LECTURE NOTES
MTH 8207

An Introduction to the Finite Element Method

with

Comsol Multiphysics

Serge PRUDHOMME
Department of Mathematics
École Polytechnique de Montréal, Canada

Régis COTTEREAU
Centrale-Supélec, France

École Polytechnique de Montréal
Fall 2016

Contents

Chapter1

Introduction to the Finite Element Method

1.1 Introduction

"The origins of the finite element method can be traced back to the 1950s when engineers started solving structural mechanics problems in aeronautics using numerical tools. Since then, the field of applications has steadily widened and encompasses nowadays nonlinear solid mechanics, fluid-structure interactions, turbulent flows in industrial or geophysical settings, multicomponent reactive flows, mass transfer in porous media, viscoelastic flows in medical sciences, electromagnetism, or option pricing (to list just a few). Numerous commercial and academic codes based on the finite element method have been developed over the years. The method has been so successful for the solution of partial differential equations (PDEs) that the term "finite element method" now refers not only to the mere interpolation technique that it is beforehand, but also to the approximation techniques to solve differential equations in general. The efficiency of the finite element method relies on two distinct ingredients: the interpolation capability of finite elements in the approximation of scalar- and vector-valued functions, as well as the ability to approximate a mathematical model given in terms of partial differential equations within a proper mathematical framework" (adapted from Ern and Guermond [2]).

The objective of these lectures is to give a brief introduction of the finite element method. In particular, we will explain how to construct finite element spaces of (parametric) functions that can be used as trial or admissible functions for the solution of partial differential equations. We will also show how PDEs can be discretized within the finite element framework. The course will be concerned with the description of key mathematical concepts related to the finite element method as well as with the learning of Comsol Multiphysics^{TM} [1] for the application of finite element techniques to engineering problems. In this introductory chapter, only 1D problems will be considered. Problems defined on 2D and 3D geometries, of higher interest in practice, will be introduced in the next chapter.

1.2 Model problem

Most problems in physics and mechanics are described as a set of partial differential equations and initial/boundary conditions. This set is called the strong form of the problem. Finite elements, however, are based on an alternative form, the weak form, which is equivalent to the former. We describe in this section both the strong and weak forms, and we prove their equivalence. Examples of physical problems are also presented.

1.2.1 Strong formulation and examples

Let $\Omega = (0,1)$ be an open bounded interval in $\mathbb{R}$. Suppose that we are interested in solving for the function $u = u(x) \in C^2(\overline{\Omega})$, that satisfies the differential equation:

$$-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu = f, \quad \text{in } \Omega \tag{1.1}$$

and subjected to the boundary conditions

$$u = u_d, \quad \text{at } x = 0 \quad \text{and} \quad a\frac{du}{dx} = g, \quad \text{at } x = 1 \tag{1.2}$$

where $a = a(x) > 0$, $b = b(x)$, and $c = c(x) \geq 0$ are "material" data, $f = f(x)$ represents "loading" data, and $u_d \in \mathbb{R}$ and $g \in \mathbb{R}$ are "boundary" data. We simply assume here that a, b, c, and f are "smooth" functions on Ω. The smoothness of these functions, as well as the positivity of a and c is important when one tries to prove the existence and uniqueness of a solution to the equations (see Sec. 3.2).

Equation (1.1) is the general example of second-order elliptic differential equations[1] expressed here in strong form. The first equation in (1.2) is called a Dirichlet or essential boundary condition while the second is a Neumann or natural boundary condition. A third type of boundary conditions, named Robin, could also be considered as a generalization of the Neumann boundary condition:

$$a\frac{du}{dx} + \alpha u = g$$

where $\alpha \in \mathbb{R}$.

Problem (1.1)–(1.2) is frequently encountered in engineering applications such as heat transfer, linear elasticity, etc.

Example 1 ***(Linear elasticity)*** *The displacement $u = u(x)$ along a rod of length L with variable cross-sectional area $A(x)$ and homogeneous Young's modulus E is governed by the 1D differential equation:*

$$-\frac{d}{dx}\left(EA\frac{du}{dx}\right) = 0, \quad \textit{in } (0, L).$$

The boundary conditions are given in terms of the displacements (Dirichlet condition) and/or the loads $EA\frac{du}{dx}$ (Neumann condition) at the extremities 0 and L. Alternatively, an imperfect connection to the support at one of the extremities (for instance, at $x = 0$) can be modeled using a relation of the form $\frac{du}{dx}(0) = K_0 u(0)$, where K_0 controls the flexibility of the connection (Robin condition).

Example 2 ***(Stationary heat transfer)*** *Consider a rod of length L with thermal conductivity k held at constant temperature θ_0 at both ends and subjected to a heat source $q(x)$. Assuming that the walls of the rod are adiabatic (no heat flux), the temperature $\theta = \theta(x)$ is uniform in each cross-section and is modeled by the strong form differential equation:*

$$-\frac{d}{dx}\left(k\frac{d\theta}{dx}\right) = q, \quad \textit{in } (0, L)$$

[1] Partial differential equations are classified into elliptic (*e.g.*, Laplace and Poisson equations), parabolic (*e.g.*, the heat equation) and hyperbolic equations (*e.g.*, the wave equation). Grossly speaking, elliptic equations relate to stationary problems, parabolic equations to dissipative evolution problems, and hyperbolic to conservative evolution problems. The mathematical properties of these equations and the smoothness of the corresponding solutions lead to the use of different numerical techniques for each type of equations [3].

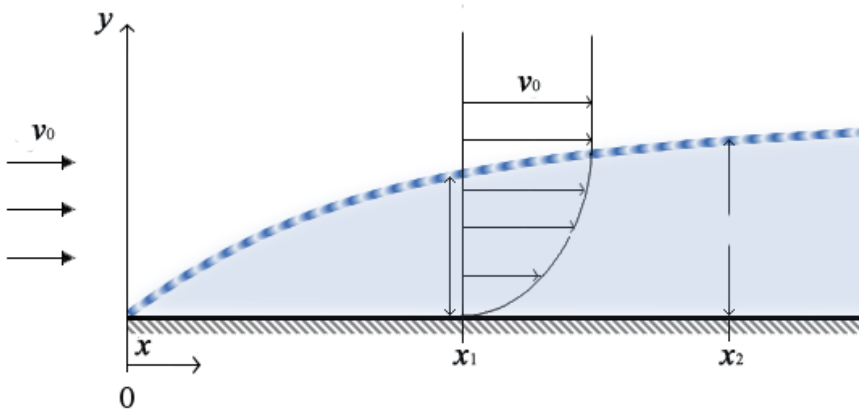


Figure 1.1: Description of a boundary layer problem.

It is moreover subjected to the Dirichlet boundary conditions $\theta(0) = \theta(L) = \theta_0$.

Example 3 ***(Boundary layer)*** *The shape of a boundary layer (see Fig. 1.1) can be modeled in non-dimensional form by the 1D differential equation*

$$-\epsilon \frac{d^2 v_0}{dy^2} + v_0 = 0, \quad in\ (0,1)$$

when subjected to the Dirichlet boundary conditions $v_0(0) = 0$ *and* $v_0(1) = 1$.

1.2.2 Weak formulation

Let $v = v(x)$ be an arbitrary function in $C^\infty(\overline{\Omega})$ (infinitely smooth), multiply (1.1) by v, and integrate over Ω. We get:

$$\int_\Omega \left[-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu \right] v dx = \int_\Omega f v dx$$

Using integration by parts, i.e. $(uv)' = u'v + uv'$, we observe that the first term on the left-hand side can be rewritten:

$$\int_\Omega -\frac{d}{dx}\left(a\frac{du}{dx}\right) v dx = -\int_\Omega \frac{d}{dx}\left(a\frac{du}{dx}v\right) dx + \int_\Omega a\frac{du}{dx}\frac{dv}{dx}dx = \int_\Omega a\frac{du}{dx}\frac{dv}{dx}dx - \left[a\frac{du}{dx}v\right]_0^1$$

Inserting the latter into the former equation, we have:

$$\int_\Omega \left[a\frac{du}{dx}\frac{dv}{dx} + b\frac{du}{dx}v + cuv \right] dx - a(1)\frac{du}{dx}(1)v(1) + a(0)\frac{du}{dx}(0)v(0) = \int_\Omega f v dx$$

and upon applying the Neumann boundary condition, one gets:

$$\int_\Omega \left[a\frac{du}{dx}\frac{dv}{dx} + b\frac{du}{dx}v + cuv \right] dx + a(0)\frac{du}{dx}(0)v(0) = \int_\Omega f v dx + g v(1)$$

We observe that the Neumann boundary condition comes naturally into the formulation, hence the name “natural” B.C.

At $x = 0$, the value of the first derivative $u'(0)$ (more specifically $a(0)u'(0)$) is unknown. However, we have the choice on how to select the test function v, and in particular, its value at $x = 0$. Let us take $v(0) = 0$. This choice actually makes sense as the value of u is known at $x = 0$ and thus does not need to be "tested" by the test function. Thus it becomes essential to enforce the kinematic condition $u(0) = u_d$, hence the name "essential" B.C. Then, one gets:

$$\int_\Omega \left[a\frac{du}{dx}\frac{dv}{dx} + b\frac{du}{dx}v + cuv \right] dx = \int_\Omega fvdx + gv(1), \quad \forall v \in C^\infty(\overline{\Omega}),\ v(0) = 0$$

Note that integration and integration by parts has allowed to eliminate the second derivative from the formulation and to consider derivatives in the distributional sense only[2]. Constraints on u have been weakened; for example, whereas the second and first derivatives of a piecewise continuous function do not exist, the above integrals with u and v piecewise continuous are well defined and the formulation makes sense. Since our goal is to find an approximation of u, the weak setting gives us more freedom when choosing the form of the trial functions.

