11} floating-point operations (flops) per second (100Gflops/sec), the cost of Cramer's rule is of about 5 minutes of CPU time if $n = 15$, 16 years if $n = 20$ and 10^{141} years if $n = 100$. More generally, the asymptotical cost as a function of matrix size n increases as $(n+1)!$ (in mathematical terms, $\mathcal{O}((n+1)!)$, see Fig. 2.1).

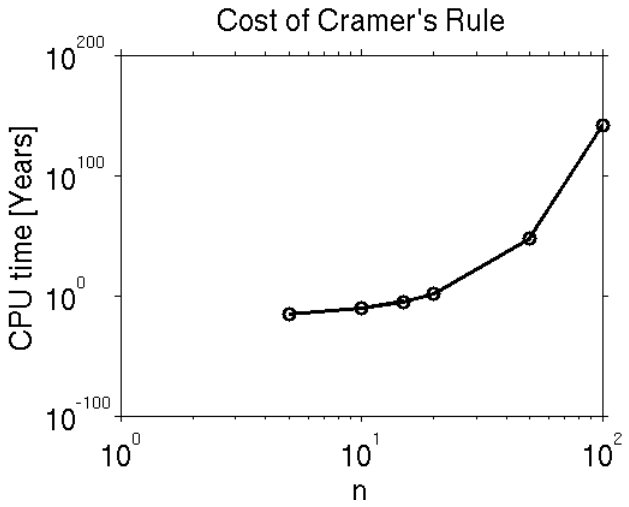


Figure 2.1: Cost of Cramer's rule (CPU time in years) as a function of n on a machine performing 100Gflops/sec.

Matlab coding. The following `Matlab` script can be used to compute the cost of Cramer's rule as a function of n.

```
SecYear=365*24*3600;
n=[5, 10, 15, 20, 50, 100];
Clock=100*1e9;
Cost=factorial((n+1))/Clock;
CostYear=Cost/SecYear;
loglog(n, CostYear,'ko-')
xlabel('n'); ylabel('CPU Time [Years]'); title('Cost of Cramer''s Rule')
```

The computational effort required by Cramer's rule is clearly not affordable, so that a remedy is urgently in order. Numerical linear algebra comes to rescue, providing two main kinds of techniques for working around the problem associated with Cramer's rule: direct methods and iterative methods. In essence, direct methods compute the solution of (2.1) in a finite number of steps, while iterative

methods compute the solution of (2.1) as the limit of a sequence, therefore requiring (theoretically) an infinite number of steps. The first approach is discussed in Sect. 2.2.

2.2 Direct methods for linear systems

Assume that there exist a lower triangular matrix L and an upper triangular U, such that

$$A = L \cdot U. \tag{2.3}$$

Relation (2.3) is referred to as the *LU factorization* of A.

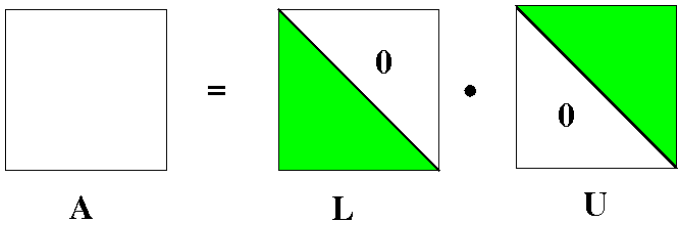


Figure 2.2: LU factorization of a matrix.

Since

$$\det(A) = \det(L) \cdot \det(U)$$

it immediately turns out that $l_{ii} \neq 0$ and $u_{ii} \neq 0$, $i = 1, \ldots, n$, because A is nonsingular.

The LU factorization can be used as a solution method for (2.1) as follows. From (2.3) we have

$$(L \cdot U)\mathbf{x} = \mathbf{b}$$

so that solving (2.1) amounts to solving the two linear triangular systems:

$$L\mathbf{y} = \mathbf{b} \tag{2.4a}$$

$$U\mathbf{x} = \mathbf{y} \tag{2.4b}$$

Matlab coding. The `Matlab` coding of (2.4a) is reported below and takes the name of *forward substitution algorithm*.

```
function y = forward_substitution(L,b)
n = length(b);
y = zeros(n,1);
y(1) = b(1)/L(1,1);
for i=2:n
    y(i) = (b(i)-L(i,1:i-1)*y(1:i-1))/L(i,i);
end
```

Matlab coding. The `Matlab` coding of (2.4b) is reported below and takes the name of *backward substitution algorithm*.

```
function x = backward_substitution(U,y)
n = length(y);
x = zeros(n,1);
x(n) = y(n)/U(n,n);
for i=n-1:-1:1
    x(i) = (y(i)-U(i,i+1:n)*x(i+1:n))/U(i,i);
end
```

Both forward and backward substitution algorithms have a computational cost of n^2 flops. To compute the triangular matrices L and U we need to solve the nonlinear system (2.3) for the $2(n^2-(n^2-n)/2) = n^2+n$ unknown coefficients l_{ij} and u_{ij}, which reads

$$\sum_{k=1}^{k} l_{ik}u_{kj} = a_{ij} \qquad i,j = 1,\dots,n \tag{2.5}$$

However, the number of equations (2.5) is only n^2, so that we need n additional equations to close the problem. These latter are found by enforcing the conditions

$$l_{ii} = 1 \qquad i = 1,\dots,n. \tag{2.6}$$

Then, the remaining coefficients can be efficiently computed using the Gauss algorithm, with a cost of $\mathscr{O}(2n^3/3)$. Summarizing, the total cost of the solution of the linear system (2.1) using the method of LU factorization of matrix A is equal to $2n^3/3+2n^2$.

We can draw two relevant conclusions from this result. The first conclusion is that the LU factorization is a computationally efficient alternative to Cramer's rule. The second conclusion is that, as n increases, the cost of forward and backward system solving becomes less important than the cost of the factorization itself.

Matlab coding. The `Matlab` coding of the solution of the nonlinear system (2.5) through Gauss algotithm is reported below.

```
function [L,U] = lufact(A,n)
for k=1:n-1
    W(k+1:n)=A(k,k+1:n);
```

```
    for i=k+1:n
        A(i,k)=A(i,k)/A(k,k);
        for j=k+1:n
            A(i,j)=A(i,j)-A(i,k)*W(j);
        end
    end
end
L=tril(A,-1)+eye(n);
U=triu(A);
```

Example 2.2.1. Let us solve (2.1) using the LU factorization in the case where A is the "magic" matrix of order $n = 3$ (a matrix constructed from the integers 1 to n^2 with equal row, column, and diagonal sums), and $\mathbf{b}$ is chosen in such a way that $\mathbf{x} = [1,1,1]^T$.

Matlab coding. The sequence of `Matlab` commands for the present example is reported below.

```
>> A=magic(3)

A =

     8     1     6
     3     5     7
     4     9     2

>> b=A*ones(3,1)

b =

    15
    15
    15

>> [L,U]=lufact(A,3)

L =

    1.0000         0         0
    0.3750    1.0000         0
    0.5000    1.8378    1.0000


U =

    8.0000    1.0000    6.0000
         0    4.6250    4.7500
         0         0   -9.7297

>> y=forward_substitution(L,b)

y =

   15.0000
    9.3750
```

```
   -9.7297

>> x=backward_substitution(U,y)

x =

     1
     1
     1
```

The following result gives a necessary and sufficient condition for the existence and uniqueness of the LU factorization.

Theorem 2.2.2. *Given a matrix $A \in \mathbb{R}^{n\times n}$, its LU factorization with $l_{ii} = 1$, $i = 1,\ldots,n$ exists and is unique iff* $\det(A_i) \neq 0$ *for* $i = 1,\ldots,n-1$.

The previous theorem is of little practical use. The following sufficient conditions are more helpful.

Proposition 2.2.3. *If A is s.p.d. or if A is diagonally dominant, then its LU factorization with $l_{ii} = 1$, $i = 1,\ldots,n$ exists and is unique. In the first case, we have $L = U^T$ (i.e., the factorization is symmetric) and its cost is $\mathcal{O}(n^3/3)$ instead of $\mathcal{O}(2n^3/3)$ (reduced of one half). The LU factorization takes the name of Cholesky factorization.*

It is easy to find cases for which the conditions of Thm. 2.2.2 are not satisfied.

Example 2.2.4. Let

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix}.$$

We have $\det(A_1) = 1$, $\det(A_2) = 0$ and $\det(A_3) = \det(A) = -6$, so that A is nonsingular but fails to satisfy condition $\det(A_2) \neq 0$ that is required to admit the LU factorization.

To accomodate the inconvenience manifested by the previous example, the LU factorization is modified as follows. Instead of (2.3), we assume that there exist a lower triangular matrix L, an upper triangular U and a permutation matrix P, such that

$$P \cdot A = L \cdot U. \tag{2.7}$$

Relation (2.7) is referred to as the LU factorization of *A with pivoting*. The role of matrix P is just to perform a suitable exchange of certain rows of A in such a way

that *all* the principal minors up to order $n-1$ of $P \cdot A$ turn out to be non-singular, as required by Thm. 2.2.2. Going back to Ex. 2.2.4, the choice

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

has the effect of transforming A into

$$P \cdot A = \begin{bmatrix} 7 & 8 & 9 \\ 2 & 4 & 5 \\ 1 & 2 & 3 \end{bmatrix} \equiv \widetilde{A}$$

that is, the first and third rows have been exchanged. Now, we have $\det(\widetilde{A}_1) = 7$ and $\det(\widetilde{A}_2) = 12$ so that $\widetilde{A}$ satisfies the conditions required by Thm. 2.2.2 and thus admits a unique LU factorization (notice that $\det(\widetilde{A}_3) = \det(\widetilde{A}) = \det(P)\det(A) = +6$, so that also $\widetilde{A}$ is nonsingular as A was). According to (2.7), we have

$$(L \cdot U)\mathbf{x} = P \cdot \mathbf{b}$$

so that solving (2.1) amounts to solving the two linear triangular systems:

$$L\mathbf{y} = P \cdot \mathbf{b} \tag{2.8a}$$

$$U\mathbf{x} = \mathbf{y}. \tag{2.8b}$$

In other words, everything remains the same as in the case of LU factorization, except the fact that the right-hand-side is subject to a reordering which corresponds to the exchange of rows of A during factorization through the Gauss algorithm.

Matlab coding. The `Matlab` command for computing L, U and P is reported below.

```
>> [L,U,P]=lu(A);
```

More synthetically, the solution of (2.1) using the LU factorization with pivoting can be obtained in `Matlab` with the following command.

```
>> x = A \ b;
```

Remark 2.2.5 (Inverse of a matrix). The LU factorization with pivoting (2.7) and the triangular systems (2.8a)- (2.8b) can be used as an efficient tool for computing the inverse of a given matrix by solving the n systems

$$A\mathbf{x}_i = \mathbf{e}_i \qquad i = 1, \ldots, n$$

where the i-th unknown vector $\mathbf{x}_i$ represents the i-th column of A^{-1}.

2.3 Stability analysis

In this section, we study the effect of round-off error in the numerical solution of problem (2.1).

Theorem 2.3.1. *Let δA and $\delta\mathbf{b}$ denote two perturbations of the problem data (A and $\mathbf{b}$, respectively), so that the perturbed system associated with* (2.1)

$$(A+\delta A)(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b}+\delta\mathbf{b} \tag{2.9}$$

is still uniquely solvable. Then, we have

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{K(A)}{1-K(A)\dfrac{|||\delta A|||}{|||A|||}}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{|||\delta A|||}{|||A|||}\right), \tag{2.10}$$

for a matrix norm $|||\cdot|||$ induced by a vector norm $\|\cdot\|$ and where $K(A) := |||A|||||A^{-1}|||$.

The estimate (2.10) tells us that the relative error on the exact solution is bounded by the sum of the relative errors on the data, modulo the effect of finite machine precision which is represented by the amplification constant

$$\frac{K(A)}{1-K(A)\dfrac{|||\delta A|||}{|||A|||}}.$$

Assuming for simplicity $\delta A = 0$, we see that if A is ill-conditioned ($K(A) \gg 1$), then the solution $\widehat{\mathbf{x}}_h$ that is *actually* produced by the computational process can be affected by a serious inaccuracy. A refinement of (2.10) is provided by the following result.

Theorem 2.3.2. *Assume that $|||\delta A||| \leq \gamma|||A|||$ and $\|\delta\mathbf{b}\| \leq \gamma\|\mathbf{b}\|$, for a positive constant $\gamma = \mathcal{O}(\mathtt{u})$, and that $\gamma K(A) < 1$. Then, we have*

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{2\gamma}{1-\gamma K(A)}K(A). \tag{2.11}$$

The estimate (2.11) tells us that the round-off error which inevitably affects every operation on a machine computer (cf.Sect. 1.5) is amplified by the condition number of matrix A. To obtain a more quantitative information on the error, take

$\gamma = \mathtt{u} \simeq \beta^{1-t}$, $\beta = 10$, $t = 16$ and $\|\cdot\| = \|\cdot\|_\infty$. Assume also $K_\infty(A) = 10^m$, for a nonnegative integer m, with $\gamma K_\infty(A) \leq 1/2$. Then, (2.11) yields

$$\frac{\|\delta \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \simeq \mathtt{u} K_\infty(A) = 10^{m-16}. \tag{2.12}$$

Roughly speaking, the above estimate tells us that the expected accuracy of $\widehat{\mathbf{x}}_h$ with respect to the solution $\mathbf{x}$ of (2.1) is, at least, of $16-m$ exact decimal digits.

Example 2.3.3 (Hilbert matrix). In this example, we study the disastrous effects of the combined occurrence of matrix ill-conditioning *and* machine round-off in the solution of a linear system. We warn the reader to consider this case not as a general paradigm, rather as a warning against "blind-faith" computations.

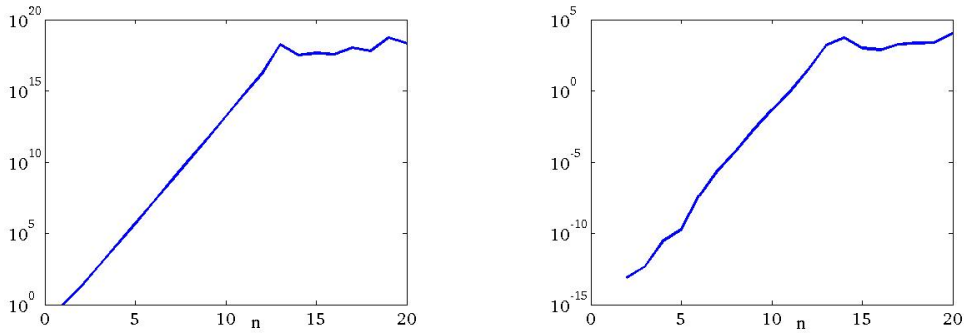
For $n \geq 1$, let $A \in \mathbb{R}^{n\times n}$ be the Hilbert matrix of order n, defined as

$$a_{ij} = \frac{1}{i+j-1} \qquad i,j = 1,\ldots,n.$$

Matlab coding. Matrix A is symmetric and positive definite for every n. To examine the stability with respect to perturbations in the solution of the linear system $A\mathbf{x} = \mathbf{b}$, we compute with the following `Matlab` script the condition number of A as a function of n.

```
n=[1:20];
for i=1:numel(n), A=hilb(n(i)); K(i)=cond(A); end
semilogy(n, K)
xlabel('n')
```

Fig. 2.3(a) shows that $K(d)$ grows exponentially as a function of n, in such a way that, even for a matrix of small size ($n = 13$), the value of K is exceedingly large (10^{15} and more). According to estimate (2.12), the expected number of exact digits becomes null, making the solution on the computer of the linear system irremediably useless.



(a) $K(d)$ (b) Percentual error

Figure 2.3: Log-plot of the condition number of the Hilbert matrix of order n as a function of n.

Chapter 3

Essentials of Numerical Approximation

Abstract

In this chapter, we review the basic elements of approximation theory of a given continuous function $f : \overline{\Omega} = [a,b] \to \mathbb{R}$ using piecewise polynomials of degree $r \geq 1$.

3.1 Interpolation

Let us introduce a partition $\mathscr{T}_h$ of $\overline{\Omega}$ into a number $M_h \geq 1$ of 1-simplex (intervals) $K_i = [x_i, x_{i+1}]$, $i = 1, \ldots, M_h$, in such a way that $x_1 := a$ and $x_{M_h+1} := b$. We denote by $h_i := x_{i+1} - x_i$ the length of each interval and set $h := \max_{\mathscr{T}_h} h_i$. The partition $\mathscr{T}_h$ takes the name of *triangulation* of the domain Ω. Each K_i is an *element* of the triangulation, while the quantities x_i, $i = 1, \ldots, M_h + 1$ are the *vertices* of the triangulation. The same terminology is adopted when Ω is a bounded set of $\mathbb{R}^d$, $d \geq 2$, and in such a case 2-simplices are triangular elements ($d = 2$) while 3-simplices are tetrahedral elements ($d = 3$). We associate with $\mathscr{T}_h$ the following space of functions

$$X_h^r(\mathscr{T}_h) := \left\{ v_h \in C^0(\overline{\Omega}) : \; v_h|_{K_i} \in \mathbb{P}_r(K_i) \; \forall K_i \in \mathscr{T}_h \right\}. \tag{3.1}$$

Unless strictly necessary, we write X_h^r instead of $X_h^r(\mathscr{T}_h)$. The space X_h^r is the set of piecewise continuous polynomials over $\overline{\Omega}$, of degree $\leq r$ over each element K_i of

$\mathcal{T}_h$, $i = 1, \ldots, M_h$. X_h^r is called the *finite element space of degree r associated with* $\mathcal{T}_h$, and its dimension depends on both h and r. A count of degrees of freedom plus the continuity requirement at each internal vertex yields

$$\dim(X_h^r) = M_h(r+1) - (M_h + 1 - 2) = rM_h + 1 \equiv N_h.$$

Fig. 3.1 shows an example of $\mathcal{T}_h$ (with $M_h = 4$) and of a function v_h of the finite element space in the cases $r = 1$ and $r = 2$.

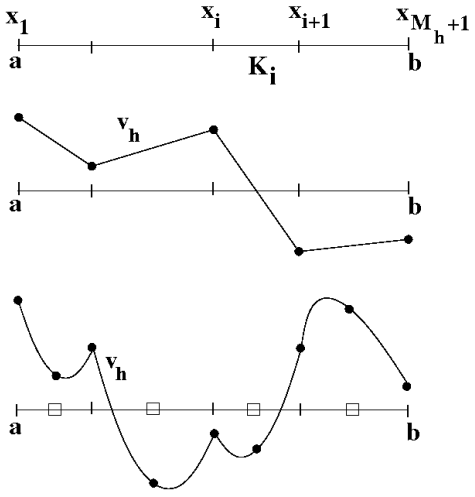


Figure 3.1: Triangulation in 1D and finite element functions ($r = 1$ and $r = 2$). Black bullets denote nodal values, while black ticks and white squares are the positions of the degrees of freedom of X_h^r. In the case $r = 1$ the vertices and the nodes coincide, while in the case $r = 2$ the nodes are more than the vertices.

Remark 3.1.1. We notice that it is necessary to introduce an increasingly larger number of degrees of freedom when the polynomial degree becomes larger than 1. In general, for a given r, we need $r - 1$ internal nodes for each element in order to construct the restriction of the finite element function $v_h|_{K_i}$. For example, in the case $r = 2$, we can choose as internal node the midpoint of K_i for each $i = 1, \ldots, M_h$, and so on, for $r \geq 3$.

3.1.1 Basis functions

By definition of vector space, any function $v_h \in X_h^r$ can be written in the following form

$$v_h(x) = \sum_{j=1}^{N_h} v_j \varphi_j(x) \tag{3.2}$$

where:

- φ_j, $j = 1, \ldots, N_h$: basis functions of V_h;
- v_j: degrees of freedom of v_h, that is, the coordinates of v_h with respect to the basis $\{\varphi_j\}$.

A particularly interesting choice for the basis of X_h^r are the so-called *Lagrangian basis functions*, such that

$$\varphi_i(x_j) = \delta_{ij} \qquad i, j = 1, \ldots, N_h. \tag{3.3}$$

This property allows us to interpret the quantities v_i as the *nodal values* of v_h at the nodes of $\mathcal{T}_h$, i.e.

$$v_h(x_i) = \sum_{j=1}^{N_h} v_j \varphi_j(x) = \sum_{j=1}^{N_h} v_j \delta_{ji} = v_i \qquad i = 1, \ldots, N_h.$$

Example 3.1.2 (Basis for $r = 1$ and $r = 2$). In the case $r = 1$, two basis functions associated with the triangulation $\mathcal{T}_h$ are plotted in Fig. 3.2 (top), while Fig. 3.2 (bottom) refers to the case $r = 2$.

Remark 3.1.3 (Support of basis functions). For any basis function φ_i, we define the *support* of φ_i as the subset of $\overline{\Omega}$ on which the function is nonvanishing. We notice that in the case $r = 1$ the support of a basis function associated with an internal node x_i is made by the union of K_{i-1} and K_i. In the case $r = 2$, instead, we have two kinds of basis functions, those associated with the midpoint of each element (white square) and those associated with a vertex (black bullet). The first kind of function (called "bubble" function) is localized within the element and its support is given by K_i, the second kind of function has a support made by the union of K_{i-1} and K_i.

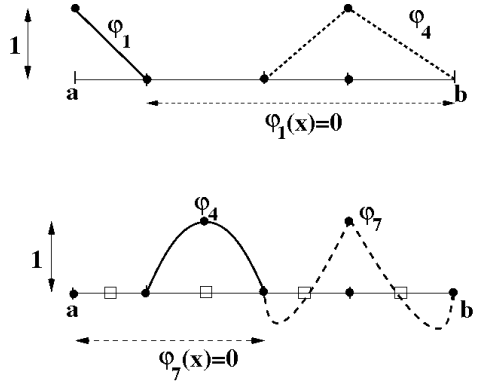


Figure 3.2: Basis functions. Top: $r=1$, bottom: $r=2$. In the first case, $\dim(X_h^r)=5$ while in the second case we have $\dim(X_h^r)=9$.

3.1.2 Finite element interpolation and error analysis

We are now ready to use the finite element basis functions of X_h^r to construct a piecewise polynomial approximation of a given continuous function f. Precisely, let us intoduce the *interpolation operator* $\Pi_h^r f : X_h^r \to \mathbb{R}$ defined as

$$\Pi_h^r f(x) := \sum_{j=1}^{N_h} f(x_j)\varphi_j(x) \tag{3.4}$$

and such that

$$\Pi_h^r f(x_i) := \sum_{j=1}^{N_h} f(x_j)\varphi_j(x_i) = \sum_{j=1}^{N_h} f(x_j)\delta_{ji} = f(x_i) \qquad i=1,\dots,N_h. \tag{3.5}$$

The N_h relations (3.5) are the *interpolation conditions* that allow to uniquely characterize the operator $\Pi_h^r f \in X_h^r$ associated with the given function f.

Example 3.1.4. Let $f(x)=\sin(x)$, $[a,b]=[0,10]$, $r=1$ and $M_h=10$.

Matlab coding. The following `Matlab` commands can be used to compute $\Pi_h^r f$ and to plot it together with the function f.

```
xn = 0:10;       yn = sin(xn);
xi = 0:.25:10;   yi = interp1(xn,yn,xi);
xx =0:0.0001:10; yy = sin(xx);
plot(xn,yn,'*',xi,yi,xx,yy)
xlabel('x');
legend('f(x_i)','\Pi_h^1 f(x)','f(x)')
```

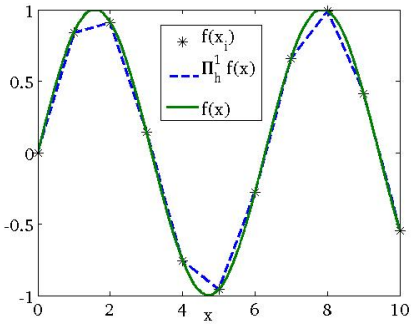


Figure 3.3: Piecewise linear continuous interpolation of the function $f(x) = \sin(x)$ over $[0, 10]$.

The *interpolation error* is defined as

$$E_h^r f(x) := f(x) - \Pi_h^r f(x). \tag{3.6}$$

By definition, we have

$$E_h^r f(x_i) = 0 \qquad i = 1, \ldots, N_h.$$

The next question is to characterize the accuracy of the piecewise polynomial approximation over *the whole domain* Ω as a function of h. For any continuous function $g = g(x)$, we define the maximum norm as (cf. (B.12) in the case $k = 0$)

$$\|g\|_\infty := \max_{x \in [a,b]} |g(x)|. \tag{3.7}$$

Theorem 3.1.5 (Interpolation error)**.** *Let us assume that $f \in C^{r+1}(\overline{\Omega})$, $r \geq 1$. Then, there exists a positive constant C independent of h such that*

$$\|E_h^r f\|_\infty \leq C h^{r+1} \|f^{(r+1)}\|_\infty, \tag{3.8}$$

where $f^{(s)}$ denotes the s-th derivative of f with respect to x, $s \geq 0$.

According to Def. 1.9, Thm. 3.1.5 tells us that $\Pi_h^r f$ converges uniformly to f with order $r+1$ with respect to the discretization parameter h, provided that f is sufficiently smooth. Moreover, the following corollary holds.

Corollary 3.1.6 (Interpolation error for the derivative of f)**.** *Under the same regularity assumption as in Thm. 3.1.5, there exists a positive constant C independent of h such that*

$$\|(E_h^r f)'\|_\infty \leq C h^r \|f^{(r+1)}\|_\infty. \tag{3.9}$$

This result tells us that the interpolation operator can be used also for the approximation of the derivative of a function f, but that, in such a case, the order of convergence is decreased by one with respect to the approximation of the function.

3.2 Quadrature

The interpolation polynomial can be profitably used for the approximate evaluation of the integral $I(f) := \int_a^b f(x)\,dx$, obtaining the following *quadrature formula*

$$I_h^r(f) = \int_a^b \Pi_h^r f(x)\,dx \simeq I(f). \tag{3.10}$$

As computing the integral of a polynomial is easy, the above relation gives an explicit formula for $I(f)$. For simplicity, we assume that the triangulation of $[a,b]$ is uniform, i.e., $h = (b-a)/M_h$. By doing so, depending on the polynomial degree r, formula (3.10) yields the so-called *Newton-Cotes quadrature rules*.

Example 3.2.1. The lowest-order case $r = 1$ corresponds to the trapezoidal rule. The approximate area computed by the quadrature rule is geometrically represented by the sum of the areas of the trapezoidal scaloids S_i, $i = 1,\ldots,M_h$ (see Fig. 3.4).

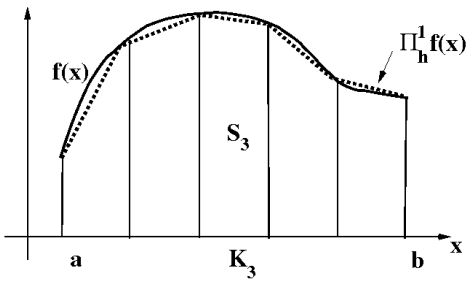


Figure 3.4: Trapezoidal quadrature rule.

In the case $r = 2$, the resulting formula is called Cavalieri-Simpson quadrature rule.

It is also interesting to consider the case of a zero-order approximation ($r = 0$). In such a case, the resulting formula is called midpoint quadrature and is geometrically represented in Fig. 1.2.

In view of the error analysis of (3.10), some definitions are in order.

Definition 3.2.2 (Quadrature error)**.** *If the error associated with a quadrature formula can be written in the form*

$$e \leq Ch^p \|f^{(q)}\|_\infty,$$

then we say that:

- *the error is* infinitesimal of order *p with respect to the discretization parameter h;*
- *the formula has a* degree of exactness *(or precision) equal to* $s := q-1$*. As a matter of fact, if* $f \in \mathbb{P}_{q-1}$ *(i.e., if f is a polynomial of degree* $\leq q-1$*), then* $f^{(q)}(x) = 0$ *for all* $x \in [a,b]$ *and* $e = 0$*. This means that the formula is* exact *if applied to polynomials of degree up to* $q-1$*.*

Theorem 3.2.3 (Quadrature error)**.** *Let* $f \in C^2([a,b])$*. Then*

$$\begin{aligned} |I(f) - I_h^0(f)| &\leq \frac{b-a}{24} h^2 \|f^{(2)}\|_\infty \quad \Rightarrow \quad p = 2, \quad s = 1 \\ |I(f) - I_h^1(f)| &\leq \frac{b-a}{12} h^2 \|f^{(2)}\|_\infty \quad \Rightarrow \quad p = 2, \quad s = 1. \end{aligned} \tag{3.11}$$

If $f \in C^4([a,b])$*, then*

$$|I(f) - I_h^2(f)| \leq \frac{b-a}{90 \cdot 2^5} h^4 \|f^{(4)}\|_\infty \quad \Rightarrow \quad p = 4, \quad s = 3. \tag{3.12}$$

Remark 3.2.4 (Gaussian quadratures). More in general, a quadrature formula for the approximate evaluation of $I(f)$ can be written as a weighted sum

$$I_h(f) = \sum_{i=1}^{N_h} w_i f(x_i)$$

where x_i and w_i are called *nodes* and *weights* of the considered quadrature formula. The optimal choice of such nodes and weights is a classical (and nontrivial) problem. If the number of nodes over each element K_i is a fixed quantity, say equal to $n+1$, for $n \geq 0$ (not too large), then it is possible to uniquely determine the nodes $x_j \in K_i$ in such a way that $s = 2n+1$, i.e., *maximum* degree of exactness. The resulting formulae are called *Gaussian quadrature rules*. An important property of Gaussian quadratures is that $w_i > 0$, $i = 1, \ldots, N$, which has the remarkable consequence to yield numerically stable formulae. This property does not hold for Newton-Cotes formulae of high order, for which some weights w_i turn out to be negative if $n > 5$.

Example 3.2.5. Two simple examples of Gaussian quadratures are those with:

- $n = 0$ (one quadrature node for each K_i), for which $N_h = M_h$, $p = 2$ and $s = 1$;
- $n = 1$ (two quadrature nodes for each K_i), for which $N_h = 2M_h$, $p = 4$ and $s = 3$.

Part II

Elliptic Problems

This part illustrates the weak formulation of elliptic model problems and their numerical approximation using the Galerkin Finite Element Method.

Chapter 4

Weak Formulation of Elliptic Boundary Value Problems

Abstract

In this chapter, we introduce the concept of weak solution of an elliptic boundary value problem and we illustrate the Lax-Milgram Lemma, a theoretical tool for the analysis of the well-posedness of the associated weak formulation.

4.1 Elliptic boundary value problems

Let Ω be a bounded set of $\mathbb{R}^d$, $d \leq 3$, with Lipschitz boundary $\partial\Omega$ on which a unit normal vector $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is defined almost everywhere, $\mathbf{x} = (x_1, \ldots, x_d)^T$ being the coordinate position vector. In this chapter, we start the mathematical study of an elliptic boundary value problem (BVP) of the form:
find $u = u(\mathbf{x}) : \Omega \to \mathbb{R}$ such that:

$$(P) \qquad \begin{cases} Lu = f & \text{in } \Omega \\ Bu = g & \text{on } \partial\Omega, \end{cases}$$

where L is the linear differential elliptic operator defined in (C.2), f is a given function and B is a linear operator associating the value of u and/or of its conormal derivative $\partial u/\partial n_L$ on the boundary $\partial\Omega$ with a given boundary datum g.

Example 4.1.1 (The Dirichlet problem for the Laplace operator). The simplest example of problem (P) corresponds to setting $A = 1$, $\mathbf{b} = \mathbf{0}$, $c = 0$, $Bu = u$ (identity

operator) and $g = 0$. The resulting elliptic problem is the so-called *Dirichlet BVP* associated with the Poisson equation with homogeneous boundary conditions:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

The BVP (4.1), in the 2D case ($d = 2$) represents in Mechanics the model of an elastic membrane clamped at the boundary and subject to the action of its own weight f. The solution $u(x,y)$ is the vertical displacement of any point $P = (x,y)^T$ of the membrane.

More in general, problem (4.1) is the mathematical model for a wide range of physical applications, including:

- Electrostatics and Magnetostatics: u is the electrostatic/magnetostatic potential, f is the space charge density;
- Thermal Physics: u is the spatial distribution of temperature in a body subject to a heat thermal source f;
- Hydraulics: u is the piezometric head in a porous medium subject to a piezometric load f.

Definition 4.1.2 (Classical solution)**.** *A classical (or strong) solution of the BVP (P) is any function $u \in C^2(\Omega)$ satisfying $(P)_1$ in the interior of Ω and the boundary conditions $(P)_2$ on $\partial\Omega$.*

In realistic applications (like those mentioned in Ex. 4.1.1), Def. 4.1.2 does not turn out to be the most appropriate, as shown in the following example.

Example 4.1.3 (Weak solution)*.* Let us consider the homogeneous Dirichlet BVP (4.1) in the case $d = 1$ where $\Omega = (0,1)$ and $f(x) = -P\delta(x - 1/2)$, P being a given positive constant and $\delta(x - x_0)$ is the Dirac function centered at $x = x_0$. The differential problem at hand represents the mathematical model of the transversal deformation of an elastic rod fixed at the endpoints and subject to a load applied at its center.

The deformed configuration, shown in Fig. 4.1, is certainly not a twice-differentiable function, rather, it is only piecewise linear and continuous, in such a way that its derivative is piecewise constant over $(0,1)$ and the second derivative is exactly the delta-function at $x = 1/2$ whose value is equal to $-P$. In this sense, comparing the smoothness of u in this case with that of a strong solution, we can speak, in a natural manner, of a "weak" solution.

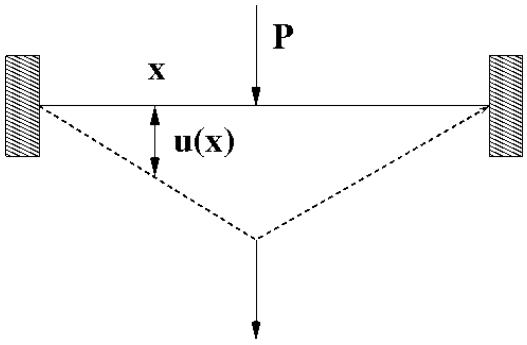


Figure 4.1: Deformed configuration of an elastic rod.

4.2 Weak solution of a BVP: the 1D case

To account for cases like that in Ex. 4.1.3, we obviously need to extend the notion of a solution of a BVP given in Def. 4.1.2. With this aim, we consider problem (4.1) in the 1D case ($d = 1$): find $u : \Omega = (0,1) \to \mathbb{R}$ such that:

$$\begin{cases} -u'' = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{4.2}$$

where f is a given function in $L^2(0,1)$. Then, we proceed by moltiplying both sides of the equation $(4.2)_1$ by an arbitrary non-vanishing function $v = v(x)$ and by integrating over $(0,1)$, obtaining

$$\int_0^1 -u''v\,dx = \int_0^1 fv\,dx \qquad \forall v.$$

We use the formula of integration by parts (C.1) applied to $w = u'$ to transform the above integral identity into

$$-\int_{\partial\Omega} vu'nd\sigma + \int_0^1 u'v'\,dx = \int_0^1 fv\,dx \qquad \forall v.$$

The boundary $\partial\Omega$ is the set made by the two end-points $x = 0$ and $x = 1$, while the outward unit normal vector n is given by $n(0) = -1$ and $n(1) = +1$, so that the previous identity becomes

$$-\big[v(1)u'(1) - v(0)u'(0)\big] + \int_0^1 u'v'\,dx = \int_0^1 fv\,dx \qquad \forall v.$$

To get rid of the boundary term in the previous relation, we choose v such that

$$v(0) = v(1) = 0$$

exactly as u does in $(4.2)_2$.

Theorem 4.2.1. *Let φ, ψ be two functions in $L^2(0,1)$, and set $g := \varphi \cdot \psi$. Then, $g \in L^1(0,1)$.*

Proof. It suffices to see that

$$\int_0^1 g\,dx \leq \left| \int_0^1 g\,dx \right|$$

and then to apply the Cauchy-Schwarz inequality (B.16) to $g \equiv \varphi \cdot \psi$. □

Using Thm. 4.2.1 we immediately see that the above described formal procedure can be synthetically written as:
find $u \in V$ such that

$$\int_0^1 u'v'\,dx = \int_0^1 fv\,dx \qquad \forall v \in V \tag{4.3}$$

where

$$V = H_0^1(\Omega) = \left\{ v \in L^2(0,1),\, v' \in L^2(0,1) \text{ such that } v(0) = v(1) = 0 \right\}. \tag{4.4}$$

Comparing (4.3)-(4.4) with the original BVP (4.2), we can make the following considerations:

1. relation (4.3) holds in an *integral sense* and not in a *pointwise sense*, as $(4.2)_1$;
2. however, to determine the solution of (4.3)-(4.4) we need to make an infinite choice of test functions $v \in V$;
3. the solution u of (4.3)-(4.4) has a lower differentiability requirements than a classical solution. For this reason, we qualify problem (4.3)-(4.4) as the *weak formulation* of the BVP.

Remark 4.2.2. To appreciate the extraordinary enlargement of the solution space from C_0^2 (classical solution) to H_0^1 (weak solution), it is useful to go back to Ex. B.6.5 and Fig. B.7.

Theorem 4.2.3 (Variational formulation of (4.2))**.** *The minimization problem: find $u \in V$ such that*

$$J(u) \leq J(v) \qquad \forall v \in V \tag{4.5}$$

where

$$J(v) := \int_0^1 \frac{1}{2}(v')^2\,dx - \int_0^1 fv\,dx \qquad \forall v \in V, \tag{4.6}$$

is completely equivalent to the weak problem (4.3)-(4.4).

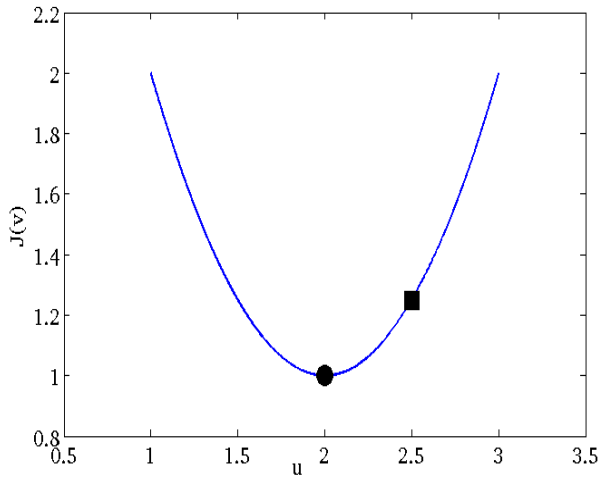


Figure 4.2: Minimization of the total potential energy. The black bullet is $J(u)$, while the black square is $J(v) \geq J(u)$ for all $v \in V$.

Proof. Let us prove that (4.3)-(4.4) implies (4.5). For each $\varepsilon \in \mathbb{R}$ and for any arbitrary function $v \in V$ we have

$$\begin{aligned} J(u+\varepsilon v) &= \int_0^1 \frac{1}{2}(u+\varepsilon v)')^2\,dx - \int_0^1 f(u+\varepsilon v)\,dx \\ &= \int_0^1 \frac{1}{2}(u')^2\,dx + \varepsilon \int_0^1 u'v'\,dx + \varepsilon^2 \int_0^1 \frac{1}{2}(v')^2\,dx \\ &- \int_0^1 fu\,dx - \varepsilon \int_0^1 fv\,dx \\ &= J(u) + \varepsilon \left[\int_0^1 u'v'\,dx - \int_0^1 fv\,dx \right] + \varepsilon^2 \int_0^1 \frac{1}{2}(v')^2\,dx. \end{aligned}$$

Then, if u is a solution of (4.3)-(4.4), we obtain that

$$J(u+\varepsilon v) = J(u) + \varepsilon^2 \int_0^1 \frac{1}{2}(v')^2\,dx \geq J(u) \qquad \forall \varepsilon \in \mathbb{R}, \forall v \in V,$$

that is, the weak solution u is also the minimizer of (4.6). Let us now prove that solving (4.5) in the space (4.4) implies (4.3). With this purpose, we need to enforce the so-called Euler condition

$$\frac{\partial J(u+\varepsilon v)}{\partial \varepsilon}\Big|_{\varepsilon=0} = 0, \tag{4.7}$$

where the above derivative, denoted *Frèchet derivative of the functional* J at $u \in V$, is defined as

$$\frac{\partial J(u+\varepsilon v)}{\partial \varepsilon}\Big|_{\varepsilon=0} := \lim_{\varepsilon \to 0} \frac{J(u+\varepsilon v) - J(u)}{\varepsilon}. \tag{4.8}$$

Computing the numerator yields

$$J(u+\varepsilon v) - J(u) = \varepsilon \left[\int_0^1 u'v'\,dx - \int_0^1 fv\,dx \right] + \varepsilon^2 \int_0^1 \frac{1}{2}(v')^2\,dx$$

and substituting the above result into (4.7) we get

$$\begin{aligned} 0 &= \lim_{\varepsilon \to 0} \frac{\varepsilon \left[\int_0^1 u'v'\,dx - \int_0^1 fv\,dx \right] + \varepsilon^2 \int_0^1 \frac{1}{2}(v')^2\,dx}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \left[\int_0^1 u'v'\,dx - \int_0^1 fv\,dx + \varepsilon \int_0^1 \frac{1}{2}(v')^2\,dx \right] \\ &= \int_0^1 u'v'\,dx - \int_0^1 fv\,dx, \end{aligned}$$

which yields

$$\int_0^1 u'v'\,dx = \int_0^1 fv\,dx \qquad \forall v \in V,$$

that is, the minimizer u of (4.6) is also a solution of the weak problem (4.3)-(4.4). This concludes the proof. □

Remark 4.2.4. In the language of Continuum Mechanics, problem (4.5)-(4.6) is referred to as *the principle of minimization of the total potential energy*. In particular, the term

$$\int_0^1 \frac{1}{2}(v')^2\,dx$$

represents the total elastic strain energy stored in the elastic body in correspondance to a generic deformed configuration $v = v(x)$, while

$$-\int_0^1 fv\,dx$$

represents the work done by the deformed system against the external body forces

The weak formulation (4.3)-(4.4), instead, is the counterpart of the *principle of virtual works*. In particular, the term

$$\int_0^1 u'v'\,dx$$

represents the scalar product bewteen the stress associated with the actual body deformation ($\sigma(u) = u'$) and the strain associated with the virtual displacement $v = v(x)$ ($\varepsilon(v) = v'$), while

$$\int_0^1 fv\,dx$$

represents the work done by the external forces in correspondance of a virtual displacement v.

So far, we have shown that u is a weak solution of the BVP iff it is also a minimizer of the total potential energy (4.6). By construction, if u solves (4.2), then it is also a solution of the weak problem (4.3). To close this chain of implications, we need to show that a weak solution is also a strong solution of the BVP.

Theorem 4.2.5 (Regularity of a weak solution)**.** *Let u be a solution of* (4.3)-(4.4)*. Assume also that*

$$u'' \in L^2(0,1). \tag{4.9}$$

Then, u is also a solution of the BVP (4.2) *a.e. in* $(0,1)$.

Proof. Let us apply Green's formula (C.1) to (4.3), to obtain

$$-\int_0^1 vw'\,dx + [v(1)w(1) - v(0)w(0)] = \int_0^1 fv\,dx \qquad \forall v \in V,$$

where $w = u'$. Since $v(0) = v(1) = 0$, we get

$$\int_0^1 -(u'' + f)v\,dx = 0 \qquad \forall v \in V. \tag{4.10}$$

Two remarks are in order with (4.10). The first remark is that the function $g := u'' \cdot v$ belongs to $L^1(0,1)$ due to Thm. 4.2.1, so that each term in the integral is well defined. The second remark is that, using the definition (B.18) of scalar product in the Hilbert space $L^2(0,1)$, the integral equation (4.10) tells us that the element $-(u''+f) \in L^2(0,1)$ is orthogonal to the whole $H_0^1(0,1)$. Since the embedding of $H_0^1(0,1)$ in $L^2(0,1)$ is completely continuous, it follows that the element $-(u''+f) \in L^2(0,1)$ is orthogonal to the whole $L^2(0,1)$ space. Therefore, $-(u''+f)$ coincides with the null function a.e. in $(0,1)$ which concludes the proof. □

Remark 4.2.6. The conclusion of Thm. (4.2.5) is that the additional regularity assumption (4.9) is a necessary requirement on the solution of the weak problem (4.3)-(4.4) in order to prove that it is actually a solution of the original BVP problem (4.2). Such extra regularity can be inferred from $(4.2)_1$, by identifying u'' with the element $f \in L^2(0,1)$ in the Lebesgue sense (i.e., almost everywhere in $(0,1)$). Therefore, we see *a posteriori* that the solution of the weak problem (4.3)-(4.4) belongs to $H^2(0,1) \cap H_0^1(0,1)$.

4.3 Weak solution of a BVP: the 2D case

In this section we extend the construction and characterization of the weak formulation of the 1D BVP (4.2) to the corresponding BVP (4.1) where the computational domain Ω is an open subset of $\mathbb{R}^2$ with Lipschitz boundary $\Gamma := \partial\Omega$ on which a unit outward normal vector $\mathbf{n}$ is defined almost everywhere.

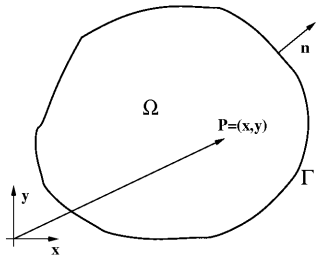


Figure 4.3: Computational domain in 2D.

Proceeding formally as in the 1D case, we multiply $(4.1)_1$ by an arbitrary test function v and integrate both sides over Ω. Then, we apply the formula of

integration by parts (C.1) to the vector field $\mathbf{w} = \nabla u$ to obtain

$$-\int_{\partial\Omega} v\nabla u\cdot\mathbf{n}\,d\Sigma+\int_{\Omega}\nabla u\cdot\nabla v\,d\Omega=\int_{\Omega} fv\,d\Omega \qquad \forall v.$$

The boundary term identically vanishes if we choose v in the space $V = H_0^1(\Omega)$ defined in (B.34). By doing so, the *weak formulation* of the BVP (4.1) is:
find $u \in V$ such that

$$\int_{\Omega}\nabla u\cdot\nabla v\,d\Omega=\int_{\Omega} fv\,d\Omega \qquad \forall v\in V \tag{4.11}$$

where

$$V = H_0^1(\Omega) = \left\{v\in L^2(\Omega),\,\nabla v\in (L^2(\Omega))^2 \text{ such that } v = 0 \text{ on } \partial\Omega\right\}. \tag{4.12}$$

All the considerations and properties discussed in Sect. 4.2 immediately extend to the 2D case. In particular, we have the following results.

Theorem 4.3.1 (Variational formulation of (4.1) in 2D)**.** *The minimization problem: find $u \in V$ such that*

$$J(u)\le J(v) \qquad \forall v\in V \tag{4.13}$$

where

$$J(v) := \int_{\Omega}\frac{1}{2}|\nabla v|^2\,d\Omega-\int_{\Omega} fv\,d\Omega \qquad \forall v\in V, \tag{4.14}$$

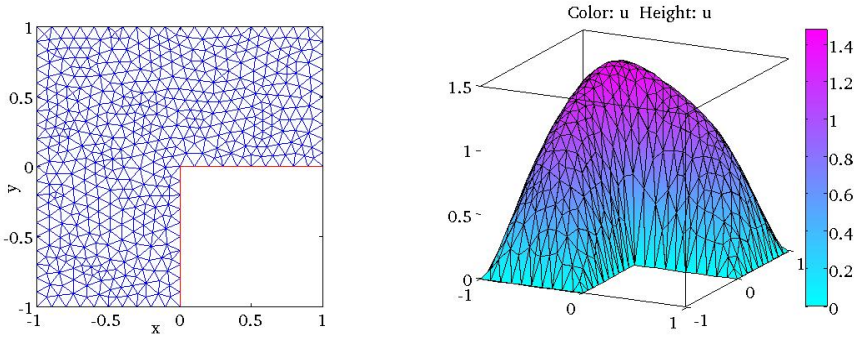
is completely equivalent to the weak problem (4.11)-(4.12).

Theorem 4.3.2 (Regularity of a weak solution)**.** *Let u be a solution of* (4.11)-(4.12). *Assume also that Ω is a convex open bounded set of $\mathbb{R}^2$, and that*

$$D^{\alpha}u\in L^2(\Omega) \qquad \text{with } |\alpha| = 2. \tag{4.15}$$

Then, u is also a solution of the BVP (4.1) *in 2D a.e. in Ω, and we have $u \in H^2(\Omega)\cap H_0^1(\Omega)$ exactly as in the one-dimensional weak problem* (4.3)-(4.4).

Example 4.3.3 (BVP in a L-shaped domain). This example serves to clarify the importance of Ω to be a convex domain in order to ensure the global H^2-regularity of the solution. Consider the homogeneous Dirichlet problem (4.1) with $d = 2$, $f = 10$ and Ω given by the L-shaped domain of Fig. 4.4(a). It can be proved that the unique weak solution u, shown in Fig. 4.4(b), belongs only to the space $H^{5/3}(\Omega)$ because of the reentrant corner at $(0,0)$.



(a) Ω and finite element mesh (b) u

Figure 4.4: Dirichlet problem in a L-shaped domain. Geometrical discretization and numerical solution have been performed using the `Matlab` function `pdetool`.

4.3.1 Non-homogeneous Dirichlet problem in 2D

We consider here the following *non-homogeneous* Dirichlet problem for the Laplace operator:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ u = g & \text{on } \Gamma \equiv \partial\Omega, \end{cases} \tag{4.16}$$

where Ω is a 2D bounded Lipschitz domain and where g is a given function in the trace space $H^{1/2}(\Gamma)$ (see Def. B.7.4). Clearly, the BVP studied in Sect. 4.3 is a special case of (4.16) when g is identically equal to zero. To repeat the steps followed previously to obtain the weak formulation of the problem, we only need to transform (4.16) into an equivalent problem with homogeneous boundary conditions. This, because it is easily checked that the space

$$V_g := \left\{ v \in H^1(\Omega) \text{ such that } \gamma_0 v = g \right\}$$

is *not* linear (does $v_1 + v_2$ belong to V_g, for $v_1, v_2 \in V_g$??). With this purpose, let us pick up in $H^1(\Omega)$, a function R_g such that $\gamma_0 R_g = g$. We call R_g a *lifting* of g to all Ω. Then, we assume that the solution of (4.16) can be written as

$$u = u_0 + R_g \tag{4.17}$$

with $\gamma_0 u_0 = 0$. Having done this, we multiply $(4.16)_1$ by a test function $v \in V = H_0^1(\Omega)$, integrate over all Ω and apply Green's formula, to get

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \qquad \forall v \in V.$$

Using (4.17), we obtain the weak formulation of the BVP (4.16):
find $u_0 \in V$ such that

$$\int_\Omega \nabla u_0 \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega - \int_\Omega \nabla R_g \cdot \nabla v \, d\Omega \qquad \forall v \in V. \tag{4.18}$$

Remark 4.3.4 (Physical interpretation). The BVP (4.16) represents in Thermal Physics the model of a plate, whose temperature u is fixed at the boundary and equal to g, and subject to a thermal source f (normalized to the thermal conductivity of the plate).

Remark 4.3.5. The weak problem (4.18) is of the same form as (4.11), except for a modification of the right-hand side. The extra term

$$-\int_\Omega \nabla R_g \cdot \nabla v \, d\Omega$$

is well defined because of Thm. 4.2.1.

4.3.2 Non-homogeneous Neumann problem in 2D

Let us now consider the *Neumann* problem for the Laplace operator:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ \nabla u \cdot \mathbf{n} = h & \text{on } \Gamma \equiv \partial\Omega, \end{cases} \tag{4.19}$$

where h is a given function in the trace space $H^{-1/2}(\Gamma)$ (see Def. B.7.10).

Remark 4.3.6 (Existence of a solution). Multiplying $(4.19)_1$ by $v = 1$, integrating over Ω and applying Green's formula (C.1) with $\mathbf{w} = -\nabla u$, yields

$$\int_\Omega f \, d\Omega + \int_\Gamma h \, d\Sigma = 0. \tag{4.20}$$

This condition must be satisfied by the data in order problem (4.19) to admit a solution, and for this reason it is often called *compatibility condition* on the data of the Neumann problem.

Remark 4.3.7 (Uniqueness of the solution). If u is a solution of the BVP (4.19), then also $u+K$ is a solution, K being an arbitrary constant. Thus, to ensure unique solvability of the Neumann problem, we need to enforce in some way a value of u in Ω. This can be done, for instance, by introducing the further request that u has null mean integral value

$$\int_\Omega u \, d\Omega = 0.$$

To derive the weak formulation of the BVP (4.19) we proceed as usual and obtain:
find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_\Gamma h v \, d\Sigma \qquad \forall v \in V \tag{4.21}$$

where

$$V = H^1(\Omega) \setminus \mathbb{R} = \left\{ v \in H^1(\Omega) \text{ such that } v \text{ is not a constant} \right\}. \tag{4.22}$$

The space V is an Hilbert space endowed with the norm

$$\|v\|_V := \|\nabla v\|_{L^2(\Omega)} \qquad v \in V. \tag{4.23}$$

Remark 4.3.8 (Physical interpretation). The BVP (4.19) represents in Electrostatics the Gauss law for a dielectric medium whose space charge density (normalized to the dielectric permittivity of the medium) is f, and whose outward flux of the electric field across the boundary is equal to $-h$.