Let us introduce the vector spaces of functions:

$$\begin{aligned}
L^2(\Omega) &= \{v :\ x \in \Omega \longrightarrow v(x) \in \mathbb{R} \text{ (or } \mathbb{C});\ \int_\Omega |v(x)|^2 dx < \infty\} \\
H^1(\Omega) &= \{v \in L^2(\Omega);\ v' \in L^2(\Omega)\} \\
H^1_0(\Omega) &= \{v \in H^1(\Omega);\ \gamma v(0) = \gamma v(1) = 0\} \\
H^2(\Omega) &= \{v \in H^1(\Omega);\ v'' \in L^2(\Omega)\}
\end{aligned}$$

where the integrals have to be understood in the Lebesgue sense ("L" in "$L^2(\Omega)$" actually stands for Lebesgue) and where γv denotes the trace of the function v on the boundary. Recall that the function is only defined on the open set $\Omega = (0, 1)$. The trace operator allows to extend the function to the boundary. We also introduce the sets of functions U and V such that:

$$\begin{aligned}
U &= \{u \in H^1(\Omega);\ \gamma u(0) = u_d\} \\
V &= \{v \in H^1(\Omega);\ \gamma v(0) = 0\}
\end{aligned}$$

A function $u \in U$ is called an admissible or trial function while a function $v \in V$ is referred to as a test function.

The weak formulation of Problem (1.1)–(1.2) is:

$$\boxed{\begin{array}{l}\text{Given } f \text{ and } g, \text{ find } u \in U, \text{ such that} \\ \qquad \displaystyle\int_\Omega \left[a\frac{du}{dx}\frac{dv}{dx} + b\frac{du}{dx}v + cuv \right] dx = \int_\Omega fvdx + gv(1), \quad \forall v \in V\end{array}} \tag{1.3}$$

or, in a more compact form,

$$\boxed{\text{Find } u \in U, \text{ such that} \quad B(u, v) = F(v), \quad \forall v \in V} \tag{1.4}$$

where we have defined the bilinear form $B(\cdot, \cdot)$ on $H^1(\Omega) \times H^1(\Omega)$ and linear form $F(\cdot)$ on $H^1(\Omega)$ as:

$$\begin{aligned}
B(u, v) &= \int_\Omega \left[a\frac{du}{dx}\frac{dv}{dx} + b\frac{du}{dx}v + cuv \right] dx \\
F(v) &= \int_\Omega fvdx + gv(1)
\end{aligned}$$

[2]Some details on the derivatives of a function in the distributional sense can be found in [2, Appendix B]

Remark 1 *Let us introduce the function $\tilde{u}$ on Ω (called the lift) such that $\tilde{u}(x) = u_d$ and the function w such that $u = w + \tilde{u}$. Observe that $\gamma w = \gamma u - \gamma \tilde{u} = u_d - u_d = 0$. Then $B(u, v) = B(w + \tilde{u}, v) = B(w, v) + B(\tilde{u}, v) = F(v)$, that is, $B(w, v) = F(v) - B(\tilde{u}, v)$. The weak form of the problem can then be recast as:*

$$\boxed{\begin{array}{l}\text{Find } w \in H_0^1(\Omega), \text{ such that} \\ \quad \displaystyle\int_\Omega \left[a\frac{dw}{dx}\frac{dv}{dx} + b\frac{dw}{dx}v + cwv\right] dx = \int_\Omega fv dx - \int_\Omega c\tilde{u}v dx + gv(1), \quad \forall v \in H_0^1(\Omega)\end{array}} \tag{1.5}$$

or, in compact form, as

$$\text{Find } w \in H_0^1(\Omega), \text{ such that} \quad B(w, v) = F(v) - B(\tilde{u}, v), \quad \forall v \in H_0^1(\Omega) \tag{1.6}$$

Since a function in $C^2(\overline{\Omega})$ is also in $H^1(\Omega)$, we have shown so far that if u is solution of Problem (1.1)–(1.2), it is also a solution of Problem (1.3). We would like to know now whether the reverse is true in order to make sure that by solving (1.3), we actually solve the problem of interest. This question is addressed in the next section.

1.2.3 Equivalence of the strong and weak formulations

Let u be the solution of Problem (1.3) and let the data f be sufficiently smooth so that u is twice-differentiable (in the distributional sense). It can actually be shown that $f \in L^2(\Omega)$ guarantees that $u \in U \cap H^2(\Omega) \subset U$. Then, integrating (1.3) by parts, we have for all $v \in V$,

$$\int_\Omega \left[-\frac{d}{dx}\left(a\frac{du}{dx}\right)v + b\frac{du}{dx}v + cuv\right] dx + \int_\Omega \frac{d}{dx}\left(a\frac{du}{dx}v\right) dx = \int_\Omega fv dx + gv(1)$$

that is

$$\int_\Omega \left[-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu - f\right] v dx + \left[a(1)\frac{du}{dx}(1) - g\right] v(1) - a(0)\frac{du}{dx}(0)v(0) = 0$$

Since $v(0) = 0$, the above equation reduces to:

$$\int_\Omega \left[-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu - f\right] v dx + \left[a(1)\frac{du}{dx}(1) - g\right] v(1) = 0 \tag{1.7}$$

The proof then proceeds in three steps:

1. We denote by $\mathcal{D}(\Omega)$ the set of infinitely differentiable functions defined on Ω with compact support in Ω ($\mathcal{D}(\Omega)$ is often called the space of test functions). Obviously, if $\phi \in \mathcal{D}(\Omega)$, then $\phi \in V$ and $\phi(0) = \phi(1) = 0$. Therefore:

   $$\int_\Omega \phi(x) \left[-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu - f\right] dx = 0, \quad \forall \phi \in \mathcal{D}(\Omega)$$

   which, by using Fourier's theorem, implies that

   $$\boxed{-\frac{d}{dx}\left(a\frac{du}{dx}\right) + b\frac{du}{dx} + cu = f \quad \text{in } \Omega} \tag{1.8}$$

2. Using (1.8) in (1.7) allows us to write:

$$\left[a(1)\frac{du}{dx}(1) - g\right] v(1) = 0$$

Choose a test function $v \in V$ such that $v(1) = 1$ (for example, $v(x) = x$). Then

$$\boxed{a\frac{du}{dx} = g, \quad \text{at } x = 1} \tag{1.9}$$

3. Finally, since $u \in U$, it immediately follows that $\gamma u = u_d$ at $x = 0$.

We conclude that if u is solution of the weak form of the problem and u is sufficiently regular (depending on the regularity of f), then u is also solution of the strong form of the problem.

We note however that if the data f is not smooth enough, the weak form of the problem may have a solution while the strong form of the problem does not. This is the case for instance when f is a Dirac function. The Dirac function does not belong to $L^s(\Omega)$ so that u is not in $H^2(\Omega)$. In 2D and 3D, the regularity of the solution also depends on the shape of the domain.

Example 4 *Let $a = 1$, $b = c = 0$, $f = 2$, $u_d = 0$, and $g = 0$, so that the model problem in the strong form reads: Find $u \in C^2(\overline{\Omega})$ such that*

$$-\frac{d^2u}{dx^2} = 2, \quad \text{in } \Omega, \qquad u(0) = 0, \qquad \frac{du}{dx}(1) = 0$$

The exact solution of this problem is $u(x) = x(2 - x)$ while the weak form reads:

$$\text{Find } u \in V \text{ such that} \quad \int_\Omega \frac{du}{dx}\frac{dv}{dx}dx = \int_\Omega 2v dx, \quad \forall v \in V \tag{1.10}$$

Remark 2 *We observe that the weak formulation is not amenable to solve the problem analytically. Indeed the above equation needs to be tested for every test function in V, which is infinite dimensional. To make things clearer, imagine that we consider trial functions in the form:*

$$u(x) = \sum_{i=1}^{\infty} u_i x^i$$

where the coefficients $u_i \in \mathbb{R}$ need to be determined. The above expansion actually corresponds to Taylor expansions for infinitely smooth functions in which the monomials $\{x^i\}$ form a basis of the subspace $V \cap C^\infty(\Omega)$ of V. Since the number of monomials is infinite, the idea is to search for approximations of u by truncating the expansion, i.e.

$$u(x) \approx \tilde{u}(x) = \sum_{i=1}^{N} \tilde{u}_i x^i$$

Replacing u by $\tilde{u}$ in the weak formulation and taking $v = x^i$, for $i = 1, \ldots, N$, we are then able to define N equations for the N unknowns u_i, $i = 1, \ldots, N$. This yields a system of N linear equations that can be solved using direct or iterative solvers. Note that $\tilde{u} \approx u$ and $\tilde{u}_i \approx u_i$.

This is essentially the concept of the finite element method in which the trial function u is approximated by piecewise continuous or discontinuous functions constructed by means of finite elements.