4.3.3 Mixed problem in 2D

We conclude the presentation of BVPs associated with the Laplace operator in a 2D domain by considering the *mixed* problem:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \\ u = g & \text{on } \Gamma_D \\ \nabla u \cdot \mathbf{n} = h & \text{on } \Gamma_N, \end{cases} \tag{4.24}$$

where Γ_D and Γ_N are mutually disjoint partitions of the boundary Γ such that $\Gamma = \Gamma_D \cup \Gamma_N$, while g and h are given data in the trace spaces $H^{1/2}(\Gamma_D)$ and $H^{-1/2}(\Gamma_N)$, respectively. The definitions of these latter spaces are the obvious generalization of Defs. B.7.4 and B.7.10, provided to replace Γ with Γ_D and Γ_N, respectively. Clearly, BVPs (4.16) and (4.19) can be recovered as special cases of (4.24) by setting $\Gamma_N = \emptyset$ and $\Gamma_D = \emptyset$, respectively. In what follows, we assume that $\text{meas}(\Gamma_D) > 0$. This avoids the need of enforcing an extra condition on u to ensure uniqueness of the solution (see Rem. 4.3.7).

To construct the weak formulation of (4.24) we replicate the steps followed in previous cases. In particular, we introduce a lifting $R_g \in H^1(\Omega)$ of g, such that $R_g|_{\Gamma_D} = g$, and decompose the solution u as

$$u = u_0 + R_g$$

where u_0 is such that $u_0|_{\Gamma_D} = 0$, and then we obtain:
find $u_0 \in V$ such that

$$\int_\Omega \nabla u_0 \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\Gamma_N} h v \, d\Sigma - \int_\Omega \nabla R_g \cdot \nabla v \, d\Omega \qquad \forall v \in V \quad (4.25)$$

where

$$V = H^1_{0,\Gamma_D}(\Omega) = \left\{ v \in H^1(\Omega) \text{ such that } v|_{\Gamma_D} = 0 \right\}. \quad (4.26)$$

The space V is an Hilbert space endowed with the norm (B.36) in virtue of Rem. B.7.8.

Remark 4.3.9 (Physical interpretation). The BVP (4.24) represents in Thermal Physics the Fourier law for thermal diffusion in a medium under the action of a distributed thermal source f (normalized to the thermal conductivity of the medium), of a fixed temperature g on the boundary portion Γ_D and of a given thermal flux $-h$ on the boundary portion Γ_N, respectively.

4.4 Well-posedness analysis: the Lax-Milgram Lemma

All of the weak formulations considered in the previous sections can be written in the following abstract form:
find $u \in V$ such that

$$B(u,v) = F(v) \qquad \forall v \in V, \quad (4.27)$$

where:

- V is a given Hilbert space endowed with norm $\|\cdot\|_V$;
- $B(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a given *bilinear form*, i.e., a linear real-valued functional over the pair $V \times V$ such that:

  $$\begin{aligned} B(\lambda u + \mu w, v) &= \lambda B(u,v) + \mu B(w,v), \\ B(u, \lambda v + \mu w) &= \lambda B(u,v) + \mu B(u,w), \end{aligned}$$

  for every $u, v, w \in V$ and for every real numbers λ and μ;
- $F(\cdot) : V \to \mathbb{R}$ is a given *linear form*, i.e., a linear functional over the space V such that

  $$F(\lambda v + \mu w) = \lambda F(v) + \mu F(w).$$

Theorem 4.4.1 (Lax-Milgram (LM) Lemma)**.** *Assume that there exist three positive constants M, β and Λ such that:*

$$|B(u,v)| \leq M\|u\|_V\|v\|_V \qquad u,v \in V \tag{4.28a}$$

$$B(u,u) \geq \beta\|u\|_V^2 \qquad u \in V \tag{4.28b}$$

$$|F(v)| \leq \Lambda\|v\|_V \qquad v \in V. \tag{4.28c}$$

Then, problem (4.27) *admits a unique solution $u \in V$ and the following stability estimate holds*

$$\|u\|_V \leq \frac{\Lambda}{\beta}. \tag{4.29}$$

Proof. The proof of the Lax-Milgram Lemma requires the use of advanced tools of Functional Analysis and for this reason is omitted here. To prove (4.29), we simply need to take $v = u$ in (4.27). This yields, using (4.28c)

$$B(u,u) = F(u) \leq |F(u)| \leq \Lambda\|u\|_V,$$

from which, using (4.28b)

$$\beta\|u\|_V^2 \leq B(u,u) \leq \Lambda\|u\|_V$$

which gives (4.29). □

Remark 4.4.2 (Meaning of the conditions in the LM Lemma). Conditions (4.28a) and (4.28c) express the *continuity* of the bilinear form B and of the linear form F, respectively. Continuity has the same meaning as the concept of continuous dependence on the data that was introduced in Def. 1.1.4. Therefore, it is clearly desirable M and Λ to be as small as possible in such a way that perturbations on problem data are not eventually amplified. Condition (4.28b) expresses the *coercivity* of the bilinear form B. Roughly speaking, it tells us that the energy of the system (represented by the quantity $\|u\|_V^2$) is uniformly bounded from below. Therefore, it is clearly desirable β to be as large as possible in such a way to ensure a good control of the energy of the solution.

Tab. 4.1 gathers in a synthetic format all the considered BVPs for the Laplace operator in a two-dimensional domain Ω with Lipschitz boundary Γ. In all cases, the norm for V is given by (B.36).

Example 4.4.3 (Verification of the LM Lemma). We use here the LM Lemma to verify that the weak problem (4.11) is well-posed.

BVP	V	B	F
Dirichlet (hom.)	$H^1_0(\Omega)$	$\int_\Omega \nabla u \cdot \nabla v \, d\Omega$	$\int_\Omega f v \, d\Omega$
Dirichlet (non-hom.)	$H^1_0(\Omega)$	$\int_\Omega \nabla u_0 \cdot \nabla v \, d\Omega$	$\int_\Omega f v \, d\Omega - a(R_g, v)$
Neumann (non-hom.)	$H^1(\Omega) \setminus \mathbb{R}$	$\int_\Omega \nabla u \cdot \nabla v \, d\Omega$	$\int_\Omega f v \, d\Omega + \int_\Gamma h v \, d\Sigma$
Mixed	$H^1_{0,\Gamma_D}(\Omega)$	$\int_\Omega \nabla u_0 \cdot \nabla v \, d\Omega$	$\int_\Omega f v \, d\Omega + \int_{\Gamma_N} h v \, d\Sigma - a(R_g, v)$

Table 4.1: Examples of BVPs for the Laplace operator.

- Continuity of B: we have

$$|B(u,v)| = \Big| \int_\Omega \nabla u \cdot \nabla v \, d\Omega \Big| \leq \int_\Omega |\nabla u||\nabla v| \, d\Omega \leq \|u\|_V \|v\|_V$$

  because of CS inequality (B.16) and (B.36). Therefore, we have $M = 1$.

- Coercivity of B: we immediately have

$$B(u,u) = \int_\Omega |\nabla u|^2 \, d\Omega = \|u\|_V^2$$

  so that $\beta = 1$.

- Continuity of F: we have

$$|F(v)| = \Big| \int_\Omega f v \, d\Omega \Big| \leq \int_\Omega |f||v| \, d\Omega \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq C_P \|f\|_{L^2(\Omega)} \|v\|_V$$

  having used again (B.16) and Poincarè inequality (B.35). Therefore, we have $\Lambda = C_P \|f\|_{L^2(\Omega)}$, and the stability estimate for the solution of the Dirichlet problem for the Laplace operator is

$$\|u\|_V \leq C_P \|f\|_{L^2(\Omega)}. \tag{4.30}$$

Remark 4.4.4 (Mechanical interpretation). The estimate (4.30) tells us that the membrane deformation energy is proportional to the energy of the vertical load (its weight, for instance) in agreement with Hooke's law of Elasticity.

Chapter 5

Galerkin Finite Element Approximation of Elliptic Boundary Value Problems

Abstract

In this chapter, we construct the numerical approximation of an elliptic boundary value problem in weak form by means of the Galerkin Finite Element Method. In particular, we study the consistency, stability and convergence properties of the associated discrete formulation, and we illustrate the numerical performance of the method in the study of simple two-point boundary value problems with smooth and non-smooth solutions.

5.1 The Galerkin method

The numerical approximation of the weak formulation (4.27) of a BVP like those studied in Chapter 4 consists of the following steps:

1. construct a family of subspaces $V_h \subset V$ that depend on a *discretization parameter* $h > 0$ and such that

$$\dim V_h = N_h < +\infty;$$

2. seek the solution of the weak problem (4.27) within V_h, that is:

find $u_h \in V_h$ such that

$$B(u_h, v_h) = F(v_h) \qquad \forall v_h \in V_h. \tag{5.1}$$

Definition 5.1.1 (Nomenclature)**.** *The equation* (5.1) *is called* the Galerkin Problem *associated with the weak problem* (4.27)*, and u_h is the* Galerkin approximation *of the solution $u \in V$ of* (4.27)*.*

Since V_h is a finite-dimensional space, we have

$$V_h = \text{span}\,\{\varphi_i\}_{i=1}^{N_h}$$

where the functions $\varphi_i = \varphi_i(\mathbf{x})$ constitute a *basis* for V_h (cf. Sect. A.1). Of course, we can write

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

where the N_h real numbers u_j, $j = 1, \ldots, N_h$ are the *degrees of freedom* (dofs) of u_h with respect to the basis of V_h. Replacing the previous expansion into (5.1) and using the linearity of $B(\cdot,\cdot)$ with respect to the first argument, yields the problem: find u_j, $j = 1, \ldots, N_h$, such that

$$\sum_{j=1}^{N_h} u_j B(\varphi_j, v_h) = F(v_h) \qquad \forall v_h \in V_h.$$

Noting that the mathematical phrase "$\forall v_h \in V_h$" is equivalent to "$\forall \varphi_i$, $i = 1, \ldots, N_h$", the Galerkin problem amounts to solving the following *linear algebraic system*

$$\mathbf{Bu} = \mathbf{f} \tag{5.2}$$

where:

- $\mathbf{B}$ is a square matrix of dimension N_h, called *stiffness matrix*, whose entries are defined as
$$B_{ij} = B(\varphi_j, \varphi_i) \qquad i, j = 1, \ldots, N_h; \tag{5.3}$$
- $\mathbf{u} = [u_1, \ldots, u_{N_h}]^T$ is the column vector of dimension N_h containing the unknown dofs of u_h;

- **f** is a column vector of dimension N_h, called *load vector*, whose compoments are defined as

$$f_i = F(\varphi_i) \qquad i = 1, \ldots, N_h. \tag{5.4}$$

In conclusion, once the basis $\{\varphi_i\}_{i=1}^{N_h}$ is properly selected, we only have to use the methodologies investigated in Sect. 2 to efficiently and accurately solve system (5.2), and we are done. Of course, it remains to prove that:

1. the solution u_h exists and is unique;
2. the solution u_h is a *good approximation* of u, that is, using the terminology of Sect. 1.2, u_h converges to u, i.e.

$$\lim_{h \to 0} \|u - u_h\|_V = 0. \tag{5.5}$$

The answer to the first question is given by the following result.

Theorem 5.1.2. *The stiffness matrix* **B** *is positive definite, so that the Galerkin problem* (5.1) *(equivalently, the linear system* (5.2)*) is uniquely solvable.*

Proof. We prove the result in the simple case $N_h = 2$, leaving to an exercise the extension to $N_h > 2$. Computing explicitly the quantity $B(u_h, u_h)$ we get

$$B(u_h, u_h) = u_1^2 B(\varphi_1, \varphi_1) + u_1 u_2 B(\varphi_1, \varphi_2) + u_2 u_1 B(\varphi_2, \varphi_1) + u_2^2 B(\varphi_2, \varphi_2) = \mathbf{u}^T \mathbf{B} \mathbf{u}.$$

Using the coercivity of B, we have

$$\mathbf{u}^T \mathbf{B} \mathbf{u} = B(u_h, u_h) \geq \beta \|u_h\|_V^2 > 0 \qquad \forall u_h \in V_h$$

that is, **B** is positive definite (see Sect. A.3 for the definition). □

Theorem 5.1.3 (A priori estimate for u_h)**.** *The unique solution $u_h \in V_h$ of* (5.1) *satisfies the following a priori estimate*

$$\|u_h\|_V \leq \frac{\Lambda}{\beta}. \tag{5.6}$$

Proof. Note that $V_h \in V$; then, apply Lax-Milgram lemma to (5.1). □

To answer the second question, we need to check that the Galerkin formulation (5.1) is stable and consistent (cf. Thm. 1.2.2).

Definition 5.1.4 (Discretization error)**.** *The discretization error introduced by the Galerkin method is defined as*

$$e_h := u - u_h.$$

Theorem 5.1.5 (Stability)**.** *The Galerkin method is stable, i.e., e_h satisfies the estimate*

$$B(e_h, e_h) \geq \beta \|e_h\|_V^2,$$

where β is the coercivity constant appearing in the Lax-Milgram Lemma 4.4.1.

Proof. Since $V_h \subset V$, the function $e_h \in V$, and then we can apply the coercivity assumption on B to get the result. □

Theorem 5.1.6 (Consistency)**.** *The Galerkin method is consistent, i.e., the discretization error satisfies the relation*

$$B(e_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{5.7}$$

Proof. Since $V_h \subset V$, we can choose $v = v_h \in V_h$ in (4.27). Then, subtracting (5.1) from (4.27), we get (5.7). □

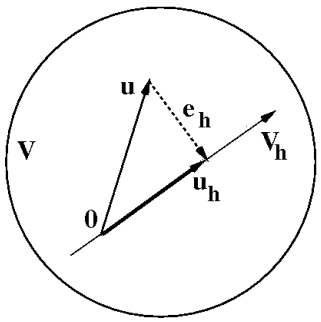


Figure 5.1: The property of orthogonality of the Galerkin method.

Remark 5.1.7 (Galerkin orthogonality). Thm. 5.1.6 has an interesting geometrical interpretation in the case where the bilinear form $B(\cdot,\cdot)$ is *a scalar product* on the Hilbert space V (cf. Def. B.5.1). In such an event, relation (5.7) tells us that the error is *orthogonal* to the solution space V with respect to the metrics induced by $B(\cdot,\cdot)$. This means, in particular, that the discrete solution u_h is the *orthogonal projection* of the exact solution u onto V_h.

Theorem 5.1.8 (Ceà's Lemma)**.** *The discretization error satisfies the following estimate*

$$\|u-u_h\|_V \leq \frac{M}{\beta} \inf_{v_h \in V_h} \|u-v_h\|_V. \tag{5.8}$$

Proof. Using coercivity, we get

$$B(u-u_h, u-u_h) \geq \beta \|u-u_h\|_V^2.$$

On the other hand, we also have, for every $v_h \in V_h$

$$B(u-u_h, u-u_h) = B(u-u_h, u-v_h) + \underbrace{B(u-u_h, v_h-u_h)}_{=0 \text{ because of (5.7)}}.$$

Thus, using continuity and combining the previous two relations, we get

$$\beta \|u-u_h\|_V^2 \leq M \|u-u_h\|_V \|u-v_h\|_V \qquad \forall v_h \in V_h,$$

from which we get (5.8). □

Remark 5.1.9. Ceà's Lemma has a remarkable conceptual interpretation

$$\boxed{discretization\,error} \leq \boxed{amplification\,factor} \times \boxed{approximation\,error}.$$

The amplification factor M/β is the effect of the continuous problem on its corresponding approximation. This, a posteriori, supports the importance of having a "small" M and a "big" β in Lax-Milgram Lemma.

Theorem 5.1.10 (Convergence)**.** *Let $v \in V$ be an arbitrary element of the function space V. Assume that the approximation space V_h is chosen in such a way that*

$$\lim_{h \to 0} \inf_{v_h \in V_h} \|v-v_h\|_V = 0. \tag{5.9}$$

Then, the Galerkin method is convergent.

Proof. It suffices to use (5.9) in (5.8) with $v = u$ and apply the definition of convergence (5.5). □

5.2 The Galerkin Finite Element Method

In this section, we specify one particular choice of V_h that satisfies (5.9). To make things simple, we assume here that $\Omega = (0,1)$ and $V = H_0^1(\Omega)$, and denote from now on by C a positive constant, independent of h, not necessarily having the same value in all its occurrences.

Let $\mathscr{T}_h$ denote a family of partitions of $\overline{\Omega}$ into $M_h \geq 2$ subintervals $K_i := [x_{i-1}, x_i]$, $i = 1, \ldots, M_h$, of length $h_i = x_i - x_{i-1}$, with $h := \max_{K_i \in \mathscr{T}_h} h_i$. For a given $r \geq 1$, we set

$$V_h := X_{h,0}^r(\mathscr{T}_h) = \{v_h \in X_h^r, \text{ such that } v_h(0) = v_h(1) = 0\}, \tag{5.10}$$

where X_h^r is the piecewise polynomial space introduced in Sect. 3.1. The number of lineary independent basis functions spanning V_h, denoted by $N_h(r)$, gives the dimension of V_h and reads

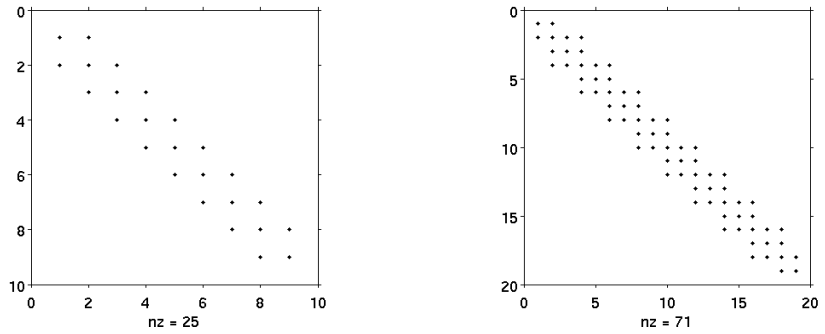
$$N_h(r) = M_h(r+1) - (M_h - 1) - 2 = M_h r - 1. \tag{5.11}$$

The Galerkin method with V_h as in (5.10) is called *Galerkin Finite Element Method of order r* (GFEM), or, shortly, the *Finite Element Method* (FEM).

Remark 5.2.1 (Sparsity pattern of A). An important consequence of the choice of V_h is that the stiffness matrix $\mathbf{B}$ is *sparse*. This means that only a few of the N_h^2 entries B_{ij} is actually non-vanishing. In the 1D case, it is easy to characterize the *sparsity pattern* of $\mathbf{B}$, i.e., the subset of node indices such that $B_{ij} \neq 0$. As a matter of fact, looking at the structure of Lagrangian basis functions (see Sect. 3.1.1), we see that for each row $i \in [1, N_h(r)]$, the column indices $j(i)$ belonging to the pattern of i are those such that $j(i) \in [i-r, i+r]$. For example, if $r = 1$, for each row i, we have (a priori) at most three nonzero entries: $B_{i,i-1}$, $B_{i,i}$ and $B_{i,i+1}$, so that $\mathbf{B}$ is a tridiagonal matrix. If $r = 2$, the sparsity pattern is larger, and we have a priori at most five nonzero entries per each row: $B_{i,i-2}$, $B_{i,i-1}$, $B_{i,i}$, $B_{i,i+1}$ and $B_{i,i+2}$, and so on. Fig. 5.2 shows these two sparsity patterns on a mesh of $M_h = 10$ elements. The property of generating a sparse matrix structure maintains its validity also in multi-dimensional problems, and represents one of the strongest advantages of using the GFEM. Sparsity allows to minimize storage occupation and to optimize computing resources in the solution of the linear algebraic system (5.2).

5.2.1 Error analysis

The following interpolation error estimates generalize (3.8) and (3.9) to the case of the Sobolev norm (B.36).



(a) $r = 1$ (b) $r = 2$

Figure 5.2: Sparsity patterns in the cases $r = 1$ and $r = 2$, with $M_h = 10$. The graphs have been obtained using the `Matlab` commands: `spy(B)` and `nnz(B)`, respectively.

Theorem 5.2.2 (Interpolation error in the L^2 and $\|\cdot\|_V$-norms)**.** *Let f be a given function in $H^{r+1}(\Omega)$ and $\Pi_h^r f$ the V_h-interpolant of f. Then, we have*

$$\|f - \Pi_h^r f\|_{L^2(\Omega)} \leq Ch^{r+1}\|f\|_{H^{r+1}(\Omega)}. \tag{5.12}$$

Morever, we also have

$$\|f - \Pi_h^r f\|_V \leq Ch^r\|f\|_{H^{r+1}(\Omega)}. \tag{5.13}$$

Remark 5.2.3. Thm. 5.2.2 contains two important informations. The first information is that the finite element interpolant $\Pi_h^r f$ converges to f, in the topology of the space V, with order r with respect to the discretization parameter h (cf. Def. 1.4.2). The second information is the order of convergence of the approximation decreases by one passing from the L^2-norm to the H^1-norm.

Theorem 5.2.4 (GFEM Convergence (1))**.** *Let u and u_h denote the solutions of* (4.27) *and* (5.1)*, respectively. Assume also that $u \in H^{r+1}(\Omega) \cap H_0^1(\Omega)$. Then, we have*

$$\|u - u_h\|_V \leq C\frac{M}{\beta}h^r\|u\|_{H^{r+1}(\Omega)}. \tag{5.14}$$

Under the same assumptions, we also have

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{r+1}\|u\|_{H^{r+1}(\Omega)}. \tag{5.15}$$

Proof. Let us consider (5.8). Instead of determining the "optimal" v_h, i.e., the function v_h in correspondance of which the approximation error attains its infimum, we pick the "special" function $v_h := \Pi_h^r u$. By doing so, we obtain

$$\|u - u_h\|_V \leq \frac{M}{\beta} \|u - \Pi_h^r u\|_V$$

from which, using (5.13) with $f \equiv u$, we immediately get (5.14). The proof of (5.15) requires the use of sophisticated functional arguments (duality analysis, Aubin-Nitsche trick), and for this reason is omitted here. □

Remark 5.2.5. The above proof reveals another remarkable conceptual interpretation of Ceà's Lemma

$$\boxed{discretization\,error} \leq \boxed{amplification\,factor} \times \boxed{interpolation\,error}.$$

Thus, the error analysis of the GFEM method is reduced to a simple interpolation error analysis.

The error analysis of the GFEM has been carried out under the very restrictive assumption that the exact solution u of (4.27) has an *arbitrarily high regularity* (as a matter of fact, we have postulated that, for any given r, the function u belongs to H^{r+1}!). This is clearly not true, in general, so that we have to modify Thm. 5.2.4 to account for the realistic case.

Theorem 5.2.6 (GFEM Convergence (2))**.** *Let u and u_h denote the solutions of* (4.27) *and* (5.1)*, respectively. Assume that $u \in H^s(\Omega) \cap H_0^1(\Omega)$, $s \geq 2$ being the "true" regularity of the solution of the weak problem. Then, we have*

$$\|u - u_h\|_V \leq C \frac{M}{\beta} h^l \|u\|_{H^{l+1}(\Omega)}, \tag{5.16}$$

where

$$l := \min\{r, s-1\}$$

is called regularity threshold*. Under the same assumptions, we also have*

$$\|u - u_h\|_{L^2(\Omega)} \leq C h^{l+1} \|u\|_{H^{l+1}(\Omega)}. \tag{5.17}$$

Remark 5.2.7 (*h*-and-*r* refinement strategies). Thm. 5.2.6 shows that there is no convenience in using high-order polynomials to increase approximation accuracy, when the regularity of the exact solution is small (*r*-refinement). Rather, it is better to employ the maximum value of r that is allowed by the regularity threshold, and then reduce the discretization parameter h (*h*-refinement). This general phylosophical statement is summarized in Tab. 5.1.

r	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
1	conv.	$\boxed{\mathscr{O}(h)}$	$\mathscr{O}(h)$	$\mathscr{O}(h)$	$\mathscr{O}(h)$
2	conv.	$\mathscr{O}(h)$	$\boxed{\mathscr{O}(h^2)}$	$\mathscr{O}(h^2)$	$\mathscr{O}(h^2)$
3	conv.	$\mathscr{O}(h)$	$\mathscr{O}(h^2)$	$\boxed{\mathscr{O}(h^3)}$	$\mathscr{O}(h^3)$
4	conv.	$\mathscr{O}(h)$	$\mathscr{O}(h^2)$	$\mathscr{O}(h^3)$	$\boxed{\mathscr{O}(h^4)}$

Table 5.1: Order of convergence of the GFEM method as a function of r and of s. The lower triangular part of the table corresponds to sub-optimal convergence rates of the method (equal to $s-1$) because the order of convergence is limited by the regularity threshold. The diagonal part of the table (boxed terms) correspond to the best trade-off between accuracy and computational effort for a given solution regularity. The upper triangular part of the table corresponds to optimal convergence rates of the method (equal to r) when the solution is sufficiently smooth.

5.3 Experimental convergence study of the GFEM

In this section, we shortly verify the numerical performance of the GFEM in the study of two BVPs in 1D, characterized by a smooth and no-smooth solution, respectively. All computations have been performed using the `Matlab` finite element package `EF1D` developed by Marco Restelli and available at the link `http://www1.mate.polimi.it/CN/MNIC/Laboratori/EF1D.zip`.

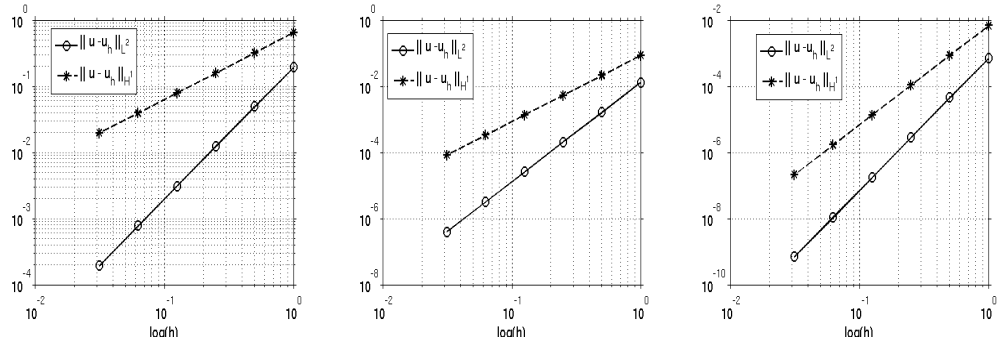
5.3.1 BVP with smooth solution

This example serves to verify Thm. 5.2.6 in the numerical solution of the following BVP:

$$\begin{cases} -u'' = \sin(x) & \text{in } (0,10) \\ u(0) = 0 \\ u(10) = \sin(10). \end{cases} \tag{5.18}$$

The exact solution is $u(x) = \sin(x)$, so that $s = +\infty$ and Thm. 5.2.6 gives us the optimal result $l = r$ for every r. This prediction is confirmed by the plots in Fig. 5.3 which show that the orders of convergence in the L^2 and H^1 norms are equal to

$r+1$ and r, respectively. The mesh size is uniform and equal to $h = 10/M_h$, $M_h = [10, 20, 40, 80, 160, 320]^T$.



(a) $r = 1$ (b) $r = 2$ (c) $r = 3$

Figure 5.3: BVP in 1D with smooth solution. Circles refer to the error in L^2, asterisks refer to the error in H^1.

5.3.2 BVP with a non-smooth solution

This example serves to verify Thm. 5.2.6 in the numerical solution of the following BVP:

$$\begin{cases} -u'' = f(x;\lambda) & \text{in } (0,2) \\ u(0) = 0 \\ u(2) = 1 \end{cases} \tag{5.19}$$

where λ is a real parameter and

$$f(x;\lambda) = \begin{cases} \sin(x) & 0 \le x \le 1 \\ \sin(x) - \lambda(\lambda-1)(x-1)^{\lambda-2} & 1 < x \le 2. \end{cases}$$

The exact solution is

$$u(x;\lambda) = \begin{cases} \sin(x) & 0 \le x \le 1 \\ \sin(x) + (x-1)^{\lambda} & 1 < x \le 2. \end{cases}$$

For a given $s \ge 2$, the source term f belongs to $L^s(0,2)$ iff

$$\int_1^2 (x-1)^{(\lambda-2)s}\, dx < +\infty.$$

Setting $\xi := (x-1)$ and $\kappa := (\lambda-2)s$, the above condition becomes

$$\int_0^1 \xi^\kappa d\xi =< +\infty,$$

which shows that $f \in L^s(0,2) \Leftrightarrow \kappa+1>0$, that is, iff $\lambda > 2-1/s$. Under this condition, the solution $u(x;\lambda)$ of (5.19) belongs to the Sobolev space $H^s(0,2)$ so that Thm. 5.2.6 tells us that (in the worst scenario) the order of convergence of the GFEM is equal to

$$l = \min\{r, s-1\}.$$

Setting $s=3$, for example, we should have a convergence order p_{H^1} in the H^1-norm equal to 1 and 2 (optimal) if r is equal to 1 and 2, respectively, while $p_{H^1} = s-1 = 2$ for every $r \geq 3$ (sub-optimal). This prediction is fully confirmed by Tab. 5.2 which shows the order of convergence computed experimentally using (1.9) on a triangulation $\mathscr{T}_h$ with uniform mesh size equal to $h = 2/M_h$, $M_h = [7,23,59,131,277,551]^T$.

M_h	$p_{H^1}(r=1)$	$p_{H^1}(r=2)$	$p_{H^1}(r=3)$	$p_{H^1}(r=4)$
23	9.9434e-01	1.9833e+00	2.1384e+00	2.0032e+00
59	9.9938e-01	1.9947e+00	2.0559e+00	2.0010e+00
131	9.9990e-01	1.9978e+00	2.0247e+00	2.0004e+00
277	9.9998e-01	1.9990e+00	2.0116e+00	2.0002e+00
551	9.9999e-01	1.9995e+00	2.0057e+00	1.9998e+00

Table 5.2: Order of convergence of the GFEM method as a function of r in the case of a solution belonging to $H^3(0,2)$.

Part III

The GFEM for Elliptic Problems in 1D and 2D

This part illustrates the weak formulation and the numerical approximation using the Galerkin Finite Element Method (GFEM) of reaction-diffusion and advection-diffusion problems in the 1D case. Then, the application of the GFEM to elliptic problems in the 2D case is addressed, and the principal issues that characterize the implementation of the method in a computational algorithm are illustrated.

Chapter 6

Elliptic Boundary Value Problems in 1D: Theory and Finite Element Approximation

Abstract

In this chapter, we apply the general machinery of weak formulation and Galerkin Finite Element (GFE) approximation to two particular, and significant, classes of elliptic boundary value problems, namely, a reaction-diffusion equation and an advection-diffusion equation. To keep the presentation as simple as possible, we consider the 1D case, with constant coefficients and homogeneous Dirichlet boundary conditions. For both model problems, the standard GFE approximation is first presented, and then an appropriate stabilized form of the method is proposed to cure the occurrence of numerical instabilities in the computed solution when the continuous problem is dominated by reaction or advection terms. Computational examples (run with the 1D finite element code `EF1D` written by Marco Restelli) are included to show how things can go wrong and how to fix problems. The code has been developed by Marco Restelli and is available at the link: `http://www1.mate.polimi.it/CN/MNIC/Laboratori/EF1D.zip`.

6.1 Reaction-diffusion model problem

Let us consider the following two-point BVP:

$$\begin{cases} -\mu u'' + \sigma u = f & \text{in } \Omega = (0,1) \\ u(0) = 0 \\ u(1) = 0, \end{cases} \tag{6.1}$$

where $\mu > 0$ is the *diffusion coefficient*, $\sigma > 0$ is the *reaction coefficient* while the right-hand side f is a given *production term*. The equation system (6.1) is commonly referred to as a *reaction-diffusion* (RD) problem, and represents a simple model of a chemical substance whose concentration is u, that diffuses in the environment (a fluid, the air) with diffusion coefficient μ and reacts with the environment according to a net reaction mechanism given by $R := f - \sigma u$. Setting for simplicity $f = 1$, the exact solution of the problem is

$$u(x) = \frac{1}{\sigma}\left[1 + e^{\alpha x}\frac{1-e^{-\alpha}}{e^{-\alpha}-e^{\alpha}} + e^{-\alpha x}\frac{e^{\alpha}-1}{e^{-\alpha}-e^{\alpha}}\right] \tag{6.2}$$

where $\alpha := \sqrt{\sigma/\mu}$. A plot of u is reported in Fig. 6.1 for three increasing values of α, corresponding to $\sigma = 1$ and $\mu = 1, 10^{-1}, 10^{-3}$.

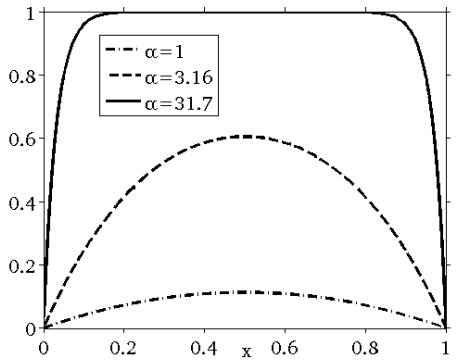


Figure 6.1: Plot of the solution of the reaction-diffusion model problem as a function of α.

Fig. 6.1 shows that u tends to a constant, equal to $1/\sigma$, as $\alpha \to \infty$. As a matter of fact, assuming $\alpha \gg 1$ and expanding the exponential terms in (6.2), we get

$$u(x) \simeq \frac{1}{\sigma}\left[1 - e^{\alpha(x-1)} - e^{-\alpha x}\right].$$

The solution behaves, asymptotically, as the sum of three contributions: a constant and two exponential terms that become significant as x gets closer to 1 and 0, respectively. These two exponential terms are called *boundary layer terms* because are responsible for the variation of u from the constant value $1/\sigma$ to the boundary conditions $u(1) = 0$ and $u(0) = 0$. The subinterval of Ω within which such a variation occurs is called *boundary layer* and is denoted by δ It can be proved that $\delta = \mathcal{O}(\alpha^{-1})$, therefore it gets smaller and smaller if diffusion is dominated by reaction in the model (6.1).

Definition 6.1.1. *Model* (6.1) *is said to be reaction-dominated if* $\alpha \gg 1$*, otherwise it is called diffusion-dominated.*

Matlab coding. The `Matlab` script for computing and plotting u is reported below.

```
x=[0:0.0001:1];
s=1; mu=[1, 1e-1, 1e-3];
a=sqrt(s./mu);
for k=1:numel(a),
    u(k,:)=1/s*(1+exp(a(k)*x)*(1-exp(-a(k)))/(exp(-a(k))-exp(a(k)))+...
    exp(-a(k)*x)*(exp(a(k))-1)/(exp(-a(k))-exp(a(k))));
end
plot(x,u)>> xlabel('x')
legend('\alpha=1','\alpha=3.16','\alpha=31.7')
```

6.1.1 Weak formulation

Let $V := H_0^1(0,1)$ with norm given by (B.28). Assume also that f is a given function in $L^2(0,1)$. Multiplying $(6.1)_1$ by an arbitrary test function $v \in V$ and integrating by parts the term $\int_0^1 -\mu u'' v\,dx$, we get the following weak formulation of the reaction-diffusion BVP:
find $u \in V$ such that

$$\underbrace{\int_0^1 \mu u'v'\,dx + \int_0^1 \sigma uv\,dx}_{B(u,v)} = \underbrace{\int_0^1 fv\,dx}_{F(v)} \qquad \forall v \in V \tag{6.3}$$

Theorem 6.1.2. *Problem* (6.3) *is uniquely solvable and u satifies the estimate*

$$\|u\|_V \leq \frac{C_P \|f\|_{L^2(0,1)}}{\mu}. \tag{6.4}$$

Proof. Let us check that B and F satisfy the conditions required by the Lax-Milgram Lemma.

- continuity of B:

$$\begin{aligned} |B(u,v)| &\leq \mu \int_0^1 |u'|\,|v'|\,dx + \sigma \int_0^1 |u||v|\,dx \\ &\leq \mu \|u\|_V \|v\|_V + \sigma \|u\|_V \|v\|_{L^2(0,1)} \\ &\leq (\mu + \sigma C_P^2) \|u\|_V \|v\|_V \qquad \forall u, v \in V, \end{aligned}$$

  from which $M = \mu + \sigma C_P^2$;

- coercivity of B:

$$B(u,u) \geq \mu \|u\|_V^2 \qquad \forall u \in V$$

  because σ is > 0, from which $\beta = \mu$;

- continuity of F:

$$\begin{aligned} |F(v)| &\leq \int_0^1 |f|\,|v|\,dx \leq \|f\|_{L^2(0,1)} \|v\|_{L^2(0,1)} \\ &\leq C_P f\|_{L^2(0,1)} \|v\|_V \forall v \in V, \end{aligned}$$

  from which $\Lambda = C_P f\|_{L^2(0,1)}$.

Since all the assumptions of the Lax-Milgram Lemma are verified, the solution u of (6.3) exists and is unique in V, and satisfies (6.4). □

6.1.2 Galerkin finite element approximation

The GFE discretization of (6.1) is (5.1) and the corresponding matrix form is (5.2) with

$$\begin{aligned} B_{ij} &= \int_0^1 \mu \varphi_j' \varphi_i'\,dx + \int_0^1 \sigma \varphi_j \varphi_i\,dx \qquad i,j = 1,\ldots,N_h \\ F_i &= \int_0^1 f \varphi_i\,dx \qquad i = 1,\ldots,N_h, \end{aligned}$$

where the approximation space V_h is the finite element space of degree r defined in (5.10), and $\{\varphi_i\}_{i=1}^{N_h}$ is the corresponding Lagrange basis introduced in Sect. 3.1. The stiffness matrix $\mathbf{B}$ is symmetric and positive definite, so that the discrete problem (5.1) admits a unique solution (see Thm. 5.1.2). Moreover, Ceà's Lemma tells us that

$$\|u - u_h\|_V \leq C \left(1 + \frac{\sigma}{\mu}\right) h^r \|u\|_{H^{r+1}(0,1)}, \tag{6.5}$$

because the exact solution $u \in C^\infty([0,1])$ so that $l = r$, while $C_P = 1$ for $\Omega = (0,1)$. Thus, we see that u_h converges to u, in the H^1-norm, as $h \to 0$, and that accuracy improves as the polynomial order is increased, because of the infinite regularity of the exact solution.

Remark 6.1.3 (The reaction-dominated case). It is interesting to take a closer look at the convergence estimate (6.5) and try to analyze it as a function of model parameters, μ and σ. For this, we set $r = 1$ (piecewise linear finite elements) and fix a tolerance ε, sufficiently small. Then, we see that in order the error to be $\simeq \varepsilon$, we need that

$$h \simeq \frac{\varepsilon}{C(1+\alpha^2)\|u\|_{H^2(0,1)}}.$$

Therefore, if the reaction-diffusion problem is reaction-dominated ($\alpha \gg 1$), the mesh size needs to be much smaller than the tolerance ε, in accordance with Fig. 6.1, because boundary layer effects become increasingly significant as α gets larger and the RD problem gets "tougher" to solve. In conclusion, we expect possible difficulties to occur in the GFE approximation of the RD two-point boundary value problem (6.1) when μ is $\ll \sigma$.

6.1.3 The linear system and the discrete maximum principle

We enter in more details with the GFE problem for the RD equation, and consider the case $r = 1$ with a uniform partition of $[0,1]$ into $M_h \geq 2$ elements of size $h = 1/M_h$. Matrix $\mathbf{B}$ can be written as the sum of a *diffusion matrix*

$$B^d_{ij} = \int_0^1 \mu \varphi'_j \varphi'_i \, dx = \begin{cases} -\dfrac{\mu}{h} & j = i-1 \\ \dfrac{2\mu}{h} & j = i \\ -\dfrac{\mu}{h} & j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

and of a *reaction matrix*

$$B^r_{ij} = \int_0^1 \sigma \varphi_j \varphi_i \, dx = \begin{cases} \dfrac{\sigma h}{6} & j = i-1 \\ \dfrac{4\sigma h}{6} & j = i \\ \dfrac{\sigma h}{6} & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

As for the load vector, in the simple case $f = 1$, we have

$$F_i = \int_0^1 \varphi_i \, dx = h, \qquad \forall i = 1, \ldots, N_h.$$

In conclusion, the explicit form of the linear algebraic system (5.2) in the case $r = 1$, $h = 1/M_h$ and $f = 1$ is

$$\begin{bmatrix} \left(\frac{2\mu}{h} + \frac{4\sigma h}{6}\right) & \left(-\frac{\mu}{h} + \frac{\sigma h}{6}\right) & & \emptyset \\ \left(-\frac{\mu}{h} + \frac{\sigma h}{6}\right) & \left(\frac{2\mu}{h} + \frac{4\sigma h}{6}\right) & \ddots & \\ & \ddots & \ddots & \left(-\frac{\mu}{h} + \frac{\sigma h}{6}\right) \\ \emptyset & & \left(-\frac{\mu}{h} + \frac{\sigma h}{6}\right) & \left(\frac{2\mu}{h} + \frac{4\sigma h}{6}\right) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N_h} \end{bmatrix} = h \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6.6}$$

Theorem 6.1.4 (Maximum principle for the RD problem)**.** *Let* $Lw := -\mu w'' + \sigma w$*. Then, L is inverse-monotone and satisfies the comparison principle C.2.5 with* $\phi(x) = 1/\sigma$*, so that the solution of* (6.1) *satisfies the maximum principle (MP)*

$$0 \le u(x) \le \max_{x \in [0,1]} \phi(x) = \frac{1}{\sigma} \qquad \forall x \in [0,1]. \tag{6.7}$$

Having established the MP property for the exact solution of (6.1), we would clearly like the corresponding finite element approximate solution u_h to inherit such a property. Should this occur, we say that u_h satisfies a *discrete maximum principle* (DMP). With this scope, the following definitions turn out to be useful.

Definition 6.1.5 (Inverse monotone matrix)**.** *An invertible square matrix* $\mathbf{A}$ *is said to be inverse monotone if*

$$\mathbf{A}^{-1} \geq 0$$

the inequality being understood in the element-wise sense. Should $\mathbf{A}$ *be inverse-monotone, then*

$$\mathbf{Aw} \leq \mathbf{Az} \quad \Rightarrow \quad \mathbf{w} \leq \mathbf{z}$$

(always in the element-wise sense), $\mathbf{w}$*,* $\mathbf{z}$ *being two vectors of* $\mathbb{R}^{N_h}$*.*

Definition 6.1.6 (M-matrix)**.** *An invertible square matrix* $\mathbf{A}$ *is an M-matrix if:*

- $A_{ij} \leq 0$ *for* $i \neq j$*;*
- $\mathbf{A}$ *is inverse-monotone.*

Then, we have the following important result.

Theorem 6.1.7 (Sufficient condition for DMP)**.** *If the stiffness matrix* $\mathbf{B} = \mathbf{B}^d + \mathbf{B}^r$ *is an M-matrix, then* u_h *satisfies the DMP, i.e.*

$$0 \leq u_h(x) \leq \frac{1}{\sigma} \qquad \forall x \in [0,1]. \tag{6.8}$$

To verify the property of being an M-matrix using directly Def. (6.1.6) is, in general, prohibitive. The following (necessary and sufficient) condition is useful.

Theorem 6.1.8 (Discrete comparison principle)**.** *Let* $\mathbf{A}$ *be a matrix with non-positive off diagonal entries (*$A_{ij} \leq 0$ *for* $i \neq j$*). Then,* $\mathbf{A}$ *is an M-matrix iff there exists a positive vector* $\mathbf{e}$ *such that* $\mathbf{Ae} > \mathbf{0}$ *(in the component-wise sense).*

Remark 6.1.9. A first choice to try for the test vector in Thm. 6.1.8 is $\mathbf{e} = [1,1,\ldots,1]^T \in \mathbb{R}^{N_h}$. Thus, computing the matrix-vector product $\mathbf{Ae}$ amounts to computing the row sum for each row of $\mathbf{A}$.

To apply Thm. 6.1.8 we need to ascertain that $B_{ij} \leq 0$. This is true iff

$$-\frac{\mu}{h} + \frac{\sigma h}{6} \leq 0,$$

that is, if the following (more conservative) condition is satisfied

$$\mathbb{P}e_{rd} < 1, \tag{6.9}$$

where

$$\mathbb{P}e_{rd} := \frac{\sigma h^2}{6\mu} \tag{6.10}$$

is the non-dimensional *Pèclet number* associated with the reaction-diffusion problem.

Theorem 6.1.10 (DMP principle)**.** *Assume that* (6.9) *is satisfied. Then* $\mathbf{B} = \mathbf{B}^d + \mathbf{B}^r$ *is an M-matrix and* u_h *satisfies the DMP* (6.8).

Proof. It suffices to apply the discrete comparison principle with $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{N_h}$ and then apply Thm. 6.1.7. □

Example 6.1.11 (Reaction-dominated problem). The upper bound on the Pèclet number (6.9) is equivalent to enforcing an upper limiting value for the mesh size h, that is

$$h < \sqrt{\frac{6\mu}{\sigma}} := h_{max} \tag{6.11}$$

The corresponding minimum number of (uniform) mesh elements is

$$M_{h,min} = \texttt{round}\left(\frac{1}{h_{max}}\right) = \texttt{round}\left(\sqrt{\frac{\sigma}{6\mu}}\right). \tag{6.12}$$

If h does not satisfy (6.11), then spurious oscillations are likely to occur in the computed solution u_h preventing the DMP from being verified. To see this, we study BVP (6.1) in the case where $\mu = 10^{-4}$, $\sigma = 1$ and $f = 1$, and start to take $M_h = 10$ uniform partitions of $[0,1]$, i.e., $h = 1/10 = 0.1$. Condition (6.11) would require $h < 2.45 \cdot 10^{-2}$, i.e., $M_h \geq 41$, to guarantee a DMP for u_h, therefore, we expect numerical difficulties to occur. This is exactly what is shown in Fig. 6.2. Wild oscillations arise in the neighbourhood of the boundary layers, at $x = 0$ and $x = 1$, but some small over-and undershoots are present also in the rest of the computational domain.

Things drastically change if we take $M_h = 41$. The result is shown in Fig. 6.3.

Oscillations have disappeared, and it can be checked that $\max_{x\in[0,1]} u_h(x) = 1$ (recall that $\sigma = 1$), so that the DMP is satisfied. However, a closer glance to u_h reveals that in the neighbourhood of the layers, the accuracy of the approximation is not very good, because h is not sufficiently small to "capture" the steep gradient of u. On the other hand, it is quite clear that the mesh size is excessively refined outside the layers, where the exact solution is almost constant.

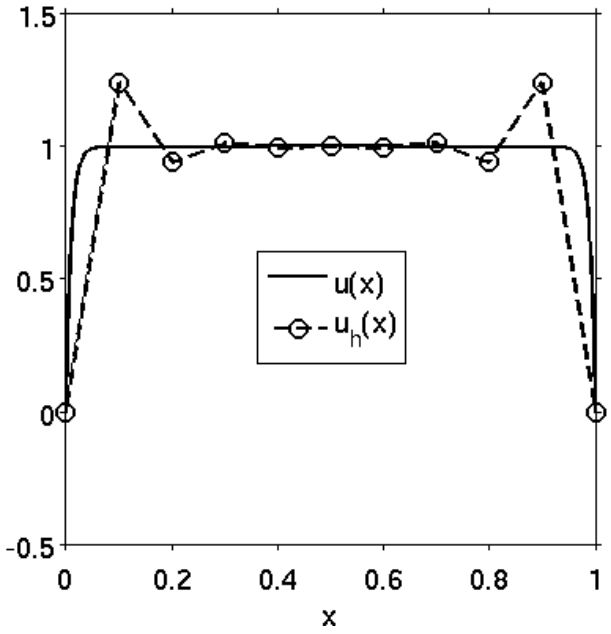


Figure 6.2: Plot of exact and approximate solutions ($r = 1$) in the case $M_h = 10$.

Definition 6.1.12 (Discrete maximum norm). *For any grid function* $\eta_h : \{x_i\}_{i=1}^{N_h} \to \mathbb{R}$, *we define the discrete maximum norm as*

$$\|\eta_h\|_{\infty,h} := \max_{i=1,\dots,N_h} |\eta_i|, \tag{6.13}$$

where the quantities η_i *are the nodal values of* η_h. *The norm* (6.13) *is the discrete version of the maximum norm* (3.7) *and coincides with the maximum norm for a vector* (A.2) *upon setting* $\mathbf{x} := [\eta_1, \eta_2, \dots, \eta_{N_h}]^T$.

Computing the discrete maximum norm of the error $u - u_h$ in the case of Fig. 6.3 yields $\|u - u_h\|_{\infty,h} = 0.086$, and it can be checked that the maximum nodal error occurs in correspondance of the very first internal node. A trade-off between numerical stability and accuracy may be obtained using a *non-uniform* partition $\mathscr{T}_h$, where the mesh size is smaller close to the boundary layers and larger elsewhere. The result of the application of this latter strategy is shown in Fig. 6.4, where the number of intervals ($M_h = 41$) is the same used in Fig. 6.3, but with a different distribution of the element size. In particular, h is equal to $0.1/15$ in $[0, 0.1]$ and $[0.9, 1]$, and equal to $0.8/11$ in $[0.1, 0.9]$. With this choice

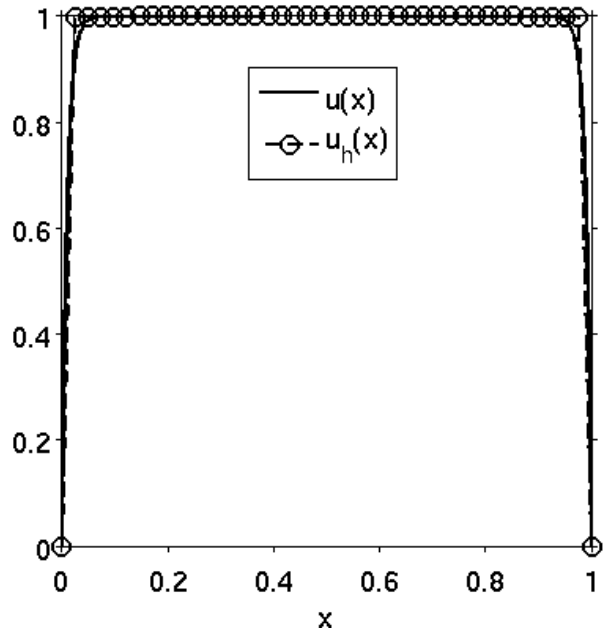


Figure 6.3: Plot of exact and approximate solutions ($r = 1$) in the case $M_h = 41$.

of the grid, the accuracy is highly improved, and the maximum nodal error is $\|u - u_h\|_{\infty,h} = 0.0067632$, i.e., more than ten times smaller than in the case of a uniform mesh.

6.1.4 Stabilization: the method of lumping of the reaction matrix

Ex. 6.1.11 has shown how to properly cope with a reaction-dominated problem. Basically, oscillations are removed by taking a sufficiently small mesh size h. In the case of a uniform mesh, however, this restriction may become too severe. For instance, assuming $\sigma = 1$, if $\mu = 10^{-4}$ then condition (6.11) yields $h_{max} = 2.45 \cdot 10^{-2}$, i.e., $M_{h,min} = 41$, while if $\mu = 10^{-6}$ we have $h_{max} = 2.45 \cdot 10^{-3}$, i.e., $M_{h,min} = 409$, and if $\mu = 10^{-8}$ we have $h_{max} = 2.45 \cdot 10^{-4}$, i.e., $M_{h,min} = 4083$. This trend may become even worse if the reaction-diffusion problem is to be solved in two or three spatial dimensions, because in such a case, condition (6.11) has to be satisfied in all spatial directions, giving rise to an explosion of the total number of mesh nodes, approximately as $M_{h,min}^d$, $d = 2, 3$.

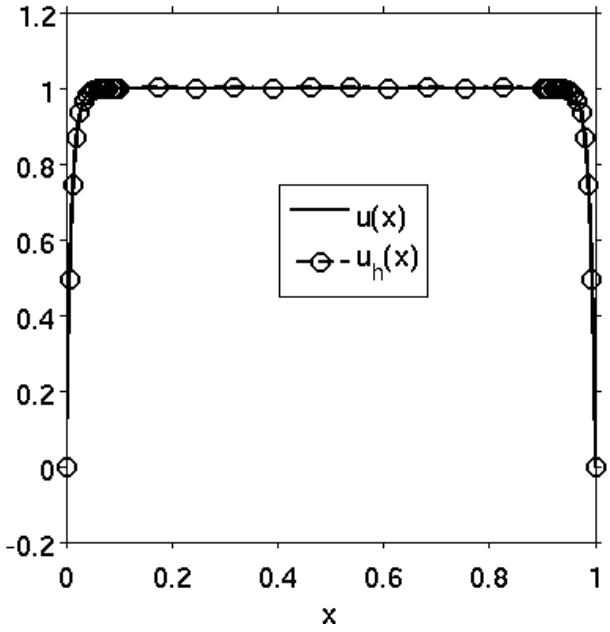


Figure 6.4: Plot of exact and approximate solutions ($r = 1$) in the case $M_h = 41$ and where the finite element mesh is non-uniform.