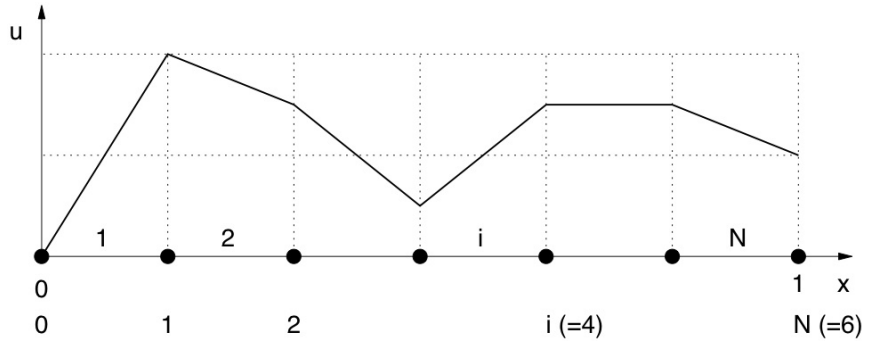


Figure 1.2: Example of piecewise linear continuous function using the partition $\{I_i\}$.

1.3 Finite Element approximations

The objective in this section is to construct piecewise linear continuous functions to be used as trial functions in order to compute approximate solutions of the problem of interest. Note that such functions belong to $H^1(\Omega)$ (but not to $C^\infty(\Omega)$ nor to $H^2(\Omega)$).

1.3.1 Mesh and elements

Let N be a positive integer and let x_i, $i = 0, \ldots, N$, define points in $\Omega = (0, 1)$ such that $0 = x_0 < x_1 < \ldots < x_i < \ldots < x_N = 1$. The points x_i are called vertices. We also introduce the intervals $I_i = [x_{i-1}, x_i]$ such that $\overline{\Omega} = \cup_{1 \leq i \leq N} I_i$. The intervals I_i are called elements. The size of the elements is given by $h_i = |x_i - x_{i-1}|$ and we denote $h = \max_i h_i$. Using such a partition of the domain Ω, it is then possible to define piecewise continuous functions (see e.g. Fig. 1.2). Let ${V_h}^3$ denote the vector space of all piecewise linear continuous functions u_h defined on $\overline{\Omega}$ such that $u_h(0) = 0$.

1.3.2 Basis functions

We can show that any function in V_h can be defined as a linear combination of the basis functions $\phi_i \in V_h$ (hat functions) shown in Fig. 1.3, i.e. $\forall u_h \in V_h$, there exists one and only one set of N coefficients u_i in $\mathbb{R}$ such that:

$$u_h(x) = \sum_{i=1}^{N} u_i \phi_i(x)$$

where the basis function can be explicitly written as:

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & \text{if } x \in I_i \\ (x_{i+1} - x)/h_{i+1}, & \text{if } x \in I_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{1.11}$$

[3]For notational simplicity, we restrict ourselves here to the case where a homogeneous Dirichlet boundary condition is enforced at $x = 0$ and a Neumann boundary condition at $x = 1$. More generally, V_h would represent the vector space of all piecewise linear continuous functions u_h defined on $\overline{\Omega}$ such that the Dirichlet boundary conditions are verified.

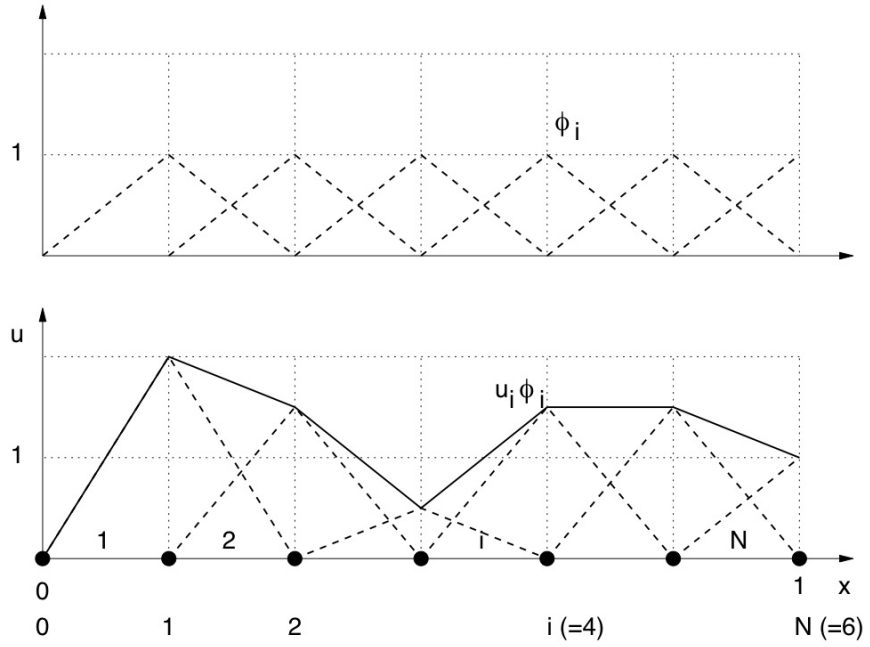


Figure 1.3: Basis functions for piecewise linear continuous approximations.

Note that the domain of the basis functions (hat functions) is actually the whole interval $\overline{\Omega}$. For the function shown in Fig. 1.2, we clearly have:

$$u_h = 2.0\phi_1 + 1.5\phi_2 + 0.5\phi_3 + 1.5\phi_4 + 1.5\phi_5 + 1.0\phi_6$$

so that the vector of unknowns $U = \{u_i\} = [2.0, 1.5, 0.5, 1.5, 1.5, 1.0]^T$. See Fig. 1.3. The coefficients u_i are referred to as the degrees of freedom.

1.3.3 Shape functions

The basis functions ϕ_i can easily be constructed element by element by introducing shape functions $N_{j,e}$, $j = 1, 2$, defined on each element I_e. In 1D, the correspondence between the indices i and e is trivial with $i = e$. This is not the case in higher dimensions. From Fig. 1.4, we see that, if the support of ϕ_i corresponds to the elements I_e and I_{e+1}:

$$\phi_i(x) = \begin{cases} N_{2,e}(x), & \text{if } x \in I_e \\ N_{1,e+1}(x), & \text{if } x \in I_{e+1} \\ 0, & \text{otherwise} \end{cases} \tag{1.12}$$

with $N_{2,e}(x) = (x - x_{i-1})/h_e$ and $N_{1,e+1}(x) = (x_{i+1} - x)/h_{e+1}$.

1.3.4 Galerkin method and system of equations

The Galerkin method is a method by which one can derive approximations of problems given in weak forms. The method consists of two steps:

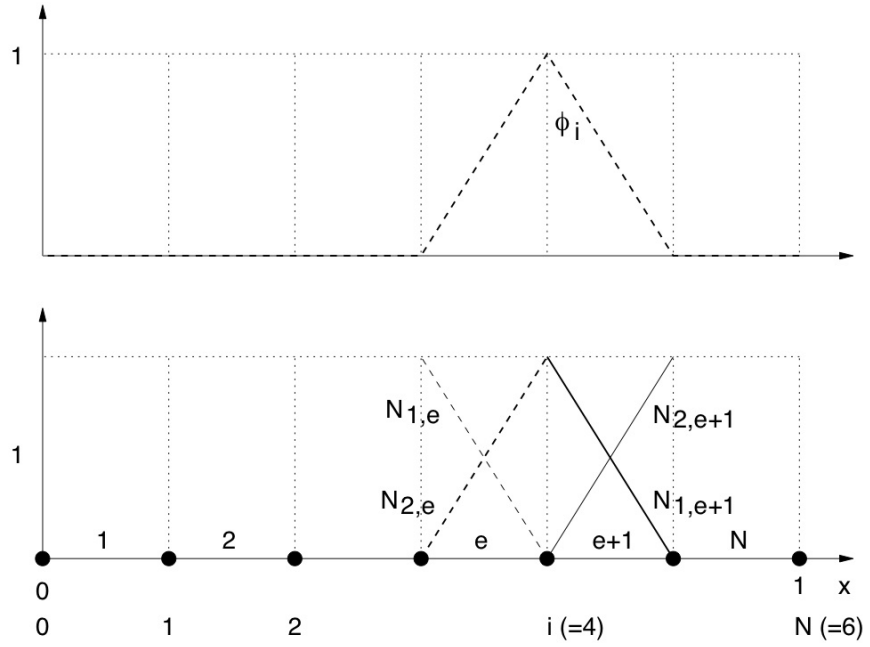


Figure 1.4: Shape functions in elements e and $e+1$.

1. Replace u by $u_h = \sum_{i=1}^{N} u_i \phi_i$ for the trial function.
2. Test the equation with all test functions $v_h \in V_h$.

We now apply the Galerkin method to Problem (1.10) in Example 4 using $N = 2$ elements, $x_0 = 0$, $x_1 = 1/2$, and $x_2 = 1$. Note that the two elements are not necessarily of the same size. In this case, we have two basis functions ϕ_1 and ϕ_2 associated with the degrees of freedom u_1 and u_2. The discrete trial functions are written $u_h = u_1\phi_1 + u_2\phi_2 \in V_h$.

The finite element problem then reads: Find $u_h \in V_h$ such that

$$\int_\Omega \frac{du_h}{dx}\frac{dv_h}{dx}dx = \int_\Omega 2v_h dx, \quad \forall v_h \in V_h$$

which is equivalent to: Find $[u_1, u_2] \in \mathbb{R}^2$ such that

$$\int_\Omega \frac{d}{dx}(u_1\phi_1 + u_2\phi_2)\frac{d\phi_i}{dx}dx = \int_\Omega 2\phi_i dx, \quad \forall i = 1, 2$$

Rearranging,

$$u_1 \int_\Omega \frac{d\phi_1}{dx}\frac{d\phi_i}{dx}dx + u_2 \int_\Omega \frac{d\phi_2}{dx}\frac{d\phi_i}{dx}dx = \int_\Omega 2\phi_i dx, \quad \forall i = 1, 2$$

We readily observe that the two equations above can be written in matrix form

$$KU = F \tag{1.13}$$

where

$$K = \begin{bmatrix} \int_\Omega \frac{d\phi_1}{dx}\frac{d\phi_1}{dx}dx & \int_\Omega \frac{d\phi_2}{dx}\frac{d\phi_1}{dx}dx \\ \int_\Omega \frac{d\phi_1}{dx}\frac{d\phi_2}{dx}dx & \int_\Omega \frac{d\phi_2}{dx}\frac{d\phi_2}{dx}dx \end{bmatrix} \qquad U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \qquad F = \begin{bmatrix} \int_\Omega 2\phi_1 dx \\ \int_\Omega 2\phi_2 dx \end{bmatrix}$$

where K is referred to as the stiffness matrix, U the vector of unknowns, and F the loading vector.

We now show how to compute the elements of the matrix K and of the vector F. We start with K_{11}. Let denote $I_1 = [x_0, x_1]$ and $I_2 = [x_1, x_2]$ the first and second elements. We have:

$$\begin{aligned} K_{11} = \int_\Omega \frac{d\phi_1}{dx}\frac{d\phi_1}{dx}dx &= \int_{I_1} \frac{d\phi_1}{dx}\frac{d\phi_1}{dx}dx + \int_{I_2} \frac{d\phi_1}{dx}\frac{d\phi_1}{dx}dx \\ &= \int_{I_1} \frac{dN_{2,1}}{dx}\frac{dN_{2,1}}{dx}dx + \int_{I_2} \frac{dN_{1,2}}{dx}\frac{dN_{1,2}}{dx}dx \\ &= \int_{I_1} \frac{1}{h_1}\frac{1}{h_1}dx + \int_{I_2} \left(-\frac{1}{h_2}\right)\left(-\frac{1}{h_2}\right)dx = \frac{1}{h_1} + \frac{1}{h_2} \end{aligned}$$

In the same manner, we have:

$$\begin{aligned} K_{12} = \int_\Omega \frac{d\phi_2}{dx}\frac{d\phi_1}{dx}dx &= \int_{I_1} \frac{d\phi_2}{dx}\frac{d\phi_1}{dx}dx + \int_{I_2} \frac{d\phi_2}{dx}\frac{d\phi_1}{dx}dx = \int_{I_2} \frac{dN_{2,2}}{dx}\frac{dN_{1,2}}{dx}dx \\ &= \int_{I_2} \left(\frac{1}{h_2}\right)\left(-\frac{1}{h_2}\right)dx = -\frac{1}{h_2} \end{aligned}$$

and

$$\begin{aligned} K_{22} = \int_\Omega \frac{d\phi_2}{dx}\frac{d\phi_2}{dx}dx &= \int_{I_1} \frac{d\phi_2}{dx}\frac{d\phi_2}{dx}dx + \int_{I_2} \frac{d\phi_2}{dx}\frac{d\phi_2}{dx}dx = \int_{I_2} \frac{dN_{2,2}}{dx}\frac{dN_{2,2}}{dx}dx \\ &= \int_{I_2} \left(\frac{1}{h_2}\right)\left(\frac{1}{h_2}\right)dx = \frac{1}{h_2} \end{aligned}$$

Finally, by symmetry, we have:

$$K_{21} = K_{12} = -\frac{1}{h_2}$$

For the right-hand side vector, we get:

$$\begin{aligned} F_1 = \int_\Omega 2\phi_1 dx &= \int_{I_1} 2\phi_1 dx + \int_{I_2} 2\phi_1 dx = \int_{I_1} 2N_{2,1}dx + \int_{I_2} 2N_{1,2}dx \\ &= \int_{I_1} 2\frac{x - x_0}{h_1}dx + \int_{I_2} 2\frac{x_2 - x}{h_2}dx = \left[\frac{(x-x_0)^2}{h_1}\right]_{x_0}^{x_1} - \left[\frac{(x_2-x)^2}{h_2}\right]_{x_1}^{x_2} \\ &= h_1 + h_2 \\ F_2 = \int_\Omega 2\phi_2 dx &= \int_{I_1} 2\phi_2 dx + \int_{I_2} 2\phi_2 dx = \int_{I_2} 2N_{2,2}dx \\ &= \int_{I_2} 2\frac{x - x_1}{h_2}dx = \left[\frac{(x-x_1)^2}{h_1}\right]_{x_1}^{x_2} \\ &= h_2 \end{aligned}$$

so that the system of equations is given by:

$$\begin{bmatrix} \left(\frac{1}{h_1} + \frac{1}{h_2}\right) & -\frac{1}{h_2} \\ -\frac{1}{h_2} & +\frac{1}{h_2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} h_1 + h_2 \\ h_2 \end{bmatrix}$$

The solution of this 2×2 system of linear equations is:

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} h_1^2 + 2h_1h_2 \\ (h_1 + h_2)^2 \end{bmatrix}$$

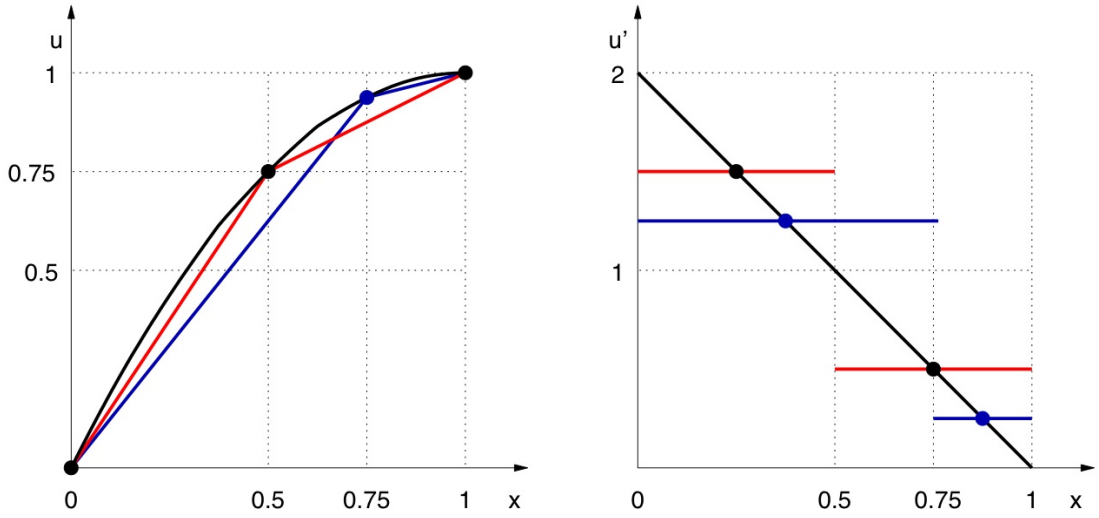


Figure 1.5: Finite element approximations and exact solution (left), as well as "first derivatives" (right), for case 1 (red line) and case 2 (blue line) of the problem presented in Example 5.

Example 5 *Recall that the exact solution is given by* $u(x) = 2x - x^2$*, which yields* $u'(x) = 2(1-x)$*. We consider two approximations for* u_h*:*

1. *Case 1: Set* $h_1 = h_2 = 1/2$*. Then* $U = [3/4; 1]^T$*. The derivative of* u_h *is given by* $u_h'|_{I_1} = (3/4 - 0)/(1/2) = 3/2$ *in* I_1 *and* $u_h'|_{I_2} = (1 - 3/4)/(1/2) = 1/2$ *in* I_2*. The solution for that case is plotted in red in Fig. 1.5.*

2. *Case 2: Set* $h_1 = 3/4$ *and* $h_2 = 1/4$*. Then* $U = [15/16; 1]^T$*. The derivative of* u_h *is then given by* $u_h'|_{I_1} = (15/16 - 0)/(3/4) = 5/4$ *in* I_1 *and* $u_h'|_{I_2} = (1 - 15/16)/(1/4) = 1/4$ *in* I_2*. The solution for that case is plotted in blue in Fig. 1.5.*

The exact and finite element solutions, as well as their first derivatives (inside each element for the finite element approximations) are shown in Fig. 1.5. One question one may ask is which of the two finite element solutions is more accurate. This answer is widely subjective as it depends on the goal of the simulations. On one hand, if one is interested in the first derivative in the region [0.75,1] then the approximation given by case 2 is more accurate. On the other hand, if one is interested in the solution itself in the region [0,0.5], then the approximation given by case 1 is definitely better. It is also clear that with the same number of degrees of freedom we may reach different degrees of accuracy depending on the position and size of the elements. The design of optimal meshes is crucial in Finite Element methods, can be very much time-consuming, and is the subject of adaptive meshing procedures. Attention: the fact that the values of the finite element solutions at the nodes are exact in this example are coincidental. This is actually an exception.