In order to ensure a numerically stable solution and, at the same time, maintain the computational effort to an acceptable level, a possible alternative is to introduce a modification to the standard GFE approximation (5.1) of (6.1). The modification belongs to the more general technique of *stabilization*, and consists, in the specific case of the reaction-diffusion BVP, of the following modified discrete problem:
find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = F(v_h) \qquad \forall v_h \in V_h, \tag{6.14}$$

where V_h is the finite element space $X^1_{h,0}$ and the *modified* bilinear form B_h is defined as

$$B_h(u_h, v_h) = \int_0^1 \mu u_h' v_h' \, dx + I_h^1(\sigma u_h v_h) \qquad \forall u_h, v_h \in V_h. \tag{6.15}$$

The proposed modification is represented by the use of the *trapezoidal quadrature rule* (cf. (3.10) with $r = 1$ and $[a, b] = [0, 1]$) to compute in an approximate manner

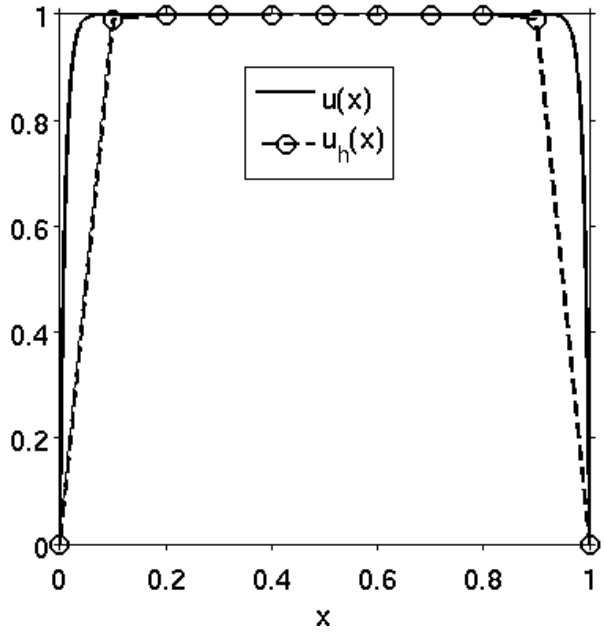


Figure 6.5: Plot of exact and approximate solutions ($r = 1$) in the case $M_h = 10$. The lumping stabilization has been used.

the integral

$$\int_0^1 \sigma u_h v_h \, dx.$$

The effect of the use of the trapezoidal numerical quadrature can be seen in Fig. 6.5, which shows the computed solution u_h over the same mesh as in Fig. 6.2. No spurious oscillations are present, $u_h(x)$ satisfies the DMP and is always bounded from above by the exact solution $u(x)$, unlike the case of Fig. 6.3 (with $M_h = 41$). Also, the maximum nodal error is equal to 0.0097595, which is far better than that of Fig. 6.3.

Remark 6.1.13 (The linear system). The use of trapezoidal quadrature is called *reduced integration* of the reaction matrix, because instead of computing exactly the entries B^r_{ij}, we are deliberately introducing a quadrature error, given by $(3.11)_2$. The advantage of such reduced integration is that the modified reaction matrix,

denoted $\widetilde{\mathbf{B}}^r$, is diagonal and positive

$$\widetilde{B}^r_{ij} = I^1_h(\sigma\varphi_j\varphi_i) = \begin{cases} \sigma\left(\frac{h}{2}+\frac{h}{2}\right) = \sigma h & j = i \\ 0 & \text{otherwise,} \end{cases}$$

so that the explicit form of the linear algebraic system (5.2) in the (stabilized) case is ($r = 1$, $f = 1$ and $h = 1/M_h$)

$$\begin{bmatrix} \left(\frac{2\mu}{h}+\sigma h\right) & -\frac{\mu}{h} & & \emptyset \\ -\frac{\mu}{h} & \left(\frac{2\mu}{h}+\sigma h\right) & \ddots & \\ & \ddots & \ddots & -\frac{\mu}{h} \\ \emptyset & & -\frac{\mu}{h} & \left(\frac{2\mu}{h}+\sigma h\right) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N_h} \end{bmatrix} = h \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6.16}$$

Theorem 6.1.14 (DMP principle (revisited)). *Problem* (6.14) *admits a unique solution that satisfies the DMP* (6.8) *for every value of the Pèclet number* $\mathbb{P}e_{ad}$.

Proof. By inspection, the modified stiffness matrix $\widetilde{\mathbf{B}} := \mathbf{B}^d + \widetilde{\mathbf{B}}^r$ is a s.d.p. matrix because it is the sum of a s.d.p. matrix with a positive diagonal matrix. Moreover, it is a strictly diagonally dominant matrix, with off-diagonal entries ≤ 0. Thus, the comparison principle shows that $\widetilde{\mathbf{B}}$ is an M-matrix irrespective of the Pèclet number. Thm. 6.1.7 then ensures the DMP to hold. □

Remark 6.1.15 (The lumping procedure). The diagonalization of the reaction matrix is often called *lumping* because the entries $\widetilde{B}^r_{ii}$ can be interpreted as obtained by summing by row the matrix $\mathbf{B}^r$. This is equivalent to "lump" the weight of the reaction term into each mesh node x_i. Such approach can be extended to higher-degree finite elements ($r \geq 2$), but no theoretical proof of DMP can be easily established.

Remark 6.1.16 (The error). The integration error $(3.11)_2$ associated with the use of the trapezoidal quadrature formula is of the order of h^2. This allows to show (using a generalized form of Ceà's Lemma known as Strang's Lemma) that

$$\begin{aligned} \|u-u_h\|_V &\leq Ch\|u\|_{H^2(\Omega)} \\ \|u-u_h\|_{L^2(\Omega)} &\leq Ch^2\|u\|_{H^2(\Omega)} \end{aligned} \tag{6.17}$$

provided $u \in H^1_0(\Omega) \cap H^2(\Omega)$.

6.2 Advection-diffusion model problem

Let us consider the following two-point BVP:

$$\begin{cases} -\mu u'' + a u' = f & \text{in } \Omega = (0,1) \\ u(0) = 0 \\ u(1) = 0, \end{cases} \tag{6.18}$$

where a is the *advection coefficient*. Throughout the section, unless otherwise stated, we assume $a > 0$, although completely similar analysis and results hold in the case $a < 0$. The equation system (6.18) is commonly referred to as an *advection-diffusion* (AD) problem, and represents a simple model of a chemical substance whose concentration is u, that diffuses in a fluid with diffusion coefficient μ and velocity a, and interacts with the environment according to a net production term given by f. Setting for simplicity $f = 1$, the exact solution of the problem is

$$u(x) = \frac{1}{a}\left[x - \frac{e^{\alpha x} - 1}{e^{\alpha} - 1}\right] \tag{6.19}$$

where $\alpha := a/\mu$. A plot of u is reported in Fig. 6.6 for three increasing values of α, corresponding to $a = 1$ and $\mu = 1, 10^{-1}, 10^{-3}$.

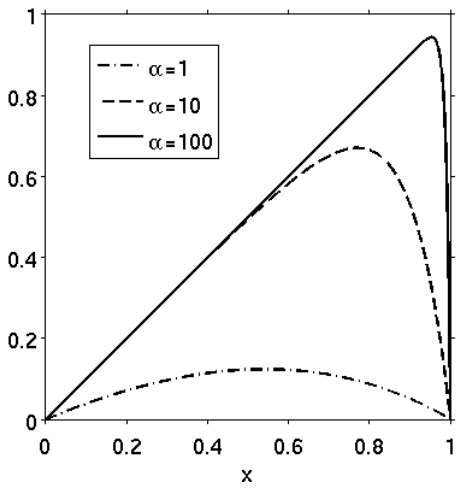


Figure 6.6: Plot of the solution of the advection-diffusion model problem as a function of α.

Fig. 6.6 shows that u tends to the linearly varying function $y(x) = x/a$, as $\alpha \to \infty$, except for an interval close to $x = 1$, denoted boundary layer, where a steep variation of u occurs to satisfy the boundary condition $u(1) = 0$. As a matter of fact, assuming $\alpha \gg 1$ and expanding the exponential terms in (6.19), we get

$$u(x) \simeq \frac{1}{a}(x - e^{\alpha(x-1)}).$$

The solution behaves, asymptotically, as the sum of two contributions: a linear function and an exponential term that becomes significant as x gets closer to 1. The exponential term is called *boundary layer term* because it is responsible for the variation of u from the value $1/a$ to the boundary condition $u(1) = 0$. As in the reaction-diffusion case, it can be proved that the width δ of the boundary layer is of the order of α^{-1}, therefore it gets smaller and smaller if diffusion is dominated by advection in the model (6.18).

Definition 6.2.1. *Model* (6.18) *is said to be advection-dominated if* $\alpha \gg 1$*, otherwise it is called diffusion-dominated.*

Remark 6.2.2 (Inflow/outflow boundaries). The solution of the advection-diffusion model problem is a *non symmetric* function of x. This is due to the fact that the advection term a introduces a preferential direction to the motion of u. In this case, a is positive so that the barycenter of the front u is drifted from left to right, as α increases. We distinguish between *inflow boundary* (the point $x = 0$), where we have $a \cdot n < 0$ (here, $n = -1$) and *outflow boundary* (the point $x = 1$), where we have $a \cdot n > 0$ (here, $n = +1$), having denoted by n the outward unit normal on $\partial\Omega$.

Matlab coding. The `Matlab` script for computing and plotting u is reported below.

```
close all
x=[0:0.0001:1];
b=1; mu=[1, 1e-1, 1e-2];
a=b./mu;
for k=1:numel(a),
    u(k,:)=1/b*(x- (exp(a(k)*x)-1)/(exp(a(k))-1));
end
plot(x,u(1,:),'k-.',x,u(2,:),'k--',x,u(3,:),'k-')
xlabel('x')
legend('\alpha=1','\alpha=10','\alpha=100')
```

Remark 6.2.3 (The conservative form of the AD problem). In many applications,

the BVP (6.18) is written in the following form of a first-order system:

$$\begin{cases} (J(u))' = f & \text{in } \Omega = (0,1) \\ J(u) = -\mu u' + au & \text{in } \Omega \\ u(0) = 0 \\ u(1) = 0. \end{cases} \tag{6.20}$$

The above system is called the *conservative form* of the AD problem. In particular, Eq. $(6.20)_1$ represents the law of conservation of the *advective-diffusive flux* $J(u)$ defined in Eq. $(6.20)_2$. The two formulations (6.18) and (6.20) are completely equivalent in the case where μ and a are constant coefficients.

6.2.1 Weak formulation

Let $V := H_0^1(0,1)$ with norm given by (B.28). Assume also that f is a given function in $L^2(0,1)$. Multiplying $(6.18)_1$ by an arbitrary test function $v \in V$ and integrating by parts the term $\int_0^1 -\mu u'' v\,dx$, we get the following weak formulation of the advection-diffusion BVP:
find $u \in V$ such that

$$\underbrace{\int_0^1 \mu u' v'\,dx + \int_0^1 au' v\,dx}_{B(u,v)} = \underbrace{\int_0^1 fv\,dx}_{F(v)} \qquad \forall v \in V \tag{6.21}$$

Theorem 6.2.4. *Problem* (6.21) *is uniquely solvable and u satifies the same estimate* (6.4) *as for the RD problem.*

Proof. Let us check that B satisfies the conditions required by the Lax-Milgram Lemma, because $F(\cdot)$ is the same as in the case of the RD problem.

- continuity of B:

$$\begin{aligned} |B(u,v)| &\le \mu \textstyle\int_0^1 |u'|\,|v'|\,dx + a\int_0^1 |u'||v|\,dx \\ &\le \mu \|u\|_V \|v\|_V + a\|u\|_V \|v\|_{L^2(0,1)} \\ &\le (\mu + aC_P)\|u\|_V \|v\|_V \qquad \forall u, v \in V, \end{aligned}$$

  from which $M = \mu + aC_P$;

- coercivity of B:

$$B(u,u) \quad = \quad \mu\|u\|_V^2 + a\int_0^1 u'u\,dx \qquad \forall u \in V.$$

We have

$$u'u = \frac{1}{2}(u^2)'$$

so that

$$\int_0^1 u'u\,dx = \frac{1}{2}\int_0^1 (u^2)'\,dx = 0$$

because $u \in H_0^1(0,1)$. Thus $\beta = \mu$.

Since all the assumptions of the Lax-Milgram Lemma are verified, the solution u of (6.21) exists and is unique in V, and satisfies (6.4). □

6.2.2 Galerkin finite element approximation

The GFE discretization of (6.18) is (5.1) and the corresponding matrix form is (5.2) with

$$\begin{aligned} B_{ij} &= \int_0^1 \mu\varphi_j'\varphi_i'\,dx + \int_0^1 a\varphi_j'\varphi_i\,dx \qquad i,j = 1,\ldots,N_h \\ F_i &= \int_0^1 f\varphi_i\,dx \qquad i = 1,\ldots,N_h. \end{aligned}$$

The stiffness matrix **B** is positive definite in virtue of Thm. 5.1.2, so that the discrete problem (5.1) admits a unique solution. Moreover, Ceà's Lemma tells us that, assuming $u \in H^{r+1}(0,1)$

$$\|u - u_h\|_V \leq C\left(1 + \frac{a}{\mu}\right) h^r \|u\|_{H^{r+1}(0,1)}. \tag{6.22}$$

Thus, we see that u_h converges to u, in the H^1-norm, as $h \to 0$, and that accuracy improves as the polynomial order is increased, because of the optimal regularity of the exact solution.

Remark 6.2.5 (The advection-dominated case). Let us examine the estimate (6.22) as a function of model parameters, μ and a. For this, we set $r = 1$ (piecewise linear finite elements) and fix a tolerance ε, sufficiently small. Then, we see that in order the error to be $\simeq \varepsilon$, we need that

$$h \simeq \frac{\varepsilon}{C(1+\alpha)\|u\|_{H^2(0,1)}}.$$

Therefore, if the advection-diffusion problem is advection-dominated ($\alpha \gg 1$), the mesh size needs to be much smaller than the tolerance ε, in accordance with Fig. 6.6, because the boundary layer effect at $x = 0$ becomes increasingly significant as α gets larger and the AD problem gets "tougher" to solve. In conclusion, we expect possible difficulties to occur in the GFE approximation of the AD two-point boundary value problem (6.18) when μ is $\ll |a|$.

6.2.3 The linear system and the discrete maximum principle

We enter in more details with the GFE problem for the AD equation, and consider the case $r = 1$ with a uniform partition of $[0,1]$ into $M_h \geq 2$ elements of size $h = 1/M_h$. Matrix $\mathbf{B}$ can be written as the sum of the diffusion matrix $\mathbf{B}^d$ and of the *advection matrix*

$$B^a_{ij} = \int_0^1 a\varphi'_j\varphi_i\,dx = \begin{cases} -\dfrac{a}{2} & j = i-1 \\ 0 & j = i \\ +\dfrac{a}{2} & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

In conclusion, the explicit form of the linear algebraic system (5.2) in the case $r = 1$, $h = 1/M_h$ and $f = 1$ is

$$\begin{bmatrix} \dfrac{2\mu}{h} & \left(-\dfrac{\mu}{h}+\dfrac{a}{2}\right) & & \emptyset \\ \left(-\dfrac{\mu}{h}-\dfrac{a}{2}\right) & \dfrac{2\mu}{h} & \ddots & \\ & \ddots & \ddots & \left(-\dfrac{\mu}{h}+\dfrac{a}{2}\right) \\ \emptyset & & \left(-\dfrac{\mu}{h}-\dfrac{a}{2}\right) & \dfrac{2\mu}{h} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N_h} \end{bmatrix} = h \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6.23}$$

We notice that in the case of the AD problem the stiffness matrix $\mathbf{B}$ is no longer symmetric as in the case of the RD problem. As a matter of fact, the diffusion matrix $\mathbf{B}^d$ is symmetric (and positive definite), while the advection matrix $\mathbf{B}^a$ is not symmetric.

In the case $f = 1$, we immediately see that a barrier function for the AD problem is $\phi(x) = x/a$, so that the application of the comparison principle C.2.5 allows us to conclude the following result.

Theorem 6.2.6 (Maximum principle for the AD problem)**.** *The solution of* (6.18) *satisfies the maximum principle (MP)*

$$0 \le u(x) \le \frac{1}{a} \qquad \forall x \in [0,1]. \tag{6.24}$$

Remark 6.2.7. The analysis of Thm. 6.2.6 is confirmed by Fig. 6.6 where $a = 1$. The conclusions of Rem. C.2.7 apply obviously also to the case of the AD problem.

Let us now investigate under which conditions the finite element approximate solution u_h of (6.18) enjoys a DMP, by applying the discrete comparison principle 6.1.8. As in the RD case, we need to ascertain that $B_{ij} \le 0$. The entries $B_{i,i-1}$ are ≤ 0 because $a > 0$, while the coefficients $B_{i,i+1}$ are nonpositive iff

$$-\frac{\mu}{h} + \frac{a}{2} \le 0,$$

that is, if the following (more conservative) condition is satisfied

$$\mathbb{P}e_{ad} < 1, \tag{6.25}$$

where

$$\mathbb{P}e_{ad} := \frac{ah}{2\mu} \tag{6.26}$$

is the non-dimensional Pèclet number associated with the advection-diffusion problem.

Remark 6.2.8 (Definition of Pèclet number). If $a < 0$, the definition of the Pèclet number becomes

$$\mathbb{P}e_{ad} := \frac{|a|h}{2\mu}. \tag{6.27}$$

Theorem 6.2.9 (DMP principle)**.** *Assume that* (6.25) *is satisfied. Then* $\mathbf{B} = \mathbf{B}^d + \mathbf{B}^a$ *is an M-matrix and* u_h *satisfies the DMP*

$$0 \le u_h(x) \le \frac{1}{a} \qquad \forall x \in [0,1]. \tag{6.28}$$

Proof. It suffices to apply the discrete comparison principle with $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{N_h}$ and then apply Thm. 6.1.7. □

Example 6.2.10 (Advection-dominated problem). The upper bound on the Pèclet number (6.25) is equivalent to enforcing an upper limiting value for the mesh size h, that is

$$h < \frac{2\mu}{a} := h_{max} \tag{6.29}$$

The corresponding minimum number of (uniform) mesh elements is

$$M_{h,min} = \texttt{round}\left(\frac{1}{h_{max}}\right) = \texttt{round}\left(\frac{a}{2\mu}\right). \tag{6.30}$$

If h does not satisfy (6.29), then spurious oscillations are likely to occur in the computed solution u_h preventing the DMP from being verified. To see this, we study BVP (6.18) in the case where $\mu = 5 \cdot 10^{-3}$, $a = 1$ and $f = 1$, and start to take $M_h = 10$ uniform partitions of $[0,1]$, i.e., $h = 1/10 = 0.1$. Condition (6.29) would require $h < 10^{-2}$, i.e., $M_h \geq 100$, to guarantee a DMP for u_h, therefore, we expect numerical difficulties to occur. This is exactly what is shown in Fig. 6.7. Wild oscillations arise in the neighbourhood of the boundary layer, at $x = 1$, and propagate throughout the whole computational domain, polluting completely the correct behaviour of the exact solution.

Things drastically change if we take $M_h = M_{h,min} = 100$. The result is shown in Fig. 6.8. Oscillations have disappeared, and it can be checked that $\max_{x\in[0,1]} u_h(x) = 0.99$ in agreement with Thm. 6.2.9 which predicts $u_h(x) \leq 1/a = 1$ for all $x \in [0,1]$, so that the DMP is satisfied.

A closer glance to u_h reveals that in the neighbourhood of the outflow boundary layer, the accuracy of the approximation is not very good, because h is not sufficiently small to "capture" the steep gradient of u. On the other hand, it is quite clear that the mesh size is excessively refined outside the layers, where the exact solution is practically linear. To clarify this issue in a quantitative manner, we compute the discrete maximum norm of the error $u - u_h$ in the case of Fig. 6.3 obtaining $\|u - u_h\|_{\infty,h} = 0.13534$, and it can be checked that the maximum nodal error occurs in correspondance of the last internal node. A trade-off between numerical stability and accuracy may be obtained using a *non-uniform* partition $\mathscr{T}_h$, where the mesh size is smaller close to the boundary layers and larger elsewhere. The result of the application of this latter strategy is shown in Fig. 6.9, where the number of intervals ($M_h = 100$) is the same used in Fig. 6.8, but with a different

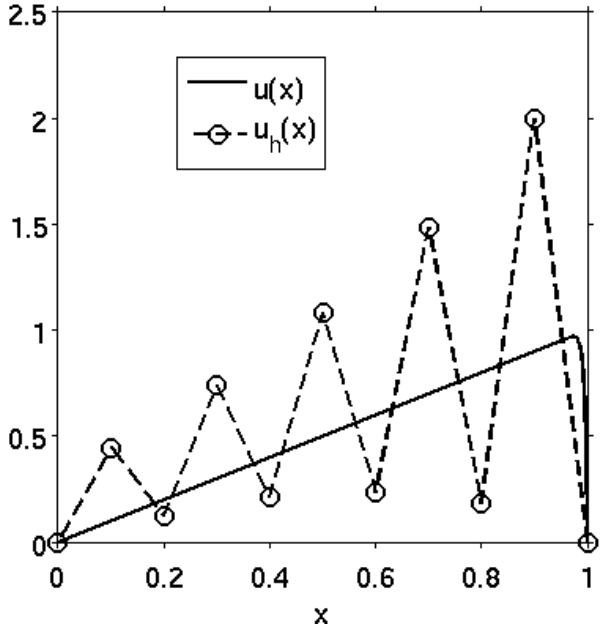


Figure 6.7: Plot of exact and approximate solutions ($r = 1$) in the case $\mu = 5 \cdot 10^{-3}$ and $M_h = 10$

distribution of the element size. In particular, h is equal to 0.1 in $[0, 0.9]$ and equal to $0.1/90$ in $[0.9, 1]$. With this choice of the grid, the accuracy is highly improved, and the maximum nodal error is $\|u - u_h\|_{\infty,h} = 0.001513$, i.e., almost 100 times smaller than in the case of a uniform mesh.

6.2.4 Stabilization: the method of artificial diffusion

Ex. 6.2.10 has shown how to properly cope with an advection-dominated problem. Basically, oscillations are removed by taking a sufficiently small mesh size h. In the case of a uniform mesh, however, this restriction may become too severe. For instance, assuming $a = 1$, if $\mu = 5 \cdot 10^{-4}$ then condition (6.29) yields $h_{max} = 10^{-3}$, i.e., $M_{h,min} = 1000$, while if $\mu = 5 \cdot 10^{-6}$ we have $h_{max} = 10^{-5}$, i.e., $M_{h,min} = 100000$, and if $\mu = 5 \cdot 10^{-8}$ we have $h_{max} = 10^{-7}$, i.e., $M_{h,min} = 10^7$ (ten millions of elements!). This trend may become even worse if the advection-diffusion problem is to be solved in two or three spatial dimensions, as already commented in the case of the RD equation.

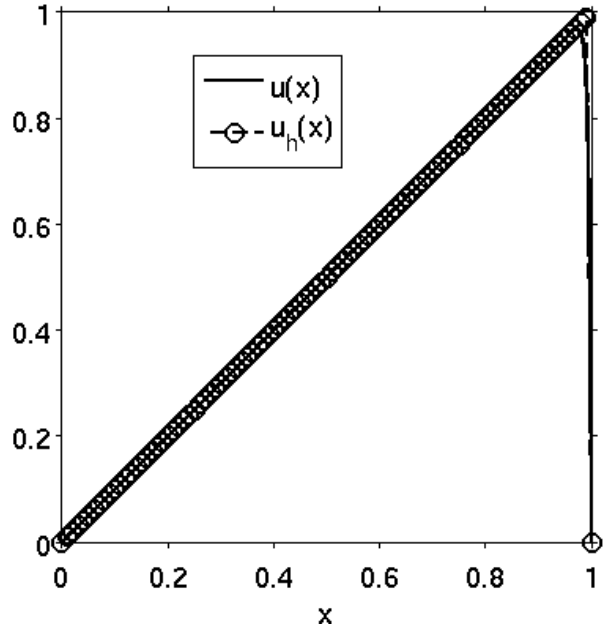


Figure 6.8: Plot of exact and approximate solutions ($r = 1$) in the case $\mu = 5 \cdot 10^{-3}$ and $M_h = 100$..

In order to ensure a numerically stable solution and, at the same time, maintain the computational effort to an acceptable level, a possible alternative is to resort to a stabilized form of the GFEM as in (6.14), where the modified bilinear form B_h is defined as

$$B_h(u_h, v_h) = B(u_h, v_h) + \underbrace{\int_0^1 \mu \Phi(\mathbb{P}e_{ad}) u_h' v_h' \, dx}_{b_h(u_h, v_h)} \qquad \forall u_h, v_h \in V_h. \tag{6.31}$$

The additional contribution b_h represents the discrete weak form of the *artificial diffusion term*

$$-\mu \Phi(\mathbb{P}e_{ad}) u''$$

so that the generalized GFEM (6.14), with B_h as in (6.31), can be regarded as the

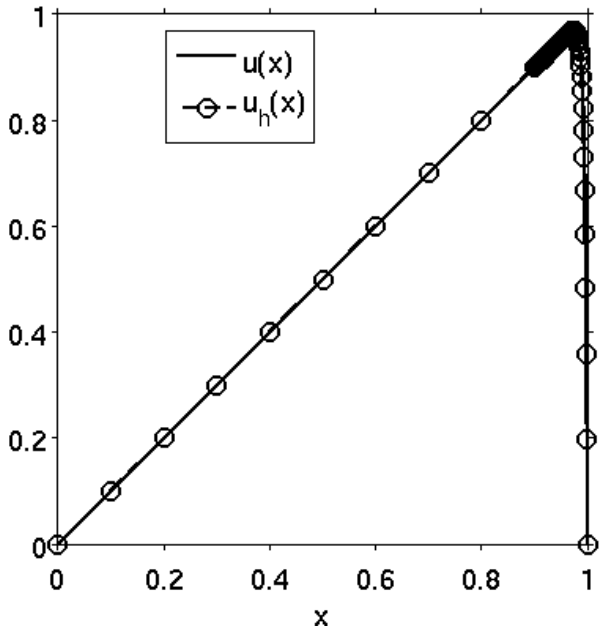


Figure 6.9: Plot of exact and approximate solutions ($r = 1$) in the case $\mu = 5 \cdot 10^{-3}$ and $M_h = 100$, where the finite element mesh is non-uniform.

standard GFE approximation of the *modified* advection-diffusion two-point BVP:

$$\begin{cases} -\mu(1+\Phi(\mathbb{P}e_{ad}))u'' + au' = f & \text{in } \Omega = (0,1) \\ u(0) = 0 \\ u(1) = 0. \end{cases} \tag{6.32}$$

Clearly, in order to ensure numerical stability, the *stabilization function* Φ has to be designed with care. In particular, it is required that:

$$\Phi(t) \geq 0 \qquad \text{if } t \geq 0; \tag{6.33a}$$

$$\lim_{t \to 0^+} \Phi(t) = 0. \tag{6.33b}$$

Condition (6.33a) amounts to requiring that if no advection term is present in the model ($a = 0$) then $\Phi = 0$, so that no artificial diffusion is introduced. Condition (6.33b) tells us that the amount of artificial diffusion has to be small when the advective field strength $|a|$ is small, in order the modified AD problem (6.32) to

be a small perturbation of the original problem (6.18) and, consequently, the computed solution u_h to be a sufficiently accurate approximation of the exact solution of (6.18).

Definition 6.2.11 (Pèclet number of the modified AD problem)**.** *Let*

$$\mu_h := \mu(1+\Phi(\mathbb{P}e_{ad})) \tag{6.34}$$

denote the modified diffusion coefficient. Then, the Pèclet number of the modified AD problem is

$$\widetilde{\mathbb{P}e}_{ad} := \frac{ah}{2\mu_h} = \frac{ah}{2\mu(1+\Phi(\mathbb{P}e_{ad}))} = \frac{\mathbb{P}e_{ad}}{1+\Phi(\mathbb{P}e_{ad})}. \tag{6.35}$$

Proposition 6.2.12 (Choice of Φ)**.** *Let $\mathbb{P}e_{ad}$ be the Pèclet number associated with the AD problem* (6.18) *on a given triangulation $\mathscr{T}_h$. Then, we have*

$$\Phi(\mathbb{P}e_{ad}) > \mathbb{P}e_{ad} - 1 \qquad \Rightarrow \qquad \widetilde{\mathbb{P}e}_{ad} < 1. \tag{6.36}$$

Using the above proposition we have the following result.

Theorem 6.2.13 (Stability of the GFEM with artificial diffusion)**.** *If the stabilization function Φ satisfies* (6.36)*, then the solution u_h computed by the GFEM with artificial diffusion satisfies the DMP for every value of $\mathbb{P}e_{ad}$.*

Among the possible choices for Φ, we consider here two alternatives:

- Upwind (UP) stabilization function

$$\Phi^{UP}(t) := t \qquad t \geq 0 \tag{6.37}$$

- Scharfetter-Gummel (SG) stabilization function

$$\Phi^{SG}(t) := t - 1 + \mathscr{B}(2t) \qquad t \geq 0 \tag{6.38}$$

  where

$$\mathscr{B}(t) := \frac{t}{e^t - 1}$$

  is the inverse of the Bernoulli function, such that $\mathscr{B}(-t) = t + \mathscr{B}(t)$. Expanding e^t in Mc Laurin series, we see that $\mathscr{B}(0) = 1$. Moreover, we have the following asymptotical behavior:

$$\lim_{t\to\infty} \mathscr{B}(t) = \begin{cases} 0 & t > 0 \\ -t & t < 0 \end{cases} \qquad \lim_{t\to\infty} \mathscr{B}(-t) = \begin{cases} t & t > 0 \\ 0 & t < 0. \end{cases}$$

  A plot of $\mathscr{B}(t)$ and $\mathscr{B}(-t)$ is reported in Fig. 6.10.

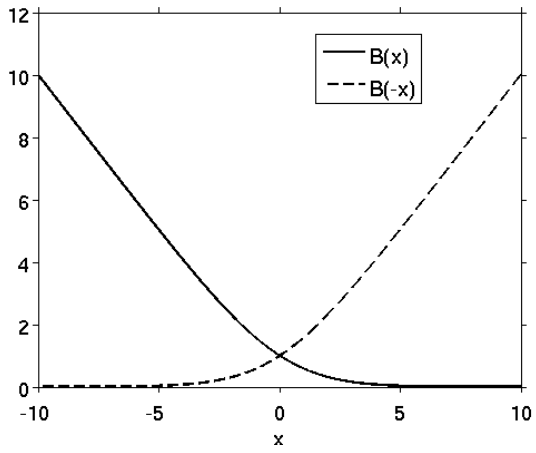


Figure 6.10: Plot of $\mathscr{B}(x)$.

Theorem 6.2.14 (DMP principle (revisited))**.** *Problem* (6.14) *with the UP and SG stabilizations admits a unique solution that satisfies the DMP* (6.8) *for every value of the Pèclet number* $\mathbb{P}e_{rd}$.

Proof. The stiffness matrix $\widetilde{\mathbf{B}} = \widetilde{\mathbf{B}}^d + \mathbf{B}^a$ is positive definite, so that (6.14) is uniquely solvable. From definitions (6.37) and (6.38), we see that both stabilization functions satisfy Prop. 6.2.12, so that the UP and SG generalized GFE formulations both satisfy the DMP due to Thm. 6.2.13. □

Using the asymptotical properties of (6.38), we can also see that $\Phi^{SG}(\mathbb{P}e_{ad}) \simeq \Phi^{UP}(\mathbb{P}e_{ad})$ for large values of the Pèclet number, so that the two schemes produce the same approximation. Things go very differently if $\mathbb{P}e_{ad}$ is small. As a matter of fact, we have

$$\lim_{t\to 0}\Phi^{SG}(t) \simeq t - 1 + \frac{2t}{1+2t+\dfrac{4t^2}{2}-1} = t - 1 + \frac{1}{1+t} = \frac{t^2}{1+t} = t^2,$$

which shows that the amount of artificial diffusion introduced by the SG stabilization is far smaller than that of the UP stabilization, for which we have

$$\Phi^{UP}(t) = t \qquad t \to 0.$$

Therefore, in a moderately to weakly advection-dominated regime, we expect the SG and UP stabilized GFEMs to produce substantially different approximations,

that computed by the SG method being characterized by a better accuracy than that of the UP method. As a matter of fact we have the following result.

Theorem 6.2.15 (Convergence in the discrete maximum norm)**.** *Assume that μ, a and f are given continuous functions on $[0,1]$, and let h be sufficiently small. Then, the UP and SG stabilized GFEMs are both convergent in the discrete maximum norm and we have*

$$\|u-u_h\|_{\infty,h} \leq Ch \qquad \textit{Upwind stabilization}, \tag{6.39a}$$

$$\|u-u_h\|_{\infty,h} \leq Ch^2 \qquad \textit{SG stabilization}. \tag{6.39b}$$

A more striking result holds for the SG stabilization.

Theorem 6.2.16 (Nodal exactness)**.** *Let $r=1$. Assume that $\mu>0$, a and f are given constants (with $f=1$), and that $\mathscr{T}_h$ is a uniform partition of $\overline{\Omega}$. Then the solution $u_h \in V_h = X^1_{h,0}$ of the SG GFEM satisfies the following relation*

$$u_h(x_i) = u(x_i) = \frac{1}{a}\left[x_i - \frac{e^{\alpha x_i}-1}{e^{\alpha}-1}\right] \qquad i=1,\ldots,N_h. \tag{6.40}$$

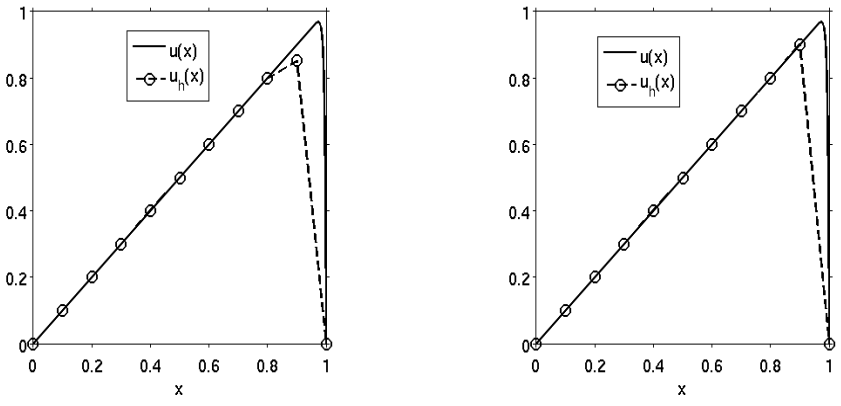
Remark 6.2.17. Relation (6.40) expresses the property of *nodal exactness* of the SG GFEM, and tells us that the approximate solution u_h computed by the SG stabilized formulation coincides with the Π^1_h interpolant of the exact solution of the AD BVP (6.18). Because of the above particularly exceptional accuracy of the SG method, this latter is also known as *exponential fitting method* or *optimal artificial diffusion method*.

The convergence performance in the V-norm of the stabilized formulations is illustrated by the following result.

Theorem 6.2.18 (Convergence in the V-norm)**.** *Let $u \in H^{r+1}(\Omega)$ be the solution of* (6.18) *and $u_h \in V_h = X^r_{h,0}$ be the corresponding approximation computed by the generalized GFEM* (6.14) *with artificial diffusion. Then, for h sufficiently small, we have*

$$\|u-u_h\|_V \leq C\frac{h^r}{\mu(1+\Phi(\mathbb{P}e_{ad}))}\|u\|_{H^{r+1}(\Omega)} + \frac{\Phi(\mathbb{P}e_{ad})}{1+\Phi(\mathbb{P}e_{ad})}\|u\|_{H^{r+1}(\Omega)}. \tag{6.41}$$

Remark 6.2.19. Three important conclusions can be drawn from (6.41):



(a) $\Phi = \Phi^{UP}$ (b) $\Phi = \Phi^{SG}$

Figure 6.11: Plot of exact and approximate solutions ($r = 1$) in the case $\mu = 5 \cdot 10^{-3}$ and $M_h = 10$. Left: upwind stabilization. Right: SG stabilization. The SG scheme is nodally exact in the special case of constant coefficients and uniform grid.

1. the stabilized GFEM converges to the exact solution of (6.18) because of property (6.33b);
2. the stabilized GFEM with $\Phi = \Phi^{UP}$ has order of convergence equal to 1 irrespective of the polynomial degree r;
3. the choice $\Phi = \Phi^{SG}$ is optimal for $r = 1, 2$, while it is sub-optimal if $r \geq 3$.

Fig. 6.11 shows the numerical solutions computed on a very coarse grid by the UP and SG methods. We see that both approximations satisfy the DMP, the UP approximation being not very accurate in the layer region while the SG approximation is nodally exact. The L^2 error, the H^1 error and the error in the discrete maximum norm are equal to 0.17371, 3.2014 and 0.047619 for the UP method, while are equal to 0.16871, 3.2559 and 7.7716e-16 for the SG method (round-off error).

Chapter 7

2D Implementation of the GFEM

Abstract

In this chapter, we describe the principal issues that characterize the implementation in a computational algorithm of the GFEM. For ease of presentation, we restrict the analysis to the case where the domain Ω is a 2D polygon, with Lipschitz boundary $\Gamma = \partial\Omega$ and outward unit normal vector $\mathbf{n}$. The model boundary value problem considered in this chapter is:

$$\begin{cases} Lu := -\mu\triangle u + \mathbf{a}\cdot\nabla u + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma, \end{cases} \tag{7.1}$$

where μ, $\mathbf{a}$, σ and f are assumed to be continuous functions over $\overline{\Omega}$. The BVP (7.1) is a representative example of a diffusion-advection-reaction differential problem. The advection field $\mathbf{a}$ is a given two-dimensional vector field with components $a_x = a_x(x,y)$, $a_y = a_y(x,y)$, having denoted with $\mathbf{x} = (x,y)^T$ the position vector in Ω. The reaction coefficient σ is ≥ 0 over $\overline{\Omega}$ while the diffusion coefficient μ is such that $0 < \mu_{min} \leq \mu(\mathbf{x}) \leq \mu_{max}$ for all $\mathbf{x} \in \overline{\Omega}$.

7.1 Weak formulation

Set $V := H_0^1(\Omega)$, endowed with the norm $\|\cdot\|_V$ given by (B.36). Then, the weak formulation of (7.1) is:

find $u \in V$ such that

$$\underbrace{\int_\Omega \mu \nabla u \cdot \nabla v + (\mathbf{a} \cdot \nabla u) v + \sigma u v \, d\Omega}_{B(u,v)} = \underbrace{\int_\Omega f v \, d\Omega}_{F(v)} \qquad \forall v \in V. \tag{7.2}$$

Assume that $\mathbf{a} \in (C^1(\Omega))^2$. Then, under the following (sufficient) condition

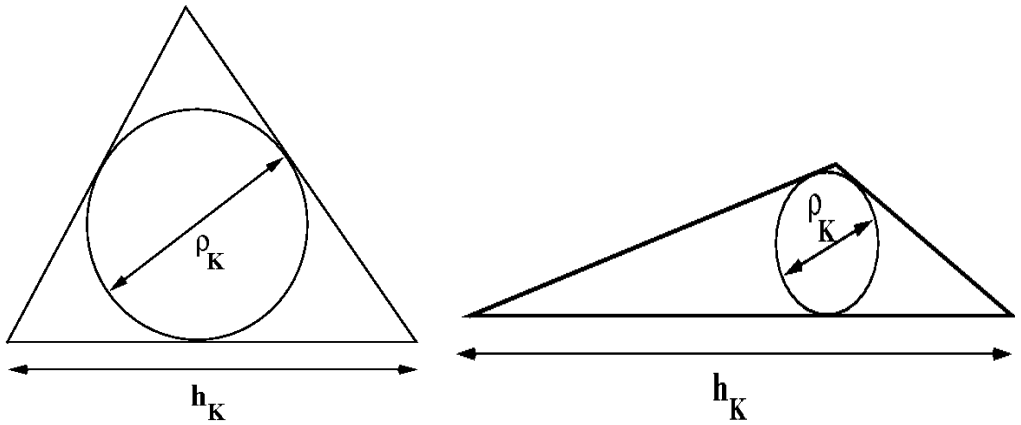
$$\sigma(\mathbf{x}) - \frac{1}{2}\mathrm{div}\mathbf{a}(\mathbf{x}) \geq 0 \qquad \forall \mathbf{x} \in \Omega, \tag{7.3}$$

it is immediate, using the Lax-Milgram Lemma, to check that (7.2) admits a unique (weak) solution such that

$$\|u\|_V \leq \frac{C_P \|f\|_{C^0(\overline{\Omega})}}{\mu_{min}}. \tag{7.4}$$

7.2 Geometrical discretization

Let us introduce a geometrical partition of $\overline{\Omega}$ into a family of triangulations $\mathscr{T}_h$, $h > 0$, made of triangular elements K, such that $h_K = \mathrm{diam}(K)$ and $h = \max_{K \in \mathscr{T}_h} h_K$. For each $K \in \mathscr{T}_h$, we denote by ρ_K the "sphericity" of K, i.e., the diameter of the largest circle that can be inscribed within K (see Fig. 7.1(a)).



(a) Regular triangle (b) Triangle with large aspect ratio

Figure 7.1: Geometrical notation of a triangulation.

Definition 7.2.1 (Aspect ratio)**.** *For each element $K \in \mathscr{T}_h$, we define the aspect ratio as*

$$R_K := \frac{h_K}{\rho_K}. \tag{7.5}$$

Definition 7.2.2 (Regular triangulation)**.** *A triangulation $\mathscr{T}_h$ is said to be regular if there exists a positive constant κ, independent of h, such that*

$$R_K \leq \kappa \qquad \forall K \in \mathscr{T}_h. \tag{7.6}$$

Remark 7.2.3. The regularity condition practically means that every triangle of $\mathscr{T}_h$ cannot have an arbitrarily large aspect ratio, as it would be the case, for example, shown in Fig. 7.1(b) where $R_K \to \infty$ as the rightmost vertex is moved along the right-hand direction. Equivalently, condition (7.6) prevents the minimum angle of $\mathscr{T}_h$ from being too small. From now on, we assume that each member of $\mathscr{T}_h$ is a regular triangulation.

7.3 Polynomial spaces in 2D

In this section, we introduce the space of polynomials that will be used in the GFE approximation of (7.1).

Definition 7.3.1 (The space $\mathbb{P}_r$ in 2D)**.** *Let $r \geq 0$ be a fixed integer. Then, the space of polynomials of degree $\leq r$ with respect to the independent variables x, y is defined as*

$$\mathbb{P}_r = \operatorname{span}\left\{x^p y^q\right\}_{p,q=0,\ldots,r}^{0\leq p+q\leq r}. \tag{7.7}$$

Example 7.3.2. In the case $r = 0$, we only have $\mathbb{P}_0 = \operatorname{span}\{1\}$. In the case $r = 1$, Def. (7.7) yields $\mathbb{P}_1 = \operatorname{span}\{1, x, y\}$, if $r = 2$, we have $\mathbb{P}_2 = \operatorname{span}\left\{1, x, y, x^2, xy, y^2\right\}$, and, finally, if $r = 3$, we have $\mathbb{P}_3 = \operatorname{span}\left\{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\right\}$.

Definition 7.3.3 (The space $\mathbb{P}_r(K)$)**.** *We set*

$$\mathbb{P}_r(K) := \{p = p(x,y) \in \mathbb{P}_r,\, (x,y) \in K\}. \tag{7.8}$$

A simple count of dofs yields the following result.

Proposition 7.3.4 (Dimension of $\mathbb{P}_r(K)$)**.** *Let $r \geq 1$ be a fixed integer. Then*

$$\dim(\mathbb{P}_r(K)) = \frac{(r+1)(r+2)}{2} \equiv N_r(K) \qquad \forall K \in \mathscr{T}_h. \tag{7.9}$$

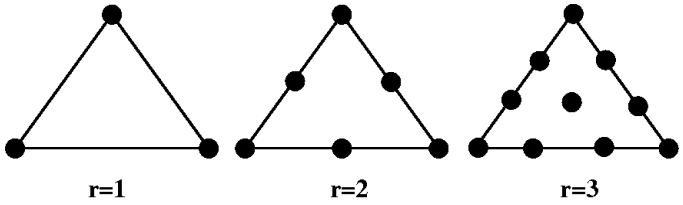


Figure 7.2: Black bullets denote the dofs of $\mathbb{P}_r(K)$.

The monomial basis used in (7.7) is, in general, not the most preferable for computations. An equivalent choice is represented by the set of Lagrangian (local) basis functions $\varphi_i = \varphi_i(\mathbf{x})$, with $\varphi_i \in \mathbb{P}_r(K)$ for $i = 1,\ldots,N_r(K)$, such that

$$\varphi_i(\mathbf{x}_j) = \delta_{ij} \qquad i,j = 1,\ldots,N_r(K), \tag{7.10}$$

where $\mathbf{x}_i$, $i = 1,\ldots,N_r(K)$ are the coordinates of the nodes over the element K. The nodal positions of the dofs for $\mathbb{P}_r(K)$ in the cases $r = 1,2,3$ are illustrated in Fig. 7.2. We see that in the case $r = 3$, dofs are also located in the *interior* of K and not only along its boundary ∂K.

7.4 The approximation space

From now on, we always assume, otherwise stated, that the polynomial degree r is a fixed integer ≥ 1. The finite dimensional space $V_h \subset V$ that is used in the GFE approximation of (7.1) is the two-dimensional generalization of the space $X^r_{h,0}$ introduced in (5.10) in the 1D case

$$V_h := X^r_{h,0}(\mathscr{T}_h) = \left\{ v_h \in C^0(\overline{\Omega}),\ v_h|_K \in \mathbb{P}_r(K)\ \forall K \in \mathscr{T}_h,\ v_h = 0 \text{ on } \Gamma \right\}. \tag{7.11}$$

By definition of vector space, any function $v_h \in V_h$ can be written as

$$v_h(\mathbf{x}) = \sum_{i=1}^{N_h(r)} v_i \varphi_i(\mathbf{x}), \tag{7.12}$$

where $N_h(r)$ is the dimension of V_h and represents the number of lineary independent basis functions spanning V_h. Such functions, still denoted by φ_i, $i = 1,\ldots,N_h(r)$, are the *global* basis functions whose restriction over each element K is one of the $N_r(K)$ *local* basis functions introduced in (7.10), while the real numbers v_i, $i = 1,\ldots,N_h(r)$, are the *nodal values* of v_h (see Fig. 7.3).

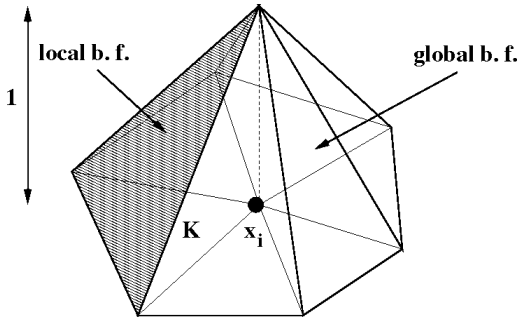


Figure 7.3: Global and local basis functions ($r = 1$).

Remark 7.4.1 (Dimension of V_h). The count of $N_h(r)$ proceeds in the same manner as in the 1D case. We first need to determine the number of dofs $N_{r,i}(K)$ that are *internal* to each element K. We have

$$N_r(K) = \frac{r^2+3r+2}{2} = \underbrace{3}_{\text{vertices}} + \underbrace{3(r-1)}_{\text{edges}} + \underbrace{N_{r,i}(K)}_{\text{internal}}$$

from which we obtain

$$N_{r,i}(K) = \frac{r^2+3r+2}{2} - 3r = \frac{r^2-3r+2}{2} \qquad r \geq 1.$$

Then, we distinguish between:

- number of internal vertices: $N_{v,i}$;
- number of internal edges: $N_{e,i}$;
- number of elements: N_K.

Finally, using the fact that any function belonging to V_h is continuous across every internal edge, we get

$$N_h(r) = N_{v,i} + (r-1)N_{e,i} + \frac{r^2-3r+2}{2}N_K. \tag{7.13}$$

Relation (7.13) is the generalization of (5.11) to the 2D case and to triangular grids.

7.5 GFE approximation

The GFE approximation of (7.1) is:
find $u_h \in V_h$ such that

$$\underbrace{\int_\Omega \mu \nabla u_h \cdot \nabla v_h + (\mathbf{a} \cdot \nabla u_h) v_h + \sigma u_h v_h \, d\Omega}_{B(u_h, v_h)} = \underbrace{\int_\Omega f v_h \, d\Omega}_{F(v_h)} \qquad \forall v_h \in V_h. \tag{7.14}$$

Under assumption (7.3) it is easy to use Lax-Milgram Lemma to check that (7.14) admits a unique solution u_h that satisfies the same a priori estimate (7.4) valid for the solution u of (7.2). Moreover, assuming that $u \in H^s(\Omega) \cap H_0^1(\Omega)$, $s \geq 2$ being a given quantity, the discretization error satisfies the estimate (5.16), which in the present case takes the following form

$$\|u - u_h\|_V \leq C \frac{\mu_{max} + C_P \|\mathbf{a}\|_{C^0(\overline{\Omega})} + C_P^2 \|\sigma\|_{C^0(\overline{\Omega})}}{\mu_{min}} h^l \|u\|_{H^{l+1}(\Omega)} \tag{7.15}$$

where $l := \min\{r, s-1\}$.

Remark 7.5.1. Estimate (7.15) tells us that the numerical solution on the computer machine of problem (7.14) may be negatively influenced by three possible sources: 1) large variation of the diffusion coefficient; 2) large dominance of advection; 3) large dominance of reaction. In each of these three cases, the choice of h that ensures a small error may be strongly penalized with respect to situations where problem coefficients vary more gently over Ω. In such cases, suitable stabilization techniques (generalizing those analyzed in Chapt. 6) have to be adopted.

7.6 The linear system

Problem (7.14) can be written in the form of linear algebraic system as

$$\mathbf{Bu} = \mathbf{f}, \tag{7.16}$$

where the entries of the stiffness matrix $\mathbf{B}$ are

$$B_{ij} = \int_\Omega \mu \nabla \varphi_j \cdot \nabla \varphi_i + (\mathbf{a} \cdot \nabla \varphi_j) \varphi_i + \sigma \varphi_j \varphi_i \, d\Omega \qquad i, j = 1, \ldots, N_h(r)$$

and the load vector is given by

$$F_i = \int_\Omega f \varphi_i \, d\Omega \qquad i = 1, \ldots, N_h(r).$$

Remark 7.6.1. The stiffness matrix $\mathbf{B}$ is positive definite, but not symmetric.

Using the property of addivity of an integral, we can write

$$\begin{aligned} B_{ij} &= \sum_{K\in\mathscr{T}_h} \underbrace{\int_K \mu\nabla\varphi_j\cdot\nabla\varphi_i + (\mathbf{a}\cdot\nabla\varphi_j)\varphi_i + \sigma\varphi_j\varphi_i\, dK}_{B_{ij}^K} \qquad i,j = 1,\ldots,N_h(r) \\ F_i &= \sum_{K\in\mathscr{T}_h} \underbrace{\int_K f\varphi_i\, dK}_{f_i^K} \qquad i = 1,\ldots,N_h(r). \end{aligned} \tag{7.17}$$

Relations (7.17) tell us that the computation of the *global* stiffness matrix and load vector is reduced to a `for-loop` over each element $K \in \mathscr{T}_h$ that computes:

- a *local stiffness matrix* $\mathbf{B}^K \in \mathbb{R}^{N_r(K)\times N_r(K)}$;
- a *local load vector* $\mathbf{f}^K \in \mathbb{R}^{N_r(K)}$.

7.7 The assembly phase

For each $K \in \mathscr{T}_h$, let us introduce:

- the local bilinear form

$$B_K(\varphi_j,\varphi_i) = B(\varphi_j|_K,\varphi_i|_K), \qquad i,j = 1,\ldots,N_h(r);$$

- the local linear form

$$F_K(\varphi_i) = F(\varphi_i|_K) \qquad i = 1,\ldots,N_h(r).$$

B_K and F_K are nothing but B and F evaluated in correspondance of the *restrictions* of the global basis functions φ_i and φ_j on the element K. We notice that two kinds of indexing quantities involved in the assembly phase are being used:

- global node numbering: $i,j = 1,\ldots,N_h(r)$;
- global element numbering: $K \in \mathscr{T}_h$.

These two global numberings must not be independent, rather they have to be mutually related in order to distribute the contributions coming from each $K \in \mathscr{T}_h$ into the right global row and colum indices i and j. With this aim, it is convenient to introduce the so-called *connectivity matrix* $\mathbf{T}^{N_r(K)\times N_K}$, whose generic entry $T_{I,J}$ contains, for each mesh element $J = 1,\ldots,N_K$, the global number of the local node $I = 1,\ldots,N_r(K)$, having adopted the convention that local node numbering is made according to a counterclockwise orientation. In the example of Fig. 7.4, we would have

$$T_{1,53} = 78, \qquad T_{2,53} = 900, \qquad T_{3,53} = 1213.$$

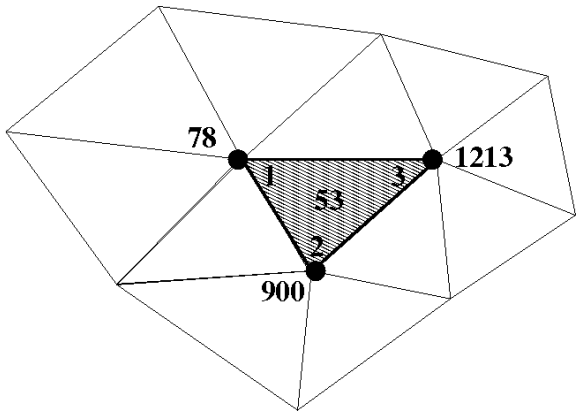


Figure 7.4: Local and global numbering of nodes and elements ($r = 1$).

Matlab coding. A `Matlab` script for assembling $\mathbf{B}$ and $\mathbf{f}$ is reported below. The list of the functions, variables and parameters is commented here:

- `compute_local_stiffness_matrix` is the function that computes the local stiffness matrix $\mathbf{B}^K$ using a suitable 2D quadrature formula;
- `compute_local_load_vector` is the function that computes the local load vector $\mathbf{f}^K$ using a suitable 2D quadrature formula;
- `N_nodes_tot` represents the total number of nodes (i.e., the interior nodes $N_h(r)$ plus the boundary nodes);
- `N_dofs_loc` is $N_r(K)$;

- the input parameters `mu`, `a`, `sigma` and `f` are function pointers to compute problem coefficients;
- `X_dofs_K` and `Y_dofs_K` contain the x and y coordinates of the $N_r(K)$ dofs over the element K.

```
B = sparse(N_nodes_tot);
f = sparse(N_nodes_tot,1);
for K=1:N_K
    B_K = compute_local_stiffness_matrix(mu,a,sigma,X_dofs_K,Y_dofs_K);
    f_K = compute_local_load_vector(f,X_dofs_K,Y_dofs_K);
    for i_loc = 1:N_dofs_loc
        row = T(i_loc,K);
        for j_loc = 1:N_dofs_loc
            col        = T(j_loc,K);
            B(row,col) = B(row,col) + B_K(i_loc,j_loc);
        end
        f(row) = f(row) + F_K(i_loc);
    end
end
```

Remark 7.7.1 (The boundary conditions). To account for the homogeneous boundary conditions $u|_\Gamma = 0$, we need to eliminate all the rows and columns of `B` and all the rows of `f` that correspond to nodes belonging to Γ.

Matlab coding. The `Matlab` commands for enforcing Dirichlet homogeneous boundary conditions are reported below. The variable `Dir_nodes` contains the list of nodes of $\mathscr{T}_h$ which belong to Γ.

```
>> B(Dir_nodes,Dir_nodes) = [];
>> f(Dir_nodes)           = [];
```

Once the above assembly loop is performed, system (7.16) is ready to be solved using a direct or iterative method as described in Sect. 2.

Example 7.7.2 (The homogeneous problem for the Laplacian). To close this short presentation of the 2D implementation of the GFEM, we solve problem (7.1) using the toolbox `pdetool` provided by the `Matlab` software environment. The GFEM is implemented in such a software using piecewise linear continuous elements ($r = 1$). We have set $\Omega = (0,1) \times (0,1)$, $\mu = 1$, $\mathbf{a} = \mathbf{0}$, $\sigma = 0$ and $f(x,y) = 32(y - y^2 + x - x^2)$ in such a way that the exact solution is the function $u(x,y) = 16xy(x-1)(y-1)$ (whose value at the center of the unit square is 1). A plot of the discrete solution u_h computed over a grid of average mesh size of 0.05 is shown in Fig. 7.5.

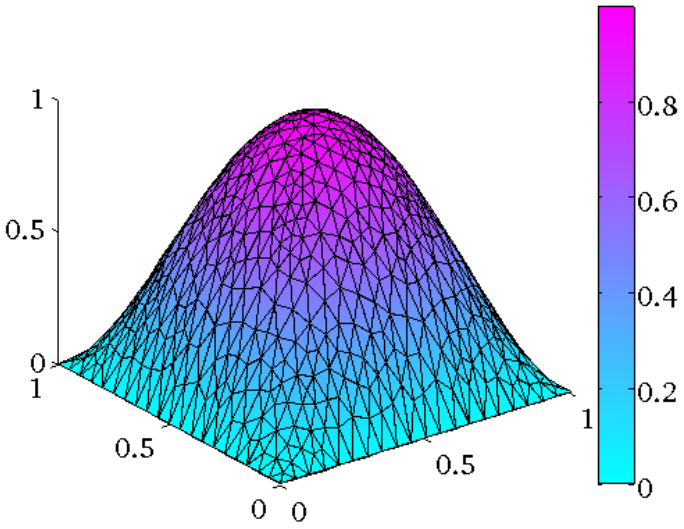


Figure 7.5: Numerical solution of the homogeneous Dirichlet problem for the Laplace operator using `pdetool` ($r = 1$).

Part IV

The GFEM for Linear Elasticity

This part addresses the study of the Navier-Lamè equations for linear elasticity in the compressible and incompressible regimes. Theory is carried out in a general three-dimensional setting while numerical approximation using the GFEM is addressed in two spatial dimensions (plane stress/plane strain).

Chapter 8

Compressible Linear Elasticity: Theory and Finite Element Approximation

Abstract

In this chapter, we apply the general machinery of weak formulation and Galerkin Finite Element approximation to the case of the Navier-Lamè equations of linear elasticity in the compressible regime. The considered approach is the classical displacement-based formulation in the two-dimensional (2D) case (plane stress and plane strain conditions). Computational tests to validate the numerical performance of the method are run in `Matlab` using the code `EF2D_el` developed by Marco Restelli and available at the link:
`http://www1.mate.polimi.it/CN/MNIC/Laboratori/EF2D_el_new.zip`.

8.1 Essentials of solid mechanics

Let $\mathscr{B}$ be a material body which occupies a volume portion $\Omega \subset \mathbb{R}^3$. The surface of the body, denoted by Γ is divided into two mutually disjoint parts, Γ_D and Γ_N, and a unit outward normal vector $\mathbf{n} = [n_1, n_2, n_3]^T$ is defined almost everywhere (a.e.) on Γ. On Γ_D, the body is constrained at ground, while on Γ_N a force per unit area $\mathbf{g} = [g_1, g_2, g_3]^T$ is applied. Moreover, at each point of the volume Ω, a body force per unit volume $\mathbf{f} = [f_1, f_2, f_3]^T$ is defined (see Fig. 8.1).

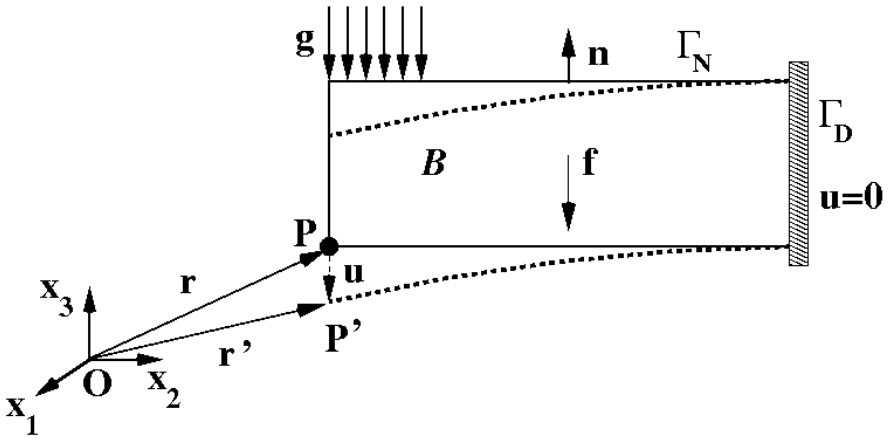


Figure 8.1: Elastic body: geometrical notation; undeformed and deformed configurations; loads and constraints.

The displacement of a material point $P = [x_1, x_2, x_3]^T$ of $\mathscr{B}$, because of the deformation consequent to the action of the loads, is expressed by the vector

$$\mathbf{r}' - \mathbf{r} := \mathbf{u} = \begin{bmatrix} x_1' - x_1 \\ x_2' - x_2 \\ x_3' - x_3 \end{bmatrix},$$

where the coordinates of the point P in the deformed state are denoted with a $(\cdot)'$ superscript. The local measure of the deformation of the body is given by the *Green-Lagrange deformation tensor*

$$\mathscr{U}_{ik} := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} + \frac{\partial u_l}{\partial x_i}\frac{\partial u_l}{\partial x_k}\right) \tag{8.1}$$

where the Einstein convention of repeated indices has been used in the third term at the right-hand side.

Definition 8.1.1 (The strain tensor)**.** *Assume that deformations are small, i.e.*

$$\left|\frac{\partial u_l}{\partial x_i}\right| \ll 1. \tag{8.2}$$

Then, the Green-Lagrange tensor can be approximated by the (small) deformation tensor *or* strain tensor*, defined as*

$$\varepsilon_{ik} := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right). \tag{8.3}$$

Using (C.9) into definition (8.3), we see that

$$\varepsilon = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right) \equiv \nabla_S\mathbf{u}$$

where $\nabla_S\mathbf{u}$ denotes the *symmetric gradient* of the displacement $\mathbf{u}$. The strain tensor is symmetric and its trace is given by

$$\mathrm{Tr}(\varepsilon) = \sum_{i=1}^{3} \varepsilon_{ii} = \mathrm{div}\mathbf{u} = \mathrm{Tr}(\nabla\mathbf{u}), \tag{8.4}$$

having used (C.10) in the last equality. The last identity is consistent with the fact that

$$\mathrm{Tr}(\nabla\mathbf{u}) = \mathrm{Tr}(\nabla_S\mathbf{u}) + \mathrm{Tr}(\nabla_{SS}\mathbf{u}) = \mathrm{Tr}(\nabla_S\mathbf{u}) = \mathrm{Tr}(\varepsilon),$$

because the trace of a skew-symmetric tensor is identically zero, by definition.

Remark 8.1.2 (Small deformations/small displacements). The assumption of small deformations does not necessarily imply that also displacements are small. To see this, consider the case of a long thin rod under large deflection conditions, in which the motion of the end points of the rod may be non-negligible but the extensions and compressions in the rod are small. In practice, however, the displacement vector $\mathbf{u}$ for a three-dimensional body under small deformation is itself small. This amounts to adding, from now on, to (8.2) the so-called *small displacement assumption*

$$|u_i| \ll \ell, \tag{8.5}$$

where ℓ is a characteristic length of the 3D body.

8.1.1 The relation $\sigma - \varepsilon$

Consider the elastic body $\mathscr{B}$ under the uniaxial loading condition of Fig. 8.2. Let

$$\varepsilon := \frac{\ell - \ell_0}{\ell_0}, \qquad \varepsilon_t := \frac{d - d_0}{d_0}$$

denote the *longitudinal* and *transveral* deformation, respectively. Define also the *normal stress* (in N/mm^2, or MPa)

$$\sigma := \frac{F}{A_0}$$

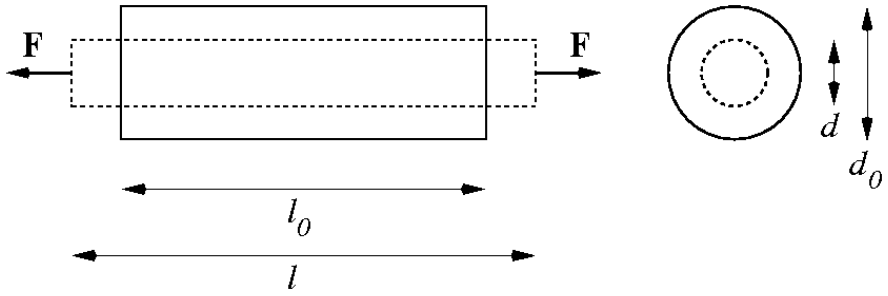


Figure 8.2: Elastic body subject to an uniaxial load.

where A_0 is the areal cross-section of the body, and the *transversal contraction coefficient*

$$\nu := -\frac{\varepsilon_t}{\varepsilon}.$$

Definition 8.1.3 (Reversibility)**.** *The relation $\sigma - \varepsilon$ is reversible whenever the applied force is released, the deformation is completely recovered, i.e., the body returns to the original (undeformed) configuration. In mathematical terms, this corresponds to assuming that the relation $\sigma = \sigma(\varepsilon)$ is invertible.*

Definition 8.1.4 (The linear relation: Hooke's law)**.** *Within a certain deformation range*

$$0 \leq |\varepsilon| \leq \varepsilon_{max},$$

the relation $\sigma - \varepsilon$ is linear, i.e.

$$\sigma = E\varepsilon \tag{8.6}$$

where E is the so-called Young modulus (in MPa). Relation (8.6) *is universally known as Hooke's law ("Ut tensio, sic vis"). In the general 3D case,* (8.6) *becomes*

$$\sigma_{ij} = \mathscr{C}_{ijkl}\varepsilon_{kl} \qquad i,j,k,l = 1,2,3 \tag{8.7}$$

or, in tensor notation

$$\sigma = \mathscr{C}\varepsilon. \tag{8.8}$$

The fourth-order tensor $\mathscr{C}$ is called the elastic matrix of the material and the $3^4 = 81$ quantities $\mathscr{C}_{ijkl}$ are the elastic constants of the material. Relation (8.7) *is the so-called generalized Hooke's law.*

The constraint of the stress and deformation tensors to be symmetric

$$\sigma_{ij} = \sigma_{ji}, \quad \text{and} \quad \varepsilon_{kl} = \varepsilon_{lk}$$

and the fact that the body energy functional is a symmetric and positive definite quadratic form

$$\frac{\partial \sigma_{ij}}{\partial \varepsilon_{kl}} = \frac{\partial \sigma_{kl}}{\partial \varepsilon_{ij}},$$

allow us to conclude that the elastic tensor $\mathscr{C}$ is symmetric, so that the 81 constants reduce to 21, and the matrix form of (8.7) reads

$$\underbrace{\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{bmatrix}}_{\sigma} = \underbrace{\begin{bmatrix} \mathscr{C}_{1111} & \mathscr{C}_{1122} & \mathscr{C}_{1133} & \mathscr{C}_{1112} & \mathscr{C}_{1123} & \mathscr{C}_{1131} \\ & \mathscr{C}_{2222} & \mathscr{C}_{2233} & \mathscr{C}_{2212} & \mathscr{C}_{2223} & \mathscr{C}_{2231} \\ & & \mathscr{C}_{3333} & \mathscr{C}_{3312} & \mathscr{C}_{3323} & \mathscr{C}_{3331} \\ & \text{Sym} & & \mathscr{C}_{1212} & \mathscr{C}_{1223} & \mathscr{C}_{1231} \\ & & & & \mathscr{C}_{2323} & \mathscr{C}_{2331} \\ & & & & & \mathscr{C}_{3131} \end{bmatrix}}_{\mathscr{C}} \underbrace{\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{23} \\ 2\varepsilon_{31} \end{bmatrix}}_{\varepsilon}.$$

8.1.2 Linear isotropic elasticity

Definition 8.1.5 (The linear isotropic elastic regime)**.** *A material body $\mathscr{B}$ in which the mechanical response is independent of the coordinate system is said to be a linear isotropic elastic material.*

Under this condition, the $\sigma - \varepsilon$ relation takes the following form

$$\sigma_{ij} = [\lambda \delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})]\varepsilon_{kl} \qquad i,j,k,l = 1,2,3, \tag{8.9}$$

or, more simply

$$\sigma_{ij} = 2\mu\varepsilon_{ij} + \lambda\varepsilon_{kk}\delta_{ij} \qquad i,j = 1,2,3. \tag{8.10}$$

Example 8.1.6. Relation (8.10) can be easily verified. Take, for instance, $i = 1$, $j = 2$ in (8.9), and obtain

$$\begin{aligned} \sigma_{12} &= \lambda\delta_{12}[\delta_{11}\varepsilon_{11} + \delta_{22}\varepsilon_{22} + \delta_{33}\varepsilon_{33}] + \mu[\delta_{11}\delta_{22}\varepsilon_{12} + \delta_{11}\delta_{22}\varepsilon_{21}] \\ &= \lambda \operatorname{div}\mathbf{u}\,\delta_{12} + 2\mu\varepsilon_{12} = 2\mu\varepsilon_{12} \end{aligned}$$

that is, relation (8.10) in the case $i = 1$, $j = 2$.

The two quantities, λ and μ, are known as the *Lamè constants* of the material and have the following definitions in terms of the material mechanical parameters E (Young modulus) and ν (Poisson modulus)

$$\lambda := \frac{\nu E}{(1+\nu)(1-2\nu)}, \qquad \mu := \frac{E}{2(1+\nu)}. \tag{8.11}$$

Remark 8.1.7 (Limits for E and ν). The Young modulus E is a strictly positive quantity. The Poisson modulus, in principle, satisfies

$$-1 < \nu < 0.5.$$

In practice, it is customary to assume

$$0 < \nu < 0.5. \tag{8.12}$$

The limiting case $\nu = 0.5$ is critical, as can be seen from definition (8.11) because λ becomes infinite. In this situation, the elastic material is said to be *incompressible* and $\mathscr{B}$ can deform under the condition that its volume remains constant. This case will be studied in detail in Chapter 9.

Using (8.10), the stress-strain relation in matrix form yields

$$\underbrace{\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{bmatrix}}_{\sigma} = \underbrace{\begin{bmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ & & \lambda+2\mu & 0 & 0 & 0 \\ & \text{Sym} & & \mu & 0 & 0 \\ & & & & \mu & 0 \\ & & & & & \mu \end{bmatrix}}_{\mathscr{C}} \underbrace{\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{23} \\ 2\varepsilon_{31} \end{bmatrix}}_{\varepsilon} \tag{8.13}$$

or, in compact tensor notation

$$\sigma = 2\mu\varepsilon + \lambda\,(\mathrm{Tr}\varepsilon)\,\delta = 2\mu\varepsilon + \lambda\,\mathrm{div}\mathbf{u}\,\delta, \tag{8.14}$$

where δ is the identity tensor and (8.4) has been used.

8.2 Mathematical model of isotropic linear elasticity

Let $\Omega \subset \mathbb{R}^3$ be the volume of a linear elastic isotropic body $\mathscr{B}$ before deformation occurs (see Fig. 8.1 for notation). Let $\mathbf{f} : \Omega \to \mathbb{R}^3$ and $\mathbf{g} : \Gamma_N \to \mathbb{R}^3$ denote two given vector fields representing forces per unit volume and unit surface, respectively. Finally, let $E > 0$ and $\nu \in (0, 1/2)$ denote the Young modulus and Poisson coefficient of the material constituting the body, respectively. The problem of linear elastic isotropic elasticity (shortly, elasticity, from now on) reads:

find the displacement $\mathbf{u} : \Omega \to \mathbb{R}^3$ such that:

$$\begin{aligned} \operatorname{div}\sigma(\mathbf{u}) + \mathbf{f} &= \mathbf{0} \quad \text{in } \Omega && (8.15a)\\ \sigma(\mathbf{u}) = \mathscr{C}\varepsilon(\mathbf{u}) &= 2\mu\nabla_S\mathbf{u} + \lambda \operatorname{div}\mathbf{u}\,\delta \quad \text{in } \Omega && (8.15b)\\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma_D && (8.15c)\\ \sigma(\mathbf{u})\mathbf{n} &= \mathbf{g} \quad \text{on } \Gamma_N, && (8.15d) \end{aligned}$$

or, using Einstein's notation:
find $u_i : \Omega \to \mathbb{R}^3$ such that:

$$\begin{aligned} \sigma_{ij,j} + f_i &= 0 \quad \text{in } \Omega && (8.16a)\\ u_i &= 0 \quad \text{on } \Gamma_D && (8.16b)\\ \sigma_{ij} n_j &= g_i \quad \text{on } \Gamma_N && (8.16c)\\ \sigma_{ij} = \mathscr{C}_{ijkl}\varepsilon_{kl} &= 2\mu\varepsilon_{ij} + \lambda\varepsilon_{kk}\delta_{ij} \quad \text{in } \Omega && (8.16d)\\ \varepsilon_{ij} &= \frac{1}{2}(u_{i,j} + u_{j,i}) \quad \text{in } \Omega. && (8.16e) \end{aligned}$$

- (8.15a): indefinite equilibrium equation (force balance);
- (8.15b): constitutive relation (stress-strain);
- (8.15c): homogeneous Dirichlet boundary condition (the body is constrained to the ground on a part of its boundary);
- (8.15d): Neumann boundary condition (the body is subject to external forces per unit area on a part of its boundary).

Remark 8.2.1. The mathematical model (8.15) is usually referred to as the *Navier-Lamè* system for linear isotropic elasticity. It is clearly possible to eliminate the strain and stress fields from the constitutive laws (8.16e)-(8.16d) in favor of the sole displacement field $\mathbf{u}$ to obtain the following form of the Navier-Lamè system:
find $u_i : \Omega \to \mathbb{R}^3$ such that:

$$\begin{aligned} -\operatorname{div}\left(2\mu\nabla_S\mathbf{u} + \lambda \operatorname{div}\mathbf{u}\delta\right) &= \mathbf{f} \quad \text{in } \Omega && (8.17a)\\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma_D && (8.17b)\\ \left(2\mu\nabla_S\mathbf{u} + \lambda \operatorname{div}\mathbf{u}\,\delta\right)\mathbf{n} &= \mathbf{g} \quad \text{on } \Gamma_N, && (8.17c) \end{aligned}$$

The BVP (8.17) is also called *displacement formulation* of elasticity because the only remaining dependent variable is the displacement field $\mathbf{u}$.

Remark 8.2.2 (Reaction forces). Integrating (8.15a) over the body volume yields

$$\int_\Omega (\mathrm{div}\boldsymbol{\sigma} + \mathbf{f})\, d\Omega = \mathbf{0}.$$

Using Green's formula to treat the first contribution, we obtain

$$\int_{\Gamma_D} \boldsymbol{\sigma}\mathbf{n}\, d\Gamma + \int_{\Gamma_N} \boldsymbol{\sigma}\mathbf{n}\, d\Gamma + \int_\Omega \mathbf{f}\, d\Omega = \mathbf{0}$$

Enforcing the Neumann boundary condition (8.15d), we obtain the following relation that expresses the *global* balance of forces that act on the material body

$$\mathbf{R}_D + \mathbf{F}_{tot,N} + \mathbf{F}_{tot,\Omega} = \mathbf{0} \tag{8.18}$$

where:

- (reaction forces, to be determined in order to satisfy global equilibrium)

$$\mathbf{R}_D := \int_{\Gamma_D} \boldsymbol{\sigma}\mathbf{n}\, d\Gamma;$$

- (total surface forces)

$$\mathbf{F}_{tot,N} := \int_{\Gamma_N} \mathbf{g}\, d\Gamma;$$

- (total volume forces)

$$\mathbf{F}_{tot,\Omega} := \int_\Omega \mathbf{f}\, d\Omega.$$

8.3 The weak formulation

Let us now apply the machinery of weak formulation of a BVP to the elasticity problem (8.15). To this purpose, from now on we assume that $\mathbf{f} \in (L^2(\Omega))^3$ and $\mathbf{g} \in (L^2(\Gamma_N))^2$, respectively. Then, we set

$$V := (H^1_{0,\Gamma_D}(\Omega))^3 \tag{8.19}$$

and multiply (8.15a) by a vector-valued test function $\mathbf{v} \in V$ to obtain

$$\int_\Omega (\mathrm{div}\boldsymbol{\sigma}(\mathbf{u}) + \mathbf{f}) \cdot \mathbf{v}\, d\Omega = 0 \qquad \forall \mathbf{v} \in V.$$

Using Green's formula to treat the first term, enforcing the Neumann boundary conditions on Γ_N and the Dirichlet boundary conditions for $\mathbf{v}$ on Γ_D, we obtain:
find $\mathbf{u} \in V$ such that

$$\int_\Omega \sigma(\mathbf{u}) : \nabla \mathbf{v}\, d\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega + \int_{\Gamma_N} \mathbf{g}\cdot\mathbf{v}\, d\Gamma \qquad \forall \mathbf{v} \in V.$$

Proposition 8.3.1 (Scalar product with a symmetric tensor)**.** *Let $q \in \mathbb{R}^{3\times 3}$ be a given tensor. For every symmetric tensor $\tau \in \mathbb{R}^{3\times 3}$, we have*

$$\tau : q = \tau : (q_S + q_{SS}) = \tau : q_S. \tag{8.20}$$

Applying Prop. 8.3.1 to the previous integral equality, where $\tau \equiv \sigma(\mathbf{u})$ and $q \equiv \nabla \mathbf{v}$, and using the genaralized Hooke's law (8.15b), we finally obtain the weak formulation of the elastic BVP:
find $\mathbf{u} \in V$ such that

$$\underbrace{\int_\Omega \mathscr{C}\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\Omega}_{B(\mathbf{u},\mathbf{v})} = \underbrace{\int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega + \int_{\Gamma_N} \mathbf{g}\cdot\mathbf{v}\, d\Gamma}_{F(\mathbf{v})} \qquad \forall \mathbf{v} \in V. \tag{8.21}$$

Theorem 8.3.2 (Variational formulation of (8.21))**.** *The minimization problem: find $\mathbf{u} \in V$ such that*

$$J(\mathbf{u}) \le J(\mathbf{v}) \qquad \forall \mathbf{v} \in V \tag{8.22}$$

where

$$\begin{aligned} J(\mathbf{v}) &= \underbrace{\frac{1}{2}\int_\Omega \sigma(\mathbf{v}) : \varepsilon(\mathbf{v})\, d\Omega}_{\textit{strain energy}} \underbrace{- \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega - \int_{\Gamma_N} \mathbf{g}\cdot\mathbf{v}\, d\Gamma}_{\textit{work against external forces}} \\ &= \int_\Omega \mu\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\Omega + \frac{1}{2}\int_\Omega \lambda (\operatorname{div}\mathbf{v})^2\, d\Omega \\ &\quad - \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega - \int_{\Gamma_N} \mathbf{g}\cdot\mathbf{v}\, d\Gamma \qquad \forall \mathbf{v} \in V, \end{aligned} \tag{8.23}$$

is completely equivalent to the weak problem (8.21).

Remark 8.3.3 (Principle of minimum of Potential Energy and Principle of Virtual Works). The functional (8.23) is the *total potential energy* stored in the elastic body $\mathscr{B}$. Thus, the minimization problem (8.22) is the *principle of minimum of the potential energy* while the weak formulation (8.21) is the *principle of virtual works*. In this regard, the test function $\mathbf{v}$ has the mechanical meaning of *virtual displacement*.

8.4 Existence and uniqueness of the weak solution

In this section, we apply the theoretical tool of the Lax-Milgram Lemma to verify that (8.21) admits a unique solution depending continuously on the data of the elastic problem.

Theorem 8.4.1 (Korn's inequality)**.** *There exists a positive constant $C_K = C_K(\Omega)$ such that*

$$\|\varepsilon(\mathbf{v})\|^2_{L^2(\Omega)} \geq C_K \|\nabla \mathbf{v}\|^2_{L^2(\Omega)} \qquad \forall \mathbf{v} \in V \tag{8.24}$$

where

$$\|\varepsilon(\mathbf{v})\|^2_{L^2(\Omega)} = \int_\Omega \varepsilon(\mathbf{v}) : \varepsilon(\mathbf{v})\, d\Omega = \sum_{i,j=1}^{3} \int_\Omega (\varepsilon_{ij}(\mathbf{v}))^2\, d\Omega \qquad \forall \mathbf{v} \in V. \tag{8.25}$$

Proposition 8.4.2 (Rigid body motions)**.** *Let*

$$\mathscr{R} := \left\{ \mathbf{v} : \Omega \to \mathbb{R}^3 \textit{ of the form } \mathbf{v} = \mathbf{a} + \mathbf{b} \times \mathbf{x}, \textit{ with } \mathbf{a}, \mathbf{b} \textit{ constant vectors} \right\}$$

be the space of rigid body motions*. Then, we have*

$$\|\varepsilon(\mathbf{v})\|_{L^2(\Omega)} = 0 \quad \Leftrightarrow \quad \mathbf{v} = \mathbf{0}, \qquad \forall \mathbf{v} \in V. \tag{8.26}$$

Proof. By definition of rigid body motion, we have

$$\varepsilon(\mathbf{v}) = 0 \qquad \forall \mathbf{v} \in \mathscr{R}.$$

Applying the homogeneous Dirichlet boundary conditions $(8.15)_3$ yields

$$\mathbf{a} + \mathbf{b} \times \mathbf{x} = \mathbf{0} \qquad \forall \mathbf{x} \in \Gamma_D$$

from which, necessarily, we must have $\mathbf{a} = \mathbf{b} = \mathbf{0}$. This proves that the only admissible rigid body motion for the elasticity problem is $\mathbf{v}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in \Omega$, which is (8.26). □

The combined use of Korn's inequality (8.24), of Prop. 8.4.2 and of the Cauchy-Schwarz inequality allow us to prove the following result.

Proposition 8.4.3. *For every $\mathbf{v} \in V$ we have*

$$\min\left\{C_K^{1/2}, 1\right\} \|\nabla \mathbf{v}\|_{L^2(\Omega)} \leq \|\varepsilon(\mathbf{v})\|_{L^2(\Omega)} \leq \max\left\{C_K^{1/2}, 1\right\} \|\nabla \mathbf{v}\|_{L^2(\Omega)}. \tag{8.27}$$

Proof. For every $\mathbf{v} \in V$, we have (omitting the explicit dependence on $\mathbf{v}$ for notational clarity)

$$\begin{aligned} \|\varepsilon\|^2_{L^2(\Omega)} &= \textstyle\sum_{i,j=1}^3 \|\varepsilon_{ij}\|^2_{L^2(\Omega)} = \sum_{i=1}^3 \|\varepsilon_{ii}\|^2_{L^2(\Omega)} \\ &+ 2\|\varepsilon_{12}\|^2_{L^2(\Omega)} + 2\|\varepsilon_{13}\|^2_{L^2(\Omega)} + 2\|\varepsilon_{23}\|^2_{L^2(\Omega)}. \end{aligned} \tag{8.28}$$

The sum of the first three terms on the right-hand side yields

$$\sum_{i=1}^3 \|\varepsilon_{ii}\|^2_{L^2(\Omega)} = \sum_{i=1}^3 \left|\left|\frac{\partial v_i}{\partial x_i}\right|\right|^2_{L^2(\Omega)}. \tag{8.29}$$

As for the sum of the other three terms, we have, for the first contribution

$$\begin{aligned} &2\|\varepsilon_{12}\|^2_{L^2(\Omega)} = 2\int_\Omega \left[\frac{1}{2}\left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1}\right)\right]^2 d\Omega \\ &\leq \frac{1}{2}\left[\left|\left|\frac{\partial v_1}{\partial x_2}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_2}{\partial x_1}\right|\right|^2_{L^2(\Omega)} + 2\left|\left|\frac{\partial v_1}{\partial x_2}\right|\right|_{L^2(\Omega)}\left|\left|\frac{\partial v_2}{\partial x_1}\right|\right|_{L^2(\Omega)}\right], \end{aligned}$$

and similarly for the other two contributions. Noting that for every $a, b \in \mathbb{R}$ we have $2ab \leq a^2 + b^2$, the previous inequality yields

$$2\|\varepsilon_{12}\|^2_{L^2(\Omega)} \leq \left|\left|\frac{\partial v_1}{\partial x_2}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_2}{\partial x_1}\right|\right|^2_{L^2(\Omega)}. \tag{8.30}$$

Replacing (8.29) and (8.30) into (8.28), we obtain

$$\begin{aligned} \|\varepsilon\|^2_{L^2(\Omega)} &\leq \left|\left|\frac{\partial v_1}{\partial x_1}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_2}{\partial x_2}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_3}{\partial x_3}\right|\right|^2_{L^2(\Omega)} \\ &+ \left|\left|\frac{\partial v_1}{\partial x_2}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_2}{\partial x_1}\right|\right|^2_{L^2(\Omega)} \\ &+ \left|\left|\frac{\partial v_1}{\partial x_3}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_3}{\partial x_1}\right|\right|^2_{L^2(\Omega)} \\ &+ \left|\left|\frac{\partial v_2}{\partial x_3}\right|\right|^2_{L^2(\Omega)} + \left|\left|\frac{\partial v_3}{\partial x_2}\right|\right|^2_{L^2(\Omega)} \\ &= \|\nabla v_1\|^2_{L^2(\Omega)} + \|\nabla v_2\|^2_{L^2(\Omega)} + \|\nabla v_3\|^2_{L^2(\Omega)} = \|\nabla \mathbf{v}\|^2_{L^2(\Omega)}. \end{aligned}$$

Combining this latter inequality with Korn's inequality (8.24), we obtain (8.27). □

Theorem 8.4.4. *The quantity*

$$\|\varepsilon(\mathbf{v})\|_{L^2(\Omega)} : V \to \mathbb{R}^+ \tag{8.31}$$

is an equivalent norm on $V = (H^1_{0,\Gamma_D}(\Omega))^3$.

Proof. Relation (8.26) shows that (8.31) is a norm. Then, applying Def. B.3.2 to the inequality (8.27), with $K_1 \equiv \min\left\{C_K^{1/2}, 1\right\}$ and $K_2 \equiv \max\left\{C_K^{1/2}, 1\right\}$, completes the proof. □

Having introduced an appropriate norm on V, we are ready to proceed.

- Continuity of $B(\cdot,\cdot)$: we have

$$|B(\mathbf{u},\mathbf{v})| \leq \int_\Omega |\mathscr{C}||\varepsilon(\mathbf{u})||\varepsilon(\mathbf{v})|\, d\Omega \leq \sup_{\mathbf{x}\in\Omega} |\mathscr{C}_{ijkl}| \|\mathbf{u}\|_V \|\mathbf{v}\|_V \qquad \forall \mathbf{u},\mathbf{v} \in V.$$

  From the definition (8.13), we immediately see that the supremum of the elasticity matrix is attained if $i = j = k = l = 1,2,3$, yielding

$$\sup_{\mathbf{x}\in\Omega} |\mathscr{C}_{ijkl}| = \lambda + 2\mu.$$

  Thus, $B(\cdot,\cdot)$ is continuous with

$$M = \lambda + 2\mu. \tag{8.32}$$

- Coercivity of $B(\cdot,\cdot)$: we have

$$B(\mathbf{u},\mathbf{u}) = \int_\Omega \mathscr{C}\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{u})\, d\Omega = \int_\Omega (2\mu\varepsilon(\mathbf{u}) + \lambda \mathrm{div}\mathbf{u}\, \delta) : \varepsilon(\mathbf{u})\, d\Omega \qquad \forall \mathbf{u} \in V,$$

  from which

$$\begin{aligned} B(\mathbf{u},\mathbf{u}) &= 2\mu\|\mathbf{u}\|_V^2 + \lambda \int_\Omega \mathrm{div}\mathbf{u}\, \delta : \varepsilon(\mathbf{u})\, d\Omega \\ &= 2\mu\|\mathbf{u}\|_V^2 + \lambda \int_\Omega (\mathrm{div}\mathbf{u})^2\, d\Omega \geq 2\mu\|\mathbf{u}\|_V^2 \qquad \forall \mathbf{u} \in V. \end{aligned}$$

  Thus, $B(\cdot,\cdot)$ is coercive with

$$\beta = 2\mu. \tag{8.33}$$

- Continuity of $F(\cdot)$: we have

$$\begin{aligned} |F(\mathbf{v})| &\leq \int_\Omega |\mathbf{f}||\mathbf{v}|\,d\Omega + \int_{\Gamma_N} |\mathbf{g}||\mathbf{v}|\,d\Gamma \\ &\leq \|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\Gamma_N)}\|\mathbf{v}\|_{L^2(\Gamma_N)} \qquad \forall \mathbf{v} \in V. \end{aligned}$$

Using Poincarè's inequality, the Trace Theorem B.7.7 and Thm. 8.4.4, we finally get

$$|F(\mathbf{v})| \leq (C_P\|\mathbf{f}\|_{L^2(\Omega)} + C_\Gamma\|\mathbf{g}\|_{L^2(\Gamma_N)})\|\mathbf{v}\|_V \qquad \forall \mathbf{v} \in V.$$

Thus, $F(\cdot)$ is continuous with

$$\Lambda = C_P\|\mathbf{f}\|_{L^2(\Omega)} + C_\Gamma\|\mathbf{g}\|_{L^2(\Gamma_N)}. \tag{8.34}$$

Theorem 8.4.5 (Well posedness of the elasticity problem)**.** *Problem* (8.21) *admits a unique solution* $\mathbf{u} \in V$ *that satisfies the a priori estimate*

$$\|\mathbf{u}\|_V \leq \frac{C_P\|\mathbf{f}\|_{L^2(\Omega)} + C_\Gamma\|\mathbf{g}\|_{L^2(\Gamma_N)}}{2\mu}. \tag{8.35}$$

Proof. To prove (8.35), notice that

$$B(\mathbf{u},\mathbf{u}) = F(\mathbf{u})$$

and then use the coercivity of B, the continuity of F and relations (8.33) and (8.34). □

8.5 Two-dimensional models in elasticity

In many practical applications of Continuum Mechanics, it is not really necessary to solve the general three-dimensional model (8.15) in order to determine the defomrmed configuration of the elastic body. As a matter of fact, under suitable mechanical assumptions, it is possible to deduce from (8.15) two simpler formulations that are appropriate for the study of problems in two spatial dimensions. These formulations are known as *plane stress* and *plane strain* models for linear elasticity, and the stress-strain relation can be written in the following general form

$$\boldsymbol{\sigma} = \mathscr{C}\boldsymbol{\varepsilon}, \qquad \boldsymbol{\sigma} := \left\{ \begin{array}{c} \sigma_{xx} \\ \sigma_{yy} \\ \tau_{xy} \end{array} \right\}, \qquad \boldsymbol{\varepsilon} := \left\{ \begin{array}{c} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \gamma_{xy} \end{array} \right\}, \tag{8.36}$$

where $\gamma_{xy} := 2\varepsilon_{xy}$ and $\mathscr{C} \in \mathbb{R}^{3\times 3}$ is the elastic matrix.

8.5.1 Plane stress

The characteristic assumptions for this case are (see Fig. 8.3):

- the thickness of the body is small compared to the dimensions of the body in the other two directions (example: plates);
- the acting loading forces are applied in the symmetry plane of the thickness and do not have other transversal components with respect to such a plane.

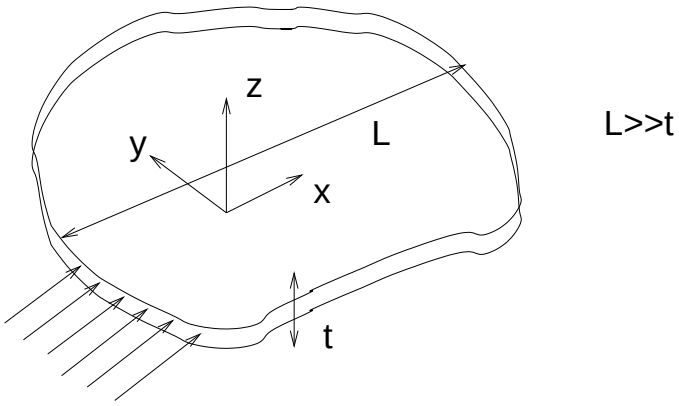


Figure 8.3: Plane stress.

Assuming that the xy plane coincides with the symmetry plane of the body thickness and that the z-axis is perpendicular to the xy plane, the *plane stress* problem is mathematically identified by the following conditions

$$\sigma_{zz} = 0, \qquad \tau_{zx} = \tau_{zy} = 0. \tag{8.37}$$

The corresponding elastic matrix is

$$\mathscr{C} = \frac{E}{1-\nu^2} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & (1-\nu)/2 \end{bmatrix}. \tag{8.38}$$

Remark 8.5.1. The strain component ε_{zz} is, in general, nonnull, and can be post-computed as

$$\varepsilon_{zz} = -\frac{\nu}{1-\nu}(\varepsilon_{xx} + \varepsilon_{yy}).$$

8.5.2 Plane strain

The characteristic assumptions for this case are (see Fig. 8.4):

- the body has a dimension much larger than the other two (example: long thin beam, cylinders, drive axles);
- the displacement field does not have components along the axis of the beam (taken equal to z) but only in directions perpendicular to z;
- the acting loading forces do not depend on the z coordinate and have null component with respect to the z axis.

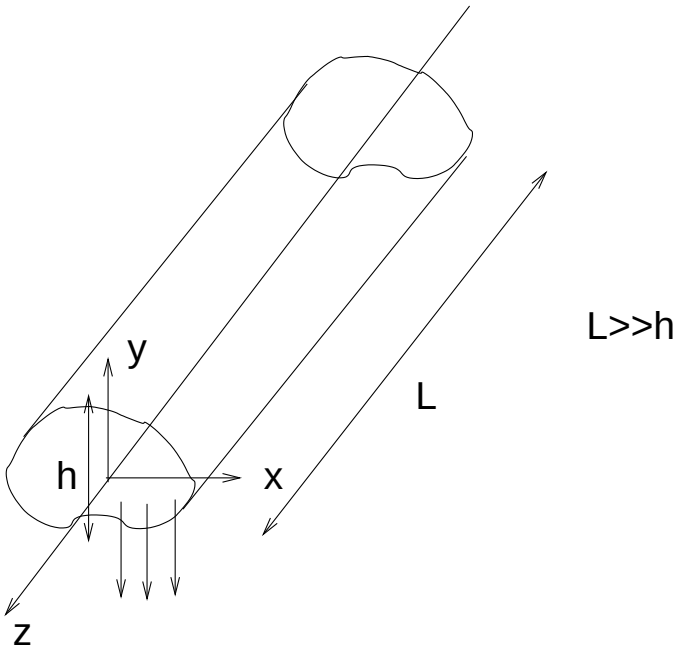


Figure 8.4: Plane strain.

Assuming that the z axis coincides with the symmetry axis of the body, the *plane strain* problem is mathematically identified by the following conditions

$$\varepsilon_{zz} = 0, \qquad \gamma_{zx} = \gamma_{zy} = 0. \tag{8.39}$$

The corresponding elastic matrix is

$$\mathscr{C} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & (1-2\nu)/2 \end{bmatrix}. \tag{8.40}$$

The elastic matrix (8.40) is the same as in the general 3D case (Eq. (8.10)), except for the elimination of the rows and columns in (8.13) corresponding to the null components of the strain tensor.

Remark 8.5.2. The stress component σ_{zz} is, in general, nonnull, and can be post-computed as

$$\sigma_{zz} = \nu(\sigma_{xx} + \sigma_{yy}).$$

8.6 The GFE approximation in the 2D case

In this section, we briefly describe the use of the GFEM for the numerical approximation of the elasticity equations (8.15) in the plane stress/plane strain regimes discussed in Sects. 8.5.1 and 8.5.2. To this purpose, we refer to Chapt. 7, for the main general concepts and implementation issues. In what follows, we assume that $\mathscr{T}_h$ is a given regular triangulation of the computational domain Ω into triangles K of diameter h_K and area K and for sake of simplicity of the presentation of the method, we restrict ourselves to the special, and widely used, case $r = 1$, corresponding to the finite element space

$$V_h = \left\{ \mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_1(K))^2 \,\forall K \in \mathscr{T}_h,\ \mathbf{v}_h = \mathbf{0} \text{ on } \Gamma_D \right\}. \tag{8.41}$$

The above choice of $V_h \subset V$ amounts to representing each component of the discrete displacement field by a linearly varying function on each mesh element, continuous across interelement edges and vanishing on the Dirichlet portion of the boundary where the elastic body is fixed to ground. For a given $\mathscr{T}_h$, we denote by N_h the dimension of V_h, so that any vector-valued function $\mathbf{v}_h = \left[v_{h,x}, v_{h,y}\right]^T \in V_h$ can be represented as

$$v_{h,x}(x,y) = \sum_{j=1}^{N_h} v_j^x \varphi_j(x,y) \qquad v_{h,y}(x,y) = \sum_{j=1}^{N_h} v_j^y \varphi_j(x,y)$$

where the real numbers $\left\{v_j^x\right\}_{j=1}^{N_h}$, $\left\{v_j^y\right\}_{j=1}^{N_h}$ are the nodal dofs of the x- and y-components of the displacement $\mathbf{v}_h$, while the N_h functions $\left\{\varphi_j(x,y)\right\}_{j=1}^{N_h}$ are

the global pyramid-like basis functions associated with the triangulation $\mathscr{T}_h$ (see Fig. 7.3).

8.6.1 The Galerkin FE problem

The Galerkin finite element problem associated with the weak formulation (8.21) of the elasticity problem in two spatial dimensions reads:
find $\mathbf{u}_h \in V_h$ such that

$$B(\mathbf{u}_h, \mathbf{v}_h) = F(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h. \tag{8.42}$$

In compact matrix form, the above problem translates into the solution of the following linear algebraic system

$$\mathbf{K}\mathbf{u} = \mathbf{f} \tag{8.43}$$

where the entries of the global stiffness matrix $\mathbf{K}$ and of the global load vector $\mathbf{f}$ are given by:

$$\begin{array}{rcll} K_{ij} & = & B(\varphi_j, \varphi_i) & i,j = 1,\ldots,N_h \\ F_i & = & F(\varphi_i) & i = 1,\ldots,N_h. \end{array}$$

The application of the Lax-Milgram lemma to (8.42) allows us to conclude the following result.

Theorem 8.6.1 (Well posedness of the discrete elasticity problem)**.** *Problem* (8.42) *admits a unique solution* $\mathbf{u}_h \in V_h$ *(equivalently, system* (8.43) *is uniquely solvable). Moreover,* $\mathbf{u}_h$ *satisfies the a priori estimate* (8.35).

Remark 8.6.2. Coercivity of $B(\cdot,\cdot)$ implies that $\mathbf{K}$ is a s.p.d. matrix.

8.6.2 Local approximations and matrices

Let K be a given triangle in $\mathscr{T}_h$. We denote from now on by $u_h = u_h(x,y)$ and $v_h = v_h(x,y)$ the two components of the local discrete displacement field $\mathbf{u}_h^K \equiv \mathbf{u}_h|_K$ in the x and y directions, respectively, so that for each $K \in \mathscr{T}_h$ we have

$$u_h(x,y) = \sum_{i=1}^{3} u_i \varphi_i^K(x,y), \qquad v_h(x,y) = \sum_{i=1}^{3} v_i \varphi_i^K(x,y),$$

where, for $i = 1,2,3$:

- φ_i^K is the restriction over K of the corresponding global basis function defined over $\mathscr{T}_h$ (see Fig. 7.3);

- u_i, v_i are the dofs of the FE approximation over K, corresponding to the nodal values of u_h and v_h, respectively (see Fig. 8.5);

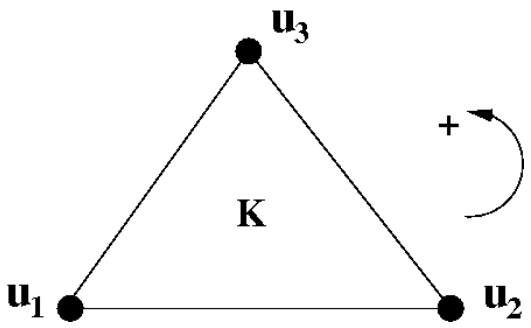


Figure 8.5: Local dofs for the CST in Linear Elasticity.

Denoting by $\mathbf{d}^K = [u_1, v_1, u_2, v_2, u_3, v_3]^T$ the vector of the local nodal displacements (numbered according to a counterclockwise orientation along the triangle boundary ∂K), we have in compact matrix form

$$\mathbf{U}^K(x,y) = \mathbf{N}^K(x,y)\mathbf{d}^K,$$

where $\mathbf{U}^K(x,y) := [u_h(x,y), v_h(x,y)]^T$ and

$$\mathbf{N}^K(x,y) := \begin{bmatrix} \varphi_1^K(x,y) & 0 & \varphi_2^K(x,y) & 0 & \varphi_3^K(x,y) & 0 \\ 0 & \varphi_1^K(x,y) & 0 & \varphi_2^K(x,y) & 0 & \varphi_3^K(x,y) \end{bmatrix}$$

is the so-called *local matrix of the shape functions*. Using the kinematic relation (8.3) we obtain the following expression for the discrete strain tensor over the element K

$$\boldsymbol{\varepsilon}^K(x,y) = \mathbf{B}^K(x,y)\mathbf{d}^K$$

where

$$\boldsymbol{\varepsilon}^K(x,y) := \begin{bmatrix} \varepsilon_{11}(x,y) \\ \varepsilon_{22}(x,y) \\ \gamma_{12}(x,y) \end{bmatrix}$$

and the *local strain matrix* is

$$\mathbf{B}^K(x,y) := \begin{bmatrix} \dfrac{\partial \varphi_1^K(x,y)}{\partial x} & 0 & \dfrac{\partial \varphi_2^K(x,y)}{\partial x} & 0 & \dfrac{\partial \varphi_3^K(x,y)}{\partial x} & 0 \\ 0 & \dfrac{\partial \varphi_1^K(x,y)}{\partial y} & 0 & \dfrac{\partial \varphi_2^K(x,y)}{\partial y} & 0 & \dfrac{\partial \varphi_3^K(x,y)}{\partial y} \\ \dfrac{\partial \varphi_1^K(x,y)}{\partial y} & \dfrac{\partial \varphi_1^K(x,y)}{\partial x} & \dfrac{\partial \varphi_2^K(x,y)}{\partial y} & \dfrac{\partial \varphi_2^K(x,y)}{\partial x} & \dfrac{\partial \varphi_3^K(x,y)}{\partial y} & \dfrac{\partial \varphi_3^K(x,y)}{\partial x} \end{bmatrix}.$$

Remark 8.6.3 (CST). Clearly, since the shape functions φ_i^K are linear in x and y, it turns out that the local strain matrix $\mathbf{B}^K(x,y)$ is constant, and consequently also $\boldsymbol{\varepsilon}^K(x,y)$ is a constant tensor. For this reason, the finite element corresponding to the choice $r = 1$ is well-known as *CST* or *Constant Strain Triangle*.

Remark 8.6.4 (Higher-order FE spaces). Higher-order approximations of displacement and strain can be obtained by taking $r > 1$ (see Fig. 7.2). All the formulas written above continue to apply provided to replace 3 with `Ndofs`, where

$$\texttt{Ndofs} = \frac{1}{2}(r+1)(r+2)$$

is the number of dofs of the finite element space $\mathbb{P}_r(K)$.

As already described in Chapt. 7, the assembly phase of the global coefficient matrix is based on the computation of:

- a local stiffness matrix;
- a local load vector.

To construct the local stiffness matrix, we simply need to define the *local bilinear form* associated with the global bilinear form $B(\mathbf{u}_h, \mathbf{v}_h)$. We have

$$\begin{aligned} B^K(\mathbf{u}_h, \mathbf{v}_h) &:= \int_K \mathscr{C}\boldsymbol{\varepsilon}(\mathbf{u}_h) : \boldsymbol{\varepsilon}(\mathbf{v}_h)\, dK = \int_K (\mathscr{C}\mathbf{B}^K\mathbf{d}^K)^T(\mathbf{B}^K\mathbf{v}^K)\, dK \\ &= (\mathbf{d}^K)^T \left[\int_K (\mathbf{B}^K)^T \mathscr{C}\mathbf{B}^K\, dK\right] \mathbf{v}^K \qquad \forall K \in \mathscr{T}_h, \end{aligned}$$

where $\mathscr{C}$ is the matrix in (8.38) in plane stress conditions or in (8.40) in plane strain conditions. From the above relation, we identify

$$\mathbf{K}^K := \int_K (\mathbf{B}^K)^T \mathscr{C}\mathbf{B}^K\, dK \in \mathbb{R}^{6\times 6}$$

as the *local stiffness matrix* associated with element K, so that we can write in a synthetic notation

$$B^K(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{d}^K)^T \mathbf{K}^K \mathbf{v}^K.$$

The vector $\mathbf{v}^K \in \mathbb{R}^6$ contains cyclically the dofs associated with the test function $\mathbf{v}_h$, and is taken equal to $\mathbf{e}_i = \underbrace{[0,0,\ldots,1,0,\ldots,0]^T}_{\text{in the } i\text{-th position}}$, $i = 1,\ldots,6$.

To construct the local load vector, we define the *local linear form* associated with the global form $F(\mathbf{v}_h)$. We have

$$\begin{aligned} F^K(\mathbf{v}_h) &:= \int_K \mathbf{f}\cdot\mathbf{v}_h\, dK + \int_{\partial K\cap\Gamma_N} \mathbf{g}\cdot\mathbf{v}_h\, d\sigma \\ &= \int_K \mathbf{F}^T\mathbf{N}^K\mathbf{v}^K\, dK + \int_{\partial K\cap\Gamma_N} \mathbf{G}^T\mathbf{N}^K\mathbf{v}^K\, d\sigma \\ &= \left[\int_K (\mathbf{N}^K)^T\mathbf{F}\, dK + \int_{\partial K\cap\Gamma_N} (\mathbf{N}^K)^T\mathbf{G}\, d\sigma\right]\mathbf{v}^K \qquad \forall K\in\mathscr{T}_h. \end{aligned}$$

From the above relation, we identify

$$\mathbf{f}^K := \int_K (\mathbf{N}^K)^T\mathbf{F}\, dK + \int_{\partial K\cap\Gamma_N} (\mathbf{N}^K)^T\mathbf{G}\, d\sigma \in \mathbb{R}^6$$

as the local load vector associated with element K, so that we can write in a synthetic manner

$$F^K(\mathbf{v}_h) = (\mathbf{f}^K)^T\mathbf{v}^K.$$

The vectors $\mathbf{F} = \mathbf{F}(x,y) = [f_x(x,y), f_y(x,y)]^T$ and $\mathbf{G} = \mathbf{G}(x,y) = [g_x(x,y), g_y(x,y)]^T$ contain the functions describing the external volume and surface applied loads. As discussed in Sect. 7.7, to evaluate $\mathbf{f}^K$ a suitable quadrature formula is needed. In the case of linear finite elements, a simple and appropriately accurate formula is the two-dimensional extension of the trapezoidal rule (see Sect. 3.2) given by

$$\int_K \eta(x,y)\, dK \simeq \sum_{i=1}^{3} \frac{|K|}{3}\eta(x_i,y_i)$$

for every continuous function $\eta : K \to \mathbb{R}$, where $(x_i,y_i)^T$, $i = 1,2,3$, are the coordinates of the vertices of K, numbered in counterclockwise sense (see Fig. 8.5) while $|K|$ denotes the area of triangle K. Notice that the above formula is exact for every $\eta \in \mathbb{P}_1(K)$ and satisfies the error estimate

$$\left|\int_K \eta(x,y)\, dK - \sum_{i=1}^{3}\frac{|K|}{3}\eta(x_i,y_i)\right| \le C h_K^2 \qquad \eta\in C^2(K)$$

$C > 0$ being a constant independent of h and depending on η.