1.3.5 Element-by-element assembly

In finite element codes, the system of equations (1.13) is rarely assembled using the global approach presented in the previous section. Rather, sub-matrices and sub-vectors of K and F, respectively, are computed at the element level and then combined together to produce K and F. Note that, in this case, the Dirichlet boundary conditions are applied only at the global level.

Using above example, shape functions in I_1 are given by $N_{1,1}$ and $N_{2,1}$ and those in I_2 by $N_{1,2}$ and $N_{2,2}$. Ignoring the Dirichlet boundary condition, we also have three basis functions such that

$$u_h = \sum_{i=0}^{2} u_i \phi_i$$

constructed from the shape functions as (see Fig. 1.4):

$$\phi_0 = \begin{cases} N_{1,1}, & \text{in } I_1 \\ 0, & \text{in } I_2 \end{cases} \qquad \phi_1 = \begin{cases} N_{2,1}, & \text{in } I_1 \\ N_{1,2}, & \text{in } I_2 \end{cases} \qquad \phi_2 = \begin{cases} 0, & \text{in } I_1 \\ N_{2,2}, & \text{in } I_2 \end{cases}$$

In practice, this construction is performed using the so-called connectivity array C which assigns for each element e and each shape function j of e the corresponding basis function i (i.e. degree of freedom u_i):

$$i = C(e, j)$$

For instance, the connectivity array for our example reads:

$$\begin{aligned} C(1,1) &= 0 \\ C(1,2) &= 1 \\ C(2,1) &= 1 \\ C(2,2) &= 2 \end{aligned}$$

The element stiffness matrix and load vector are then computed in element e as:

$$K^e = \begin{bmatrix} \int_\Omega \frac{dN_{1,e}}{dx}\frac{dN_{1,e}}{dx}dx & \int_\Omega \frac{dN_{2,e}}{dx}\frac{dN_{1,e}}{dx}dx \\ \int_\Omega \frac{dN_{1,e}}{dx}\frac{dN_{2,e}}{dx}dx & \int_\Omega \frac{dN_{2,e}}{dx}\frac{dN_{2,e}}{dx}dx \end{bmatrix} \qquad F^e = \begin{bmatrix} \int_\Omega 2N_{1,e}dx \\ \int_\Omega 2N_{2,e}dx \end{bmatrix}$$

We see from previous calculations that:

$$K^1 = \begin{bmatrix} \frac{1}{h_1} & -\frac{1}{h_1} \\ -\frac{1}{h_1} & \frac{1}{h_1} \end{bmatrix} \quad F^1 = \begin{bmatrix} h_1 \\ \\ h_1 \end{bmatrix} \quad K^2 = \begin{bmatrix} \frac{1}{h_2} & -\frac{1}{h_2} \\ -\frac{1}{h_2} & \frac{1}{h_2} \end{bmatrix} \quad F^2 = \begin{bmatrix} h_2 \\ \\ h_2 \end{bmatrix}$$

Now, each entry of the sub-matrices K^e and of the sub-vectors F^e are added to the global matrix K and global vector F, respectively, according to the following rules:

1. K^e_{kl} is added to the entry K_{ij}, where $i = C(e,k)$ and $j = C(e,l)$,
2. F^e_k is added to the entry F_i, where $i = C(e,k)$.

It readily follows that:

$$K = \begin{bmatrix} \frac{1}{h_1} & -\frac{1}{h_1} & 0 \\ -\frac{1}{h_1} & \left(\frac{1}{h_1}+\frac{1}{h_2}\right) & -\frac{1}{h_2} \\ 0 & -\frac{1}{h_2} & \frac{1}{h_2} \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} h_1 \\ \\ h_1 + h_2 \\ \\ h_2 \end{bmatrix}$$

Dirichlet boundary conditions are then applied by modifying the systems of equations in two steps. Suppose that $u = u_d$ is prescribed at $x = 0$. Consider the vector $U_0 = [u_d, 0, 0]^T$. The first step consists in subtracting KU_0 from F and setting the entries of the column in K corresponding to the degree of freedom u_0 to zero. The second step consists in replacing the equation in the system corresponding to the test function associated with the first degree of freedom (i.e the first equation in this example) by the equation $u_0 = u_d$. The system of equations then read:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \left(\frac{1}{h_1} + \frac{1}{h_2}\right) & -\frac{1}{h_2} \\ 0 & -\frac{1}{h_2} & \frac{1}{h_2} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} u_d \\ h_1 + h_2 \\ h_2 \end{bmatrix} - \begin{bmatrix} 0 \\ -\frac{u_d}{h_1} \\ 0 \end{bmatrix}$$

If $u_d = 0$, then the system reduces to:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \left(\frac{1}{h_1} + \frac{1}{h_2}\right) & -\frac{1}{h_2} \\ 0 & -\frac{1}{h_2} & \frac{1}{h_2} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ h_1 + h_2 \\ h_2 \end{bmatrix}$$

whose solution is simply:

$$U = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ h_1^2 + 2h_1h_2 \\ (h_1 + h_2)^2 \end{bmatrix}$$

as before.

1.4 Conclusions and outline

We have presented in this chapter some of the main issues of Finite Element modeling in 1D: (a) the derivation of the weak formulation from the strong one, (b) the discretization of the weak formulation using a particular basis of functions, and (c) the reduction of the finite element formulation into a finite system of linear equations. There is not much change when going to 2D or 3D. In particular, the weak formulation is derived in exactly the same manner, although the mathematical formalism seems more complex. Likewise, the choice of basis functions is driven by the same concerns, although these functions are sampled over higher dimensional spaces. However, the implementation must be performed in a more rigorous manner. The connectivity array, only hinted at in this chapter, then becomes an essential tool. Finally, we only brushed the two related questions of error estimation and meshing, which are in fact essential. We will come back to that extensively.

1.5 Problems

1.5.1 Exercise 1

Suppose that you are interested in knowing the displacement u along a rod of length L with variable cross-sectional area $A(x)$. We assume that the rod is homogeneous with constant Young's modulus

E, that it is held fixed at $x = 0$, and that it is submitted to a longitudinal traction T at $x = L$. In this case, we know that the displacement $u = u(x)$ in the rod is governed by the 1D differential equation:

$$-\frac{d}{dx}\left(EA\frac{du}{dx}\right) = 0, \quad \text{in } (0, L)$$

and subjected to the Dirichlet and Neuman boundary conditions:

$$\begin{aligned} u &= 0, \quad \text{at } x = 0 \\ E\frac{du}{dx} &= T, \quad \text{at } x = L \end{aligned}$$

Derive the weak formulation of the problem.

1.5.2 Exercise 2

Consider a rod of length L held at constant temperature θ_0 at both ends and subjected to a heat source $q(x) = q_0$ constant along the rod. Assume that the walls of the rod are adiabatic (no heat flux) and that the thermal conductivity k is constant. In that case, the temperature $\theta = \theta(x)$ is uniform in each cross-section and is modeled by the strong form differential equation:

$$-\frac{d}{dx}\left(k\frac{d\theta}{dx}\right) = q_0, \quad \text{in } (0, L)$$

and boundary conditions:

$$\begin{aligned} \theta &= \theta_0, \quad \text{at } x = 0 \\ \theta &= \theta_0, \quad \text{at } x = L \end{aligned}$$

1. Solve for the exact solution $\theta = \theta(x)$ to this problem in terms of L, k, q_0, and θ_0 (the data of the problem).

2. Derive the weak formulation of the problem.

Suppose now that the source term is a pointwise source, i.e. $q = 0$ everywhere in the domain, except at a point $x_0 \in (0, L)$ where $q = q_0$. In terms of the Dirac function δ, the source term can be rewritten as $q(x) = q_0\delta(x - x_0)$. Recall that the Dirac function is defined as

$$\int_0^L \delta(x - x_0)v(x)dx = v(x_0)$$

for any "smooth" function v.

3. Write the weak form of the problem.

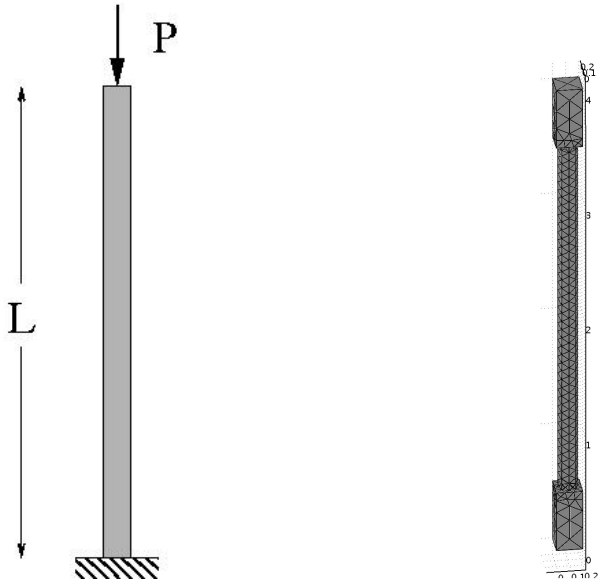
4. Recover the strong form of this new problem (note here that the second derivative of θ is not defined anymore at x_0).

1.5.3 Exercise 3

Consider the following problem written in weak form as: Find $u \in V$, V being a space of smooth functions with $u = u_L$ at $x = L$, such that:

$$\int_0^L \left(2\frac{du}{dx}\frac{dv}{dx} + 5\frac{du}{dx}v + 10uv \right) dx = \int_0^L v dx - 2v(0)$$

for all v smooth such that $v = 0$ at $x = L$. Write the strong form of this problem.