8.6.3 Convergence analysis

Let us assume to deal with the two-dimensional elasticity problem (plane stress or plane strain conditions) and that $\Omega \subset \mathbb{R}^2$ is a polygonal domain covered by a family of regular triangulations $\{\mathscr{T}_h\}_{h>0}$ (see (7.6)). Assume also that the finite dimensional subspace of V is the space of (vector-valued) finite elements of degree $\leq r$, $r \geq 1$ being a given integer

$$V_h = \left\{ \mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_r(K))^2 \forall K \in \mathscr{T}_h,\ \mathbf{v}_h = \mathbf{0} \text{ on } \Gamma_D \right\}. \tag{8.44}$$

Theorem 8.6.5 (Error estimate for the GFEM applied to Linear Elasticity)**.** *Let us denote by* $\mathbf{u}$ *and* $\mathbf{u}_h \in V_h$ *the solution of* (8.21) *and of the GFE approximation* (8.42)*, respectively. Assume also that* $\mathbf{u} \in (H^s(\Omega))^2 \cap V$*,* $s \geq 2$ *being a given quantity representing the regularity of the exact solution. Then, there exists a positive constant* $C_{\mathscr{T}_h}$ *is a depending on* κ *but not on the discretization parameter h, such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_V \leq C_{\mathscr{T}_h} \underbrace{\left(1 + \frac{\lambda}{2\nu}\right)}_{M/\beta} h^{\ell} \|\mathbf{u}\|_{(H^{\ell+1}(\Omega))^2} \tag{8.45}$$

where $\ell := \min\{r, s-1\}$ *is the usual regularity threshold.*

Proof. The estimate (8.45) is the result of the straightforward application of Ceà's Lemma with M and β given by (8.32) and (8.33). □

Remark 8.6.6 (The 3D case). Using similar arguments, one can prove a 3D analogue of Thm. 8.6.5 in the case where $\{\mathscr{T}_h\}_{h>0}$ is a family of regular triangulations made of tetrahedral elements.

Remark 8.6.7 (The dependence on ν). Replacing relations (8.11) in (8.45), we can express the ratio M/β as a function of the material elastic parameters as

$$\frac{M}{\beta} = \left(1 + \frac{\lambda}{2\nu}\right) = 1 + \frac{\dfrac{E\nu}{(1+\nu)(1-2\nu)}}{\dfrac{2E}{2(1+\nu)}} = 1 + \frac{\nu}{1-2\nu}.$$

This relation shows that M/β is a bounded quantity (of the order of unity) as long as ν is sufficiently less than 0.5 (for example, $\nu = 0.33$, as in the case of steel). In such a situation, the choice of the mesh size in order to obtain a small error is determined only by the regularity of the exact solution. This is no longer true

when ν becomes close to 0.5, i.e., as the elastic material becomes *incompressible*. In such a situation, the ratio $M/\beta \to +\infty$ and the choice of the mesh size in order to obtain a small error is furtherly penalized by the incompressibility condition. The practical result of this analysis is that the computed deformation is severely affected by the so-called *volumetric locking*, unless the mesh size is extremely small. This introduces a strong limitation in the use of the Navier-Lamè model equations (8.15) in the incompressible limit, and an alternative approach to overcome such a limitation will be the object of Chapt. 9.

8.7 Numerical examples

In this section we study two simple problems for which the exact solution is available, in order to verify the numerical performance of the GFEM. The CST is adopted in the choice of V_h.

8.7.1 Example 1: patch test (constant stress)

In this case we consider a unit square plate (plane stress conditions) with $E = 1$ and $\nu = 0.3$ subject to a pure boundary traction force $\mathbf{g}_N$ as shown in Fig. 8.6, while the body force $\mathbf{f}$ is set equal to zero.

The exact solution of the elasticity problem is the displacement field

$$u = \frac{1-\nu}{E}x, \quad v = \frac{1-\nu}{E}y \tag{8.46}$$

while the corresponding stress field is

$$\sigma_{xx} = 1, \quad \sigma_{yy} = 1, \quad \tau_{xy} = 0. \tag{8.47}$$

Since $\mathbf{u}$ is linear and the stress is constant, we expect the GFEM to be *nodally exact* (up to machine precision, of course) in this elementary case. This particular application is a *consistency test* of the numerical method, as it aims at verifying the ability of the finite element subspace to represent a solution that *actually belongs* to it. In the computational mechanics literature, such a test problem is called *patch test*, because the numerical solution should be capable to capture exactly the solution of the continuous problem at the mesh nodes, irrespective of the choice of the triangulation. The result of the computation is illustrated in Fig. 8.7 which shows the deformed configuration of the elastic body.

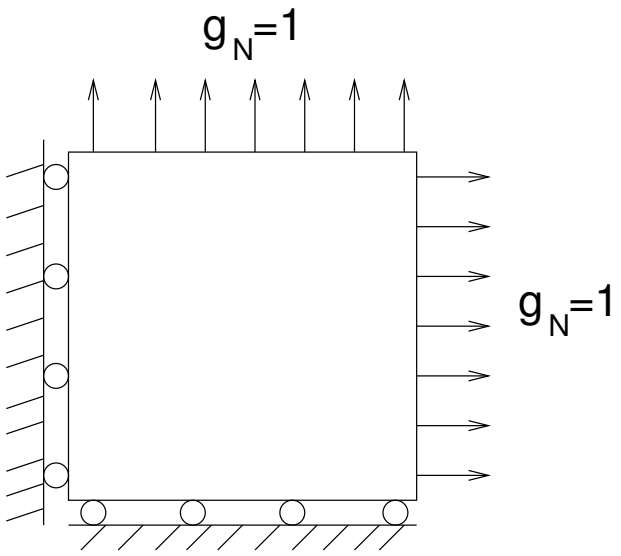


Figure 8.6: Constant stress patch test: geometry, boundary conditions and applied loads.

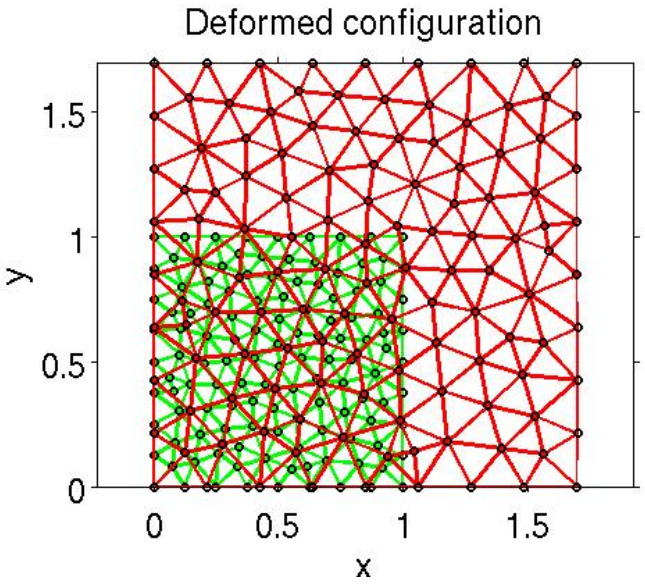


Figure 8.7: Constant stress patch test: deformed configuration.

Matlab coding. The following `Matlab` commands can be used to compute the me-

chanical response of the elastic body and the H^1 and L^2 norms of the discretization error $\mathbf{u} - \mathbf{u}_h$.

```
>> EF2D_elinc

 ********* test case ----> patchtest *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
solving the linear system
plot of solution

||u-u_h||_{H^1} -->   3.63706e-15
||u-u_h||_{L^2} -->   7.92168e-16
```

The results clearly demonstrate that the solution is nodally exact (up to machine precision), as expected.

Matlab coding. The following `Matlab` commands can be used to evaluate the maximum displacement after application of the external loads.

```
>> max(abs(U_ux))

ans =

    0.7000

>> max(abs(U_uy))

ans =

    0.7000
```

We see that the maximum deformation is of 70%, in accordance with the exact solution (8.46).

8.7.2 Example 2: experimental convergence analysis

In this case we consider the unit square section of a beam (plane strain conditions) shown in Fig. 8.8, with $E = 1$, $\nu = 0.33$ and body force $\mathbf{f} = [-9y/20, -3x/10]^T$. On the right and top sides of the beam, a force $\mathbf{g}$ is applied while the displacement $\mathbf{u}$ is constrained on the other remaining sides.

The exact solution of the elasticity problem is the displacement field

$$u = \frac{y^3}{10}, \quad v = \frac{xy^2}{10}.$$

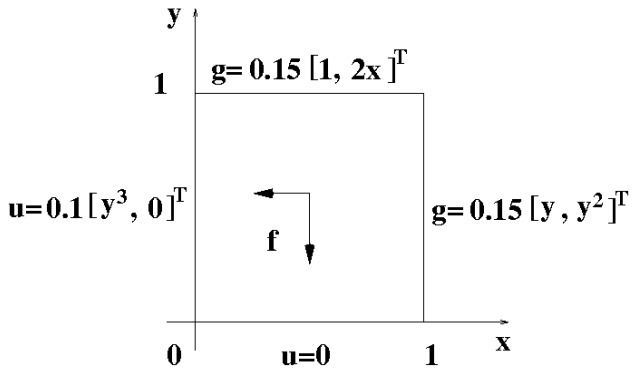


Figure 8.8: Cross-section of a beam: geometry, boundary conditions and applied loads.

Since $\mathbf{u}$ is a cubic function of (x,y), we expect the GFEM (with $r=1$) to introduce a discretization error. Fig. 8.9 shows the deformed configuration of the beam computed by the GFEM on a mesh of right-angled triangles of size $h=1/5$.

Matlab coding. The following `Matlab` script allows to generate the computational mesh of Fig. 8.9. The command `pdemesh(p,e,t)` allows to visualize the grid.

```
N=5;
[p,e,t]=poimesh('squareg',N,N);
p(1,:)=0.5*(p(1,:)+1);
p(2,:)=0.5*(p(2,:)+1);
pdemesh(p,e,t)
```

An experimental convergence analysis as a function of the mesh size h allows to verify the theoretical conclusions of Thm. 8.6.5. To this purpose, we use the above `Matlab` coding to generate 5 grids of right-angled triangles of successively refined size $h=1/N$, where $N=[5,10,20,40,80]^T$.

Tab. 8.1 reports the computed errors in the H^1 and L^2 norms as a function of h. Results clearly indicate a linear reduction of $\|\mathbf{u}-\mathbf{u}_h\|_V$ (for each row, the next error is halved with respect to that of the considered row), while $\|\mathbf{u}-\mathbf{u}_h\|_{(L^2(\Omega))^2}$ decreases quadratically vs. h (for each row, the next error is divided by a factor of $2^2=4$ with respect to that of the considered row). These results agree with (8.45) (in the case $r=1$ and $s=+\infty$) and with Thm. 5.2.6. Fig. 8.10 shows the log-log plot of the error $\mathbf{u}-\mathbf{u}_h$, measured in the H^1 and L^2 norms, respectively.

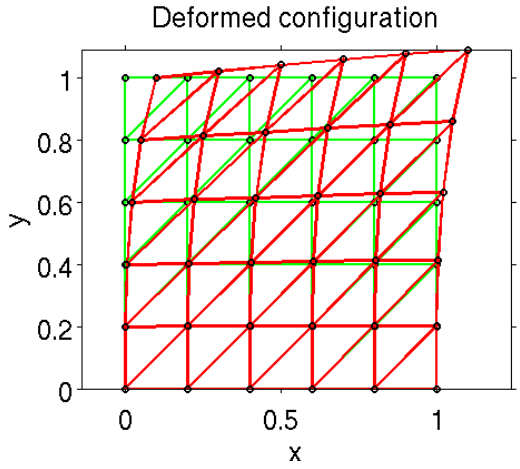


Figure 8.9: Cross-section of a beam: deformed configuration.

h	$\|\mathbf{u}-\mathbf{u}_h\|_V$	$\|\mathbf{u}-\mathbf{u}_h\|_{(L^2(\Omega))^2}$
0.2	2.71368e-02	2.16674e-03
0.1	1.33554e-02	5.63020e-04
0.05	6.61045e-03	1.42246e-04
0.025	3.28992e-03	3.56057e-05
0.0125	1.64201e-03	8.89275e-06

Table 8.1: Discretization error in the H^1 and L^2 norms as a function of h.

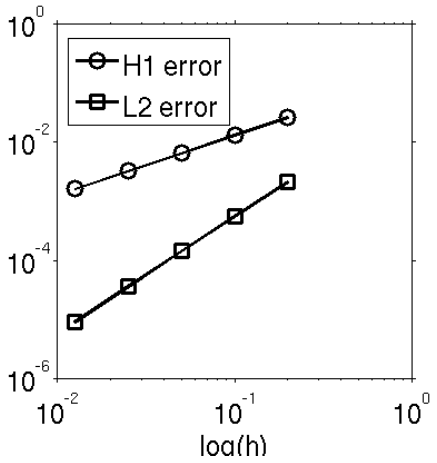


Figure 8.10: Cross-section of a beam: convergence analysis.

Chapter 9

Incompressible Linear Elasticity: Theory and Finite Element Approximation

Abstract

In this chapter, we apply the general machinery of weak formulation and Galerkin Finite Element approximation to the case of the Navier-Lamè equations of linear elasticity in the incompressible regime. To deal with the incompressibility condition (and avoid the volumetric locking phenomenon), we introduce the so-called Hermann formulation of the elasticity problem, through the introduction of an additional unknown, the pressure parameter, and we discuss the compatibility condition that needs be satisfied in order the resulting discrete scheme to be stable. Several choices of stable and unstable pairs of finite element spaces (for displacement and pressure, respectively) are considered and analyzed in the 2D case. Computational tests to validate the numerical performance of the method are run in `Matlab` using the code `EF2D_elinc` developed by Marco Restelli and available at the link:
`http://www1.mate.polimi.it/CN/MNIC/Laboratori/EF2D_el_new.zip`.

9.1 The incompressible regime

Rubber, polymers, aluminum alloys, water: these materials have in common one special feature. Whenever subject to an external load, they deform without chang-

ing their volume, that is to say, they are *incompressible*. This corresponds to the physical evidence that producing a change of volume in a (nearly) incompressible material requires an extremely high amount of energy (water is a limit example!).

Definition 9.1.1 (Incompressibility (mathematical definition))**.** *A material is said to be incompressible if its displacement field is solenoidal, i.e.*

$$\operatorname{div}\mathbf{u} = 0. \tag{9.1}$$

In the case of a fluid, the above condition means that the fluid velocity $\mathbf{u}$ *is solenoidal.*

Recalling (C.10), we see that an alternative definition of incompressibility is

$$\operatorname{Tr}\varepsilon(\mathbf{u}) = \frac{\Delta V}{V} = 0, \tag{9.2}$$

where $\Delta V/V$ represents the *volumetric strain* occurring in the elastic body due to the applied stress.

From a mechanical point of view, it is useful to express condition (9.2) in terms of the representative parameters of the material. Using definitions (8.11), we first introduce the *shear modulus*

$$G := \frac{E}{2(1+\nu)} \equiv \mu$$

and the *bulk modulus*

$$K := \lambda + \frac{2}{3}\mu = \frac{E\nu}{(1+\nu)(1-2\nu)} + \frac{2}{3}\mu.$$

Then, we introduce the following ratio

$$\rho := \frac{K}{G} = \frac{2}{3} + \frac{\lambda}{\mu} \tag{9.3}$$

which expresses the resistance of the material to volume changes. Replacing in (9.3) the definitions of G and K we get

$$\rho = \frac{2}{3} + \frac{\lambda}{\mu} = \frac{2}{3} + \frac{\dfrac{E\nu}{(1+\nu)(1-2\nu)}}{\dfrac{E}{2(1+\nu)}} = \frac{2}{3} + \frac{2\nu}{(1-2\nu)} = \frac{2(1+\nu)}{3(1-2\nu)}$$

which shows that

$$\lim_{\nu \to 0.5^-} \rho = \lim_{\nu \to 0.5^-} \frac{1}{1-2\nu} = +\infty. \tag{9.4}$$

Therefore, consistently with physical expectation, as long as the Poisson ratio gets closer and closer to the limiting value 0.5, resistance to change of volume increases without control. Noting that

$$\lambda = \frac{E}{1+\nu}\frac{\nu}{1-2\nu} = 2\mu\frac{\nu}{1-2\nu},$$

it immediately follows that

$$\lim_{\nu \to 0.5^-} \lambda = +\infty. \tag{9.5}$$

In conclusion, we have the following alternative definition of incompressibility.

Definition 9.1.2 (Incompressibility (mechanical definition))**.** *A material is said to be incompressible if condition* (9.4) *(equivalently,* (9.5)*) is satisfied.*

9.2 Volumetric locking: examples

Consider the study of the deformation of the 2D cross-section of a long thin beam (plane strain conditions) shown in Fig. 9.2.

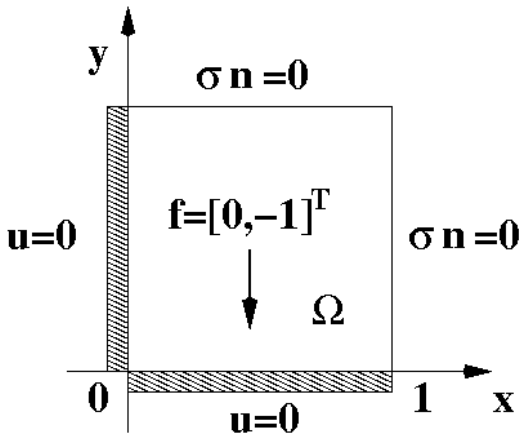


Figure 9.1: 2D cross-section of a beam: geometrical notation, loads and constraints.

The solution of the elastic problem (8.15) is a function of the elastic Lamè parameters λ and μ, so that we can write

$$\mathbf{u} = \mathbf{u}(\lambda, \mu).$$

To the purpose of analysis, let us assume that $\mu = 1$ and perform a study as a function of the sole remaining parameter $\lambda \in [0, +\infty]$. A computation of the solution of the elastic problem yields the results reported in Tab. 9.1, which indicate that as the material becomes incompressible, the deformation tends to a limit (non null) value of almost 18%.

λ	maximum deformation
10	0.21662
10^4	0.17676
10^7	0.17716

Table 9.1: Maximum deformation as a function of λ.

In correspondance of these data, we compute the approximate solution of problem (8.15) using the GFEM with $r = 1$ as explained in detail in Chapt. 8. The obtained results are reported in Tab. 9.2, which indicate that the approximate deformation accurately represents the exact one *only* when λ is very small. The percentage error is in such a case

$$\varepsilon_{rel} = \frac{|0.21662 - 0.212|}{0.21662} \times 100\,\% \simeq 2\,\%.$$

Things are very different, however, when the material approaches the incompressible regime ($\lambda = 10^7$). As a matter of fact, despite the mesh size is halved with respect to the case $\lambda = 10$, the maximum deformation is practically null, and $\varepsilon_{rel} \simeq 100\,\%$. The fact that the material does not exhibit an appreciable deformation is a manifestation of the so-called *volumetric locking*.

The reasons for such an unphysical performance of the numerical formulation were anticipated in Rem. 8.6.7. As a matter of fact, in the case where $\lambda = 10^7$, we have $(1 + \lambda/(2\mu)) = \mathscr{O}(10^7)$ in the a-priori error estimate (8.45). This implies that to have a small error, we need to take h much smaller than $1/64$, at least of the order of $h_{min} = 10^{-7}$, if we make the (non precise, but reasonable) assumption that $\|\mathbf{u}\|_{(H^2(\Omega))^2} = \mathscr{O}(1)$. Should h be chosen signficantly larger than h_{min} (as in the present example), it is no surprise that the corresponding error is very large and volumetric locking occurs.

λ	maximum deformation	h
10	0.212	1/32
10^7	0.0003	1/64

Table 9.2: Maximum approximate deformation as a function of λ and for given values of h.

Example 9.2.1 (Kinematic interpretation of locking). From the results of the previous example, one is tempted to draw the conclusion that the only discrete solution in the space V_h (with $r = 1$) that is also compatible with the incompressibility constaint (9.1), is

$$\mathbf{u}_h(x,y) = \mathbf{0} \qquad \forall (x,y) \in \Omega. \tag{9.6}$$

To check the correctness of this claim, let us consider the geometrical representation of Fig. 9.2.1 which shows a (very coarse) triangulation of the domain Ω, and assume $\lambda = +\infty$ (exact incompressibility). Since the body has to deform at

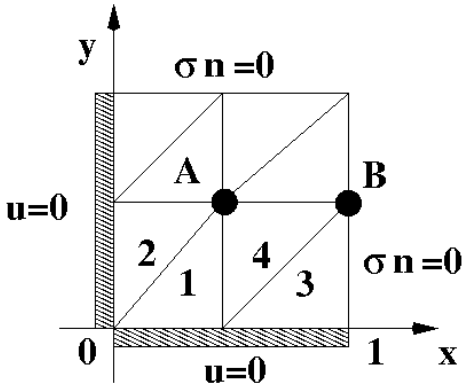


Figure 9.2: Kinematic interpretation of volumetric locking.

constant volume, the area of all mesh triangles must remain the same. Referring in particular to triangle 1, the constraint of constant area prevents node A from moving along the y direction, but allows only for an horizontal motion. At the same time, node A belongs also to triangle 2, so that, in order to maintain the area constant, no horizontal motion is permitted to node A, but only vertical. This, however, is prohibited by the previous analysis, so that the conclusion is that the only admissible motion for node A is $\mathbf{u}_A = \mathbf{0}$. This argument applies also to node B

and to all the remaining other nodes, from which it follows that the only piecewise linear continuous deformed configuration compatible with (9.1) is that of (9.6).

9.3 A two-field model for compressible and incompressible elasticity

In this section, we introduce a unified mathematical formulation of the elasticity problem in both compressible and incompressible regimes. To this purpose, we start with the following definition.

Definition 9.3.1 (The hydrostatic pressure)**.** *The hydrostatic pressure $\wp$ is defined as*

$$\wp := -\frac{1}{3}\mathrm{Tr}\sigma = -\frac{1}{3}(\sigma_{11}+\sigma_{22}+\sigma_{33}). \tag{9.7}$$

Using the pressure, the generalized Hooke's law can be written in the following equivalent form

$$\sigma = \sigma + \wp\delta - \wp\delta = (2\mu\varepsilon + \lambda \mathrm{div}\mathbf{u}\,\delta + \wp\delta) - \wp\delta = 2\mu\varepsilon + (\lambda\mathrm{div}\mathbf{u} - \frac{1}{3}\mathrm{Tr}\sigma) - \wp\delta.$$

Using the Navier-Lamè stress-strain relation (8.14), we have

$$\mathrm{Tr}\sigma = 2\mu\mathrm{Tr}\varepsilon + 3\lambda\mathrm{div}\mathbf{u} = (2\mu+3\lambda)\mathrm{div}\mathbf{u},$$

so that

$$\lambda\mathrm{div}\mathbf{u} - \frac{1}{3}\mathrm{Tr}\sigma = \lambda\mathrm{div}\mathbf{u} - \frac{2\mu}{3}\mathrm{div}\mathbf{u} - \lambda\mathrm{div}\mathbf{u} = -\frac{2\mu}{3}\mathrm{div}\mathbf{u}.$$

This suggests the following (equivalent) representation of the stress tensor

$$\sigma = 2\mu S - \wp\delta, \tag{9.8}$$

where

$$S := \varepsilon - \frac{1}{3}\mathrm{div}\mathbf{u}\,\delta \tag{9.9}$$

is the *deviatoric* part of the stress tensor, and the hydrostatic pressure is

$$\wp = -\frac{1}{3}\mathrm{Tr}\,\sigma = -\frac{2\mu}{3}\mathrm{Tr}\,\varepsilon - \lambda\mathrm{div}\mathbf{u} = -\lambda\left(1+\frac{2}{3}\frac{\mu}{\lambda}\right)\mathrm{div}\mathbf{u}. \tag{9.10}$$

By construction, we have

$$\mathrm{Tr}\, S = S_{11} + S_{22} + S_{33} = \mathrm{Tr}\, \varepsilon - \mathrm{div}\mathbf{u} = 0.$$

The representation (9.8)- (9.9) has the important mechanical effect of decomposing the state of stress in a material into a deviatoric part and an hydrostatic part. From the mathematical point of view, the decomposition (9.8), unlike (8.14), does not fail in the incompressible case, provided that the pressure is bounded. This observation can be profitably exploited to construct a mathematical formulation of the linear elasticity problem that is *robust* with respect to the compressibility parameter λ.

With this aim, let us introduce the *new dependent variable*

$$p := -\lambda\, \mathrm{div}\, \mathbf{u}. \tag{9.11}$$

Comparing (9.11) and (9.10), we see that

$$\wp = \left(1 + \frac{2}{3}\frac{\mu}{\lambda}\right) p,$$

from which we can express p as a function of $\wp$ as

$$p = \frac{3\lambda}{2\mu + 3\lambda}\wp. \tag{9.12}$$

The new variable p is referred to as the *pressure parameter*, because (9.12) shows that

$$\lim_{\lambda \to +\infty} p = \wp, \tag{9.13}$$

that, is, in the incompressible limit the pressure parameter coincides with the hydrostatic pressure.

Remark 9.3.2. For brevity, we shall refer to the pressure parameter p as the pressure. As noted above, there should not be confusion between pressure and hydrostatic pressure, these two quantities being related through (9.12).

Replacing (9.11) into the Navier-Lamè model (8.15), we obtain the following novel formulation of the linear elasticity problem:

find the displacement $\mathbf{u} : \Omega \to \mathbb{R}^3$ and the pressure $p : \Omega \to \mathbb{R}$ such that:

$$\begin{aligned}
\operatorname{div}\sigma(\mathbf{u},p) + \mathbf{f} = \mathbf{0} \quad & \text{in } \Omega && (9.14a)\\
\sigma(\mathbf{u},p) = 2\mu\nabla_S\mathbf{u} - p\,\delta \quad & \text{in } \Omega && (9.14b)\\
\frac{p}{\lambda} + \operatorname{div}\mathbf{u} = 0 \quad & \text{in } \Omega && (9.14c)\\
\mathbf{u} = \mathbf{0} \quad & \text{on } \Gamma_D && (9.14d)\\
\sigma(\mathbf{u},p)\mathbf{n} = \mathbf{g} \quad & \text{on } \Gamma_N. && (9.14e)
\end{aligned}$$

Remark 9.3.3 (Two-field model). The Navier-Lamè system (8.15) is a displacement formulation of elasticity (cf. Rem. 8.2.1), while the novel system (9.14) is a *two-field formulation (displacement-pressure)*. System (9.14) is also known as the *Herrmann formulation* of linear elasticity, because of the introduction of the pressure variable.

Remark 9.3.4 (A unified model). Except for the special case $\lambda = 0$ (corresponding to a material that does not exhibit trasverse contraction in its mechanical response and that can be treated by using the standard Navier-Lamè model), system (9.14) is well defined for every $\lambda \in (0, +\infty]$. For this reason, its use is by no means restricted to the (nearly) incompressible regime, but can be adopted also for the study of compressible materials. In this sense, system (9.14) can be regarded as a *unified* formulation of linear elasticity.

9.4 Weak formulation of the two-field model

The weak formulation of the two-field system (9.14) is the natural extension of that considered in Sect. 8.3 for the Navier-Lamè model. Let V and Q be the function spaces for displacement and pressure, respectively. The definition of V is the same as in (8.19), so that we proceed by multiplying (9.14a) by a test function $\mathbf{v} \in V$ and integrating by parts the term

$$\int_\Omega \operatorname{div}\sigma(\mathbf{u},p)\cdot\mathbf{v}\,d\Omega \qquad \forall \mathbf{v} \in V$$

to obtain

$$\int_\Gamma \mathbf{v}\cdot\sigma(\mathbf{u},p)\mathbf{n}\,d\Gamma - \int_\Omega \sigma(\mathbf{u},p) : \varepsilon(\mathbf{v})\,d\Omega \qquad \forall \mathbf{v} \in V.$$

The boundary integral can be treated as done in Sect. 8.3. The volume integral becomes

$$\begin{aligned}\int_\Omega \sigma(\mathbf{u},p):\varepsilon(\mathbf{v})\,d\Omega &= \int_\Omega (2\mu\varepsilon(\mathbf{u})-p\delta):\varepsilon(\mathbf{v})\,d\Omega\\ &= \int_\Omega 2\mu\varepsilon(\mathbf{u}):\varepsilon(\mathbf{v})\,d\Omega - \int_\Omega p\,\mathrm{Tr}\,\varepsilon(\mathbf{v})\,d\Omega\\ &= \int_\Omega 2\mu\varepsilon(\mathbf{u}):\varepsilon(\mathbf{v})\,d\Omega - \int_\Omega p\,\mathrm{div}\,\mathbf{v}\,d\Omega \qquad \forall \mathbf{v}\in V.\end{aligned}$$

The fact that $\mathbf{v}\in V$ implies that the function $\mathrm{div}\,\mathbf{v}$ belongs to $L^2(\Omega)$. Thus, using Thm. 4.2.1, we have that a sufficient condition for the function $p\,\mathrm{div}\,\mathbf{v}$ to belong to $L^1(\Omega)$ is that p belongs to $L^2(\Omega)$. Concluding, the appropriate function space where to seek the pressure is

$$Q := L^2(\Omega), \tag{9.15}$$

and the weak formulation of the Herrmann model for elasticity reads:
find $\mathbf{u}\in V$ and $p\in Q$ such that, for all $\mathbf{v}\in V$ and for all $q\in Q$ we have:

$$\begin{cases} \displaystyle\int_\Omega 2\mu\varepsilon(\mathbf{u}):\varepsilon(\mathbf{v})\,d\Omega - \int_\Omega p\,\mathrm{div}\,\mathbf{v}\,d\Omega &= \displaystyle\int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega + \int_{\Gamma_N}\mathbf{g}\cdot\mathbf{v}\,d\Gamma \\ \displaystyle -\int_\Omega q\,\mathrm{div}\,\mathbf{u}\,d\Omega - \frac{1}{\lambda}\int_\Omega p\,q\,d\Omega &= 0. \end{cases} \tag{9.16}$$

9.5 Matrix block form of the Herrmann system

The weak form of the Herrmann formulation can be written in synthetic structure as a two-by-two block linear abstract system of integral equations:
find $\mathbf{u}\in V$ and $p\in Q$ such that, for all $\mathbf{v}\in V$ and for all $q\in Q$ we have:

$$\begin{bmatrix} \mathscr{A} & \mathscr{B}^T \\ \mathscr{B} & \mathscr{E} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathscr{F} \\ 0 \end{bmatrix}. \tag{9.17}$$

The symbols $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{E}$ are the operators corresponding to the various integrals in (9.16), while $\mathscr{F}$ represents the right-hand side of the weak form of the force balance equation. The action of these operators on the corresponding dependent variables is a *matrix-vector* multiplication, so that $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{E}$ can be interpreted as matrices, as is the case in the FE discretization of system (9.16).

Theorem 9.5.1 (Elimination of the pressure)**.** *If $\lambda < +\infty$, system* (9.17) *is equivalent to the following displacement-based formulation:*

find $\mathbf{u} \in V$ *such that*

$$\mathscr{K}\mathbf{u} = \mathscr{F} \tag{9.18}$$

where

$$\mathscr{K} = \mathscr{A} - \mathscr{B}^T \mathscr{E}^{-1} \mathscr{B}$$

is a symmetric positive definite operator (equiv., matrix) representing the abstract stiffness matrix of the Herrmann formulation.

Proof. If $\lambda < +\infty$, the operator (equiv., matrix) $-\mathscr{E}$ is positive definite (take $q = p$ to see that!), so that it is invertible over Q. Thus, we can eliminate the variable p from $(9.17)_2$ in favor of $\mathbf{u}$ to obtain

$$p = -\mathscr{E}^{-1}\mathscr{B}\mathbf{u}.$$

Then, replacing the above relation into $(9.17)_1$, we immediately get (9.18). To prove that $\mathscr{K}$ is a symmetric positive definite (s.p.d.) operator (equiv., matrix), it suffices to notice that $-\mathscr{B}^T\mathscr{E}^{-1}\mathscr{B}$ is s.p.d. and that the sum of two s.p.d. operators (equiv., matrices) is a s.p.d. operator (equiv., matrix). □

Remark 9.5.2 (The role of the pressure in the incompressible case). From the proof of Thm. 9.5.1, we immediately see that in the exactly incompressible case ($\lambda = +\infty$), the elimination of the pressure from $(9.17)_2$ is no longer possible, so that the pressure has to be determined by solving the full two-by-two coupled system (9.17). Therefore, while in the compressible regime the pressure is an *auxiliary* variable (not strictly necessary for the solution of the elasticity problem), in the incompressible limit the pressure is *necessary* for the solution of the problem. To overcome this fundamental difficulty of the exactly incompressible regime, a popular approach in Computational Mechanics consists of introducing a (small) *additional compressibility* to the mechanical behavior of the solid, in order to end up with a *modified* problem of the form (9.17) to which Thm. 9.5.1 can be applied. This approach is known as the *B-bar* method.

9.6 Well-posedness analysis of the two-field model

Let us now address the issue of assessing whether the two-field formulation (9.16) is well-posed. To this purpose, we set $\lambda = +\infty$ (incompressible regime) and slightly simplify the problem by setting $\Gamma_N = \emptyset$ so that the elasticity problem

becomes:
find $\mathbf{u} : \Omega \to \mathbb{R}^3$ and $p : \Omega \to \mathbb{R}$ such that:

$$-2\mu \operatorname{div} \varepsilon(\mathbf{u}) + \nabla p = \mathbf{f} \quad \text{in } \Omega \tag{9.19a}$$
$$\operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega \tag{9.19b}$$
$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma. \tag{9.19c}$$

Remark 9.6.1. From (9.19a) and the (mechanical) fact that no information is given on the normal stress (and thus, on p) on the boundary, it immediately follows that if p^* is a solution, also $p^* + C$ is a solution, C being an arbitrary constant.

Because of (9.19c), the function space where to seek the displacement is

$$V = (H_0^1(\Omega))^3.$$

Because of Rem. 9.6.1, an appropriate choice for the function space where to seek the pressure is

$$Q = L_0^2(\Omega) := \left\{ q \in L^2(\Omega) \,|\, \int_\Omega q(\mathbf{x})\, d\Omega = 0 \right\}.$$

Adding the further requirement on the pressure to have zero mean over Ω allows to fix in a unique manner the arbitrary constant C introduced in Rem. 9.6.1. This approach closely resembles what already done in the case of the non-homogeneous Neumann problem (see Rem. 4.3.7).

The weak problem associated with (9.19) is:
find $\mathbf{u} \in V$ and $p \in Q$ such that, for all $\mathbf{v} \in V$ and for all $q \in Q$ we have:

$$\begin{cases} \displaystyle\int_\Omega 2\mu\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\Omega - \int_\Omega p \operatorname{div} \mathbf{v}\, d\Omega &= \displaystyle\int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega \\ \displaystyle -\int_\Omega q \operatorname{div} \mathbf{u}\, d\Omega &= 0. \end{cases} \tag{9.20}$$

The following result is crucial in the analysis of well-posedness of problem (9.20).

Theorem 9.6.2 (Compatibility condition relating V and Q)**.** *There exists a positive constant γ such that, for every $q \in Q$, there exists $\mathbf{v} \in V$ such that*

$$\int_\Omega q \operatorname{div} \mathbf{v}\, d\Omega \geq \gamma \|q\|_Q \|\mathbf{v}_q\|_V. \tag{9.21}$$

Remark 9.6.3 (Inf-sup condition). The inequality (9.21) expresses a *compatibility condition*, through the action of the bilinear form $b(\cdot,\cdot)$, between the two function spaces V and Q. It is also called *inf-sup condition* because it can be reformulated as

$$\underbrace{\inf_{q\in Q}}_{\text{for all } q\in Q} \quad \underbrace{\sup_{\mathbf{v}\in V}}_{\text{there exists } \mathbf{v}\in V} \quad \frac{\int_\Omega q\,\mathrm{div}\,\mathbf{v}\,d\Omega}{\|q\|_Q\|\mathbf{v}\|_V} \geq \gamma. \tag{9.22}$$

The inf-sup condition is also known as *LBB condition*, from the names of the various mathematicians that, over the last 40 years, have given fundamental contributions to its development and analysis: Olga Ladinszeskaya, Franco Brezzi and Ivo Babuska.

Remark 9.6.4 (Weak coercivity). The inf-sup condition can be regarded as a property of *weak coercivity* of the bilinear form

$$b(\mathbf{v},q) := \int_\Omega q\,\mathrm{div}\,\mathbf{v}\,d\Omega \;:(V\times Q)\to\mathbb{R}. \tag{9.23}$$

The constant γ plays here the role of β in the Lax-Milgram Lemma 4.4.1.

Theorem 9.6.5 (Existence, uniqueness and a priori estimate)**.** *Let* $\mathbf{f}$ *be a given function in* $(L^2(\Omega))^3$*, and assume that the compatibility condition* (9.21) *is satisfied. Then, the weak problem* (9.20) *admits a unique solution pair* $(\mathbf{u},p)\in(V\times Q)$*, such that*

$$\|\mathbf{u}\|_V \leq \frac{C_P}{2\mu}\|\mathbf{f}\|_{(L^2(\Omega))^3} \tag{9.24a}$$

$$\|p\|_Q \leq \frac{2C_P}{\gamma}\|\mathbf{f}\|_{(L^2(\Omega))^3}. \tag{9.24b}$$

Proof. To prove the uniqueness of the solution pair $(\mathbf{u},p)$, we proceed by assuming that there exist two distinct pairs that satisfy (9.20), namely, $(\mathbf{u}_1,p_1)$ and $(\mathbf{u}_2,p_2)$, and then try to show that, necessarily, one must have $\mathbf{u}_1=\mathbf{u}_2$ and $p_1=p_2$. Subtracting the two weak formulations in correspondance of the two distinct solution pairs we obtain

$$\begin{cases} \displaystyle\int_\Omega 2\mu\,\varepsilon(\mathbf{u}_1-\mathbf{u}_2):\varepsilon(\mathbf{v})\,d\Omega - \int_\Omega (p_1-p_2)\,\mathrm{div}\,\mathbf{v}\,d\Omega &= 0 \qquad \forall \mathbf{v}\in V \\ \displaystyle -\int_\Omega q\,\mathrm{div}\,(\mathbf{u}_1-\mathbf{u}_2)\,d\Omega &= 0 \qquad \forall q\in Q. \end{cases} \tag{9.25}$$

Now, take $\mathbf{v} = \mathbf{u}_1 - \mathbf{u}_2$ and $q = -(p_1 - p_2)$ and sum $(9.25)_1$ and $(9.25)_2$, to get

$$\int_\Omega 2\mu\varepsilon(\mathbf{u}_1 - \mathbf{u}_2) : \varepsilon(\mathbf{u}_1 - \mathbf{u}_2)\, d\Omega = 0,$$

that is

$$\|\mathbf{u}_1 - \mathbf{u}_2\|_V^2 = 0 \Rightarrow \mathbf{u}_1 = \mathbf{u}_2.$$

This proves that the displacement is unique.

To prove now that the pressure is unique, set $P := p_1 - p_2$. Then, from $(9.25)_1$, we get (recall that $\mathbf{u}_1 = \mathbf{u}_2$!)

$$\int_\Omega P \operatorname{div} \mathbf{v}\, d\Omega = 0 \qquad \forall \mathbf{v} \in V.$$

Using (9.21), we have

$$0 = \int_\Omega P \operatorname{div} \mathbf{v}^*\, d\Omega \geq \gamma \|P\|_Q \|\mathbf{v}^*\|_V$$

for, at least, one particular $\mathbf{v}^* \in V$ such that $\mathbf{v}^* \neq \mathbf{0}$. Thus, we get

$$\|P\|_Q = 0 \Rightarrow p_1 = p_2.$$

This proves uniqueness also of the pressure.

Let us now prove the a-priori estimate (9.24). To this purpose, take $\mathbf{v} = \mathbf{u}$ in $(9.20)_1$ and $q = p$ in $(9.20)_2$, and then subtract the second equation from the first, obtaining

$$\int_\Omega 2\mu\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{u})\, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{u}\, d\Omega.$$

Using the Cauchy-Schwarz inequality and the Poincarè inequality to treat the right-hand side, and the definition of equivalent norm in V to treat the left-hand side, we immediately get (9.24a).

Let us now come to the a-priori bound for the pressure. From $(9.20)_1$, we find

$$\int_\Omega p \operatorname{div} \mathbf{v}\, d\Omega = \int_\Omega 2\mu\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\Omega - \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega \qquad \forall \mathbf{v} \in V,$$

from which we obtain

$$\begin{aligned}
\int_\Omega p \operatorname{div} \mathbf{v}\, d\Omega &\leq 2\mu \|\mathbf{u}\|_V \|\mathbf{v}\|_V + \|\mathbf{f}\|_{(L^2(\Omega))^3} \|\mathbf{v}\|_Q \\
&\leq 2\mu \frac{C_P \|\mathbf{f}\|_{(L^2(\Omega))^3}}{2\mu} \|\mathbf{v}\|_V + \|\mathbf{f}\|_{(L^2(\Omega))^3} C_P \|\mathbf{v}\|_V \\
&= 2C_P \|\mathbf{f}\|_{(L^2(\Omega))^3} \|\mathbf{v}\|_V \qquad \forall \mathbf{v} \in V.
\end{aligned}$$

Thus, using (9.21) with $q = p$, we have, for, at least, one particular $\mathbf{v}^* \in V$ such that $\mathbf{v}^* \neq \mathbf{0}$

$$\gamma \|\mathbf{v}^*\|_V \|p\|_Q \leq \int_\Omega p \operatorname{div} \mathbf{v}^* \, d\Omega \leq 2C_P \|\mathbf{f}\|_{(L^2(\Omega))^3} \|\mathbf{v}^*\|_V,$$

from which (9.24b) immediately follows. □

9.7 Energy formulation of incompressible elasticity

Thm. 8.3.2 expresses the relation between the weak solution of the Navier-Lamè equations of linear elasticity (Principle of Virtual Works) with the minimizer of a suitable energy functional (Principle of Minimum of the Potential Energy). Such a relation breaks down in the incompressible limit, because of the presence in the energy functional (8.23) of the term

$$\frac{1}{2} \int_\Omega \lambda (\operatorname{div} \mathbf{v})^2 \, d\Omega. \tag{9.26}$$

As a matter of fact, in the incompressible regime, we have, *simultaneously*, $\lambda \to +\infty$ and $\operatorname{div} \mathbf{v} \to 0$, so that the integrand in (9.26) gives rise to an undetermined form. A possible way out to this deadend is suggested by the following result.

Proposition 9.7.1. *We have*

$$\frac{1}{2} \int_\Omega \lambda (\operatorname{div} \mathbf{v})^2 \, d\Omega = \sup_{q \in Q} \Phi_v(q) \tag{9.27}$$

where

$$\Phi_v(q) := -\frac{1}{2\lambda} \int_\Omega q^2 \, d\Omega - \int_\Omega q \operatorname{div} \mathbf{v} \, d\Omega \qquad \mathbf{v} \in V. \tag{9.28}$$

Proof. Enforcing the stationarity condition on Φ_v

$$\frac{\partial \Phi_v(q)}{\partial q} = \lim_{\delta \to 0} \frac{\Phi_v(q + \delta \eta) - \Phi_v(q)}{\delta} = 0 \qquad \forall \eta \in Q$$

yields

$$-\int_\Omega \left(\frac{1}{\lambda} q + \operatorname{div} \mathbf{v} \right) \eta \, d\Omega = 0 \qquad \forall \eta \in Q,$$

from which we conclude that the element

$$z := \frac{1}{\lambda} q + \operatorname{div} \mathbf{v} \in Q$$

is orthogonal (with respect to the inner product in $L^2(\Omega)$) to *all* elements of Q, i.e.

$$z = 0 \; a.e. \text{in } \Omega \qquad \Rightarrow \qquad q = -\lambda \operatorname{div} \mathbf{v} \equiv q^* \; a.e. \text{in } \Omega.$$

To check whether q^* is a mimimizer of maximizer of Φ_v, we have to compute the second partial derivative of Φ_v with respect to q, obtaining

$$\frac{\partial^2 \Phi_v(q)}{\partial q^2} = -\frac{1}{\lambda} \int_\Omega \eta^2 \, d\Omega < 0 \qquad \forall \eta \in Q.$$

This shows that $\Phi_v(q)$ is maximized at $q = q^*$, i.e.

$$\begin{aligned} \sup_{q \in Q} \Phi_v(q) = \Phi_v(q^*) &= -\frac{1}{2\lambda} \int_\Omega (-\lambda \operatorname{div} \mathbf{v})^2 \, d\Omega - \int_\Omega (-\lambda \operatorname{div} \mathbf{v}) \operatorname{div} \mathbf{v} \, d\Omega \\ &= \frac{1}{2} \int_\Omega \lambda (\operatorname{div} \mathbf{v})^2 \, d\Omega \end{aligned}$$

which is (9.27). □

Using Prop. 9.7.1, we can prove the following result.

Theorem 9.7.2 (Saddle-point formulation of (9.16))**.** *Let V and Q be defined as in* (8.19) *and* (9.15)*, respectively. Then, the saddle-point problem:*
find the pair $(\mathbf{u}, p) \in (V \times Q)$ such that

$$\Pi_\lambda(\mathbf{u}, q) \le \Pi_\lambda(\mathbf{u}, p) \le \Pi_\lambda(\mathbf{v}, p) \qquad \forall (\mathbf{v}, q) \in (V \times Q) \tag{9.29}$$

where

$$\begin{aligned} \Pi_\lambda(\mathbf{v}, q) &= \int_\Omega \mu \varepsilon(\mathbf{v}) : \varepsilon(\mathbf{v}) \, d\Omega - \frac{1}{2\lambda} \int_\Omega q^2 \, d\Omega \\ &\quad - \int_\Omega q \operatorname{div} \mathbf{v} \, d\Omega - \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\Omega - \int_{\Gamma_N} \mathbf{g} \cdot \mathbf{v} \, d\Gamma \end{aligned} \tag{9.30}$$

is completely equivalent to the weak problem (9.16)*.*

Remark 9.7.3. Thm. 9.7.2 is the counterpart of Thm. 8.3.2 in the incompressible limit. For a given $\lambda \in (0,+\infty]$, the functional $\Pi_\lambda(\mathbf{v},q) : V \times Q \to \mathbb{R}$ is the *Lagrangian* associated with the elasticity problem. Notice that, unlike in the variational principle of Thm. 8.3.2 (search of the minimizer of the total potential energy $J(\mathbf{v})$ stored in the elastic body), here we are looking for the *saddle-point* of the Lagrangian $\Pi_\lambda(\mathbf{v},q)$. This makes the nature of problem (9.29) very different from that of (8.22). The main disadvantage related to the solution of the saddle-point formulation is the introduction of an auxiliary unknown (the pressure parameter) which makes the method more computationally expensive, and, above all, the need of satisfying the inf-sup condition (9.21) *also* on the discrete level (this issue will be thouroughly addressed in Sect. 9.8). The main advantage provided by the use of the saddle-point formulation compared to the minimum energy principle (or the B-bar method) is represented by the fact that in the incompressible limit, $\Pi_\lambda(\mathbf{v},q)$ *does not* break down, unlike $J(\mathbf{v})$, rather it tends to the limiting value

$$\Pi_{+\infty}(\mathbf{v},q) = \int_\Omega \mu\,\boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v})\,d\Omega - \int_\Omega q\,\mathrm{div}\,\mathbf{v}\,d\Omega - \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega - \int_{\Gamma_N} \mathbf{g}\cdot\mathbf{v}\,d\Gamma.$$

In this sense, the saddle-point formulation is *robust* with respect to the compressibility parameter λ over the whole range $\lambda \in (0,+\infty]$.

9.8 GFE approximation of the two-field model

Throughout this section, we assume that the computational domain Ω representing the elastic body is a polygon in $\mathbb{R}^2$. We also set $\lambda = +\infty$ (exactly incompressible case) and $\Gamma = \Gamma_D$ (body constrained to ground over all its boundary $\partial\Omega \equiv \Gamma$). To address the numerical approximation of (9.20) using the GFEM, we introduce a family $\{\mathscr{T}_h\}_{h>0}$ of triangulations made of triangular elements K and for a given member of the family, we consider two finite dimensional subspaces of polynomial scalar functions

$$V_h \subset V, \qquad Q_h \subset Q.$$

Then, the GFE approximation of (9.20) is:
find $\mathbf{u}_h \in V_h$ and $p_h \in Q_h$ such that, for all $\mathbf{v}_h \in V_h$ and for all $q_h \in Q_h$ we have:

$$\left\{\begin{array}{lcl} \displaystyle\int_\Omega 2\mu\,\boldsymbol{\varepsilon}(\mathbf{u}_h) : \boldsymbol{\varepsilon}(\mathbf{v}_h)\,d\Omega - \int_\Omega p_h\,\mathrm{div}\,\mathbf{v}_h\,d\Omega &=& \displaystyle\int_\Omega \mathbf{f}\cdot\mathbf{v}_h\,d\Omega \\ \displaystyle -\int_\Omega q_h\,\mathrm{div}\,\mathbf{u}_h\,d\Omega &=& 0. \end{array}\right. \tag{9.31}$$

To characterize the linear system of algebraic equations emanating from (9.31), we introduce two sets of polynomial basis functions, ϕ_j, $j = 1,\ldots,N_h$, and ψ_k, $k = 1,\ldots,M_h$, such that

$$V_h = (\text{span}\left\{\phi_j\right\}_{j=1}^{N_h} \times \text{span}\left\{\phi_j\right\}_{j=1}^{N_h}) \qquad Q_h = \text{span}\left\{\psi_j\right\}_{j=1}^{M_h},$$

in such a way that $\dim V_h = 2N_h$, $\dim Q_h = M_h$ and the approximations $\mathbf{u}_h \in V_h$ and $p_q \in Q_h$ to $\mathbf{u} \in V$ and $p \in Q$, respectively, can be written as

$$\begin{aligned} \mathbf{u}_h(\mathbf{x}) &= \sum_{j=1}^{N_h} \mathbf{u}_j \phi_j(\mathbf{x}) = \begin{bmatrix} \sum_{j=1}^{N_h} u_j\,\phi_j(\mathbf{x}) \\ \sum_{j=1}^{N_h} v_j\,\phi_j(\mathbf{x}) \end{bmatrix} \in V_h \\ p_h(\mathbf{x}) &= \sum_{j=1}^{M_h} p_j \psi_j(\mathbf{x}) \in Q_h, \end{aligned} \tag{9.32}$$

where $\left\{\mathbf{u}_j\right\}_{j=1}^{N_h} \equiv \left\{[u_j, v_j]^T\right\}_{j=1}^{N_h}$ and $\left\{p_j\right\}_{j=1}^{M_h}$ are the sets of dofs for $\mathbf{u}_h$ and p_h, respectively. Replacing the expressions (9.32) into (9.31) and taking $\mathbf{v}_h = [\phi_i, 0]^T$ and $\mathbf{v}_h = [0, \phi_i]^T$, respectively, for $i = 1,\ldots,N_h$, in $(9.31)_1$, and $q_h = \psi_i$, $i = 1,\ldots,M_h$ in $(9.31)_2$, yields the following linear algebraic system

$$\mathscr{K}\,\mathbf{U} = \mathscr{F} \tag{9.33}$$

where

$$\mathscr{K} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix}, \quad \mathscr{F} = \begin{bmatrix} \mathbf{F} \\ \mathbf{0} \end{bmatrix}. \tag{9.34}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ have the following block-form structure

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{uu} & \mathbf{A}^{uv} \\ \mathbf{A}^{vu} & \mathbf{A}^{vv} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{B}^{qu} & \mathbf{B}^{qv} \end{bmatrix}$$

where $\mathbf{A}^{uu}$, $\mathbf{A}^{uv}$, $\mathbf{A}^{vu}$ and $\mathbf{A}^{vv}$ are square matrices each of size N_h, while $\mathbf{B}^{qu}$ and $\mathbf{B}^{qv}$ are rectangular matrices each of size $M_h \times N_h$. The entries of each sub-block matrix in $\mathbf{A}$ are given by

$$\begin{aligned} A_{ij}^{uu} &= \int_\Omega 2\mu\varepsilon([\phi_j, 0]^T) : \varepsilon([\phi_i, 0]^T)\,d\Omega \qquad i,j = 1,\ldots,N_h \\ A_{ij}^{uv} &= \int_\Omega 2\mu\varepsilon([0, \phi_j]^T) : \varepsilon([\phi_i, 0]^T)\,d\Omega \qquad i,j = 1,\ldots,N_h \\ A_{ij}^{vu} &= \int_\Omega 2\mu\varepsilon([\phi_j, 0]^T) : \varepsilon([0, \phi_i]^T)\,d\Omega \qquad i,j = 1,\ldots,N_h \\ A_{ij}^{uu} &= \int_\Omega 2\mu\varepsilon([0, \phi_j]^T) : \varepsilon([0, \phi_i]^T)\,d\Omega \qquad i,j = 1,\ldots,N_h \end{aligned}$$

while those in matrix **B** are given by

$$B_{ij}^{qu} = -\int_\Omega \psi_i \frac{\partial \phi_j}{\partial x}\, d\Omega \qquad i = 1,\ldots,M_h, \quad j = 1,\ldots,N_h$$
$$B_{ij}^{qv} = -\int_\Omega \psi_i \frac{\partial \phi_j}{\partial y}\, d\Omega \qquad i = 1,\ldots,M_h, \quad j = 1,\ldots,N_h.$$

A similar stucture represents the load vector **F**

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^u \\ \\ \mathbf{F}^v \end{bmatrix},$$

where the entries of each sub-block vector are given by

$$\begin{aligned} F_i^u &= \int_\Omega \mathbf{f}\cdot[\phi_i, 0]^T\, d\Omega + \int_{\Gamma_N} \mathbf{g}\cdot[\phi_i, 0]^T\, d\Gamma \\ &= \int_\Omega f_x\,\phi_i\, d\Omega + \int_{\Gamma_N} g_x\,\phi_i\, d\Gamma \qquad\qquad i = 1,\ldots,N_h \\ F_i^v &= \int_\Omega \mathbf{f}\cdot[0, \phi_i]^T\, d\Omega + \int_{\Gamma_N} \mathbf{g}\cdot[0, \phi_i]^T\, d\Gamma \\ &= \int_\Omega f_y\,\phi_i\, d\Omega + \int_{\Gamma_N} g_y\,\phi_i\, d\Gamma \qquad\qquad i = 1,\ldots,N_h. \end{aligned}$$

Matrix **A** is symmetric and positive definite, and has $2N_h$ rows and $2N_h$ columns, while matrix **B** is rectangular and has M_h rows and $2N_h$ columns. The right-hand side **F** is a vector with $2N_h$ rows. The construction of these matrices and vector proceeds in the same fashion as described in Sect. 8.6.1.