Description of the problem in 1D (left) and example of mesh for the 3D model (right, problem 2).

Problem 1. We are interested in the analysis of a column of length L, cross-sectional area A, and Young's modulus E. We assume that the column stands on a support at $x = 0$, that it is subjected to a longitudinal compression force P at $x = L$ and to the gravitational force density g. The displacement $u = u(x)$ in the column is governed by the 1D differential equation:

$$-\frac{d}{dx}\left(EA\frac{du}{dx}\right) = -\rho g A, \quad \text{in } (0, L)$$

and subjected to the Dirichlet and Neuman boundary conditions:

$$u = 0, \quad \text{at } x = 0, \qquad \text{and} \quad EA\frac{du}{dx} = -P, \quad \text{at } x = L$$

The following data will be the same for all questions: $L = 4$ m, $g = 9.81$ m/s^2, $P = 40$ kN.

1.1) In this question, take E, A, and ρ constant along x: $E = 20$ GPa, $\rho = 2300$ kg/m^3 (concrete), and $A = A_0 = 0.0341$ m^2.

1. Solve for the exact solution and derive the weak formulation of the problem.

2. Develop an application in Comsol Multiphysics to model the problem.

3. Compute the stress $\sigma = Edu/dx$ and the relative error in the stress at $x = 0$ when using 1, 2, 4, 8, and 16 linear elements of uniform size.

4. Using non uniform linear elements, design, by trial and error, a mesh that yields the minimal number of degrees of freedom while reaching a relative error in the stress at $x = 0$ smaller than half a percent.

1.2) Keep here E and ρ constant ($E = 20$ GPa, $\rho = 2300$ kg/m^3), and consider A such that

$$A = A_0 \left[1 - \frac{x(L-x)}{L^2} \right]$$

with $A_0 = 0.0341$ m^2. Repeat questions 2), 3), 4) of Question 1.1).

1.3) Suppose that the column is made of two different materials: in regions $(0, l)$ and $(L-l, L)$, with $l = 50$ cm, the column is made of a material with properties $E = 10$ GPa and $\rho = 500$ kg/m^3, and in these two regions, the column has a constant square cross-section with width $a_0 = 0.20$ m; in region $(l, L-l)$, the column has material properties $E = 20$ GPa, $\rho = 2300$ kg/m^3, and a constant circular cross-section with diameter $d_0 = 0.15$ m. Find the location x_s where the stress is maximal in the column. Design a mesh that should give a relative error in the maximal stress smaller than one percent.

Problem 2. Develop a 3D FE model using linear elasticity to simulate problem 1.3. Suppose that the different components of the column are perfectly aligned along the centerline and that the force P is equally distributed at $x = L$. Find the maximal stress σ_s and corresponding location x_s in the column (make sure that the mesh is sufficiently refined to provide accurate solution).

Problem 3. Suppose now that the circular column was imperfectly aligned with respect to the two other blocks by $\delta = 0.02$ m. Using 3D linear elasticity and assuming that the force P is equally distributed at $x = L$, compute the maximal deflection of the column.

Problem 4. Write a report in which you concisely describe and comment your results.

Chapter2

Finite Element Method in 2D and 3D

2.1 Introduction

The objective of these lecture notes is to generalize the concepts introduced in the case of one-dimensional problems to two and three-dimensions. In particular, we introduce the formal definition of finite elements and the concept of the reference element (or master element). We also show how the finite element system of equations is assembled and how integrals are computed using special integration rules. At the end of this chapter, all the basic features of a finite element software will have been presented, in the case of a scalar elliptic equation.

2.2 Model problem: strong and weak formulation

This section is very similar in concept to the corresponding one in the previous chapter. The main differences lie in the format, that is slightly more involved in high dimensional problems, and in the fact that the geometry may create singularities that distort the equivalence between strong and weak formulations.

2.2.1 Strong formulation

Let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$, $d = 2$ or 3, with boundary $\partial\Omega$, composed of two parts, Γ_d and Γ_n, such that $\partial\Omega = \overline{\Gamma_d \cup \Gamma_n}$. We are interested in solving for the scalar function $u = u(x)$, $x \in \overline{\Omega}$, that satisfies the differential equation:

$$-\nabla \cdot (k\nabla u) + cu = f, \quad \text{in } \Omega \subset \mathbb{R}^d \tag{2.1}$$

and subjected to the boundary conditions

$$\begin{aligned} &u = u_d, \quad \text{on } \Gamma_d \\ &n \cdot (k\nabla u) = g, \quad \text{on } \Gamma_n \end{aligned} \tag{2.2}$$

where n is the unit outward normal vector to the boundary, $f = f(x)$, $c = c(x)$, $x \in \Omega$, $g = g(x)$, $x \in \Gamma_n$, and $u_d = u_d(x)$, $x \in \Gamma_d$, are scalar functions. In the most general case, $k = k(x) = \{k_{ij}(x)\}$, $1 \leq i, j \leq d$, is a tensor-valued function, characterizing material properties derived from constitutive relations (e.g., the elasticity tensor E in linear elasticity). For now, we assume that all data are such that the problem is well-posed.

2.2.2 Weak formulation

As before, the weak formulation is obtained by multiplying the equation by an arbitrary smooth test function v and integrating over the domain Ω:

$$\int_\Omega [-\nabla \cdot (k\nabla u) + cu]v \, dx = \int_\Omega fv \, dx$$

We then integrate by parts using the relation $\nabla \cdot [(k\nabla u)v] = (\nabla \cdot k\nabla u)v + k\nabla u \cdot \nabla v$, to get:

$$\int_\Omega k\nabla u \cdot \nabla v + cuv \, dx - \int_\Omega \nabla \cdot [(k\nabla u)v] \, dx = \int_\Omega fv \, dx,$$

and making use of the divergence theorem (or Green-Ostrogradski), we arrive at:

$$\int_\Omega k\nabla u \cdot \nabla v + cuv \, dx - \int_{\partial\Omega} n \cdot (k\nabla u)v \, ds = \int_\Omega fv \, dx$$

or

$$\int_\Omega k\nabla u \cdot \nabla v + cuv \, dx - \int_{\Gamma_n} n \cdot (k\nabla u)v \, ds - \int_{\Gamma_d} n \cdot (k\nabla u)v \, ds = \int_\Omega fv \, dx$$

Applying the Neumann boundary condition $n \cdot (k\nabla u) = g$ on Γ_n and choosing the test function v to vanish on Γ_d (in other words, we do not need to test the integral on the boundary Γ_d since the solution is known to be $u = u_d$), the above integral equation is simplified as:

$$\int_\Omega k\nabla u \cdot \nabla v + cuv \, dx = \int_\Omega fv \, dx + \int_{\Gamma_n} gv \, ds, \quad \forall v \text{ smooth, } v = 0 \text{ on } \Gamma_d$$

Introducing the bilinear form $B(\cdot, \cdot)$ defined on $H^1(\Omega) \times H^1(\Omega)$ and the linear from $F(\cdot)$ defined on $H^1(\Omega)$ as:

$$B(u, v) = \int_\Omega k\nabla u \cdot \nabla v + cuv \, dx$$
$$F(v) = \int_\Omega fv \, dx + \int_{\Gamma_n} gv \, ds$$

the weak formulation of the problem (2.1-2.2) reads:

$$\boxed{\text{Find } u \in U, \text{ such that} \quad B(u, v) = F(v), \quad \forall v \in V} \tag{2.3}$$

where the spaces are defined as:

$$L^2(\Omega) = \{v : \ x \in \Omega \longrightarrow v(x) \in \mathbb{R} \text{ (or } \mathbb{C}); \ \int_\Omega |v(x)|^2 dx < \infty\}$$
$$H^1(\Omega) = \{v \in L^2(\Omega); \ \nabla v \in (L^2(\Omega))^d\}$$
$$H_0^1(\Omega) = \{v \in H^1(\Omega); \ \gamma v(0) = \gamma v(1) = 0\}$$
$$U = \{u \in H^1(\Omega); \gamma u(0) = u_d\}$$
$$V = \{v \in H^1(\Omega); \gamma v(0) = 0\}$$

Remark 3 *If we introduce the function $\tilde{u} \in H^1(\Omega)$, called the lift, on Ω such that $\tilde{u} = u_d$ on Γ_d as well as the function w such that $u = w + \tilde{u}$, the weak form can be recast as:*

$$\textit{Find } w \in V, \textit{ such that} \quad B(w, v) = F(v) - B(\tilde{u}, v), \quad \forall v \in V \tag{2.4}$$

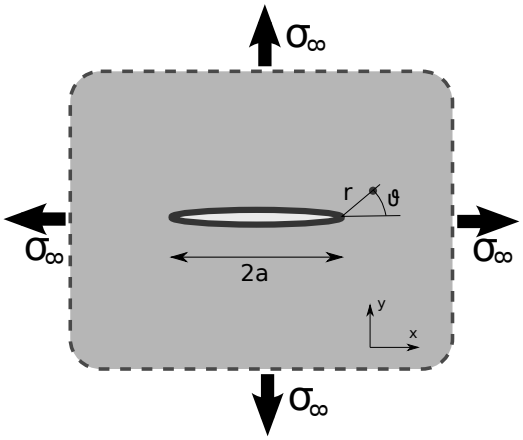


Figure 2.1: A cracked bi-axially loaded infinite material

It is straightforward to show that the weak and strong forms of the problem are equivalent if we assume, for example, that w is sufficiently regular (i.e. $w \in H^2(\Omega) \cap V$). However, it is not straightforward to show that w is regular: the regularity of w strongly depends on the data of the problem and on domain Ω when $d = 2$ or 3. This is quite different from what happened in 1D problems, because geometry there had no influence.