9.9 Unique solvability and error analysis

The linear algebraic system (9.33) emanating from the two-field formulation of linear elasticity has a remarkably different structure compared to the linear algebraic system (8.43) emanating from the classical displacement-based approach. The difference is twofold. First, the stiffness matrix **K** is symmetric and positive definite, while the corresponding (generalized) stiffness matrix $\mathscr{K}$ is symmetric *but* indefinite, because of the zero block in the (2,2) position. Second, the rectangular matrix **B** has no definite rank until we do not specify the degree of the polynomial basis functions of V_h and Q_h. The consequences of these two differences are:

- system (8.43) admits a unique solution that can be efficiently computed by using one of the most appropriate methods (direct/iterative) illustrated in Sect. 2;
- system (9.33) has no guarantee to be uniquely solvable, and should this hold true, its solution is, in general, more difficult than that of (8.43) because of the larger size of $\mathscr{K}$ compared to $\mathbf{K}$.

In this sense, the displacement-based formulation is "better " than the displacement-pressure formulation. However, we have already seen that this latter approach is *indispensable* whenever a robust treatment of the incompressibility constraint is in order, so that its adoption is out of question in such a case.

In conclusion, we need now to characterize the choice of the finite element spaces V_h and Q_h in such a way that system (9.33) admits a unique solution. To this purpose, we need a discrete counterpart of the inf-sup condition (9.21).

Proposition 9.9.1 (Discrete inf-sup condition)**.** *The two spaces V_h and Q_h are said to be compatible (or, LBB compatible) if there exists a positive constant γ^* independent of h, such that for every $q_h \in Q_h$, there exists $\mathbf{v}_h \in V_h$ such that*

$$\int_\Omega q_h \operatorname{div} \mathbf{v}_h \, d\Omega \geq \gamma^* \|q_h\|_Q \|\mathbf{v}_h\|_V . \tag{9.35}$$

The discrete inf-sup condition has an important algebraic interpretation.

Theorem 9.9.2 (Rank condition)**.** *Assume that* (9.35) *is satisfied. Then*

$$\operatorname{rank}(\mathbf{B}) = \dim Q_h = M_h. \tag{9.36}$$

By Def. A.3.2, a necessary condition for (9.36) *to hold is that*

$$\underbrace{\dim Q_h}_{M_h} \leq \underbrace{\dim V_h}_{2N_h}. \tag{9.37}$$

Remark 9.9.3. Thm. 9.9.2 tells us that satisfying the discrete inf-sup condition is equivalent to stating that matrix $\mathbf{B}$ has *full rank*. For this reason, the discrete inf-sup condition is often referred to as the *rank condition*. From (9.37), it turns out that the rank condition introduces the constraint on the approximation space for the displacement field of being *richer* than that of the pressure field. This, in particular, does not allow to use *equal-order* polynomial spaces for *both* V_h and Q_h.

Having introduced the discrete version of the inf-sup condition, we can proceed as in Thm. 9.6.5, to prove the following result.

Theorem 9.9.4 (Unique solvability of the discrete Herrmann formulation)**.** *Assume that V_h and Q_h satisfy the discrete inf-sup condition* (9.35)*. Then, problem* (9.31) *admits a unique solution pair $(\mathbf{u}_h, p_h) \in (V_h \times Q_h)$, such that*

$$\|\mathbf{u}_h\|_V \leq \frac{C_P}{2\mu}\|\mathbf{f}\|_{(L^2(\Omega))^3} \tag{9.38a}$$

$$\|p_h\|_Q \leq \frac{2C_P}{\gamma^*}\|\mathbf{f}\|_{(L^2(\Omega))^3}. \tag{9.38b}$$

We can also prove the following result, that represents the extension of Ceà's Lemma to the two-field formulation.

Theorem 9.9.5 (Ceà's Lemma for the Herrmann formulation)**.** *Assume that V_h and Q_h satisfy the discrete inf-sup condition* (9.35)*. Then, there exists a positive constant C independent of h such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_V + \|p-p_h\|_Q \leq C\left[\inf_{\mathbf{v}_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_V + \inf_{q_h\in Q_h}\|p-q_h\|_Q\right] \tag{9.39}$$

where $(\mathbf{u}, p)$ and $(\mathbf{u}_h, p_h)$ are the solution pairs of problems (9.20) *and* (9.31)*, respectively.*

The previous result tells us that the discretization error (left-hand side) is bounded by the approximation error (right-hand side).

Theorem 9.9.6 (Convergence)**.** *Assume that V_h and Q_h satisfy the discrete inf-sup condition* (9.35) *and that the property of "good approximation" holds, i.e.*

$$\lim_{h\to 0}\|\mathbf{z}-\mathbf{v}_h\|_V = 0 \qquad \forall \mathbf{z}\in V$$
$$\lim_{h\to 0}\|\eta - q_h\|_Q = 0 \qquad \forall \eta \in Q.$$

Then, the approximate pair $(\mathbf{u}_h, p_h)$ converges to the exact solution $(\mathbf{u}, p)$ of (9.20)*, i.e.*

$$\lim_{h\to 0}\|\mathbf{u}-\mathbf{u}_h\|_V = 0 \qquad \lim_{h\to 0}\|p-p_h\|_Q = 0.$$

Moreover, if

$$\lim_{h\to 0}\|\mathbf{z}-\mathbf{v}_h\|_V = \mathscr{O}(h^p) \qquad \forall \mathbf{z}\in V$$
$$\lim_{h\to 0}\|\eta - q_h\|_Q = \mathscr{O}(h^p) \qquad \forall \eta \in Q,$$

for a certain $p \geq 1$, then

$$\lim_{h\to 0} \|\mathbf{u}-\mathbf{u}_h\|_V = \lim_{h\to 0} \|p-p_h\|_Q = \mathscr{O}(h^p)$$

and the convergence of the GFE approximation is said to be optimal.

9.10 The importance of satisfying the discrete inf-sup condition: spurious pressure modes

In this section, we provide an example that supports the importance of satisfying the discrete inf-sup in the FE approximation of the two-field model (9.20).

Definition 9.10.1 (Spurious pressure modes)**.** *A function $p_h^* \in Q_h$, with $p_h^* \neq 0$ is called a* spurious pressure mode *if*

$$\int_\Omega p_h^* \operatorname{div} \mathbf{v}_h \, d\Omega = 0 \qquad \mathbf{v}_h \in V_h. \tag{9.40}$$

Should a function p_h^* exist, then this would immediately imply that if p_h is a solution of (9.31), then also $p_h + p_h^*$ is a solution. In other words, there would be *no way* for filtering out (from the correct solution) the presence of a parasitic (i.e., unphysical) solution component that adds to the right one. This is the reason for calling p_h^* a "spurious pressure mode".

Theorem 9.10.2. *Let V_h and Q_h are such that the discrete inf-sup condition* (9.35) *is satisfied. Then, $p_h^* = 0$, i.e., no spurious mode can arise in the solution of* (9.31)*.*

Proof. We proceed by contradiction, and assume that a function p_h^*, with $p_h^* \neq 0$, exists such that (9.40) holds. Then, we have

$$0 = \int_\Omega p_h^* \operatorname{div} \mathbf{v}_h \, d\Omega \leq \gamma^* \|p_h^*\|_Q \|\mathbf{v}_h\|_V \qquad \text{with } \mathbf{v}_h \neq \mathbf{0}$$

which implies $\|p_h^*\|_Q = 0$ and thus, by definition of norm, $p_h^* = 0$. □

Remark 9.10.3. The previous result tells us that satisfying the discrete inf-sup condition is an *automatic* guarantee of avoiding the occurrence of spurious pressure modes in the numerical solution of incompressible elasticity problems.

9.11 Finite elements for incompressible elasticity

In this section, we provide a short overview of the FE pairs most commonly employed for the approximation spaces V_h and Q_h. The two-dimensional case is considered here, on a triangular partition $\mathscr{T}_h$ of the polygon Ω into triangles K. Two main classes of finite elements are available, depending on the choice for Q_h:

- discontinuous pressures on $\mathscr{T}_h$;
- continuous pressures on $\mathscr{T}_h$.

These two different approaches are allowed because Q_h is a subset of $Q = L^2(\Omega)$ and therefore *no continuity* requirement between two neighbouring elements K_1, K_2 needs be enforced a-priori in the construction of the pressure space. The same possibility is, instead, not allowed in the construction of V_h, because this latter is a subspace of $(H^1(\Omega))^2 \subset V$ and therefore discrete functions of V must be continuous across neighbouring elements.

In the error estimates that are discussed in the following, C denotes a positive constant, independent of h, not taking in general the same value at each occurrence, and the exact solutions $\mathbf{u}$ and p of (9.16) are assumed to be sufficiently regular in order the norms appearing in the right-hand side of each considered estimate to make sense. In all figures of this section, a black bullet identifies a dof for a scalar-valued function, while a black square identifies a dof for a vector-valued function.

9.11.1 Discontinuous pressures

The basic FE pair is the so-called $\mathbb{P}_1 - \mathbb{P}_0$ element, where

$$V_h = \left\{ \mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_1(K))^2 \; \forall K \in \mathscr{T}_h \right\}$$

and

$$Q_h = \left\{ q_h \in L^2(\Omega) \,|\, q_h|_K \in \mathbb{P}_0(K) \; \forall K \in \mathscr{T}_h \right\}.$$

The dofs for V_h and Q_h over each element $K \in \mathscr{T}_h$ are shown in Fig. 9.3.

The $\mathbb{P}_1 - \mathbb{P}_0$ FE pair does not satisfy the discrete inf-sup condition and is typically affected by severe locking phenomena. To see this, let us consider the incompressibility constraint

$$\int_\Omega q_h \operatorname{div} \mathbf{u}_h \, d\Omega = 0 \qquad \forall q_h \in Q_h.$$

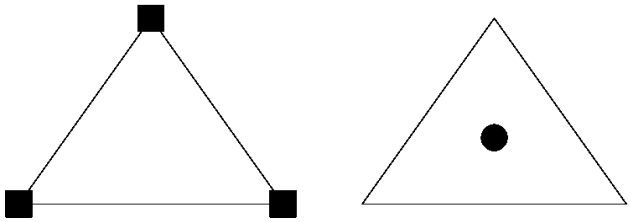


Figure 9.3: Dofs for the $\mathbb{P}_1 - \mathbb{P}_0$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

From the definition of the space Q_h, this condition amounts to requiring

$$\int_K \operatorname{div} \mathbf{u}_h \, dK = 0 \qquad \forall K \in \mathscr{T}_h.$$

Since $\mathbf{u}_h$ is linear over K, the above equation yields the strong condition

$$\operatorname{div} \mathbf{u}_h = 0 \qquad \forall K \in \mathscr{T}_h.$$

Going back to Ex. 9.2.1 (cf. Fig. 9.2.1), it is immediate to see that this latter condition implies that each triangle has to deform maintaining a constant area, so that $\mathbf{u}_h = \mathbf{0}$ in $\overline{\Omega}$ and the structure goes in complete locking.

A variant of the $\mathbb{P}_1 - \mathbb{P}_0$ element that passes the discrete inf-sup condition (thus avoiding the locking problem) is the so-called $(\text{cross} - \text{grid} - \mathbb{P}_1) - \mathbb{P}_0$ FE pair, whose dofs are depicted in Fig. 9.4.

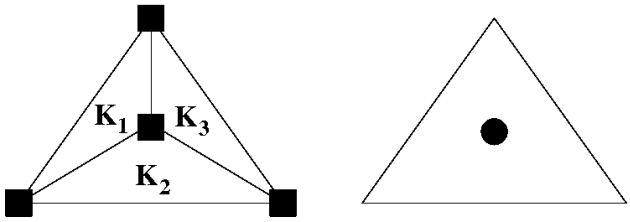


Figure 9.4: Dofs for the $(\text{cross} - \text{grid} - \mathbb{P}_1) - \mathbb{P}_0$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

The approximation space for the displacement consists of functions that are linear over the sub-elements K_1, K_2 and K_3, and are continuous across interelement edges. In this case, the incompressibility constraint becomes

$$\sum_{i=1}^{3} \operatorname{div} \mathbf{u}_h|_{K_i} |K_i| = 0 \qquad \forall K \in \mathscr{T}_h,$$

which does not admit as unique solution the displacement field $\mathbf{u}_h = \mathbf{0}$, so that the solid body is free to deform, maintaining the area of K constant as required. The $(\text{cross}-\text{grid}-\mathbb{P}_1)-\mathbb{P}_0$ FE pair satisfies the discrete inf-sup condition and the following optimal error estimate can be proved to hold

$$\|\mathbf{u}-\mathbf{u}_h\|_V + \|p-p_h\|_Q \leq Ch(\|\mathbf{u}\|_{(H^2(\Omega))^2} + \|p\|_{H^1(\Omega)}).$$

Passing to higher-order elements, we have the $\mathbb{P}_2-\mathbb{P}_1^{\text{disc}}$ FE pair, where

$$V_h = \left\{\mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_2(K))^2 \;\forall K \in \mathscr{T}_h\right\}$$

and

$$Q_h = \left\{q_h \in L^2(\Omega) \,|\, q_h|_K \in \mathbb{P}_1(K) \;\forall K \in \mathscr{T}_h\right\}.$$

The dofs for V_h and Q_h over each element $K \in \mathscr{T}_h$ are shown in Fig. 9.5.

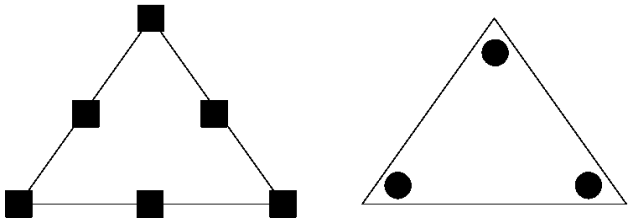


Figure 9.5: Dofs for the $\mathbb{P}_2-\mathbb{P}_1^{\text{disc}}$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

The $\mathbb{P}_2-\mathbb{P}_1^{\text{disc}}$ FE pair does not satisfy the discrete inf-sup condition. To overcome this problem, it is enough to *enrich* the space of displacements by adding an extra (vectorial) degree of freedom at the center of gravity of the element K, as depicted in Fig. 9.6.

The basis function b_K that is added to those of the space $(\mathbb{P}_2(K))^2$ is a cubic polynomial called *bubble function*, because it vanishes along all the boundary ∂K of the element and is identically equal to zero outside K. The locality of b_K can be advantageously exploited in the computer implementation of the scheme, by eliminating each interior dof corresponding to b_K in favor of those located on the boundary of K. Such a procedure is called *static condensation*.

The $(\mathbb{P}_2 \oplus b_K)-\mathbb{P}_1^{\text{disc}}$ FE pair was proposed by Crouzeix and Raviart in 1973. This element satisfies the discrete inf-sup condition and the following optimal error estimate can be proved

$$\|\mathbf{u}-\mathbf{u}_h\|_V + \|p-p_h\|_Q \leq Ch^2(\|\mathbf{u}\|_{(H^3(\Omega))^2} + \|p\|_{H^2(\Omega)}).$$

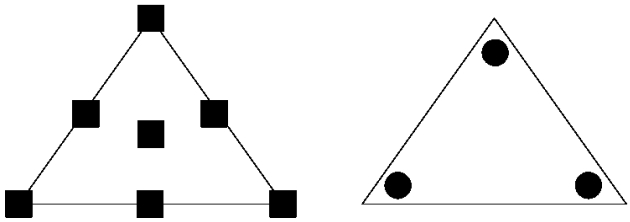


Figure 9.6: Dofs for the $(\mathbb{P}_2 \oplus b_K) - \mathbb{P}_1^{\text{disc}}$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

Remark 9.11.1 (Enriching the displacement FE space). The above presentation shows that the approach used to transform a FE pair that does not satisfy the discrete inf-sup condition into a pair the is LBB-stable consists in *enriching* the space V_h with respect to the space Q_h. In both cases (linear and quadratic elements for the displacement), the remedy consists in adding an internal degree of freedom to $\mathbf{u}_h$ which has the mechanical role to introduce an additional "flexibility" to the geometrically discretized structure, thus allowing it to deform maintaining the volume (area, in the 2D case) constant. This strategy can be extended to higher-order elements by introducing the $(\mathbb{P}_k \oplus B_{k+1}) - \mathbb{P}_{k-1}^{\text{disc}}$ FE pair, for $k \geq 2$, where

$$B_{k+1}(K) := \{v \in \mathbb{P}_{k+1}(K) \,|\, v = p_{k-2}\, b_K,\ p_{k-2} \in \mathbb{P}_{k-2}(K)\} \qquad k \geq 2 \quad (9.41)$$

where b_K is the cubic bubble function defined above. The $(\mathbb{P}_2 \oplus b_K) - \mathbb{P}_1^{\text{disc}}$ is the lowest order element of the above family, corresponding to setting $k = 2$. The $(\mathbb{P}_k \oplus B_{k+1}) - \mathbb{P}_{k-1}^{\text{disc}}$ FE pair satisfies the discrete inf-sup condition and the following optimal error estimate can be proved

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|p - p_h\|_Q \leq Ch^k(\|\mathbf{u}\|_{(H^{k+1}(\Omega))^2} + \|p\|_{H^k(\Omega)}) \qquad k \geq 2.$$

9.11.2 Continuous pressures

When using continuous pressures, the natural choice is the $\mathbb{P}_k - \mathbb{P}_k^{\text{cont}}$ FE pair, $k \geq 1$, where

$$V_h = \left\{\mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_k(K))^2 \ \forall K \in \mathscr{T}_h\right\}$$

and

$$Q_h = \left\{q_h \in C^0(\overline{\Omega}) \,|\, q_h|_K \in \mathbb{P}_k(K) \ \forall K \in \mathscr{T}_h\right\}.$$

The above choice for V_h and Q_h goes under the name *equal-order interpolation* and has, in principle, the considerable advantage of allowing the use of the *same*

shape functions for displacement and pressure in coding. Unfortunately, it leads to elements that do not satisfy the discrete inf-sup condition.

To construct FE pairs with continuous pressures that are LBB-stable, we need to use the $\mathbb{P}_k - \mathbb{P}_r^{\text{cont}}$ FE pair, with $k \geq 2$ and $1 \leq r \leq k-1$, where

$$V_h = \left\{ \mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_k(K))^2 \; \forall K \in \mathscr{T}_h \right\} \qquad k \geq 2$$

and

$$Q_h = \left\{ q_h \in C^0(\overline{\Omega}) \,|\, q_h|_K \in \mathbb{P}_r(K) \; \forall K \in \mathscr{T}_h \right\} \qquad 1 \leq r \leq k-1.$$

The lowest order element of the family, corresponding to $k = 2$ and $r = 1$, is well-known as the *Taylor-Hood* element, and its dofs are depicted in Fig. 9.7.

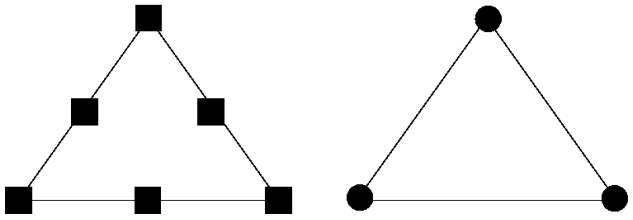


Figure 9.7: Dofs for the $\mathbb{P}_2 - \mathbb{P}_1^{\text{cont}}$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

Notice that in the case of the Taylor-Hood element the dofs for the pressure are located in correspondance of the vertices of K, so that they are single-valued over the whole triangulation, while in the case of the $\mathbb{P}_2 - \mathbb{P}_1^{\text{disc}}$ FE pair of Fig. 9.5 the dofs for the pressure are not single-valued at each vertex of $\mathscr{T}_h$ (see Fig. 9.8).

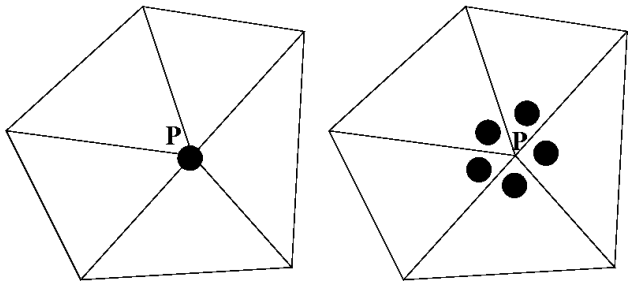


Figure 9.8: Dofs for the pressure on $\mathscr{T}_h$. Left: $\mathbb{P}_2 - \mathbb{P}_1^{\text{cont}}$; right: $\mathbb{P}_2 - \mathbb{P}_1^{\text{disc}}$.

The $\mathbb{P}_k - \mathbb{P}_r^{\text{cont}}$ FE pair satisfies the discrete inf-sup condition and the following optimal error estimate can be proved ($r = k-1$)

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|p - p_h\|_Q \leq Ch^k(\|\mathbf{u}\|_{(H^{k+1}(\Omega))^2} + \|p\|_{H^k(\Omega)}) \quad k \geq 2, \qquad r = k-1.$$

Remark 9.11.2 (How to enrich the $\mathbb{P}_k - \mathbb{P}_k$ FE pair). The enrichment strategy discussed in the case of discontinuous pressure spaces can be used also in the case where Q_h is made of continuous functions over $\overline{\Omega}$. To see this, we consider the lowest order case $k = 1$, and introduce the so-called $(\mathbb{P}_1 \oplus B_3) - \mathbb{P}_1^{\text{cont}}$ FE pair where

$$V_h = \left\{ \mathbf{v}_h \in (C^0(\overline{\Omega}))^2 \,|\, \mathbf{v}_h|_K \in (\mathbb{P}_1(K))^2 \oplus B_3 \; \forall K \in \mathscr{T}_h \right\}$$

and

$$Q_h = \left\{ q_h \in C^0(\overline{\Omega}) \,|\, q_h|_K \in \mathbb{P}_1(K) \; \forall K \in \mathscr{T}_h \right\}$$

the space B_3 being that defined in (9.41) in the case $k = 2$. The $(\mathbb{P}_1 \oplus B_3) - \mathbb{P}_1^{\text{cont}}$ FE pair is also well-known as *Mini-element*, and has been proposed by Arnold, Brezzi and Fortin in 1984. Its dofs are depicted in Fig. 9.9. The name Mini is due to the fact that the $(\mathbb{P}_1 \oplus B_3) - \mathbb{P}_1^{\text{cont}}$ FE pair is the most economical and LBB stable element with continuous pressures. The following optimal error estimate can be proved to hold for the Mini element

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|p - p_h\|_Q \leq Ch(\|\mathbf{u}\|_{(H^2(\Omega))^2} + \|p\|_{H^2(\Omega)}).$$

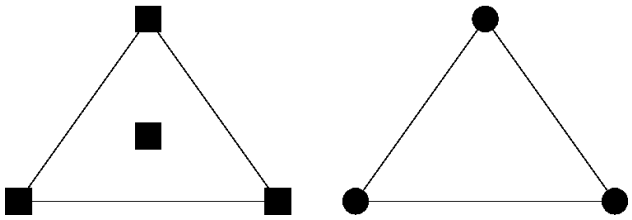


Figure 9.9: Dofs for the Mini element over K. Left: dofs for the displacement. Right: dof for the pressure.

Another possibility to enrich the $\mathbb{P}_1 - \mathbb{P}_1^{\text{cont}}$ FE pair is the so-called $(\mathbb{P}_1 - iso - \mathbb{P}_2) - \mathbb{P}_1^{\text{cont}}$ element, proposed by Bercovier and Pironneau in 1979 and whose dofs are shown in Fig. 9.10. The name "iso" is due to the fact that the geometrical location of the dofs is the same as for a standard $\mathbb{P}_2$ element for the displacement, but the functions are linear over each sub-triangle K_i, $i = 1, \ldots, 4$. An error estimate similar to that for the Mini element can be proved to hold also for the $(\mathbb{P}_1 - iso - \mathbb{P}_2) - \mathbb{P}_1^{\text{cont}}$ FE pair.

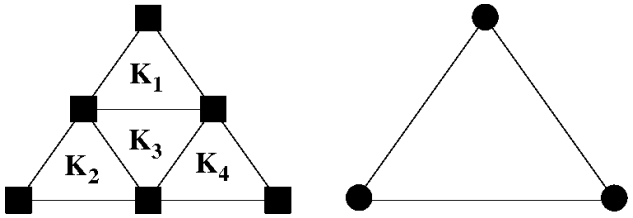


Figure 9.10: Dofs for the $(\mathbb{P}_1 - iso - \mathbb{P}_2) - \mathbb{P}_1^{cont}$ FE pair over K. Left: dofs for the displacement. Right: dof for the pressure.

9.12 Numerical example: locking and pressure spurious modes

In this section, we investigate the issue of the possible ocurrence of locking and pressure spurious modes in the numerical solution of the incompressible elastic problem (9.16) in the unit square $\Omega = (0,1)^2$ and in plane strain conditions. We assume that the body is constrained to ground along the sides $y = 0$ and $x = 0$ and that no-stress conditions are applied on the rest of the boundary. Volume forces are set equal to $\mathbf{f} = [0, -1]^T$, while Young modulus is set equal to 1. For $h > 0$, we denote by $\{\mathscr{T}_h\}_{h>0}$ a family of finite element triangulations, each member of which is a uniform partition of the unit square into $2N^2$ right-angled triangles of side $h = 1/N$. An example of $\mathscr{T}_h$ is shown in Fig. 9.11 in the case $N = 4$.

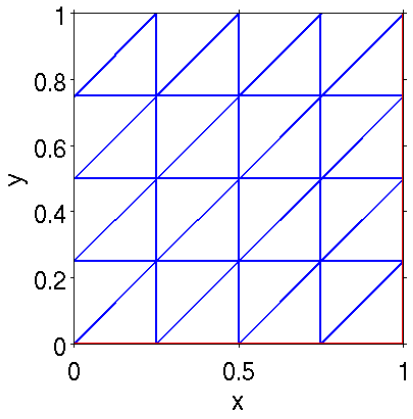


Figure 9.11: Finite element mesh.

In general, for a fixed value of N (equivalently, of h), the partition consists of

a number of:

- $\texttt{NE} = 2N^2$ elements;
- $\texttt{Nv} = (N+1)^2$ vertices;
- $\texttt{Nv\_D} = 2(N+1)-1 = 2N+1$ vertices on $\overline{\Gamma}_D$;
- $\texttt{Nv\_N} = 2(N-1)+1 = 2N-1$ vertices on Γ_N;
- $\texttt{Nv\_i} = \texttt{Nv} - (\texttt{Nv\_D} + \texttt{Nv\_N}) = (N-1)^2$ vertices in the interior of Ω;
- $\texttt{Ne\_D} = 2N$ edges on $\overline{\Gamma}_D$;
- $\texttt{Ne\_N} = 2N$ edges on $\overline{\Gamma}_N$;
- $\texttt{Ne\_i} = N^2 + 2N(N-1) = 3N^2 - 2N$ edges in the interior of Ω;
- $\texttt{Ne} = \texttt{Ne\_i} + \texttt{Ne\_D} + \texttt{Ne\_N} = 3N^2 + 2N$ edges.

In the example of Fig. 9.11, we have $\texttt{Nelem} = 32$, $\texttt{Nv} = 25$, $\texttt{Nv\_D} = 9$, $\texttt{Nv\_N} = 7$ and $\texttt{Nv\_i} = 9$, while $\texttt{Ne} = 48$, $\texttt{Ne\_D} = \texttt{Ne\_N} = 4$ and $\texttt{Ne\_i} = 2N(N-1) = 40$.

9.12.1 Discontinuous pressure FE space

We start by using the simplest choice for V_h and Q_h, the FE pair $\mathbb{P}_1 - \mathbb{P}_0$. We have

$$\dim V_h = 2N_h = 2(\texttt{Nv\_i} + \texttt{Nv\_N}) = 2((N-1)^2 + 2N - 1) = 2N^2$$

and

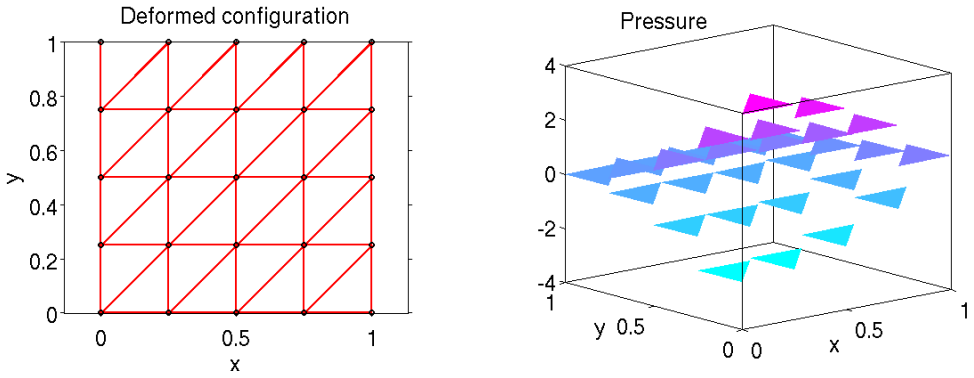
$$\dim Q_h = \texttt{NE} = 2N^2.$$

Thus, $M_h = 2N_h$ and inequality (9.37) becomes in this special case an equality. The computed deformed structure and pressure field are shown in Figs. 9.12(a) and 9.12(b).

Results clearly indicate:

- *no deformation* of the structure (locking);
- unphysical oscillations in the pressure (spurious pressure mode).

Matlab coding. The following `Matlab` commands extract from the computed solution of system (9.33) the maximum horizontal and vertical displacement components.



(a) Deformed structure (b) Pressure

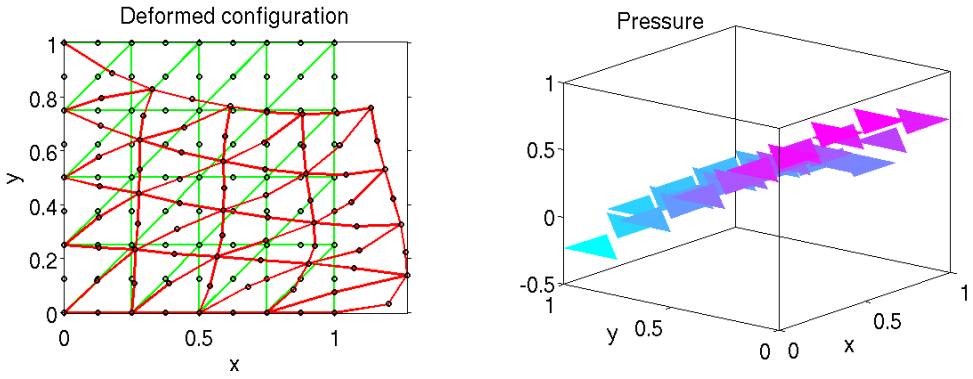
Figure 9.12: Computed solution with the $\mathbb{P}_1 - \mathbb{P}_0$ FE pair that does no satisfy the discrete inf-sup condition.

```
>> EF2D_elinc

 ********* test case ----> Locking *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
solving the linear system
plot of solution
>> figure; pdesurf(griglia_conn.p,griglia_conn.t,U_p')
>> xlabel('x')
>> ylabel('y')
>> title('Pressure')
>> max(abs(U_ux))

ans =

     0

>> max(abs(U_uy))

ans =

     0
```

To avoid the occurrence of locking, we increase the polynomial degree of the approximation for the displacement. For this, we use the FE pair $\mathbb{P}_2 - \mathbb{P}_0$. In this case we have

$$\dim V_h = 2N_h = 2(\texttt{Nv\_i} + \texttt{Nv\_N}) + 2(\texttt{Ne\_i} + \texttt{Ne\_N}) = 8N^2.$$

Thus, $M_h < 2N_h$. The computed deformed structure and pressure field are shown in Figs. 9.13(a) and 9.13(b). Results clearly indicate the absence of any locking or spurious oscillations in the pressure distribution.



(a) Deformed structure (b) Pressure

Figure 9.13: Computed solution with the $\mathbb{P}_2 - \mathbb{P}_0$ FE pair that satisfies the discrete inf-sup condition.

Matlab coding. The following `Matlab` commands extract from the computed solution of system (9.33) the maximum horizontal and vertical displacement components.

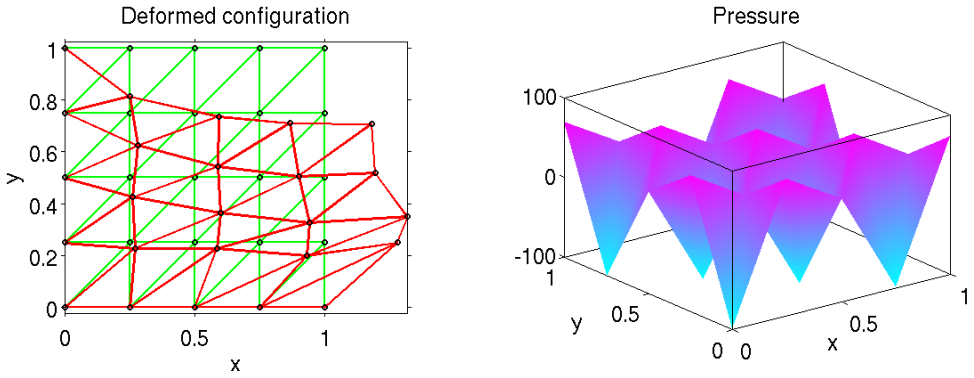
```
>> EF2D_elinc

 ********* test case ----> Locking *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
solving the linear system
plot of solution
>> figure; pdesurf(griglia_conn.p,griglia_conn.t,U_p')
>> max(abs(U_ux))

ans =

    0.2708

>> max(abs(U_uy))

ans =

    0.2642
```

9.12.2 Continuous pressure FE space

We now consider the use of a continuous pressure approximation space. The simplest element is thus the $\mathbb{P}_1 - \mathbb{P}_1^{\text{cont}}$ FE pair. We have

$$\dim Q_h = M_h = \mathtt{Nv} = (N+1)^2$$

so that condition (9.37) is satisfied for $N \geq 3$. The computed deformed structure and pressure field are shown in Figs. 9.14(a) and 9.14(b). Results clearly indicate the absence of locking and the presence of strong spurious oscillations in the pressure distribution, in accordance with the fact that the $\mathbb{P}_1 - \mathbb{P}_1^{\text{cont}}$ FE pair does not satisfy the discrete inf-sup condition.



(a) Deformed structure (b) Pressure

Figure 9.14: Computed solution with the $\mathbb{P}_1 - \mathbb{P}_1^{\text{cont}}$ FE pair that does not satisfy the discrete inf-sup condition.

Matlab coding. The following `Matlab` commands extract from the computed solution of system (9.33) the maximum horizontal and vertical displacement components. Notice the message warning us about the fact that the linear algebraic system (9.33) is singular.

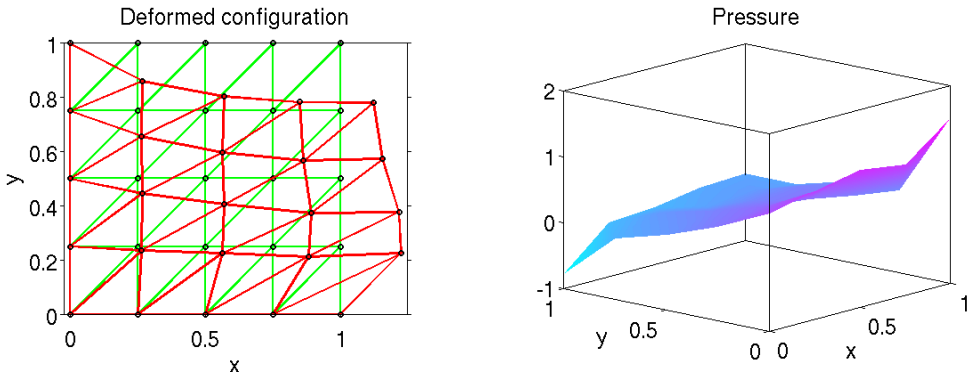
```
>> EF2D_elinc

 ********* test case ----> Locking *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
solving the linear system
Warning: Matrix is close to singular or badly scaled.
```

```
        Results may be inaccurate. RCOND = 6.938894e-18.
> In EF2D_elinc at 399
plot of solution
>> figure; pdesurf(griglia_conn.p,griglia_conn.t,U_p)
>> max(abs(U_ux))

ans =

    0.2246

>> max(abs(U_uy))

ans =

    0.2055
```

To overcome the instability of the $\mathbb{P}_1 - \mathbb{P}_1$ pair, we adopt the simplest variant of such an element that satisfies the discrete inf-sup condition, the so-called *Mini-element*, that is, the $(\mathbb{P}_1 \oplus B_3) - \mathbb{P}_1^{\text{cont}}$ FE pair. In this case, we have

$$\dim V_h = 2N_h = 2(\texttt{Nv\_i} + \texttt{Nv\_N} + \texttt{NE}) = 4N^2$$

from which it follows that condition (9.37) is satisfied for every $N \geq 1$. The computed deformed structure and pressure field are shown in Figs. 9.15(a) and 9.15(b). Results clearly indicate the absence of locking and spurious oscillations, in accordance with the fact that the Mini-element satisfies the discrete inf-sup condition.



(a) Deformed structure (b) Pressure

Figure 9.15: Computed solution with the Mini-element that satisfies the discrete inf-sup condition.

Matlab coding. The following `Matlab` commands extract from the computed solution of system (9.33) the maximum horizontal and vertical displacement components. Notice how these latter are very close to those computed by the unstable pair $\mathbb{P}_1 - \mathbb{P}_1^{\text{cont}}$, confirming the fact that not satisfying the discrete inf-sup condition does not prevent, in general, the displacement variable to be approximated correctly. It is the pressure, however, that results completely inaccurate.

```
>> EF2D_elinc

 ********* test case ----> Locking *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
solving the linear system
plot of solution
>> figure; pdesurf(griglia_conn.p,griglia_conn.t,U_p)
>> xlabel('x')
>> ylabel('y')
>> title('Pressure')
>> max(abs(U_ux))

ans =

    0.2241

>> max(abs(U_uy))

ans =

    0.2173
```

9.13 Numerical example: convergence analysis

In this section, we illustrate the convergence performance of the GFEM in the solution of a problem in incompressible elasticity (plane strain conditions) with available exact solution. The computational domain is the square $\Omega = (0,\pi)^2$, obtained as the image of the reference domain $\widehat{\Omega} = (-1,1)^2$ through the linear map

$$\begin{cases} x := \dfrac{\pi}{2}(\xi+1) & \xi \in [-1,1] \\ y := \dfrac{\pi}{2}(\eta+1) & \eta \in [-1,1]. \end{cases}$$

The value of the Young modulus is $E = 3$ and the Poisson modulus is $\nu = 0.5$. The solid body is constrained to ground on three of the four sides, as shown in

Fig. 9.16, while the remaining lateral side is subject to a normal stress condition

$$\mathbf{h}_N = \left[0, -(\sin(y))^2\right]^T.$$

The volume force is

$$\mathbf{f} = \frac{\pi}{2}\left[\sin(2y), \sin(2x)(3-8(\sin(y))^2)\right]^T,$$

in such a way that the exact displacement field is

$$\mathbf{u} = \frac{1}{\pi}\left[(\sin(x))^2\sin(2y), -\sin(2x)(\sin(y))^2\right]^T.$$

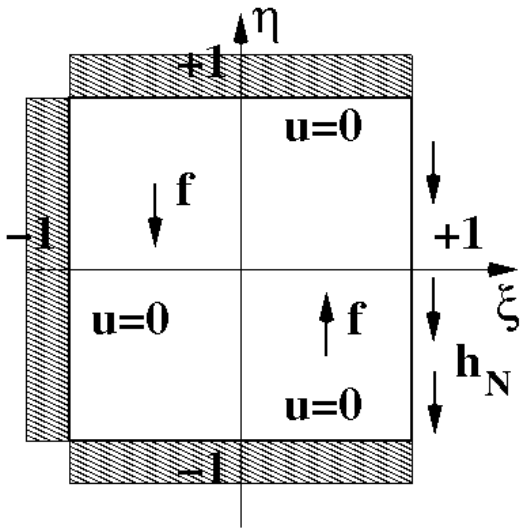


Figure 9.16: Computational domain $\widehat{\Omega}$, volume forces and boundary conditions.

A family of five uniformly refined triangulations $\{\mathscr{T}_h\}_{h>0}$, with $h = \pi/N$, is used in the nuemrical computations, with $N = [5, 10, 20, 40, 80]^T$. For sake of comparison, the following FE pairs are considered: Mini element, $\mathbb{P}_2 - \mathbb{P}_1^{\text{cont}}$, $\mathbb{P}_3 - \mathbb{P}_2^{\text{cont}}$ and $\mathbb{P}_3 - \mathbb{P}_1^{\text{cont}}$. The symbols used to identify the results obtained with each FE pair are (in the same order): circles, asterisks, squares and diamonds.

Figs. 9.13 and 9.13 show the experimental error curves for the sole displacement variable, measured in the H^1 and L^2 norm, and plotted in log-log scale as a function of h for each of the considered finite element spaces. The obtained

results are in complete agreement with the theoretical conclusions of Sect. 9.11.2. Namely, we see that the error in H^1 decreases as h^k for the Mini element ($k=1$), the $\mathbb{P}_2-\mathbb{P}_1^{\text{cont}}$ element ($k=2$, $r=1$) and the $\mathbb{P}_3-\mathbb{P}_2^{\text{cont}}$ element ($k=3$, $r=2$). For the same FE pairs, the error in L^2 decreases as h^{k+1} in agreement with Thm. 5.2.4.

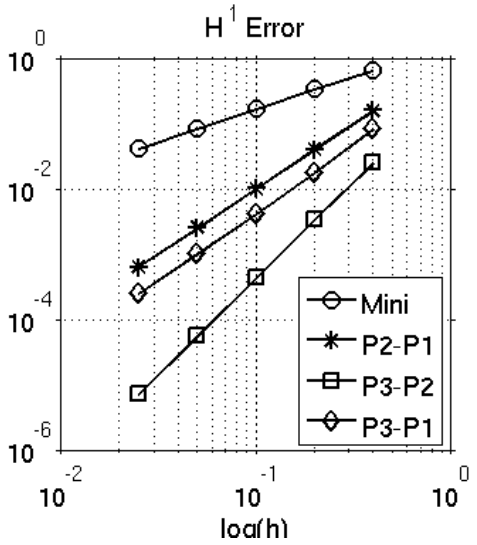


Figure 9.17: Convergence analysis: $\|\mathbf{u}-\mathbf{u}_h\|_V$.

A different asymptotic behavior as a function of h is registered for the $\mathbb{P}_3-\mathbb{P}_1^{\text{cont}}$ FE pair ($k=3$, $r=1$). In such a case, we still have convergence, because the element is LBB-stable. However, the value of r is not equal to $k-1$, so that the estimate that can be proved in this case is not optimal. As a matter of fact, because of the coupling between $\mathbf{u}_h$ and p_h in the exactly incompressible problem, the accuracy of the approximation of the displacement variable is limited by the lower degree of the pressure finite element space. Precisely, if $1 \le r < k-1$, it can be proved that

$$\|\mathbf{u}-\mathbf{u}_h\|_V + \|p-p_h\|_Q \le Ch^{\ell+1}(\|\mathbf{u}\|_{(H^{\ell+1}(\Omega))^2} + \|p\|_{H^{\ell+1}(\Omega)})$$

and

$$\|\mathbf{u}-\mathbf{u}_h\|_{(L^2(\Omega))^2} \le Ch^{\ell+1}(\|\mathbf{u}\|_{(H^{\ell+1}(\Omega))^2} + \|p\|_{H^{\ell+1}(\Omega)})$$

where $\ell := \min\{k-1, r\}$. In the present case, we have $\ell = 1$, which agrees with the result of Figs. 9.13 and 9.13 that predict a quadratic convergence in both norms.

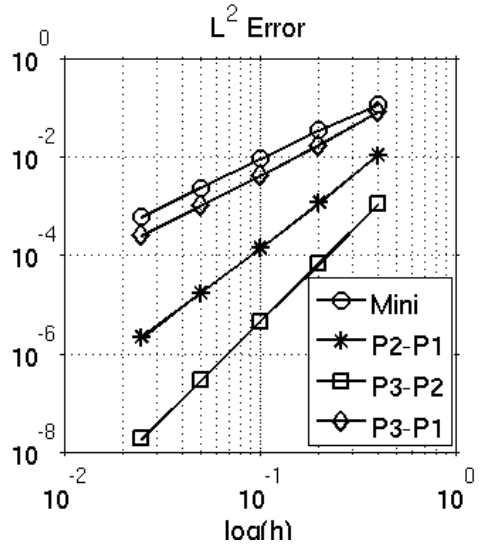


Figure 9.18: Convergence analysis: $\|\mathbf{u}-\mathbf{u}_h\|_{(L^2(\Omega))^2}$.

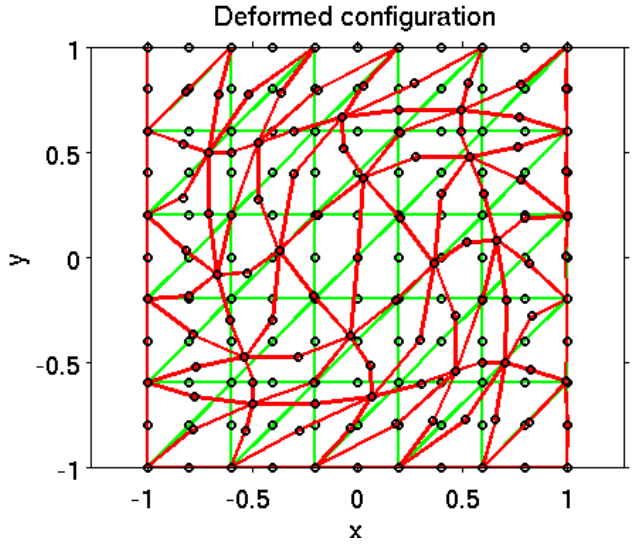


Figure 9.19: Computed solution with the $\mathbb{P}_2-\mathbb{P}_1^{\text{cont}}$ FE pair: deformed structure.

A visual plot of the deformed structure and of the pressure distribution in the elastic body are shown in Figs. 9.13 and 9.13 as computed by using the $\mathbb{P}_2-\mathbb{P}_1^{\text{cont}}$ pair on a mesh with $N=5$ (deformed structure) and $N=10$ (pressure). It can be

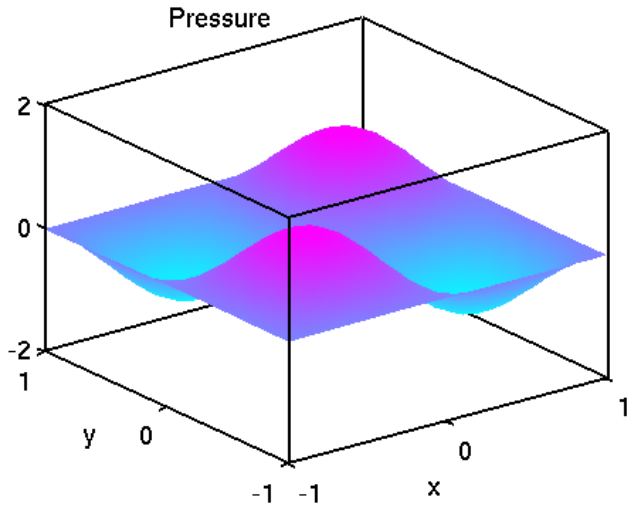


Figure 9.20: Computed solution with the $\mathbb{P}_2 - \mathbb{P}_1^{\text{cont}}$ FE pair: pressure field.

seen that the Dirichlet boundary conditions are enforced in an essential manner, i.e., the displacement of the nodes belonging to Γ_D is equal to zero. It can be also seen very clearly that the structure deforms in a counterclockwise sense, in accordance with the applied volume forces and boundary tractions, and that the area of the elastic body remains constant, in accordance with the incompressibility constraint.

Part V

Examination Problems with Solution

This part contains several examination problems with complete solution, including theoretical questions, numerical tests and `Matlab` coding implementation.

Chapter 10

Problems with Solution

Abstract

In this chapter, we deal with the detailed solution of several examination problems, including theoretical and numerical questions, as well as their coding using the `Matlab` software environment. During each exam, time left for solving the three exercises is typically 3 hours, and for each exercise a value in "points" is assigned to allow the student a preliminary self-evaluation before exam completion.

10.1 Examination of July 09, 2012

Exercise 1 (11 points). Consider the linear elasticity problem in the *incompressible regime* ($\nu = 0.5$) and in *plane strain conditions*. The computational domain is shown in Fig. 10.1, the two sides at the bottom are constrained to ground, the lateral and top sides are subject to a normal stress **h** as indicated in the figure while on the remaining sides a stress-free condition is applied. The Young modulus E is equal to 30. Solve the problem with the code `EF2D_elinc` using the *Mini* element on a triangulation with average mesh size equal to 0.05, and the following specific questions must be addressed:

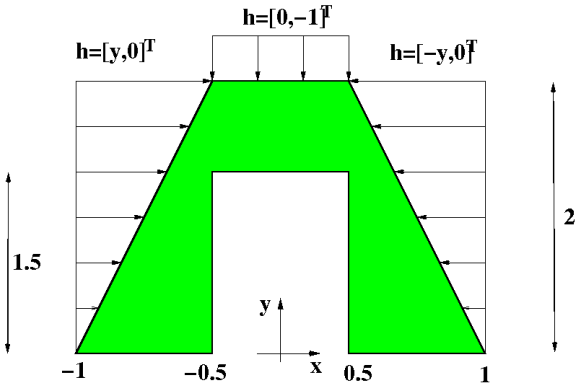


Figure 10.1: Computational domain, boundary conditions and applied loads.

(a) plot the configuration of the deformed structure superposed to the original undeformed configuration, and comment the obtained result based on the symmetry of the problem;

(b) compute the resultant of ground reacting forces and check that global force equilibrium is satisfied;

(c) compute the maximum horizontal and vertical displacements;

(d) plot the qualitative behavior of the computed pressure and give a mechanical comment to the obtained result;

(e) repeat all the points (a), (b), (c) and (d) using the $\mathbb{P}_1$-$\mathbb{P}_1$ FE pair, comparing the obtained results with those of the Mini element. Which variable between **u** and p turns out to be correctly approximated in the two cases? Why?

Exercise 2 (11 points). Consider the following two-point BVP:

$$(P) \qquad \begin{cases} -\mu\, u'' + \sigma\, u = f & \text{in } \Omega = (0,1) \\ u(0) = u_0, \quad u(1) = u_1, \end{cases}$$

where μ and σ are constants > 0, u_0 and u_1 are constants, and f is a given function in $L^2(0,1)$.

(a) Let R denote the lifting of boundary data. Write the weak formulation (W) of problem (P), specifying the space V and the corresponding norm $\|\cdot\|_V$.

(b) Use the Lax-Milgram Lemma to verify that (W) admits a unique solution and determine the a-priori estimate for $\|u\|_V$.

(c) Assuming that the solution of (W) belongs to H^s, with $s \geq 2$ given, write the error estimate for the GFEM of degree $r \geq 1$ applied to (W) and discuss the estimate as a function of the relation between r and s and of the ratio σ/μ.