For example, considering the stress field around the crack tip in a bi-axially loaded infinite material, and following the notations in figure 2.1, one derives the following formula based on linear elasticity

$$\sigma_y = \sigma_\infty \sqrt{\frac{a}{2r}} \cos\left(\frac{\theta}{2}\right)\left(1 + \sin\left(\frac{\theta}{2}\right)\sin\left(\frac{3\theta}{2}\right)\right), \tag{2.5}$$

which is singular at the tip itself ($r = 0$).

As a remark, it should be said that, although we barely discuss the geometry in this course, it is often the controlling factor of the convergence of the finite element solution in terms of mesh refinement. This will be discussed further in the next chapter.

Remark 4 (vector equations) *Although we have only mentioned here the case of scalar equations (that is when the unknown u in equation 2.1 is a scalar quantity), the results extend to the vectorial case. In particular, this is necessary for the elasticity equations. In that case, the interpretation of the gradients and divergence is correspondingly vectorial, and the integration by part slightly different.*

2.3 Finite elements

We will first give the formal definition of finite elements and finite element meshes. We will then show how finite elements can be constructed from a reference finite element (or master element), and will define finite element spaces in which one will approximate the solution u of the above problem by the Galerkin method.

2.3.1 Definitions

The formal definition of a finite element has been proposed in the seventies by Ciarlet.

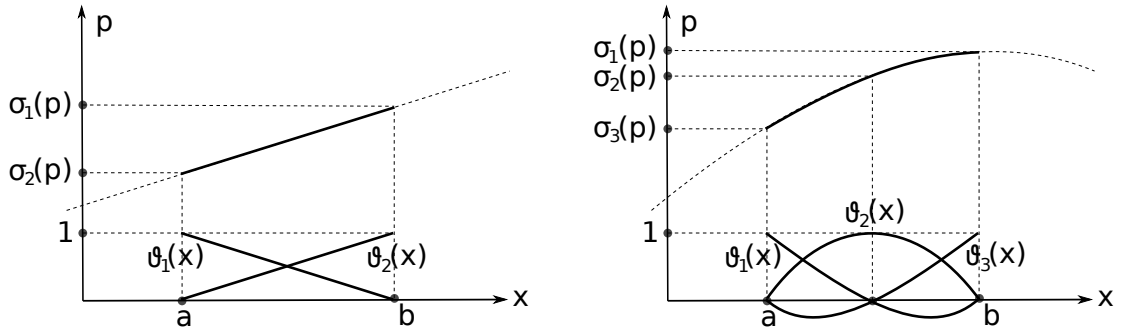


Figure 2.2: Support, sample function, degrees of freedom, and shape functions of a 1D linear (left figure) and quadratic (right figure) finite element.

Definition 1 (Finite Element) *A finite element consists of a triplet $\{K, P, \Sigma\}$ where:*

1. *K is a compact, connected, Lipschitz subset of $\mathbb{R}^d$ with non-empty interior.*
2. *P is a vector space of functions $p : K \to \mathbb{R}^m$ with m a positive integer (typically $m = 1$ for scalar-valued functions or $m = d$ for vector-valued functions).*
3. *Σ is a set of n linear forms $\sigma_i : P \to \mathbb{R}$, for $i = 1, \ldots, n$, i.e. $\sigma_i \in \mathcal{L}(P, \mathbb{R})$. Moreover, the linear mapping $\Psi : P \to \mathbb{R}^n$ such that $\Psi(p) = (\sigma_1(p), \ldots, \sigma_n(p))$ is bijective (unisolvence property). The linear forms σ_i are called the local degrees of freedom.*

Proposition 1 *Given a finite element $\{K, P, \Sigma\}$, there exists a basis $\{\theta_i\}$, $i = 1, \ldots, n$, in P such that:*

$$\sigma_i(\theta_j) = \delta_{ij}, \quad i, j = 1, \ldots, n \tag{2.6}$$

*where δ_{ij} denotes the "Kronecker delta" ($\delta_{ij} = 1$, if $i = j$, and 0, otherwise). The functions θ_i are usually called the **shape functions**.*

Example 6 (1D linear finite element) *In 1D (see figure 2.2, left), K is a closed and bounded interval $[a, b]$ where $a \neq b$. The space P consists of the linear functions $p(x) = \alpha + \beta x$, with $\alpha, \beta \in \mathbb{R}$ (in other words, P is the set of polynomial functions of degree 1 on $[a, b]$, that we shall denote by $\mathbb{P}_1$, i.e. $P = \mathbb{P}_1$). The degrees of freedom can be defined as $\sigma_1(p) = p(a)$ and $\sigma_2(p) = p(b)$. Indeed, let $p \in P$ and assume that $\sigma_1(p) = 0$ and $\sigma_2(p) = 0$. Then*

$$\left\{ \begin{array}{l} p(a) = \alpha + \beta a = 0 \\ p(b) = \alpha + \beta b = 0 \end{array} \right. \quad or \quad \left[\begin{array}{cc} 1 & a \\ 1 & b \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]$$

which necessarily implies that $\alpha = 0$ and $\beta = 0$ since $a \neq b$. This ensures that the map Ψ is in fact bijective. Note that from the unisolvence property, the dimension of the vector space P ($\dim(P) = \dim(\mathbb{P}_1) = 2$ here) is necessarily equal to the cardinality n of Σ ($n = 2$ here as well). For an arbitrary n, one could choose for P the space of polynomial functions of degree $n - 1$ defined on K and denoted by $\mathbb{P}_{n-1}$. The shape functions are defined as θ_1 and θ_2 such that:

$$\left\{ \begin{array}{l} \sigma_1(\theta_1) = \theta_1(a) = \alpha + \beta a = 1 \\ \sigma_2(\theta_1) = \theta_1(b) = \alpha + \beta b = 0 \end{array} \right. \quad or \quad \left[\begin{array}{cc} 1 & a \\ 1 & b \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] = \left[\begin{array}{c} 1 \\ 0 \end{array} \right]$$

which yields:

$$\alpha = \frac{b}{b-a}, \quad \beta = -\frac{1}{b-a} \quad \Rightarrow \quad \theta_1(x) = \frac{b-x}{b-a}$$

and

$$\begin{cases} \sigma_1(\theta_2) = \theta_2(a) = \alpha + \beta a = 0 \\ \sigma_2(\theta_2) = \theta_2(b) = \alpha + \beta b = 1 \end{cases} \quad or \quad \begin{bmatrix} 1 & a \\ 1 & b \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

which yields:

$$\alpha = -\frac{a}{b-a}, \quad \beta = \frac{1}{b-a} \quad \Rightarrow \quad \theta_2(x) = \frac{x-a}{b-a}$$

Denoting $(b-a)$ by h (the size of the element) and supposing that $a = x_{i-1}$ and $b = x_i$, we obtain:

$$\theta_1(x) = \frac{x_i - x}{h} \quad and \quad \theta_2(x) = \frac{x - x_{i-1}}{h}$$

In the particular case where $a = 0$ and $b = 1$, the shape functions simply reduce to $\theta_1(x) = 1 - x$ and $\theta_2(x) = x$.

Through the first project, we already observed that the size of the elements is an important controlling parameter for error estimation, and convergence of the approximate solution to the exact one. We give below a general definition of the size that generalizes the natural one over segments.

Definition 2 (mesh and mesh size) *Let Ω be a domain (open, bounded, connected, Lipschitz set) in $\mathbb{R}^d$. A mesh is a collection of a finite number N_{el} of compact, connected, Lipschitz sets K_m with non-empty interior such that $\{K_m\}$, $m = 1, \ldots, N_{el}$, forms a partition of Ω, i.e.*

$$\overline{\Omega} = \bigcup_{m=1}^{N_{el}} K_m, \qquad and \qquad \mathrm{Int}(K_m) \cap \mathrm{Int}(K_l) = \varnothing, \quad \forall m \neq l$$

The subsets K_m are called mesh cells or mesh elements (or simply elements, not to be confused with finite elements). A mesh $\{K_n\}$ is usually denoted by $\mathcal{T}_h$ ($\mathcal{T}$ as in triangulation) or by $\mathcal{P}_h$ ($\mathcal{P}$ as in partition). The parameter h refers to the mesh size of $\mathcal{T}_h$ and is defined as:

$$h = \max_{K \in \mathcal{T}_h} h_K$$

where h_K is the diameter of element K, i.e.

$$\forall K \in \mathcal{T}_h, \quad h_K = \mathrm{diam}\ K = \max_{x,y \in K} \|x - y\|_d$$

with $\|\cdot\|_d$ the Euclidean norm in $\mathbb{R}^d$, i.e. $\|x\|_d = \sqrt{x_1^2 + \ldots + x_d^2}$.

2.3.2 Examples of finite elements

Although other elements exist, the Lagrange finite elements certainly represent the type of finite elements most commonly used in commercial codes (e.g. Comsol Multiphysics). We therefore present several examples of Lagrange finite elements, and only present, at the end of this section, a short introduction to other types of elements.