Exercise 3 (11 points). Set $u_0 = 1$, $u_1 = 0$, $\sigma = 1$ and $f(x) = 1$, so that the exact solution of (P) is

$$u(x) = 1 + C\left(e^{-\alpha x} - e^{+\alpha x}\right),$$

with $\alpha := \sqrt{1/\mu}$ and $C := 1/(e^{\alpha} - e^{-\alpha})$.

(a) Set $\mu = 10^{-1}$. Solve (P) with the GFEM using the code `EF1D` with $r = 1, 2$ and 3 on a family of uniform grids with $N = [10, 20, 40, 80, 160]$ elements. Set $h := 1/N$, report in a table the errors in the H^1 and L^2 norms as functions of h, and determine experimentally the order of convergence p for each considered degree r. Comment the obtained results based on the conclusions drawn at point (c) of Exercise 2.

(b) Set now $\mu = 10^{-4}$, $N = 10$ and $r = 1$. Compute the solution u_h with these data and report a plot of it superposing to that of the exact solution u. Compute the error in the L^2 norm and the absolute value of the maximum nodal error, e_{max}, and check if u_h satisfies the DMP justifying the answer based on the value of the local Pèclet number.

(c) Repeat point (b) using the same values of μ, N and r but using the lumping method, and compare the obtained results.

(d) Determine the minimum number of intervals N_{min} that are needed to make u_h satisfy the DMP, repeat point (b) with $\mu = 10^{-4}$, $N = N_{min}$ and $r = 1$ (no lumping) and compare the obtained results.

10.1.1 Solution of Exercise 1

Matlab coding. The following `Matlab` functions are called by the script `EF2d_elinc` to solve Exercise 1(a). This script is the main program that implements the GFEM in 2D for the numerical treatment of the Herrmann formulation for linear elasticity.

```
%%%%%%%%%%%%%
% Ex. 1
%%%%%%%%%%%%%
function [griglia,dati_problema,soluzione] = Ex1()
% problem definition
 griglia = struct('file_griglia','griglia09072012.mat', ...
                  'ku',1.5, ...   % Mini element
                  'kp',1);
 dati_problema = struct('coeff','Es1_coeff', ...
                        'tipo',1, ... % 1 plane strain, 2 plane stress
                        'tipo_bc',[ 1  2  3  4  5  6  7  8;
                                   -2 -2 -2 -2 -2 -2 -1 -1], ...
                        'g','Es1_g', ...
                        'h','Es1_h');
 soluzione = struct('u_es',[]');
return

function [E,nu,f] = Ex1_coeff(x,y)
% problem coefficients/volume force
 E  = 30*ones(size(x));
 nu = 0.5*ones(size(x));
 f(1,:) = 0*x;
 f(2,:) = 0*x;
return

function g = Ex1_g(x,y)
% Dirichlet bcs
 g = zeros(2,size(x,2));
return

function h = Ex1_h(x,y,marker)
% Neumann bcs
 switch(marker)
  case(1)
   h(1,:) = y;
   h(2,:) = 0*x;
case(2)
   h(1,:) = 0*x;
   h(2,:) = -1;
  case(3)
   h(1,:) = -y;
   h(2,:) = 0*x;
  case(4)
```

```
   h(1,:) = 0*x;
   h(2,:) = 0*x;
  case(5)
   h(1,:) = 0*x;
   h(2,:) = 0*x;
  case(6)
   h(1,:) = 0*x;
   h(2,:) = 0*x;
  otherwise
   error('Error in Neumann bcs!!')
 end
return
```

(a) The finite element mesh used to solve numerically the problem is generated using the `pdetool` toolbox available in the `Matlab` software environment and is shown in Fig. 10.2.

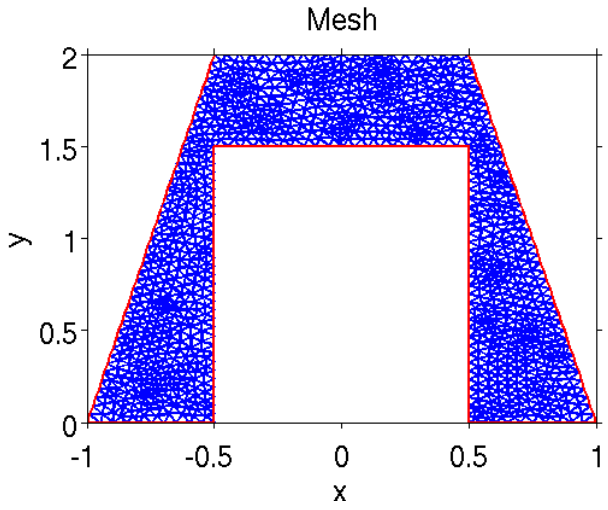


Figure 10.2: Computational mesh with average mesh size $h = 0.1$.

Matlab coding. Running the script `EF2d_elinc` yields the following output. The other `Matlab` commands allow to access the various dimensions of data structure and problem unknowns.

```
>> EF2D_elinc

 ********* test case ----> ex1 *********

reading data
assembling stiffness matrix and load vector
Neumann boundary conditions
Dirichlet boundary conditions
```

```
solving the linear system
plot of solution

>> size(griglia_conn.p)

ans =

           2        1107

>> size(griglia_conn.t)

ans =

           4        2000

>> size(U_ux), size(U_uy), size(U_p)

ans =

        3107           1


ans =

        3107           1


ans =

        1107           1
```

The number of triangles is equal to 2000 and the number of vertices is 1107. The total number of dofs for each component of the displacement $\mathbf{u}_h$ is 3107 which equals the sum of 2000 (one internal dof per mesh triangle) plus the number of vertices that do not belong to Γ_D. The total number of dofs for p_h is equal to the number of vertices.

The deformed configuration computed by the numerical scheme using the Mini element is shown in Fig. 10.3.

We can see that the structure deforms itself in a symmetric manner, according to the symmetry of the geometry and of applied external loads. In particular, the two lateral sides turn out to be subject to a compressive state that produces a significant deformation. The incompressibility constraint is responsible for the strong deflection of the internal vertical sides located at $x = -0.5$ and $x = 0.5$, respectively. The horizontal side located at $y = 1.5$ is subject to a vertical force that produces a negative displacement of the side itself along the y direction.

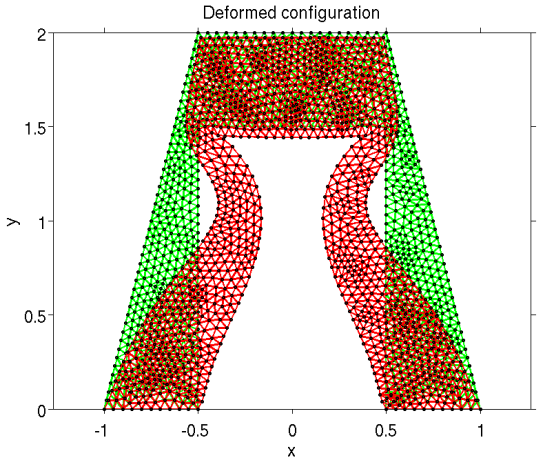


Figure 10.3: Computed deformed configuration.

(b) Enforcing global equilibrium of horizontal and vertical applied forces with the reaction forces **R** yields

$$\begin{aligned} R_x + 2*2/2 - 2*2/2 = 0 \quad &\Rightarrow \quad R_x = 0 \\ R_y + (-1)*1 = 0 \quad &\Rightarrow \quad R_y = 1. \end{aligned}$$

Matlab coding. The following `Matlab` commands allow to access the program output variables containing the components of reaction forces.

```
>> R_tot

R_tot =

   -0.0000
    1.0000
```

Results are in agreement with physical expectation.

(c)

Matlab coding. The following `Matlab` commands allow to compute the required quantities.

```
>> [Uxmax, Ix] = max(abs(U_ux))

Uxmax =

    0.3433
```

```
Ix =

   118

>> griglia_conn.p(:,118)

ans =

   -0.5888
    1.0656

>> [Uymax, Iy] = max(abs(U_uy))

Uymax =

    0.1164


Iy =

    76

>> griglia_conn.p(:,76)

ans =

   -0.6591
    1.3636
```

The maximum horizontal and vertical displacements are 0.3433 and 0.1164, respectively, and correspond to the deformed configuration of points $P = [-0.5888, 1.0656]^T$ and $Q = [-0.6591, 1.3636]^T$ (in the original undeformed structure).

(d)

Matlab coding. The following `Matlab` commands allow to plot the computed pressure field in 3D.

```
>> figure; pdesurf(griglia_conn.p,griglia_conn.t,U_p)
>> colorbar; colormap('jet'); xlabel('x'); ylabel('y'); title('Pressure')
```

The resulting picture of the pressure is displayed in Fig. 10.4. Consistently with the applied external loads and the geometry of the structure, we see that the points $[-0.5, 1.5]^T$ and $[0.5, 1.5]^T$ are subject to the highest compressive state. Concerning with the part of the structure that is constrained to ground, the point $[-1, 0]^T$ (symmetrically, $[1, 0]^T$) is subject to the highest tensile stress, while the point $[-0.5, 0]^T$ (symmetrically, $[0.5, 0]^T$) is subject

to the highest compressive state, in agreement with the clockwise (counter-clockwise) torque exerted by the horizontal linearly varying pressure force distributed along the lateral side of the structure.

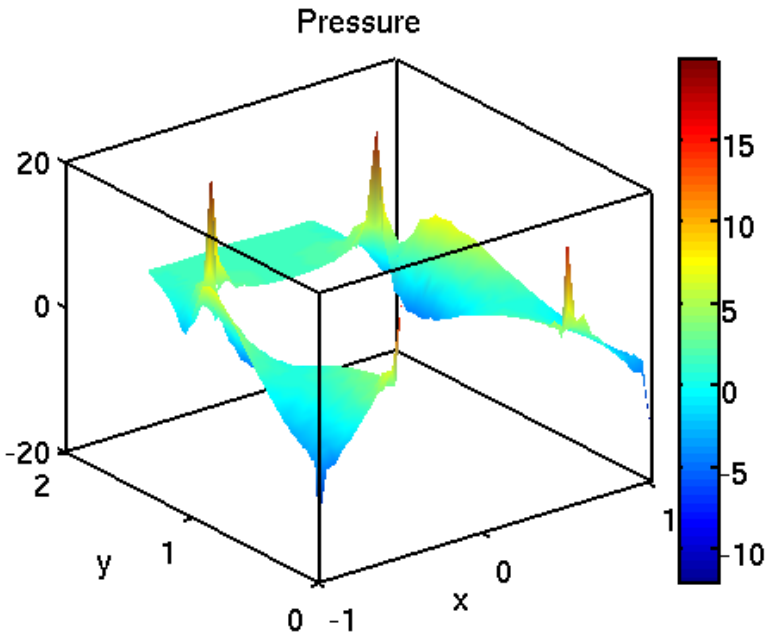


Figure 10.4: 3D plot of the pressure field.

Matlab coding. The following `Matlab` commands allow to plot the maximum and minimum values of p_h over the mesh and the corresponding nodal coordinates in the structure.

```
>> [Pmax,Ip]=max(U_p)

Pmax =

   19.8730


Ip =

     5

>> griglia_conn.p(:,5)

ans =

   -0.5000
```

```
    1.5000

>> [Pmin,Ip]=min(U_p)

Pmin =

  -12.0233


Ip =

     1

>> griglia_conn.p(:,1)

ans =

    -1
     0
```

(e) Repeating the analysis previously done with the Mini element, but using the $\mathbb{P}_1 - \mathbb{P}_1$ FE pair yields a similar result for the deformed structure configuration, with maximum horizontal and vertical displacements of 0.3383 and 0.1152, respectively (to be compared with 0.3433 and 0.1164 in the case of the Mini element). Computed reaction forces are again consistent with the theoretical value $\mathbf{R} = [0,1]^T$. However, as expected, the computed pressure field is affected by numerical instabilities (oscillations) because the $\mathbb{P}_1 - \mathbb{P}_1$ FE pair does not satisfy the LBB compatibility condition. These oscillations (wiggles) are clearly visible in Fig. 10.5, from which we can also see that the maximum variation of the pressure is much higher than in the case of the solution computed by the Mini element.

Matlab coding. The following `Matlab` commands allow to plot the maximum and minimum values of p_h over the mesh and the corresponding nodal coordinates in the structure.

```
>> [Pmax,Ip]=max(U_p)

Pmax =

   40.2087


Ip =

     6

>> griglia_conn.p(:,6)

ans =
```

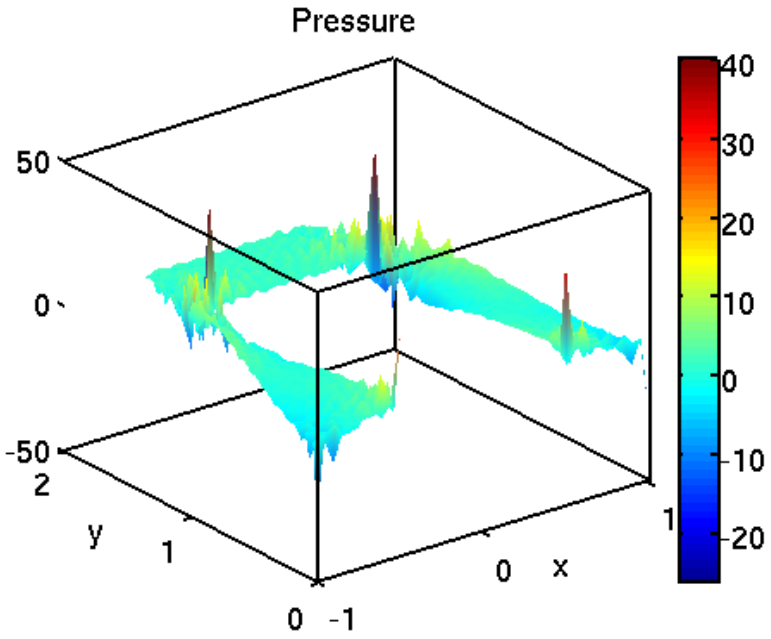


Figure 10.5: 3D plot of the pressure field.

```
    0.5000
    1.5000

>> [Pmin,Ip]=min(U_p)

Pmin =

  -26.1788


Ip =

     1

>> griglia_conn.p(:,1)

ans =

    -1
     0
```

10.1.2 Solution of Exercise 2

The two-point BVP is an example of reaction-diffusion equation with non-homogeneous Dirichlet boundary conditions.

(a) We first need to reduce problem (P) to an equivalent problem with *homogeneous* boundary conditions at $x = 0$ and $x = 1$. To do this, we introduce a function $R : \overline{\Omega} \to \mathbb{R}$ such that $R \in H^1(\Omega)$, and that $R(0) = u_0$ and $R(1) = u_1$, the so-called *lifting of boundary data*. Such a function certainly exists (is not unique), and the simplest example is

$$R(x) = u_0 + (u_1 - u_0)x.$$

Then, according to the linear superposition principle, we decompose the exact solution u as

$$u(x) = u^0(x) + R(x), \tag{10.1}$$

so that the *new* problem unknown u^0 is such that

$$u^0(0) = u^0(1) = 0.$$

After doing this, we set

$$V := H_0^1(\Omega).$$

Using Poincarè's inequality (B.26), we conclude that the equivalent norm in V can be taken as

$$\|\phi\|_V := \|\phi'\|_{L^2(0,1)} = \left(\int_0^1 (\phi')^2 \, dx \right)^{1/2} \qquad \phi \in V.$$

Defining

$$\begin{aligned} a(w,v) &:= \int_0^1 \mu \, w' \, v' \, dx + \int_0^1 \sigma \, w \, v \, dx && w, v \in V \\ F(v) &:= \int_0^1 f \, v \, dx - a(R,v) && v \in V, \end{aligned}$$

the weak formulation of (P) reads:
find $u^0 \in V$ such that

$$(W) \qquad a(u^0, v) = F(v) \qquad \forall v \in V.$$

(b) In the solution of this point of the exercise, U and W are two generic functions in V.

Continuity of $a(\cdot,\cdot)$:

$$\begin{aligned} |a(U,W)| &\leq \mu\left|\int_0^1 U'W'\,dx\right| + \sigma\left|\int_0^1 UW\,dx\right| \\ &\leq \mu\|U\|_V\|W\|_V + \sigma C_P^2\|U\|_V\|W\|_V \end{aligned}$$

so that the continuity constant is $M := \mu + \sigma C_P^2$.

Coercivity of $a(\cdot,\cdot)$:

$$a(U,U) = \mu\int_0^1 (U')^2\,dx + \sigma\int_0^1 U^2\,dx = \mu\|U\|_V^2 + \sigma\|U\|_{L^2(0,1)}^2 \geq \mu\|U\|_V^2$$

so that the coercivity constant is $\beta := \mu$.

Continuity of $F(\cdot)$:

$$\begin{aligned} |F(W)| &\leq \left|\int_0^1 fW\,dx\right| + |a(R,W)| \\ &\leq \|f\|_{L^2(0,1)}\|W\|_{L^2(0,1)} + M\|R\|_{H^1(0,1)}\|W\|_V \\ &\leq C_P\|f\|_{L^2(0,1)}\|W\|_V + M\|R\|_{H^1(0,1)}\|W\|_V \end{aligned}$$

so that the continuity constant of F is $\Lambda := C_P\|f\|_{L^2(0,1)} + M\|R\|_{H^1(0,1)}$.

Since the assumptions (4.28) of the Lax-Milgram Lemma are all satisfied, we conclude that problem (W) admits a unique solution $u^0 \in V$, such that

$$\|u^0\|_V \leq \frac{\Lambda}{\beta}.$$

Thus, using triangular inequality, we get the following a-priori estimate for the solution u

$$\begin{aligned} \|u\|_V &= \|u^0 + R\|_V \leq \|u^0\|_V + \|R\|_{H^1(0,1)} \\ &\leq \frac{\Lambda}{\beta} + \|R\|_{H^1(0,1)} \leq \frac{1}{\mu}\left(C_P\|f\|_{L^2(0,1)} + (1+\mu)\|R\|_{H^1(0,1)}\right). \end{aligned}$$

(c) Ceà's Lemma tells us that

$$\|u^0 - u_h^0\|_V \leq C_{\mathscr{T}_h}\frac{M}{\beta}h^l\|u^0\|_{H^{l+1}(\Omega)} \tag{10.2}$$

where $C_{\mathscr{T}_h}$ is a positive constant, independent of h, and related to mesh regularity, while

$$l := \min\{r, s-1\}$$

is the so-called *regularity threshold*.

Let us comment the dependence of the error on the relation between r and s. For a given value of s, $s \geq 2$, the optimal value of the degree of GFE approximation is $r = s-1$. In such a case, the discretization error tends to zero, as h tends to zero, as h^r and there is an optimal balance between regularity of the solution and choice of the approximation space (see Rem. 5.2.7 and Tab. 5.1). In particular, estimate (10.2) tells us that there is no convenience to increase accuracy in using $r \geq s$, i.e., resorting to higher-order polynomials, but it is better to refine the mesh size (i.e., for $r = s-1$, take a smaller value of h).

Let us now comment the dependence of the error on the ratio $\sigma/\mu := \mathscr{R}$. We have

$$\frac{M}{\beta} = \frac{\mu + \sigma C_P^2}{\mu} = 1 + C_P^2 \mathscr{R}.$$

In the case of the interval $(0,1)$, we have $C_P = 1$, so that (10.2) yields

$$\|u^0 - u_h^0\|_V \leq C_{\mathscr{T}_h}(1+\mathscr{R}) h^l \|u^0\|_{H^{l+1}(\Omega)}. \tag{10.3}$$

The case $\mathscr{R} \ll 1$: in this situation, problem (P) is heavily *diffusion-dominated* and the choice of h to obtain a small discretization error is only determined by the regularity of the solution. To make an example, assume $s = r+1$ for every value of the polynomial degree of the GFE approximation. Assume also that $\mathscr{T}_h$ is uniform, so that it is reasonable to have $C_{\mathscr{T}_h} = \mathscr{O}(1)$. In this case, to get $\|u^0 - u_h^0\|_V \simeq \varepsilon$, for a given tolerance ε (small), we need to take

$$h \simeq \left(\frac{\varepsilon}{\|u^0\|_{H^{r+1}(\Omega)}} \right)^{1/r}. \tag{10.4}$$

The case $\mathscr{R} \gg 1$: in this situation, problem (P) is heavily *reaction-dominated* and the choice of h is now determined in a more substantial manner by the nature of the continuous problem itself. As a matter of fact, under the same assumptions as in the previous analyzed case, to get $\|u^0 - u_h^0\|_V \simeq \varepsilon$, for a

given tolerance ε (small), we need to take

$$h \simeq \left(\frac{\varepsilon}{\mathscr{R} \|u^0\|_{H^{r+1}(\Omega)}} \right)^{1/r}. \tag{10.5}$$

The value of h predicted by (10.5) is smaller than that predicted by (10.4) by a factor of $\mathscr{R}^{-1/r}$. If $r = 1$ and $\mathscr{R} = 10^4$, this means that the mesh size in the reaction-dominated case has to be ten thousands smaller than in the diffusion-dominated case, with a very negative impact in the computational effort associated with the GFEM.

10.1.3 Solution of Exercise 3

(a)

Matlab coding. The following `Matlab` script is used to solve Exercise 3(a).

```
%%%%%%%%%%%%%%
% Ex. 3 (a)
%%%%%%%%%%%%%%
close all; clear all
global m
m = 1e-1;
% nr. of elements
N = [10 20 40 80 160];
% degree of FEs
deg = 1;
% initialize errors
El2 = [];
Eh1 = [];

for ii=1:length(N)
% uniform mesh nodes
xnod = [0 : 1/(deg*N(ii)) : 1];
% mesh
griglia = struttura_griglia(xnod,deg);
% local basis functions
base = struttura_base(deg);
% system assembling phase
K = termine_diffusione(griglia,base,'coeff_mu');
S = termine_reazione(griglia,base,'coeff_s');
bv = termine_noto(griglia,base,'coeff_f');
% boundary con ditions
dati_bordo = struct( ...
  'bc'    , [1 1] , ... % type of condition
  'gamma' , [] , ... % parameter for Robin bc
  'r'     , [] ...   % boundary datum
);
A = K+S;
b = bv;
```

```
% Dirichlet bcs and matrix system partitioning
u_incognite = [2 : 1 : griglia.dim-1];
u_note      = [1; griglia.dim];
A11 = A(u_incognite,u_incognite);
A12 = A(u_incognite,u_note     );
A21 = A(u_note     ,u_incognite);
A22 = A(u_note     ,u_note     );
b1  = b(u_incognite);
x2  = [u_ex(0); u_ex(1)];  % boundary values for u_h
% system solution
x1 = A11\(b1-A12*x2);
% inclusion of boundary values in the computed solution
x(u_incognite,1) = x1;
x(u_note     ,1) = x2;
% post-processing/plot
[x_plot y_plot] = visualizza_soluzione(griglia,base,x,20);
plot(x_plot,u_ex(x_plot),'m')
% errors
[eL2,eH1] = norme_errore(griglia,base,x,'u_ex','grad_u_ex');
disp(['L2 Error : ',num2str(eL2)]);
disp(['H1 Error : ',num2str(eH1)]);
El2 = [El2, eL2];
Eh1 = [Eh1, eH1];
end
% asymptotic convergence orders
ph1 = log(Eh1(1:end-1)./Eh1(2:end))/log(2)
pl2 = log(El2(1:end-1)./El2(2:end))/log(2)
```

Matlab coding. The following `Matlab` functions called by the previous script are listed below.

```
%%%%%%%%%%%%%%%%%%%%%%
% Ex. 3: coefficients
%%%%%%%%%%%%%%%%%%%%%%
function mu = coeff_mu(x)
% diffusion coefficient
global m
mu = m*ones(size(x));
return
%%%%%%%%%%%%%%%%%%%%%%
function s = coeff_s(x)
% reaction coefficient
s = ones(size(x));
return
%%%%%%%%%%%%%%%%%%%%%%
function f = coeff_f(x)
% source term
f = ones(size(x));
return
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Ex. 3: solution/gradient of solution
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function u = u_ex(x)
% exact solution
global m
a = sqrt(1/m);
```

```
C = (exp(a)-exp(-a))^(-1);
u = 1 + C*(exp(-a*x)-exp(a*x));
return
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function gradu = grad_u_ex(x)
% gradient of exact solution
global m
a = sqrt(1/m);
C = (exp(a)-exp(-a))^(-1);
gradu = -a*C*(exp(-a*x)+exp(a*x));
return
```

Running the `Matlab` codes shown before, we obtain the errors listed in Tab. 10.1. Errors in the H^1-norm are indicated by `eh1` and errors in L^2-norm are indicated by `el2`. Subscripts refer to the polynomial degree r that is used. The orders of convergence, estimated using (1.9), are listed in Tab. 10.2. All the obtained results are in complete agreement with the conclusions drawn at point (c) of Exercise 2. As a matter of fact, the present case corresponds to a diffusion-dominated problem $\mathscr{R} = 1$, so that, since $s = +\infty$, we expect a convergence of order h^r in H^1 and h^{r+1} in L^2, respectively.

N	$\texttt{eh1}_1$	$\texttt{eh1}_2$	$\texttt{eh1}_3$	$\texttt{el2}_1$	$\texttt{el2}_2$	$\texttt{el2}_3$
10	0.1132	$4.7197e-03$	$1.2319e-04$	$2.9958e-03$	$7.2713e-05$	$1.2967e-06$
20	0.0568	$1.1851e-03$	$1.5478e-05$	$7.5190e-04$	$9.1398e-06$	$8.1548e-08$
40	0.0284	$2.9661e-04$	$1.9373e-06$	$1.8816e-04$	$1.1441e-06$	$5.1047e-09$
80	0.0142	$7.4173e-05$	$2.4224e-07$	$4.7052e-05$	$1.4306e-07$	$3.1917e-10$
160	0.0071	$1.8544e-05$	$3.0282e-08$	$1.1764e-05$	$1.7884e-08$	$1.9950e-11$

Table 10.1: H^1-norm of the error (denoted by `eh1`) and L^2-norm of the error (denoted by `el2`) as functions of $h = 1/N$ and r.

(b)

Matlab coding. The following `Matlab` script solves Exercise 3(b).

```
%%%%%%%%%%%%%
% Ex. 3 (b)
%%%%%%%%%%%%%
close all; clear all
global m

setfonts
```

N	p_1	p_2	p_3	q_1	q_2	q_3
20	0.9954	1.9937	2.9926	1.9943	2.9920	3.9911
40	0.9989	1.9984	2.9981	1.9986	2.9980	3.9978
80	0.9997	1.9996	2.9995	1.9996	2.9995	3.9994
160	0.9999	1.9999	2.9999	1.9999	2.9999	3.9998

Table 10.2: Estimated orders of convergence in the H^1 and L^2 norms (denoted by p and q, resp.) as functions of $h = 1/N$ and r.

```
m = 1e-4;
% nr. of elements
N = 10;
% degree of FEs
deg = 1;
% mesh size
h = 1/N;
% uniform mesh nodes
xnod = [0 : h: 1];
% mesh
griglia = struttura_griglia(xnod,deg);
% local basis functions
base = struttura_base(deg);
% system assembling phase
K = termine_diffusione(griglia,base,'coeff_mu');
S = termine_reazione(griglia,base,'coeff_s');
bv = termine_noto(griglia,base,'coeff_f');
% boundary con ditions
dati_bordo = struct( ...
  'bc'    , [1 1] , ... % type of condition
  'gamma' , [] , ... % parameter for Robin bc
  'r'     , [] ...   % boundary datum
);
A = K+S;
b = bv;
% Dirichlet bcs and matrix system partitioning
u_incognite = [2 : 1 : griglia.dim-1];
u_note      = [1; griglia.dim];
A11 = A(u_incognite,u_incognite);
A12 = A(u_incognite,u_note     );
A21 = A(u_note     ,u_incognite);
A22 = A(u_note     ,u_note     );
b1  = b(u_incognite);
x2  = [u_ex(0); u_ex(1)];  % boundary values for u_h
% system solution
x1 = A11\(b1-A12*x2);
% inclusion of boundary values in the computed solution
x(u_incognite,1) = x1;
x(u_note     ,1) = x2;
% post-processing/plot
XX  = [0:0.001:1]';
Uex = u_ex(XX);
```

```
plot(XX,Uex,'k-',xnod',x,'o--');
legend('exact solution','FE solution')
xlabel('x')
% errors
[eL2,eH1] = norme_errore(griglia,base,x,'u_ex','grad_u_ex');
emax      = norm(u_ex(xnod)'-x,'inf');
Pe_loc    = h^2/(6*m);
disp(['L2 Errore : ',num2str(eL2)])
disp(['H1 Errore : ',num2str(eH1)])
disp(['Err max   : ',num2str(emax)])
disp(['Peclet    : ',num2str(Pe_loc)])
```

Running the `Matlab` code shown before, we obtain the solution shown in Fig. 10.6. The BVP is in this case strongly reaction-dominated, and the exact solution tends to behave as the solution of the "reduced" problem, corresponding to setting $\mu = 0$, that is given by the function $u_{red}(x) = 1$. However, the boundary condition in $x = 1$ forces the solution to be equal to zero, so that a steep boundary layer arises. The function u_h exhibits marked spurious oscillations in the neighbourhood of $x = 1$, and does not satisfy the DMP. This is due to the fact that the local Peclèt number is

$$\mathbb{P}e = \frac{\sigma h^2}{6\mu} = 16.6667 \gg 1.$$

The L^2 error is equal to 0.12698 while the maximum nodal error is 0.2415 and clearly occurs in correspondance of the node located immediately before $x = 1$.

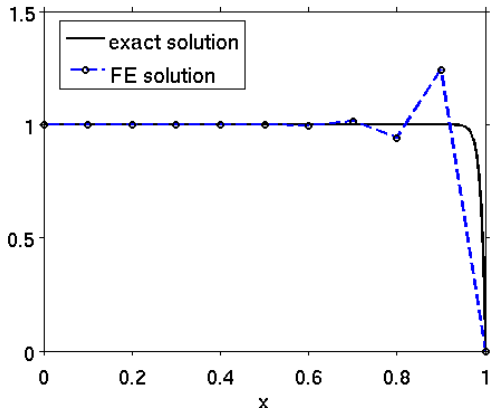


Figure 10.6: Computed solution and exact solution.

(c)

Matlab coding. The following `Matlab` script solves Exercise 3(c).

```
%%%%%%%%%%%%%
% Ex. 3 (c)
%%%%%%%%%%%%%
close all; clear all
global m

setfonts
m = 1e-4;
% nr. of elements
N = 10;
% degree of FEs
deg = 1;
% mesh size
h = 1/N;
% uniform mesh nodes
xnod = [0 : h: 1];
% mesh
griglia = struttura_griglia(xnod,deg);
% local basis functions
base = struttura_base(deg);
% system assembling phase
K = termine_diffusione(griglia,base,'coeff_mu');
S = termine_reazione_lumping(griglia,base,'coeff_s');
bv = termine_noto(griglia,base,'coeff_f');
% boundary con ditions
dati_bordo = struct( ...
  'bc'    , [1 1] , ... % type of condition
  'gamma' , [] , ... % parameter for Robin bc
  'r'     , [] ...   % boundary datum
);
A = K+S;
b = bv;
% Dirichlet bcs and matrix system partitioning
u_incognite = [2 : 1 : griglia.dim-1];
u_note      = [1; griglia.dim];
A11 = A(u_incognite,u_incognite);
A12 = A(u_incognite,u_note     );
A21 = A(u_note     ,u_incognite);
A22 = A(u_note     ,u_note     );
b1  = b(u_incognite);
x2  = [u_ex(0); u_ex(1)];  % boundary values for u_h
% system solution
x1 = A11\(b1-A12*x2);
% inclusion of boundary values in the computed solution
x(u_incognite,1) = x1;
x(u_note     ,1) = x2;
% post-processing/plot
XX  = [0:0.001:1]';
Uex = u_ex(XX);
plot(XX,Uex,'k-',xnod',x,'o--');
legend('exact solution','FE solution')
xlabel('x')
% errors
```

```
[eL2,eH1] = norme_errore(griglia,base,x,'u_ex','grad_u_ex');
emax      = norm(u_ex(xnod)'-x,'inf');
disp(['L2 Errore : ',num2str(eL2)])
disp(['H1 Errore : ',num2str(eH1)])
disp(['Err max   : ',num2str(emax)])
```

Running the `Matlab` code shown before, we obtain the solution shown in Fig. 10.7. In this case, using the lumping stabilization procedure, the computed solution u_h is no longer affected by numerical instabilities, and satisfies the DMP. The L^2 error is equal to 0.14244 while the maximum nodal error is 0.0097595, considerably smaller than in the non stabilized case.

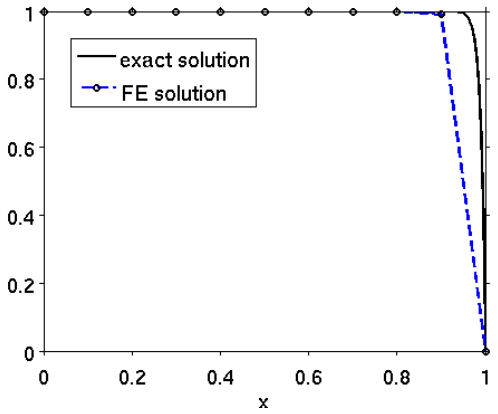


Figure 10.7: Computed solution and exact solution.

(d)

Matlab coding. The following `Matlab` script solves Exercise 3(d).

```
%%%%%%%%%%%%%
% Ex. 3 (d)
%%%%%%%%%%%%%
close all; clear all
global m

setfonts
m = 1e-4;
% compute the number of elements needed to ensure Peclet <1
Nmin = round((1/sqrt(6*m)))
% nr. of elements
N = Nmin;
% degree of FEs
deg = 1;
% mesh size
```

```
h = 1/N;
% uniform mesh nodes
xnod = [0 : h: 1];
% mesh
griglia = struttura_griglia(xnod,deg);
% local basis functions
base = struttura_base(deg);
% system assembling phase
K = termine_diffusione(griglia,base,'coeff_mu');
S = termine_reazione(griglia,base,'coeff_s');
bv = termine_noto(griglia,base,'coeff_f');
% boundary con ditions
dati_bordo = struct( ...
  'bc'    , [1 1] , ... % type of condition
  'gamma' , [] , ... % parameter for Robin bc
  'r'     , [] ...   % boundary datum
);
A = K+S;
b = bv;
% Dirichlet bcs and matrix system partitioning
u_incognite = [2 : 1 : griglia.dim-1];
u_note      = [1; griglia.dim];
A11 = A(u_incognite,u_incognite);
A12 = A(u_incognite,u_note     );
A21 = A(u_note     ,u_incognite);
A22 = A(u_note     ,u_note     );
b1  = b(u_incognite);
x2  = [u_ex(0); u_ex(1)];  % boundary values for u_h
% system solution
x1 = A11\(b1-A12*x2);
% inclusion of boundary values in the computed solution
x(u_incognite,1) = x1;
x(u_note     ,1) = x2;
% post-processing/plot
XX  = [0:0.001:1]';
Uex = u_ex(XX);
plot(XX,Uex,'k-',xnod',x,'o--');
legend('exact solution','FE solution')
xlabel('x')
% errors
[eL2,eH1] = norme_errore(griglia,base,x,'u_ex','grad_u_ex');
emax      = norm(u_ex(xnod)'-x,'inf');
Pe_loc    = h^2/(6*m);
disp(['L2 Errore : ',num2str(eL2)])
disp(['H1 Errore : ',num2str(eH1)])
disp(['Err max   : ',num2str(emax)])
disp(['Peclet    : ',num2str(Pe_loc)])
```

Running the `Matlab` code shown before, we obtain the solution shown in Fig. 10.8. In this case, the mesh is more refined than in points (b) and (c) of this exercise, since we have $N_{min} = 41$. The corresponding Peclèt number is 0.99147 (i.e., slightly less than 1), the L^2 error is 0.024579 and the maximum nodal error is 0.085817. With this combination of the parameters, the performance of the GFEM is better than that in the case (c) as far as the

L^2 error is concerned (a factor of ten smaller), while the maximum nodal error is a factor of ten larger. All in all, if computational cost is the major constraint, the method with the use of lumping of reaction matrix is the best compromise, otherwise an alternative choice might be to refine the mesh only close to the boundary layer.

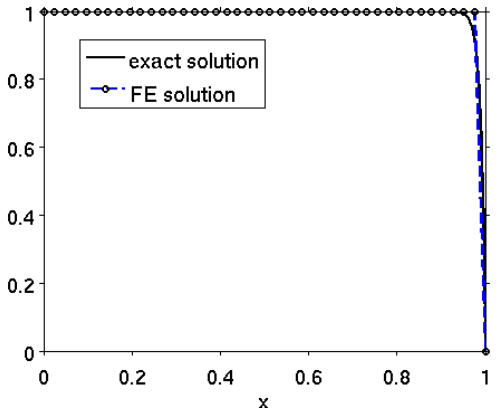


Figure 10.8: Computed solution and exact solution.

Part VI

Appendices

This part contains three appendices, devoted to a review of essential aspects of Linear Algebra, Functional Analysis and Differential Calculus, that are extensively used in the text.

Appendix A

Linear Algebra

In this appendix, we gather some basic notions of Linear Algebra, that are extensively used throughout the text.

A.1 Vector spaces

We start with some basic definitions.

Definition A.1.1 (Vector space)**.** *A vector (or linear) space V with respect to the field Λ ($\Lambda = \mathbb{R}$ or $\Lambda = \mathbb{C}$) is a nonempty set containing elements (vectors) for which the following operations are defined: the sum and the multiplication by a scalar $\lambda \in \Lambda$.*

Definition A.1.2 (Linear independence in V)**.** *A system of vectors $\{v_1, v_2, \ldots, v_n\} \subset V$ is linearly independent if the following condition is satisfied*

$$\sum_{i=1}^{n} \lambda_i v_i = 0 \Leftrightarrow \lambda_1 = \lambda_2 = \ldots = \lambda_n = 0, \qquad \lambda_i \in \Lambda, i = 1, \ldots, n.$$

Definition A.1.3 (Basis in V)**.** *Any set of linearly independent vectors u_i, $i = 1, \ldots, n$, defines a basis in V if any given element $v \in V$ can be written as*

$$v = \sum_{i=1}^{n} v_i u_i.$$

The quantities $v_i \in \Lambda$ are called the coordinates of v with respect to the basis $\{u_i\}$.

Theorem A.1.4 (Theorem of dimension)**.** *Let V be a vector space for which a basis of n vectors exists. Then, any system of linearly independent vectors in V has, at most, n elements, and the same holds for any other basis in V. In such a case the dimension of V is equal to n, that is*

$$\dim(V) = n.$$

If, instead, for every n there exists always a system of n linearly independent vectors of V, then V has infinite dimension, that is

$$\dim(V) = +\infty.$$

Example A.1.5. $V = \mathbb{R}^n$. We have $\dim(V) = n$ and the canonical basis is represented by the unit vectors

$$\mathbf{e}_i = [0,0,\ldots,1,0,\ldots,0]^T \in \mathbb{R}^n, \qquad i = 1,\ldots,n$$

where the only non-zero component of $\mathbf{e}_i$ occupies the i-th position in $\mathbf{e}_i$, that is, $(\mathbf{e}_i)_j = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

is the Kronecker symbol. On the contrary, $V = C^0([a,b])$, the space of continuous function over the interval $[a,b]$, has infinite dimension. To see this, it suffices to consider the Fourier series expansion of a periodic function over $[0,2\pi]$.

A.2 Vector and matrix norms

Given a vector space V, it is useful to introduce a mathematical instrument to give a quantitaive measure of any element of V.

Definition A.2.1 (Norm)**.** *Let v be an arbitrary element of the vector space V. The "norm" is a real functional $\|\cdot\|_V : V \to \mathbb{R}$ satisfying the following properties $\forall v, w \in V$ and $\forall \lambda \in \Lambda$:*

1. $\|v\|_V \geq 0$ *and* $\|v\|_V = 0$ *iff* $v = 0$;
2. $\|\lambda v\|_V = |\lambda| \|v\|_V$;
3. $\|v+w\|_V \leq \|v\|_V + \|w\|_V$ *(triangular inequality).*

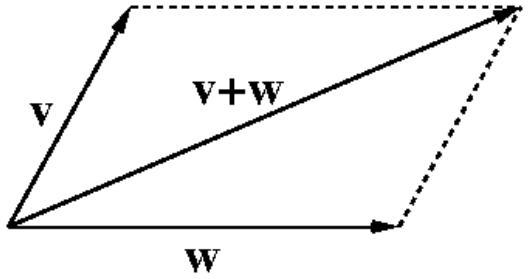


Figure A.1: Triangular inequality.

Example A.2.2 (Vector norms). Let $V = \mathbb{R}^n$. For $p \in [1,\infty)$, the Hölder (or p) norm of a vector $\mathbf{v} \in V$ is defined as

$$\|\mathbf{v}\|_p := \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}. \tag{A.1}$$

- $p = 1$: we obtain the 1-norm of a vector;
- $p = 2$: we obtain the euclidean norm of $\mathbf{v}$ (corresponding to Pythagora's theorem in $\mathbb{R}^n$);
- $p \to +\infty$: we obtain the so-called maximum norm of a vector

$$\|\mathbf{v}\|_\infty = \max_{i \in [1,n]} |v_i|. \tag{A.2}$$

Matlab coding. The `Matlab` source code for computing the norm of a vector is reported below.

```
>> v = [-10, 2, 1]';
>> norm(v,1)

ans =

    13
>> norm(v,2)

ans =

   10.2470
>> norm(v,'inf')

ans =

    10
```

Example A.2.3 (Matrix norms). Let $V = \mathbb{R}^{m\times n}$ be the space of real-valued matrices having m rows and n columns. For $p \in [1,\infty]$, we define the so-called induced p-norm of a matrix $A \in V$ (or natural p-norm of a matrix) as

$$\|A\|_p := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \qquad \mathbf{x} \in \mathbb{R}^n. \tag{A.3}$$

- $p = 1$: we obtain the 1-norm of a matrix, corresponding to taking the column sum of A
$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|;$$
- $p = 2$: we obtain the 2-norm of a matrix (also known as spectral norm)
$$\|A\|_2 = \sqrt{\rho(A^T A)}$$
where A^T is the transpose of A and for any square matrix B of size n
$$\rho(B) := \max_{i=1,\ldots,n} |\lambda_i(B)| \tag{A.4}$$
is the spectral radius of B, λ_i being its eigenvalues;
- $p \to +\infty$: we obtain the so-called maximum norm of a matrix, corresponding to taking the row sum of A
$$\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

Proposition A.2.4. *Let $A \in \mathbb{R}^{n\times n}$ be a given square matrix. Then*

$$\rho(A) \leq \|A\|_p \qquad \forall p \in [1,\infty].$$

Matlab coding. The `Matlab` source code for computing the norm of a matrix is reported below. Notice that the computed value of the spectral radius of A agrees with Prop. A.2.4.

```
>> A = rand(3)

A =

    0.8147    0.9134    0.2785
    0.9058    0.6324    0.5469
```

```
    0.1270    0.0975    0.9575

>> norm(A,1)

ans =

    1.8475

>> norm(A,2)

ans =

    1.8168

>> norm(A,'inf')

ans =

    2.0850

>> max(abs(eig(A)))

ans =

    1.7527
```

A.3 Matrices

We start with an introduction to the basic nomenclature, definitions and properties of a matrix.

Let $A \in \mathbb{R}^{m\times n}$ denote a real-valued matrix with m rows and n columns. Each entry of A is denoted by a_{ij}, $i = 1,\ldots,m$, $j = 1,\ldots,n$. If $m = n$ then A is a square matrix, otherwise it is rectangular. Unless otherwise stated, we assume henceforth $m = n$.

Definition A.3.1 (Regularity of a matrix)**.** *A is invertible (regular) iff $\exists B \in \mathbb{R}^{n\times n}$ such that*

$$AB = BA = I$$

where $I = (\delta_{ij})$ is the identity matrix of order n. B is called the inverse of A and, by definition, we set $B = A^{-1}$. If A is not invertible, then we say that A is singular.

Definition A.3.2 (Rank of a matrix)**.** *The rank of* $\mathbf{A}$ *is the maximum number of linearly independent rows (or columns) of* $\mathbf{A}$*, and is denoted* $\mathrm{rank}(\mathbf{A})$*. Matrix* $\mathbf{A}$ *is said to have maximum (or full) rank if*

$$\mathrm{rank}(\mathbf{A}) = \min\{m,n\}\,.$$

Theorem A.3.3 (Invertibility of a matrix)**.** *A is invertible iff* $\det(A) \neq 0$. *In an equivalent manner, A is invertible iff* $\text{rank}(A) = n$, *that is, iff A has maximum rank. If A is singular, then we have* $\det(A) = 0$ *and, by definition, the p-condition number $K_p(A)$ is equal to $+\infty$.*

Nomenclature and matrix properties:

- A is diagonal if $a_{ij} = 0$ for $i \neq j$;
- A is tridiagonal if $a_{ij} = 0$ for $j > i+1$ and $j < i-1$;
- A is lower triangular if $a_{ij} = 0$ for $j > i$, upper triangular if $a_{ij} = 0$ if $i > j$;
- A is symmetric if $a_{ij} = a_{ji}$, $i, j = 1, \ldots, n$, i.e., if $A = A^T$;
- A is positive definite (p.d.) if the real number $\mathbf{x}^T A\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$, is always > 0 for every $\mathbf{x} \neq \mathbf{0}$ and it is $= 0$ iff $\mathbf{x} = \mathbf{0}$.

Proposition A.3.4. *Let $A = A^T$. Then, A is (symmetric and) positive definite (s.p.d.) iff one of the following equivalent properties is satisfied:*

- $\mathbf{x}^T A\mathbf{x} > 0$ *for all* $\mathbf{x} \neq \mathbf{0}$*;*
- $\lambda_i(A) > 0$, $i = 1, \ldots, n$ *(positive eigenvalues);*
- $\det(A_i) > 0$, $i = 1, \ldots, n$, *where A_i is the submatrix formed by the first i rows and i columns of A (Sylvester criterion);*
- $\exists H \in \mathbb{R}^{n \times n}$ *such that* $A = H^T H$, *with* $\det(H) \neq 0$.

Matlab coding. The `Matlab` sequence of commands to construct A_i is reported below.

```
>> A=magic(3)

A =

     8     1     6
     3     5     7
     4     9     2

>> for i=1:3, A_i=A(1:i,1:i), end

A_i =

     8
```

```
A_i =

     8     1
     3     5


A_i =

     8     1     6
     3     5     7
     4     9     2
```

Remark A.3.5. It can be checked that if A is p.d., then $a_{ii} > 0$, $i = 1, \ldots, n$. To see this, it suffices to take $\mathbf{x} = \mathbf{e}_i$, $i = 1, \ldots, n$.

Remark A.3.6. If A is p.d., then it has positive eigenvalues, and thus, as a consequence, $\det(A) = \Pi_{i=1}^{n} \lambda_i(A) > 0$, so that A is non-singular.

Definition A.3.7 (Diagonally dominant matrices)**.** *A is diagonally dominant by rows if*

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}| \qquad i = 1, \ldots, n$$

while it is diagonally dominant by columns if

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ji}| \qquad i = 1, \ldots, n.$$

Should the $>$ operator hold in the above inequalities in place of $\geq$, then A is said to be strictly diagonally dominant.

Definition A.3.8 (M-matrix property)**.** *A nonsingular matrix A is an M-matrix if $a_{ij} \leq 0$ for $i \neq j$ and if $(A^{-1})_{ij} \geq 0$.*

Definition A.3.9. *Let $\mathbf{x}$ be a given vector in $\mathbb{R}^n$. The inequality*

$$\mathbf{x} \geq (\leq) \mathbf{0}$$

means that $x_i \geq (\leq) 0$ for each $i = 1, \ldots, n$. The same holds if $\mathbf{x}$ is replaced by a matrix $A \in \mathbb{R}^{n \times n}$.

The next result is a direct consequence of Def. A.3.8.

Theorem A.3.10 (Discrete maximum principle)**.** *Let A be an M-matrix such that $A\mathbf{x} \leq \mathbf{0}$, $\mathbf{x} \in \mathbb{R}^n$. Then, $\mathbf{x} \leq \mathbf{0}$.*

Proof. Set $\mathbf{y} := A\mathbf{x}$ and assume that $\mathbf{y} \leq \mathbf{0}$. Since A is an M-matrix, it is invertible and we have

$$x_i = \sum_{j=1}^{n} (A^{-1})_{ij} y_j \leq 0 \qquad i = 1, \ldots, n,$$

because $(A^{-1})_{ij} \geq 0$ and $y_j \leq 0$. Thus, we have proved that $\mathbf{x} \leq \mathbf{0}$. We proceed in a similar manner if $\mathbf{y} \geq \mathbf{0}$. □

Appendix B

Functional Analysis

In this appendix, we review the basic foundations of Functional Analysis that are used thoroughly in the text. We introduce the basic concepts of function spaces, associated topology and norms. We start from the notion of metric space, and we give the definition of complete metric space and of normed space. Then, we define the principal cathegory of metric complete normed spaces, the so-called Banach spaces. We close this brief review by introducing a notable member of the family of Banach spaces, the so-called Hilbert function spaces, and discuss the most relevant examples that will be used in the weak formulation of the boundary value problems considered in this text. Several examples are provided to support the theoretical presentation and definitions.

B.1 Metric spaces

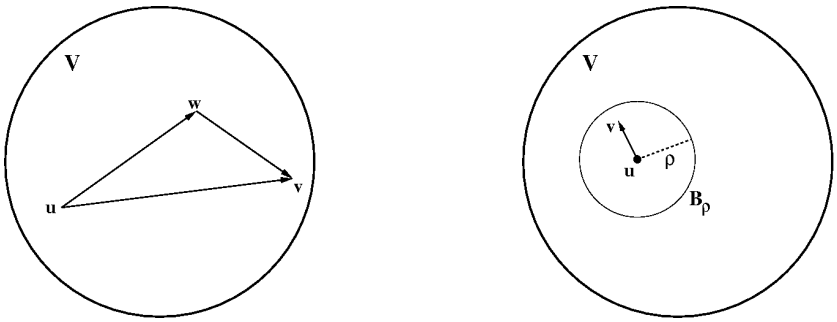
We start by giving the following general definition.

Definition B.1.1 (Metric space)**.** *V is a metric space if, for every pair $u, v \in V$, we can define a functional $\delta(u,v) : V \times V \to \mathbb{R}$, called distance, satisfying the following properties:*

1. $\delta(u,v) \geq 0$ *and* $\delta(u,v) = 0$ *iff* $u = v$;
2. $\delta(u,v) = \delta(v,u)$ *(symmetry);*
3. $\delta(u,v) \leq \delta(u,w) + \delta(v,w)$*, where* $w \in V$ *(triangular inequality).*

Definition B.1.1 introduces a metrics on V, that is, a quantitative manner to "measure" the distance between two functions.

Definition B.1.2 (Ball of radius ρ)**.** *Given $u \in V$ and $\rho \in \mathbb{R}^+$, we define the ball $B_\rho(u)$ as the subset of elements $v \in V$ such that $\delta(u,v) \leq \rho$.*



(a) Triangular inequality (b) Ball in V

Figure B.1: Geometrical representation of a metrics in V.

Example B.1.3. Let $V = \mathbb{R}$. It is immediate to see that the *euclidean metrics*

$$\delta(x,y) := |x-y| \qquad \forall x,y \in V \tag{B.1}$$

satisfies the three requirements in Def. B.1.1.

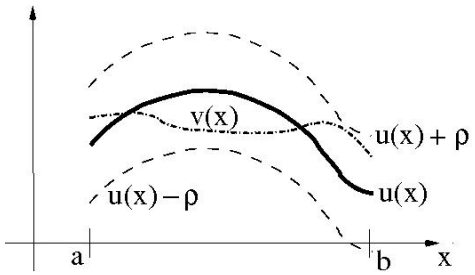


Figure B.2: Ball B_ρ in $V = C^0(\overline{\Omega})$.

Example B.1.4. Let $V = C^0([a,b])$, where $[a,b] \subset \mathbb{R}$. Setting

$$\delta(u,v) := \max_{x\in[a,b]} |u(x)-v(x)| \qquad \forall u,v \in V \tag{B.2}$$

it can be checked that the three requirements in Def. B.1.1 are satisfied. Metrics (B.2) is called Lagrangian of order zero (cf. (B.12) in the case $k = 0$), and the ball B_ρ is represented in Fig. B.2.

Having introduced the notion of metrics, we can define the *limit* in V.

Definition B.1.5 (Limit in V)**.** *Let $\{u_n\}$ be a sequence of functions in a metric space V, and z be a given element of V. We say that*

$$\lim_{n\to\infty} u_n = z \tag{B.3}$$

if the following property holds

$$\lim_{n\to\infty} \delta(u_n, z) = 0. \tag{B.4}$$

In such a case, we say that $\{u_n\}$ converges to z (or that $\{u_n\}$ is δ-convergent to z).

Remark B.1.6. If $V = C^0([a,b])$, Def. B.1.5 is equivalent to the notion of uniform convergence.

Proposition B.1.7 (Necessary condition for δ-convergence)**.** *Assume that* (B.4) *holds (i.e., $\{u_n\}$ is δ-convergent). Then,*

$$\lim_{m,n\to\infty} \delta(u_m, u_n) = 0. \tag{B.5}$$

Proof. Triangular inequality yields

$$\delta(u_m, u_n) \leq \delta(u_m, z) + \delta(u_n, z)$$

from which, using (B.4) for both terms at the right-hand side, we get (B.5). □

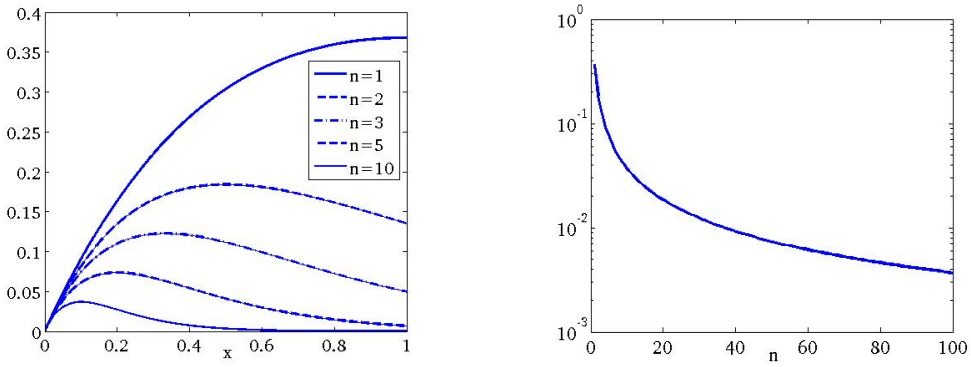
Remark B.1.8. Relation (B.5) is the extension to metric spaces of the necessary Cauchy condition for the existence of the limit for a sequence of numbers.

Example B.1.9. Let $V = C^0([0,1])$ and take $u_n(x) = xe^{-nx}$, $n \geq 1$. Let also $z(x) = 0$ for all $x \in [0,1]$. Then, we have

$$\lim_{n\to\infty} \delta(u_n, z) = \lim_{n\to\infty} \max_{x\in[0,1]} |xe^{-nx} - 0| = \lim_{n\to\infty} \max_{x\in[0,1]} xe^{-nx} = 0,$$

so that u_n converges uniformly to zero in $[0,1]$. Fig. B.3(a) shows a plot of $u_n(x)$ for $n = 1, 2, 3, 5$ and 10, while Fig. B.3(b) shows a logarithmic plot of $\delta(u_n, 0)$ as $n \to \infty$. This latter picture reveals that uniform convergence of u_n to the function $z(x) = 0$ is very slow.

Matlab coding. The `Matlab` script for generating Fig. B.3(a) is reported below.



(a) u_n (b) Maximum error

Figure B.3: Uniform convergence in $C^0([0,1])$.

```
x=[0:0.0001:1];
n=[1,2,3,5,10];
figure;
for i=1:numel(n), un=x.*exp(-n(i)*x); plot(x,un); hold on; end
xlabel('x');
legend('n=1','n=2','n=3','n=5','n=10');
```

Matlab coding. The `Matlab` script for generating Fig. B.3(b) is reported below.

```
x=[0:0.0001:1];
n=[1:100];
for i=1:numel(n), un=x.*exp(-n(i)*x); err(i)=norm(un-0,'inf'); end
figure;
semilogy(n,err);
xlabel('n');
```

B.2 Complete metric spaces

We start with the following definition.

Definition B.2.1 (Cauchy sequence)**.** *$\{u_n\}$ is a Cauchy sequence if*

$$\lim_{n,m\to\infty} \delta(u_n,u_m) = 0. \tag{B.6}$$

Prop. B.1.7 immediately implies that

Proposition B.2.2. *Every convergent sequence in a metric space V is a Cauchy sequence.*

The following example shows that Prop. B.1.7 (equivalently, relation (B.6)) is not, in general, a sufficient condition for $\{u_n\}$ to satisfy (B.4) (i.e. for $\{u_n\}$ to be δ-convergent).

Example B.2.3 (The number *e*.). Let $\mathbb{Q}$ denote the set of rational numbers, and let $u_n := (1+1/n)^n$ for $n \geq 1$. Basic Calculus analysis tells us that the numerical sequence u_n converges, as $n \to \infty$, to Nepero's number $e \simeq 2.718281828459$, i.e.

$$\lim_{n\to\infty} |u_n - e| = 0.$$

Thus, u_n converges to the finite limit e with respect to the euclidean metrics, *but* the limit is *not* a rational number (i.e., it cannot be written in the form p/q, p and q being natural numbers). does not belong to the vector space $\mathbb{Q}$

To remedy a problem like that occurring in Ex. B.2.3, the space V must satisfy the property of *completeness*.

Definition B.2.4 (Completeness)**.** *V is complete metric space if*

$$\lim_{n,m\to\infty} \delta(u_n, u_m) = 0 \qquad \Rightarrow \qquad \exists z \in V \text{ such that } \lim_{n\to\infty} u_n = z. \tag{B.7}$$

Thus, in a complete metric space, Prop. B.1.7 is also sufficient for $\{u_n\}$ to be δ-convergent.

Example B.2.5 ($\mathbb{R}$ as the completion of $\mathbb{Q}$). Taking $V = \mathbb{R}$, instead of $\mathbb{Q}$, in Ex. B.2.3, we conclude that the numerical sequence $u_n = (1+1/n)^n$, $n \geq 1$, is δ-convergent to e with respect to the euclidean metrics (B.1). This shows that the real field is the *completion* of the set of rational numbers with respect to the metrics (B.1).

Proposition B.2.6. *The space $V = C^0([0,1])$ is complete with respect to the lagrangian metrics* (B.2).

A natural question arises about the consequence of a change of metrics on the topological properties of a function space. In the case of $V = C^0([a,b])$, we can introduce the following novel metrics (called *integral metrics*)

$$\delta(u,v) := \int_a^b |u(x) - v(x)|\, dx \equiv \delta_1(u,v). \tag{B.8}$$

Remark B.2.7. The space $V = C^0([a,b])$ is *not* complete with respect to $\delta_1(\cdot,\cdot)$. In other words, there exist sequences of continuous functions $u_n = \{u_n(x)\}$ for which we have

$$\lim_{m,n\to\infty} \delta_1(u_m,u_n) = \lim_{m,n\to\infty} \int_a^b |u_m(x) - u_n(x)|\,dx = 0$$

even if there is no continuous function $z = z(x)$ such that

$$\lim_{n\to\infty} \delta_1(u_n,z) = 0.$$

Example B.2.8. Let $[a,b] = [0,1]$, and

$$u_n(x) = \begin{cases} \dfrac{1}{\sqrt{x}} & \dfrac{1}{n} \le x \le 1 \\ \sqrt{n} & 0 \le x \le \dfrac{1}{n}. \end{cases}$$

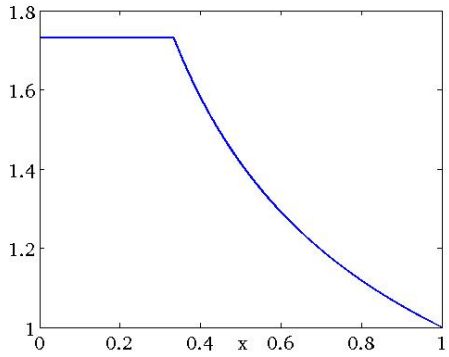


Figure B.4: Plot of $u_n(x)$ in the case $n = 3$.

Each function of the sequence is continuous, and we have

$$\lim_{n\to\infty} \delta_1(u_n, \frac{1}{\sqrt{x}}) = \lim_{n\to\infty} \int_0^{1/n} (\frac{1}{\sqrt{x}} - \sqrt{n})dx$$
$$= \lim_{n\to\infty} \left(2\sqrt{x}\Big|_0^{1/n} - \sqrt{n}\frac{1}{n} \right) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0.$$

This shows that the sequence u_n converges, in the metrics (B.8), to the function $z(x) = 1/\sqrt{x}$, so that, by Def. B.2.1, $\{u_n\}$ is a Cauchy sequence with respect to the integral metrics. However, the limit function z does not belong to $C^0([0,1]$, so that $C^0([0,1]$ is not complete with respect to the integral metrics.

Matlab coding. The `Matlab` script for generating Fig. B.4 is reported below.

```
x=[0:0.0001:1];
figure;
n=3;
un=sqrt(n)*(x<=1/n)+1./sqrt(x).*(x>1/n);
plot(x,un);
xlabel('x');
```

In the previous example, the function z is not defined at $x = 0$. More in general, to treat the case where a function is not defined in a finite number of points, the following definition allows us to extend the notion of continuity and the corresponding metrics.

Definition B.2.9. *Let $v : [a,b] \to \mathbb{R}$ be a measurable function. The space of summable functions over the closed interval $[a,b]$ is defined as*

$$L^1([a,b]) := \left\{ v : [a,b] \to \mathbb{R}, \textit{ such that } \int_a^b |v(x)|\,dx < +\infty \right\}. \tag{B.9}$$

$L^1([a,b])$ is also called the space of Lebesgue integrable functions over the closed interval $[a,b]$.

$L^1([a,b])$ does not contain only continuous functions.

Example B.2.10. Let $v(x) = 1/\sqrt{x}$. We have $v \notin C^0([0,1])$ because $\lim_{x\to 0^+} x^{-1/2} = +\infty$, but $v \in L^1([0,1])$. As a matter of fact

$$\int_0^1 x^{-1/2}\,dx = 2x^{1/2}\Big|_0^1 = 2.$$

Theorem B.2.11 (Fisher-Riesz theorem)**.** *$L^1([a,b])$ is complete with respect to the integral metrics* (B.8)*.*

Obviously, we have $C^0([a,b]) \subset L^1([a,b])$, exactly as $\mathbb{Q} \subset \mathbb{R}$. Moreover, the integral metrics δ_1 is, in general, less fine than the Lagrangian metrics (B.2) in the space $C^0([a,b])$, as shown in the next example.

Example B.2.12. Let $u_n(x) = x^n$ and $z(x) = 0$, $x \in [0,1]$, $n \geq 1$. Then

$$\lim_{n\to\infty} \delta_1(u_n, z) = \lim_{n\to\infty} \int_0^1 |x^n - 0|\,dx = \lim_{n\to\infty} \frac{1}{n+1} = 0,$$

while

$$\max_{x\in[0,1]} |x^n - 0| = 1 \qquad \forall n \geq 1.$$

B.3 Normed spaces

Def. A.2.1 extends to abstract function spaces the classical notion of the euclidean norm of a vector in $\mathbb{R}^d$.

Definition B.3.1 (Normed vector space)**.** *A real vector space V on which a norm $\|\cdot\|_V$ is introduced according to Def. A.2.1, can be made a metric space by setting*

$$\delta(u,v) := \|u-v\|_V \qquad \forall u,v \in V.$$

Again, the above definition is a consistent extension of the way for measuring the distance between two vectors in $\mathbb{R}^d$.

Definition B.3.2 (Equivalence of norms)**.** *Let $\|\cdot\|$ and $|||\cdot|||$ be two norms on V. We say that $\|\cdot\|$ and $|||\cdot|||$ are equivalent, if there exist two positive constants K_1 and K_2, with $K_1 \leq K_2$, such that*

$$K_1|||\cdot||| \leq \|\cdot\| \leq K_2|||\cdot|||. \tag{B.10}$$

B.4 Banach spaces

Gathering the properties of a vector space of being both normed and complete gives us the important cathegory of Banach spaces.

Definition B.4.1 (Banach space)**.** *V is called a Banach space if for every sequence $\{u_n\} \in V$ we have*

$$\lim_{m,n\to\infty} \|u_m - u_n\|_V = 0 \quad \Leftrightarrow \quad \exists z \in V \text{ such that } \lim_{n\to\infty} \|u_n - z\|_V = 0.$$

We notice that the above definition is nothing else the usual notion of completeness of a metric space in the special case where the distance $\delta(u,v)$ is taken as the norm of $u-v$.

Before proceeding, we introduce some notation that is useful in the presentation of function spaces suitable for the analysis of partial differential equations. Let Ω be an open bounded set of $\mathbb{R}^d$, $d \geq 1$, and $\mathbf{x} = (x_1, \ldots, x_d)^T$ be the position vector in Ω. To denote in a synthetic manner partial differentiation with respect to the i-th

coordinate x_i, we introduce the non-negative multi-index $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_d)^T$ such that, for any sufficiently smooth function $v : \Omega \to \mathbb{R}$, we set

$$D^\alpha v := \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$$

where $|\alpha| := \alpha_1 + \alpha_2 + \ldots \alpha_d$ is the length of the vector α. For example, let $d = 2$ and $|\alpha| = 2$. Then, we have the three possible cases: $\alpha = [2,0]^T$, $\alpha = [1,1]^T$ and $\alpha = [0,2]^T$, with which we can associate the following derivatives

$$D^\alpha u = \left\{ \frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \frac{\partial^2 u}{\partial x_2^2} \right\}.$$

Example B.4.2 (The space of continuously differentiable functions of order k). For $k \geq 0$, let us denote by

$$C^k(\overline{\Omega}) = \left\{ v : \overline{\Omega} \to \mathbb{R} : D^\alpha v \in C^0(\overline{\Omega}) \text{ for each } \alpha \text{ such that } 0 \leq |\alpha| \leq k \right\} \quad \text{(B.11)}$$

the space of functions having all derivatives of order $\leq k$ continuous in Ω up to the boundary $\partial\Omega$. Then, V is a Banach space with respect to the norm

$$\|v\|_V := \sum_{j=0}^{k} \sup_{\Omega} \left(\sup_{|\alpha|=j} D^\alpha v \right). \quad \text{(B.12)}$$

For example, in the case $d = 2$ and $|\alpha| = 2$, we have

$$\begin{aligned} \|u\|_{C^2(\overline{\Omega})} &= \sup_{\Omega} |u| + \sup_{\Omega} \left(\sup \left\{ \left| \frac{\partial u}{\partial x_1} \right|, \left| \frac{\partial u}{\partial x_2} \right| \right\} \right) \\ &+ \sup_{\Omega} \left(\sup \left\{ \left| \frac{\partial^2 u}{\partial x_1^2} \right|, \left| \frac{\partial^2 u}{\partial x_1 \partial x_2} \right|, \left| \frac{\partial^2 u}{\partial x_2^2} \right| \right\} \right). \end{aligned}$$

Example B.4.3 (The space of p-summable functions). For $p \in [1, \infty)$, let $V = L^p(\Omega)$ denote the set of measurable functions $u : \Omega \to \mathbb{R}$ of summable power p, that is, such that

$$\int_\Omega |u(\mathbf{x})|^p \, d\Omega < +\infty.$$

Let

$$\|u\|_{L^p} := \left(\int_\Omega |u(\mathbf{x})|^p \, d\Omega \right)^{1/p}, \quad \text{(B.13)}$$

where $d\Omega := (dx_1, \ldots dx_d)^T$ is the infinitesimal volume element centered at point $\mathbf{x} \in \Omega$. Then, L^p is a Banach space with respect to the metrics

$$\delta(u,v) := \|u-v\|_{L^p} = \Big(\int_\Omega |u(\mathbf{x}) - v(\mathbf{x})|^p \, d\Omega \Big)^{1/p}.$$

The special case $p = \infty$ deserves a separate treatment. A function $u : \Omega \to \mathbb{R}$ is said to be *essentially bounded* on Ω if there exists a set Ω' having zero measure such that the restriction of u to $\Omega \setminus \Omega'$ is bounded. The norm on the space $L^\infty(\Omega)$ is defined as follows. Let u be essentially bounded on Ω, and let $[\Omega']$ be the family of sets having zero measure such that the restriction of u over $\Omega \setminus [\Omega']$ is bounded. Setting

$$\mu_{\Omega'} := \sup_{\mathbf{x} \in \Omega \setminus [\Omega']} |u(\mathbf{x})|,$$

we have that $\mu_{\Omega'}$ is a real-valued functional over $[\Omega']$. Then,

$$\|u\|_{L^\infty} := \inf_{[\Omega']} \mu_{\Omega'} \equiv \mathrm{Ess}\sup_{\mathbf{x}\in\Omega} |u(\mathbf{x})|. \tag{B.14}$$

Again, L^∞ is a Banach space with respect to the metrics

$$\delta(u,v) := \|u-v\|_{L^\infty} = \mathrm{Ess}\sup_{\mathbf{x}\in\Omega} |u(\mathbf{x}) - v(\mathbf{x})|.$$

Theorem B.4.4 (Hölder inequality)**.** *Let $f \in L^p(\Omega)$, $g \in L^q(\Omega)$, with $p \in [1,\infty]$ and $q := p/(p-1)$ (conjugate exponent of p). Then, the following* Hölder inequality *holds*

$$\Big| \int_\Omega f\, g \, d\Omega \Big| \le \int_\Omega |f|\,|g|\, d\Omega \le \|f\|_{L^p} \|g\|_{L^q}. \tag{B.15}$$

If $p = q = 2$, (B.15) *becomes*

$$\Big| \int_\Omega f\, g \, d\Omega \Big| \le \int_\Omega |f|\,|g|\, d\Omega \le \|f\|_{L^2} \|g\|_{L^2} \qquad \forall f, g \in L^2(\Omega). \tag{B.16}$$

Relation (B.16) *is well-known as the* Cauchy-Schwarz (CS) inequality.

B.5 Hilbert spaces

Hilbert spaces are certainly the most notable members of the wider family of Banach spaces. This prominence is due to the fact that the definition of norm in a Hilbert space is *induced* by the notion of *scalar product*. This allows to extend

in a natural manner familiar concepts of vector analysis in $\mathbb{R}^d$, in particular, orthogonality among elements of a space and the definition of *energy* of a function. These structural properties make Hilbert spaces the right candidates for being the ambient function spaces for problems arising from physical applications.

Definition B.5.1 (Scalar product)**.** *A real vector space V is called pre-hilbertian if a functional $(\cdot,\cdot)_V : V \times V \to \mathbb{R}$, called* scalar product*, is defined in such a way to satisfy the following properties:*

1. $(u,u)_V \geq 0$ *and* $(u,u)_V = 0$ *iff* $u = 0$*;*
2. $(u,v)_V = (v,u)_V$*, for all* $u, v \in V$*;*
3. $(\lambda u + \mu v, z)_V = \lambda (u,z)_V + \mu (v,z)_V$.

We see that the quantity $\sqrt{(u,u)_V}$ satisfies all the properties required by Def. A.2.1, so that it is admissible to set by definition

$$\|u\|_V := \sqrt{(u,u)_V} \qquad \forall u \in V. \tag{B.17}$$

Thus, a pre-hilbertian function space V is made a normed space through (B.17), showing that the norm is directly induced by the introduction of a scalar product in V. The next step is to ensure that the pre-hilbertian space is also complete with respect to the norm (B.17). Should this happen, then V is called a Hilbert space.

Example B.5.2. Let $V = L^2(\Omega)$. In this case, V is an Hilbert space with respect to the scalar product

$$(u,v)_{L^2} := \int_\Omega u\, v\, d\Omega. \tag{B.18}$$

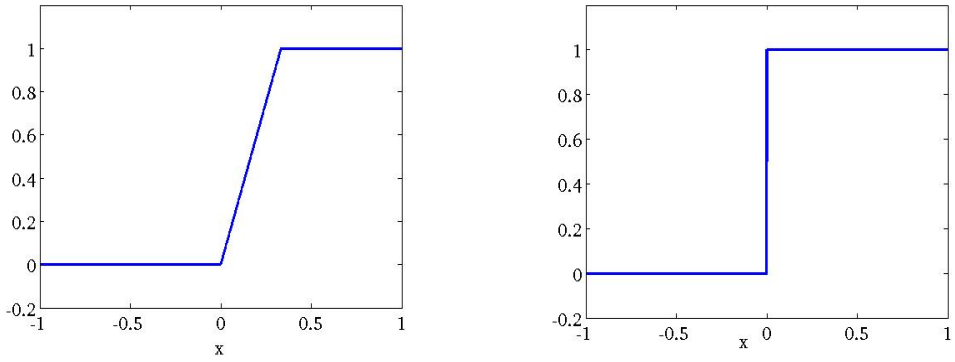
It is important to notice that $C^0(\overline{\Omega})$ is pre-hilbertian with respect to (B.18), but not complete, as the following example shows.

Example B.5.3. Let $\overline{\Omega} = [-1,1]$ and

$$u_n(x) = \begin{cases} 0 & x \leq 0 \\ nx & 0 < x \leq 1/n \\ 1 & x > 1/n. \end{cases}$$

Clearly, $u_n \in C^0([-1,1])$ for all $n \geq 1$. Moreover, with $m < n$ we have

$$\begin{aligned} \|u_m - u_n\|_{L^2}^2 &= \int_0^{1/n} (mx - nx)^2\, dx + \int_{1/n}^{1/m} (1 - mx)^2\, dx \\ &= \frac{1}{3m} + \frac{1}{3n}\left(\frac{m}{n} - 2\right) < \frac{1}{3m} - \frac{1}{3n} < \frac{1}{3m} \end{aligned}$$



(a) u_3 (b) Limit function as $n \to \infty$

Figure B.5: The space $C^0([-1,1])$ is not complete with respect to the metrics of L^2.

so that

$$\lim_{m,n\to\infty} \|u_m - u_n\|_{L^2}^2 = 0.$$

This shows that $\{u_n\}$ is a Cauchy sequence with respect to the metrics associated with (B.17). Let

$$H(x) := \begin{cases} 0 & -1 \le x < 0 \\ 1 & 0 < x \le 1 \end{cases}$$

denote the so-called *Heaviside function*. The function H has a jump discontinuity at $x = 0$ and for this reason it is also known as *step function*. We can check that for $n \to \infty$, the sequence u_n converges to $H(x)$ for almost all $x \in [0,1]$. As a matter of fact, we have

$$\lim_{n\to\infty} \|u_n - H(x)\|_{L^2}^2 = \lim_{n\to\infty} \int_0^{1/n} (nx-1)^2\,dx = \lim_{n\to\infty} \frac{(nx-1)^3}{3n}\Big|_{x=0}^{x=1/n} = \lim_{n\to\infty} \frac{1}{3n} = 0.$$

Assume now that there exists a function $v \in C^0([-1,1])$ such that $\lim_{n\to\infty} \|u_n - v\|_{L^2} = 0$. Then, applying the triangular inequality we get

$$\|v - H(x)\|_{L^2} \le \|v - u_n\|_{L^2} + \|u_n - H(x)\|_{L^2}$$

from which, passing to the limit

$$\lim_{n\to\infty} \|v - H(x)\|_{L^2} \le \lim_{n\to\infty} \|v - u_n\|_{L^2} + \lim_{n\to\infty} \|u_n - H(x)\|_{L^2} = 0 + 0 = 0$$

that implies $v(x) = H(x) \notin C^0([-1,1])$, in contradiction with the starting assumption $v \in C^0([-1,1])$. We have thus proved that u_n is a convergent sequence in the space $C^0([-1,1])$, convergence being measured with respect to the metrics of L^2, even if there is no continuous function $z = z(x)$ such that

$$\lim_{n\to\infty} \|u_n - z\|_{L^2} = 0.$$

The limit function z is the discontinuous function $H(x)$. This shows that the space $C^0([-1,1])$ is not complete with respect to the metrics of L^2.

Matlab coding. The `Matlab` script for generating Figs. B.5(a) and B.5(b) is reported below.

```
x=[-1:0.0001:1];
figure;
n=3;
un=0.*(x<=0) + n*x.*((x>0) & (x<=1/n)) + 1.*(x>1/n);
figure, plot(x,un);
xlabel('x');
ulimit=0.*(x<=0) + 1.*(x>0);
figure, plot(x,ulimit);
xlabel('x');
```

B.6 The Sobolev space of order m in one spatial dimension

Let $\Omega := (a,b)$ be an open bounded interval. For any given nonnegative integer m, we let

$$H^m(\Omega) = \left\{ v : \Omega \to \mathbb{R} : \frac{\partial^\alpha v}{\partial x^\alpha} \in L^2(\Omega),\ \alpha = 0,1,\ldots,m \right\} \tag{B.19}$$

be the space of functions whose derivative up to the m-th order is square integrable in the Lebesgue sense over Ω. The space $L^2(\Omega)$ corresponds to the special case $m = 0$. The space H^m is known as the *Sobolev space of order m*, and is an Hilbert space upon the following definition of the scalar product

$$(u,v)_{H^m(\Omega)} = \sum_{\alpha=0}^{m} \int_a^b \frac{\partial^\alpha u}{\partial x^\alpha} \frac{\partial^\alpha v}{\partial x^\alpha}\, dx. \tag{B.20}$$

The natural norm on H^m is then

$$\|u\|_{H^m(\Omega)} = \sqrt{(u,u)_{H^m(\Omega)}} = \sqrt{\sum_{\alpha=0}^{m} \left\| \frac{\partial^\alpha u}{\partial x^\alpha} \right\|^2_{L^2(\Omega)}}. \tag{B.21}$$

Theorem B.6.1 (Sobolev imbedding in 1D)**.** *Let $\Omega = (a,b) \subset \mathbb{R}$. Then*

$$H^m(\Omega) \subset C^{m-1}(\overline{\Omega}) \qquad \forall m \geq 1. \tag{B.22}$$

An important member of the family of spaces (B.19) is that corresponding to $m = 1$. In this case, we have

$$(u,v)_{H^1(\Omega)} = \int_a^b uv\,dx + \int_a^b \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,dx, \tag{B.23}$$

and

$$\|u\|_{H^1(\Omega)} = \sqrt{\|u\|^2_{L^2(\Omega)} + \left|\left|\frac{\partial u}{\partial x}\right|\right|^2_{L^2(\Omega)}}. \tag{B.24}$$

Remark B.6.2. Sobolev embedding (B.22) tells us that functions in $H^1(a,b)$ are continuous over $[a,b]$. This property does not hold, in general, when $\Omega \subset \mathbb{R}^d$, $d \geq 2$.

A useful subspace of H^1 that will be used many times in the analysis of boundary value problems is that consisting of all functions belonging to H^1 and vanishing on the boundary $\partial\Omega = \{a, b\}$

$$H^1_0(\Omega) = \left\{ v \in H^1(\Omega) : v(a) = v(b) = 0 \right\}. \tag{B.25}$$

The following important property holds.

Theorem B.6.3 (Poincaré inequality)**.** *Given the bounded interval $\Omega = (a,b)$, there exists a constant $C_P > 0$ depending only on the size of Ω such that*

$$\|v\|_{L^2(\Omega)} \leq C_P \left|\left|\frac{\partial v}{\partial x}\right|\right|_{L^2(\Omega)} \qquad \forall v \in H^1_0(\Omega). \tag{B.26}$$

Proof. Let v denote any function in $H^1_0(\Omega)$. Then, noting that $v(a) = 0$, we have $\forall x \in [a,b]$

$$\begin{aligned} |v(x)| &= |v(x) - v(a)| = \left| \textstyle\int_a^x \frac{\partial v}{\partial t}\,dt \right| \leq \int_a^b \left|\frac{\partial v}{\partial t}\right| |1|\,dt \\ &\leq \left|\left|\frac{\partial v}{\partial x}\right|\right|_{L^2(\Omega)} \|1\|_{L^2(\Omega)} = \sqrt{b-a} \left|\left|\frac{\partial v}{\partial x}\right|\right|_{L^2(\Omega)}, \end{aligned}$$

from which, squaring both sides of the inequality and integrating over (a,b), we get

$$\int_a^b |v(x)|^2\,dx \leq (b-a)^2 \left|\left|\frac{\partial v}{\partial x}\right|\right|^2_{L^2(\Omega)},$$

from which we finally obtain (B.26) with $C_P = b - a$. □

Remark B.6.4 (Equivalent norm in H_0^1). Using (B.26) in (B.24) for each $v \in H_0^1(\Omega)$, we have

$$\|v\|^2_{H^1(\Omega)} \leq (1+C_P^2)\Big|\Big|\frac{\partial v}{\partial x}\Big|\Big|^2_{L^2(\Omega)}.$$

On the other hand, from definition (B.24), we trivially have

$$\|v\|^2_{H^1(\Omega)} \geq \Big|\Big|\frac{\partial v}{\partial x}\Big|\Big|^2_{L^2(\Omega)},$$

so that we can conclude that

$$\|\frac{\partial v}{\partial x}\|^2_{L^2(\Omega)} \leq \|v\|^2_{H^1(\Omega)} \leq (1+C_P^2)\Big|\Big|\frac{\partial v}{\partial x}\Big|\Big|^2_{L^2(\Omega)} \qquad \forall v \in H_0^1(\Omega). \tag{B.27}$$

Applying Def. B.3.2 to (B.27), we see that

$$\|u\|_{H_0^1(\Omega)} = \Big|\Big|\frac{\partial u}{\partial x}\Big|\Big|_{L^2(\Omega)} \qquad u \in H_0^1(\Omega) \tag{B.28}$$

is an *equivalent norm* on $H_0^1(\Omega)$, with $K_1 = 1$ and $K_2 = (1+C_P^2)^{1/2}$.

Example B.6.5 (Completion of $C^2 \cap C_0^0$ into H_0^1). Let

$$C_0^0([-1,1]) := \big\{ v \in C^0([-1,1]) \,\text{such that}\, v(-1) = v(1) = 0 \big\}$$

and introduce the space

$$W = \big\{ v \in C^2([-1,1]) \,\text{such that}\, v(-1) = v(1) = 0 \big\} = C^2([-1,1]) \cap C_0^0([-1,1]).$$

For $n \geq 1$, let us consider the following sequence of functions in W

$$u_n(x) = \begin{cases} 1 - \dfrac{1}{n} + \dfrac{2}{n\pi}\cos\Big(\dfrac{\pi n x}{2}\Big) & |x| < \dfrac{1}{n} \\ 1 - |x| & \dfrac{1}{n} \leq |x| \leq 1 \end{cases}$$

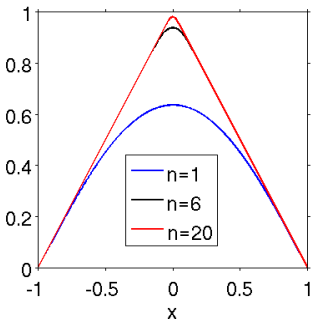
whose first and second derivatives are given by

$$u_n'(x) = \begin{cases} 1 & -1 \leq x \leq -\frac{1}{n} \\ -\sin\Big(\dfrac{\pi n x}{2}\Big) & |x| < \dfrac{1}{n} \\ -1 & \dfrac{1}{n} \leq x \leq 1 \end{cases}$$

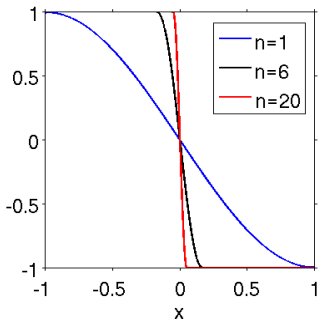
and

$$u_n''(x) = \begin{cases} 0 & -1 \le x \le -\frac{1}{n} \\ -\frac{\pi n}{2}\cos\left(\frac{\pi n x}{2}\right) & |x| < \frac{1}{n} \\ 0 & \frac{1}{n} \le x \le 1. \end{cases}$$

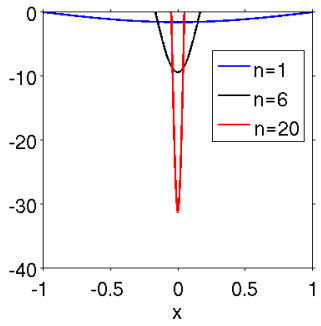
A plot of u_n, u_n' and u_n'' for $n = 1$, $n = 6$ and $n = 20$, is shown in Fig. B.6. Proceeding as in Ex. B.5.2, we can check that



(a) u_n



(b) u_n'



(c) u_n''

Figure B.6: A sequence of functions belonging to $C^2 \cap C_0^0$ whose limit is a function in H_0^1.

$$\lim_{m,n\to\infty} \|u_m - u_n\|_{H_0^1(-1,1)} = 0.$$

This shows that $\{u_n\}$ is a Cauchy sequence with respect to the metrics associated with (B.28). We can also see that for $n \to \infty$, the sequence u_n converges to $z(x) :=$

$1-|x| \in C_0^0([-1,1])$, i.e.

$$\lim_{n\to\infty} \|u_n - z(x)\|_{H_0^1(-1,1)} = 0.$$

Then, proceeding by contradiction, we show that there is no function $v \in W$ such that

$$\lim_{n\to\infty} \|u_n - v\|_{H_0^1(-1,1)} = 0.$$

The only limit function v is the continuous function $z = 1-|x|$. This shows that the space $C^2([-1,1]) \cap C_0^0([-1,1])$ is not complete with respect to the metrics of $H_0^1(-1,1)$. The limit function z belongs to $H_0^1(-1,1) \equiv V$ because $z \in L^2(-1,1)$ and the derivative z' is the Heaviside-like step function

$$z'(x) = \begin{cases} 1 & -1 \le x \le 0 \\ -1 & 0 < x \le 1. \end{cases}$$

Taking a closer glance, we see that every member u_n of the sequence (which belongs to W) actually belongs to V. However, V contains functions that do not belong to W, as is the case for the limit z. This fact is graphically documented in Fig. B.7

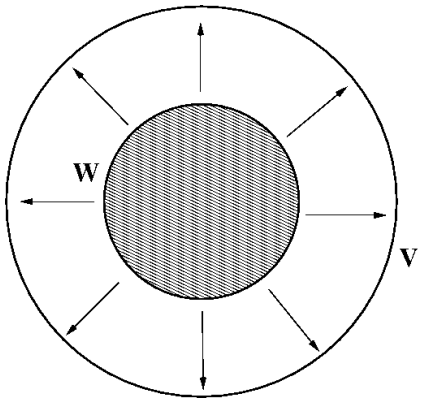


Figure B.7: The process of completion of the space $W = C^2([-1,1]) \cap C_0^0([-1,1])$. Arrows pointing from W to $V = H_0^1(-1,1)$ represent the limit for $n \to \infty$.

Matlab coding. The `Matlab` script for generating Fig. B.6 is reported below.

```
x=[-1:0.0001:1];
n=[1,6,20];
```

```
for i=1:numel(n)
  un_0(i,:) = (1-1/n(i)+2/(n(i)*pi).*cos(pi*n(i)*x/2)).*(abs(x)<1/n(i))+...
              (1-abs(x)).*((abs(x)>=1/n(i))&(abs(x)<=1));
  un_1(:,i) = 1.*((x>=-1)&(x<=-1/n(i)))+(-sin(n(i)*pi*x/2)).*...
              (abs(x)<1/n(i))+(-1).*(x>=1/n(i));
  un_2(:,i) = 0.*((x>=-1)&(x<=-1/n(i)))+ 0.*((x>=1/n(i))&(x<=1)) ...
            -pi/2*n(i)*cos(pi*n(i)*x/2).*(abs(x)<1/n(i));
end
figure, plot(x,un_0);axis('square')
xlabel('x');
legend('n=1','n=6','n=20');
figure, plot(x,un_1);axis('square')
xlabel('x');
legend('n=1','n=6','n=20');
figure, plot(x,un_2);axis('square')
xlabel('x');
legend('n=1','n=6','n=20');
```

B.7 The Sobolev space of order m in $\mathbb{R}^d$, $d \geq 2$

In the study of elliptic boundary value problems, it is important to consider the more general (and realistic) situation where the domain Ω is a bounded open set of $\mathbb{R}^d$, $d \geq 2$, whose boundary $\partial\Omega$ is Lipschitz continuous. This latter property means that a unit outward normal vector $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is defined for almost every $\mathbf{x} \in \partial\Omega$.

Then, analogously to (B.19), for $m \geq 0$ we set

$$H^m(\Omega) = \left\{ v : \Omega \to \mathbb{R} : D^\alpha v \in L^2(\Omega) \text{ for each } \alpha \text{ such that } |\alpha| \leq m \right\}. \quad \text{(B.29)}$$

The scalar product is

$$(u,v)_{H^m(\Omega)} = \sum_{|\alpha|=0}^{m} \int_\Omega D^\alpha u \cdot D^\alpha v \, d\Omega \quad \text{(B.30)}$$

and the corresponding norm is

$$\|u\|_{H^m(\Omega)} = \sqrt{(u,u)_{H^m(\Omega)}} = \sqrt{\sum_{|\alpha|=0}^{m} \|D^\alpha v\|^2_{L^2(\Omega)}}. \quad \text{(B.31)}$$

Again, the space $L^2(\Omega)$ corresponds to the case $m = 0$, while in the case $m = 1$, we have the space $H^1(\Omega)$

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) \text{ such that } \frac{\partial v}{\partial x_i} \in L^2(\Omega),\ i = 1, \ldots, d \right\}. \quad \text{(B.32)}$$

Remark B.7.1. In the applications, the quantity $\int_\Omega |\nabla v|^2 d\Omega$ has usually the meaning of an energy. For this reason, it is customary to say that a function v belonging to $H^1(\Omega)$ has *finite energy*.

The following example shows that functions in $H^1(\Omega)$ are not necessarily continuous when $d > 1$, as anticipated in Rem. B.6.2.

Example B.7.2. Let

$$\Omega = \left\{ (x_1, x_2) : \rho \equiv \sqrt{x_1^2 + x_2^2} \leq \frac{1}{2} \right\},$$

and $u(x_1, x_2) = (-\log(\rho))^\beta$, $\beta \in (0, 1/2)$. We have $\lim_{\rho \to 0^+} = +\infty$ despite the fact that $u \in H^1(\Omega)$.

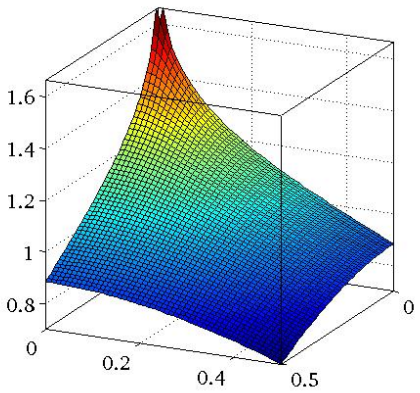


Figure B.8: A function with logarithmic singularity at the origin, which belongs to $H^1(\Omega)$ (in this case, $\beta = 1/3$).

Matlab coding. The `Matlab` script for generating Fig. B.8 is reported below.

```
beta = 1/3;
[X,Y] = meshgrid(0:0.001:1/2, 0:0.001:1/2);
rho=sqrt(X.^2+Y.^2);
Z=(-log(rho)).^(beta);
surf(X,Y,Z);
```

The following definitions extend to the functions of $H^1(\Omega)$ the notion of value of a function on the boundary of Ω that is valid in the case $d = 1$.

Definition B.7.3 (Trace operator)**.** *Given a function $v \in H^1(\Omega)$, we define the trace operator $\gamma_0 : \partial\Omega \to \mathbb{R}$ as*

$$\gamma_0 v := v|_{\partial\Omega}.$$

Definition B.7.4 (Trace space)**.** *The space of traces of all functions of $H^1(\Omega)$ is denoted $H^{1/2}(\partial\Omega)$ and is defined as*

$$H^{1/2}(\partial\Omega) = \left\{ \gamma_0 v, v \in H^1(\Omega) \right\}.$$

The space $H^{1/2}(\partial\Omega)$ is an Hilbert space by endowing it with the norm

$$\|g\|_{H^{1/2}(\partial\Omega)} := \inf_{\substack{v \in H^1(\Omega) \\ \gamma_0 v = g}} \|v\|_{H^1(\Omega)} \qquad g \in H^{1/2}(\partial\Omega). \tag{B.33}$$

The subset of $H^1(\Omega)$ of functions having vanishing trace on $\partial\Omega$ is

$$H_0^1(\Omega) = \left\{ v \in H^1(\Omega) \text{ such that } v|_{\partial\Omega} = 0 \right\}. \tag{B.34}$$

As in the one-dimensional case, we have the following results.

Theorem B.7.5 (Poincaré inequality in $\mathbb{R}^d$)**.** *There exists a positive constant C_P depending only on the domain Ω such that*

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} \qquad \forall v \in H_0^1(\Omega). \tag{B.35}$$

Theorem B.7.6 (Equivalent norm in H_0^1)**.** *The quantity $\|\nabla v\|_{L^2}$ is an equivalent norm in $H_0^1(\Omega)$, and we set*

$$\|v\|_{H_0^1(\Omega)} = \|\nabla v\|_{L^2(\Omega)} \qquad v \in H_0^1(\Omega). \tag{B.36}$$

It is often useful to relate the norm of a function belonging to $H^1(\Omega)$ to the norm of the restriction of such a function to the boundary $\partial\Omega$. The following results holds.

Theorem B.7.7 (Trace theorem)**.** *Let $v \in H^1(\Omega)$. Then there exists a positive constant $C_{\partial\Omega}$ such that*

$$\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} \leq C_{\partial\Omega} \|v\|_{H^1(\Omega)},$$

where

$$\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} = \left(\int_{\partial\Omega} |v(\mathbf{x})|^2 \, d\Gamma \right)^{1/2}.$$

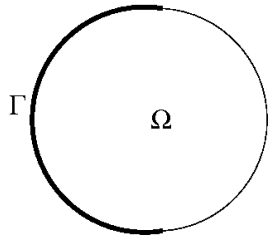


Figure B.9: A two-dimensional domain. $\Gamma \subseteq \partial\Omega$ is the portion of the boundary on which the trace of a function vanishes.

Remark B.7.8 (Poincaré-Friedrichs inequality). The Poincaré inequality and the trace theorem continue to hold if $\partial\Omega$ is replaced with a subset $\Gamma \subseteq \partial\Omega$. In such a case, $H^1_0(\Omega)$ is replaced by the space

$$H^1_{0,\Gamma}(\Omega) := \left\{ v \in H^1(\Omega) \text{ such that } v|_\Gamma = 0 \right\}$$

and the Poincaré inequality takes the name of Poincaré-Friedrichs inequality.

Having defined the space of traces of a scalar function of $H^1(\Omega)$, it is also possible to introduce the space of the traces of a vector function.

Definition B.7.9. *Let $\mathbf{v} : \Omega \to (\mathbb{R})^d$ be a given vector-valued function. A space of frequent use is*

$$H(\mathrm{div};\Omega) := \left\{ \mathbf{v} \in (L^2(\Omega))^d \textit{ such that } \mathrm{div}\mathbf{v} \in L^2(\Omega) \right\}. \tag{B.37}$$

The space $H(\mathrm{div};\Omega)$ is an Hilbert space by endowing it with the norm

$$\|\mathbf{v}\|_{H(\mathrm{div};\Omega)} := \left(\|\mathbf{v}\|^2_{L^2(\Omega)} + \|\mathrm{div}\mathbf{v}\|^2_{L^2(\Omega)} \right)^{1/2}. \tag{B.38}$$

Definitions (B.37) and (B.38) tell us that functions belonging to $H(\mathrm{div};\Omega)$ have a regularity that is intermediate between L^2 and H^1.

Definition B.7.10 (Trace space of vector functions)**.** *The space of traces of all functions of $H(\mathrm{div};\Omega)$ is denoted $H^{-1/2}(\partial\Omega)$ and is defined as*

$$H^{-1/2}(\partial\Omega) = \left\{ \gamma_0\mathbf{v} = (\mathbf{v}\cdot\mathbf{n})|_{\partial\Omega},\, \mathbf{v} \in H(\mathrm{div};\Omega) \right\}.$$

The space $H^{-1/2}(\partial\Omega)$ is an Hilbert space by endowing it with the norm

$$\|h\|_{H^{-1/2}(\partial\Omega)} := \inf_{\substack{\mathbf{v}\in H(\mathrm{div};\Omega) \\ \mathbf{v}\cdot\mathbf{n}=h}} \|\mathbf{v}\|_{H(\mathrm{div};\Omega)} \qquad h \in H^{-1/2}(\partial\Omega). \tag{B.39}$$

Appendix C

Differential Calculus

In this appendix, we gather some basic notions of Differential Calculus and related formulas that are extensively used throughout the text.

C.1 Differential operators, useful formulas and properties

Throughout the text, we deal with differential operators of first and second order, applied to scalar- and vector-valued functions. Here, we summarize the principal of them, with some useful formulas. In the following, Ω is a bounded set of $\mathbb{R}^d$, with Lipschitz boundary $\partial\Omega$ on which a unit normal vector $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is defined almost everywhere, $\mathbf{x} = (x_1, \ldots, x_d)^T$ being the coordinate position vector. We shall mainly deal with the special cases $d = 1$ (one-dimensional case, 1D) and $d = 2$ (two-dimensional case, 2D). This will ease the presentation, and its completion with supporting examples and numerical computations.

C.1.1 First-order operators

- given a differentiable function $u : \Omega \to \mathbf{R}$, we define the gradient of u as the first-order operator transforming u into a vector, $\nabla u : \Omega \to \mathbb{R}^d$, such that

$$\nabla u = \left[\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d} \right]^T;$$

- given a differentiable vector field $\mathbf{v} = [v_1, \ldots, v_d]^T : (\underbrace{\Omega \times \Omega \ldots \times \Omega}_{d \text{ times}}) \to \mathbb{R}^d$, we define the divergence of $\mathbf{v}$ as the operator transforming $\mathbf{v}$ into a scalar, $\operatorname{div} \mathbf{v} : \mathbb{R}^d \to \mathbb{R}$, such that

$$\operatorname{div} \mathbf{v} = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i};$$

C.1.2 Second-order operators

The Laplace operator (shortly, Laplacian) is the second-order differential operator obtained by application of the divergence operator to the vector field ∇u, for a given twice-differentiable function u, such that

$$\triangle u := \operatorname{div} \nabla u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}.$$

More in general, for a given differentiable function

$$A = A(\mathbf{x}) : \Omega \to \mathbb{R},$$

we define the second-order linear differential operator in *divergence form* as

$$\operatorname{div}(A \nabla u) = A \triangle u + \nabla A \cdot \nabla u$$

where the symbol $\cdot$ denotes the scalar product between two vectors.

C.1.3 Green's formula

The following result holds.

Theorem C.1.1 (Green's formula)**.** *For sufficiently smooth given scalar and vector functions v and $\mathbf{w}$, we have*

$$\int_\Omega v \operatorname{div} \mathbf{w} \, d\Omega = -\int_\Omega \nabla v \cdot \mathbf{w} \, d\Omega + \int_{\partial\Omega} v \mathbf{w} \cdot \mathbf{n} \, d\Sigma \tag{C.1}$$

where $\mathbf{n} \, d\Sigma$ denotes the oriented differential area element centered around the point $\mathbf{x}$ over the $d-1$-dimensional manifold $\partial\Omega$.

Identity (C.1) goes also under the name of *integration by parts formula*, and will play a crucial role in the following parts of the text.

C.2 Elliptic operators

Let us consider the second-order linear differential operator

$$Lu := -\mathrm{div}(A\nabla u) + \mathrm{div}(\mathbf{b}u) + cu : \Omega \to \mathbb{R} \tag{C.2}$$

where A, $\mathbf{b}$, c and u are sufficiently regular given functions.

Definition C.2.1 (Ellipticity)**.** *The operator Lu is said to be elliptic in* Ω *if there exists a positive constant* α_0 *such that*

$$\xi^T A(\mathbf{x})\xi \geq \alpha_0 |\xi|^2$$

for every $\xi = [\xi_1, \ldots, \xi_d]^T \in \mathbb{R}^d$ *and for almost every* $\mathbf{x} \in \Omega$.

Remark C.2.2. Def. C.2.1 is satisfied if A is strictly positive almost everywhere (a.e.) in Ω. In the more general case where A is a real-valued $d \times d$ matrix function, then Def. C.2.1 is satisfied if A is positive definite a.e. in Ω.

Definition C.2.3 (Conormal derivative)**.** *The conormal derivative of a function u associated with the elliptic operator L is defined as*

$$\frac{\partial u}{\partial n_L}(\mathbf{x}) := (-A(\mathbf{x})\nabla u(\mathbf{x}) + \mathbf{b}(\mathbf{x})u(\mathbf{x})) \cdot \mathbf{n}(\mathbf{x}) \qquad \textit{for a.e. } \mathbf{x} \in \partial\Omega. \tag{C.3}$$

C.2.1 Maximum principle for 2nd order elliptic operators

We assume that $c(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \overline{\Omega}$.

Definition C.2.4 (Inverse-monotonicity)**.** *Let* $w \in C^2(\Omega) \cap C^0(\overline{\Omega})$. *We say that L is inverse-monotone if the inequalities*

$$Lw(\mathbf{x}) \geq 0 \qquad \textit{for all } \mathbf{x} \in \Omega \qquad \textit{and } w(\mathbf{x}) \geq 0 \qquad \textit{for all } \mathbf{x} \in \partial\Omega \tag{C.4}$$

together imply that

$$w(\mathbf{x}) \geq 0 \qquad \textit{for all } \mathbf{x} \in \overline{\Omega}. \tag{C.5}$$

An important consequence of the inverse-monotonicity property is the following result.

Theorem C.2.5 (Comparison principle)**.** *Suppose that there exists a function* $\phi \in C^2(\Omega) \cap C^0(\overline{\Omega})$ *such that*

$$\begin{array}{rcll} Lu(\mathbf{x}) & \leq & L\phi(\mathbf{x}) & \forall \mathbf{x} \in \Omega \\ u(\mathbf{x}) & \leq & \phi(\mathbf{x}) & \forall \mathbf{x} \in \partial\Omega. \end{array}$$

Then, we have

$$u(\mathbf{x}) \leq \phi(\mathbf{x}) \qquad \forall \mathbf{x} \in \overline{\Omega}$$

and we say that ϕ *is a* barrier function *for u.*

Combining the previous two properties, we obtain the following result which is a very useful tool in the approximation process of a BVP associated with the second-order elliptic operator L.

Theorem C.2.6 (Maximum principle)**.** *Suppose that L is inverse-monotone and that the comparison principle holds for a suitable barrier function* ϕ*. Then, setting* $M_\phi := \max_{\mathbf{x} \in \overline{\Omega}} \phi(\mathbf{x})$*, we have*

$$0 \leq u(\mathbf{x}) \leq M_\phi \qquad \forall \mathbf{x} \in \overline{\Omega}$$

and we say that u satisfies a maximum principle *(MP).*

Remark C.2.7 (Physical interpretation of the MP). The MP has an important physical significance, because it mathematically expresses the obvious fact that the dependent variable of the problem, say, a concentration, a temperature or a mass density, cannot take negative values.

Example C.2.8. The reaction-diffusion and advection-diffusion operators considered in Chapt. 6 satisfy the maximum principle taking $\phi(x) = 1/\sigma$ and $\phi(x) = x/a$, respectively.

C.3 Tensors

Let $\mathbf{x} = (x_1, x_2, x_3)^T$ denote the position vector in $\mathbb{R}^3$. A tensor $\tau_{ij} = \tau_{ij}(\mathbf{x})$ is a 3×3 real-valued matrix function of the form

$$\tau \begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix} = \begin{bmatrix} \tau_1^T \\ \tau_2^T \\ \tau_3^T \end{bmatrix}$$

where $\tau_i = [\tau_{i1}, \tau_{i2}, \tau_{i3}]^T$, $i = 1,2,3$. The trace of τ is the scalar quantity defined as the sum of the diagonal entries of τ, i.e.

$$\mathrm{Tr}\tau = \sum_{i=1}^{3} \tau_{ii}.$$

A tensor is symmetric if $\tau_{ji} = \tau_{ij}$, $i, j = 1,2,3$, while if $\tau_{ji} = -\tau_{ij}$, we say that τ is skew-symmetric; the symmetric part of a given tensor τ is defined as

$$\tau_S := \frac{\tau + \tau^T}{2}$$

while the skew-symmetric part of τ is

$$\tau_{SS} := \frac{\tau - \tau^T}{2}.$$

Clearly, for every tensor τ, we have

$$\tau = \tau_S + \tau_{SS}. \tag{C.6}$$

C.3.1 Operations/operators on tensors

The product between a tensor $\tau \in \mathbb{R}^{3\times 3}$ and a vector $\mathbf{v} \in \mathbb{R}^3$ is a vector $\mathbf{w} \in \mathbb{R}^3$ that is defined as

$$w_i = \sum_{j=1}^{3} \tau_{ij} v_j = \tau_i^T \mathbf{v} \qquad i = 1,2,3. \tag{C.7}$$

The divergence of a tensor τ is a vector field defined as

$$\mathrm{div}\tau = \begin{bmatrix} \dfrac{\partial \tau_{11}}{\partial x_1} + \dfrac{\partial \tau_{12}}{\partial x_2} + \dfrac{\partial \tau_{13}}{\partial x_3} \\ \dfrac{\partial \tau_{21}}{\partial x_1} + \dfrac{\partial \tau_{22}}{\partial x_2} + \dfrac{\partial \tau_{23}}{\partial x_3} \\ \dfrac{\partial \tau_{31}}{\partial x_1} + \dfrac{\partial \tau_{32}}{\partial x_2} + \dfrac{\partial \tau_{33}}{\partial x_3} \end{bmatrix} = \begin{bmatrix} \mathrm{div}\tau_1 \\ \mathrm{div}\tau_2 \\ \mathrm{div}\tau_3 \end{bmatrix}. \tag{C.8}$$

Let $\mathbf{U} = [u_1, u_2, u_3]^T$ be a given vector field whose components $u_i = u_i(\mathbf{x})$ are sufficiently smooth scalar functions, $i = 1,2,3$. Then, the gradient of $\mathbf{U}$ is the

tensor defined as

$$\nabla \mathbf{U} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{bmatrix} = \begin{bmatrix} (\nabla u_1)^T \\ (\nabla u_2)^T \\ (\nabla u_3)^T \end{bmatrix}. \tag{C.9}$$

We notice that

$$\mathrm{Tr}(\nabla \mathbf{U}) := \sum_{i=1}^{3} (\nabla \mathbf{U})_{ii} = \sum_{i=1}^{3} \frac{\partial u_i}{\partial x_i} = \mathrm{div}\mathbf{U}. \tag{C.10}$$