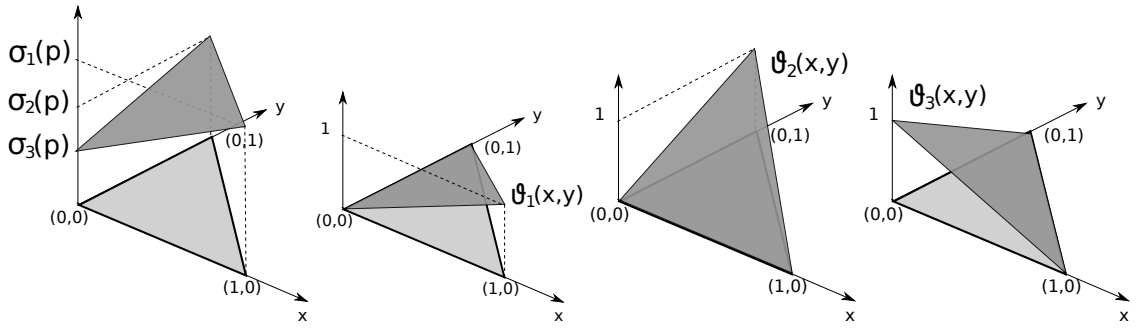


Figure 2.3: Support, sample function, degrees of freedom (left figure), and shape functions (3 right-most figures) of a 2D linear finite element.

Definition 3 (Lagrange finite element) *Let $\{K, P, \Sigma\}$ be a finite element. If there is a set of points $\{\xi_i, \ldots, \xi_n\}$ in K such that, for all $p \in P$, $\sigma_i(p) = p(\xi_i)$, $i = 1, \ldots, n$, $\{K, P, \Sigma\}$ is called a Lagrange finite element. The points $\{\xi_i, \ldots, \xi_n\}$ are called the nodes of the finite element, and the shape functions θ_i, which are such that $\theta_i(\xi_j) = \delta_{ij}$, $1 \leq i, j \leq n$, are called the nodal basis of P.*

The 1D linear finite element of Example 6 is a Lagrange finite element. If $K = [0, 1]$, the points $\xi_1 = 0$ and $\xi_2 = 1$ are the nodes and $\theta_1(x) = 1 - x$ and $\theta_2(x) = x$ are the nodal basis functions of $\mathbb{P}_1$. We see below the corresponding example for $d = 2$, and then generalize these "simplicial" type of element to higher dimension.

Example 7 *Let $d = 2$ and K be the unit triangle (we will later call it the unit simplex) with vertices $a_0 = (0, 0)$, $a_1 = (1, 0)$, and $a_2 = (0, 1)$. Let $P = \mathbb{P}_1$, i.e. $k = 1$. The nodes are given by $\xi_1 = a_0 = (0, 0)$, $\xi_2 = a_1 = (1, 0)$, and $\xi_3 = a_2 = (0, 1)$. Moreover, defining the shape functions as:*

$$\theta_1(x) = \theta_1(x_1, x_2) = 1 - x_1 - x_2$$
$$\theta_2(x) = \theta_2(x_1, x_2) = x_1$$
$$\theta_3(x) = \theta_3(x_1, x_2) = x_2$$

it is clear that they satisfy $\sigma_i(\theta_j) = \theta_j(\xi_i) = \delta_{ij}$. Therefore, $\{K, \mathbb{P}_1, \Sigma\}$ defines a Lagrange finite element.

Definition 4 (Simplicial Lagrange finite element) *Let $\{a_0, \ldots, a_d\}$ be a family of points in $\mathbb{R}^d$ such that the vectors $\{a_i - a_0\}$, $i = 1, \ldots, d$ are independent. Then the convex hull of $\{a_i\}$ is called a simplex and the points a_i are called the vertices of the simplex. The unit simplex of $\mathbb{R}^d$ is the set defined as:*

$$\{x \in \mathbb{R}^d;\ x_i \geq 0,\ i = 1, \ldots, d;\ x_1 + \ldots + x_d \leq 1\}$$

In 2D, a simplex is called a triangle, while in 3D, it is called a tetrahedron. In 1D, every interval of the line $\mathbb{R}$ satisfies the definition of a simplex.

Let K be a simplex in $\mathbb{R}^d$ with vertices $\{a_0, a_1, \ldots, a_d\}$. Let $P = \mathbb{P}_1$ be the space of polynomial functions of degree one in K; $\dim P = n = d + 1$. We define the nodes of the elements as $\xi_i = a_{i-1}$, $i = 1, \ldots, n$, and Σ the set of linear forms such that $\sigma_i(p) = p(\xi_i)$, $i = 1, \ldots, n$. Then, it is straightforward to prove that $\{K, P, \Sigma\}$ is a Lagrange finite element.

Example 8 *Let $d = 2$, $K = [-1, 1]\times[-1, 1]$, and $k = 1$. The set of nodes are given by $\xi_1 = (-1, -1)$, $\xi_2 = (1, -1)$, $\xi_3 = (1, 1)$, and $\xi_4 = (-1, 1)$. Note that the nodes coincide with the vertices of the element. Moreover, the shape functions are given by:*

$$\begin{aligned}\theta_1(x) &= \theta_1(x_1, x_2) = (1 - x_1)(1 - x_2)\\ \theta_2(x) &= \theta_2(x_1, x_2) = (1 + x_1)(1 - x_2)\\ \theta_3(x) &= \theta_3(x_1, x_2) = (1 + x_1)(1 + x_2)\\ \theta_4(x) &= \theta_4(x_1, x_2) = (1 - x_1)(1 + x_2)\end{aligned}$$

and satisfy $\sigma_i(\theta_j) = \theta_j(\xi_i) = \delta_{ij}$. The finite element $\{K, \mathbb{Q}_1, \Sigma\}$ thus defined is therefore a Lagrange finite element.

Besides the simplicial finite elements, there is another very important family of Lagrange elements, which are the tensor product finite elements. To define them, we need to introduce a tensor space of 1D polynomials. Let $\mathbb{Q}_k$ be this space of polynomial functions in the variables $x_1, \ldots, x_d$, of degree at most k in each variable. For $d = 1$, polynomials in $\mathbb{Q}_1$ and $\mathbb{Q}_2$ can be written as:

$$\begin{aligned}q \in \mathbb{Q}_1 :\ q(x) &= \alpha_{00} + \alpha_{10}x + \alpha_{01}y + \alpha_{11}xy\\ &= (\beta_{0,1} + \beta_{1,1}x)(\beta_{0,2} + \beta_{1,2}y)\\ q \in \mathbb{Q}_2 :\ q(x) &= \alpha_{00} + \alpha_{10}x + \alpha_{01}y + \alpha_{11}xy + \alpha_{20}x^2 + \alpha_{02}y^2 + \alpha_{21}x^2y + \alpha_{12}xy^2 + \alpha_{22}x^2y^2\\ &= (\beta_{0,1} + \beta_{1,1}x + \beta_{2,1}x^2)(\beta_{0,2} + \beta_{1,2}y + \beta_{2,2}y^2)\end{aligned}$$

The general definition of $\mathbb{Q}_k$, in terms of tensor product, can then be given as:

$$\mathbb{Q}_k = \left\{ q(x) = \prod_{j=1}^{d} \left(\sum_{i=0}^{k} \beta_{i,j} x_j^i \right), \quad \beta_{i,j} \in \mathbb{R} \right\}$$

from which it is clear that $\dim \mathbb{Q}_k = (k + 1)^d$. Note that in 1D, $\mathbb{Q}_k = \mathbb{P}_k$. We can then introduce the definition of tensor produce Lagrange finite elements.

Definition 5 (Tensor product Lagrange finite element) *Let $K = \Pi_{i=1}^{d}[a_i, b_i] \in \mathbb{R}^d$ where $[a_i, b_i]$ are intervals in $\mathbb{R}$. The element K is called a cuboid. For $x \in K$, there exists a unique vector $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$ such that $x_i = a_i + t_i(b_i - a_i)$, $i = 1, \ldots, d$.*

Let K be a cuboid in $\mathbb{R}^d$. Let $P = \mathbb{Q}_k$ with $k \geq 1$. Denote by n the dimension of P, i.e. $n = (k+1)^d$, and consider the set of nodes ξ_i, $i = 1, \ldots, n$, such that

$$\xi_i = \left(a_1 + (b_1 - a_1)\frac{i_1}{k}, \ldots, a_d + (b_d - a_d)\frac{i_d}{k} \right), \quad 0 \leq i_1, \ldots, i_d \leq k.$$

Finally, let Σ be the set of linear forms (degrees of freedom) such that $\sigma_i(p) = p(\xi_i)$, $i = 1, \ldots, n$. It can be proved that $\{K, P, \Sigma\}$ is a Lagrange finite element.

Example 9 *Let $d = 2$, $K = [-1, 1] \times [-1, 1]$, and $k = 2$. The set of nodes are given by*

$$\begin{array}{lll}\xi_1 = (-1, -1) & \xi_5 = (0, -1) & \xi_9 = (0, 0)\\ \xi_2 = (+1, -1) & \xi_6 = (+1, 0) & \\ \xi_3 = (+1, +1) & \xi_7 = (0, +1) & \\ \xi_4 = (-1, +1) & \xi_8 = (-1, 0) & \end{array}$$

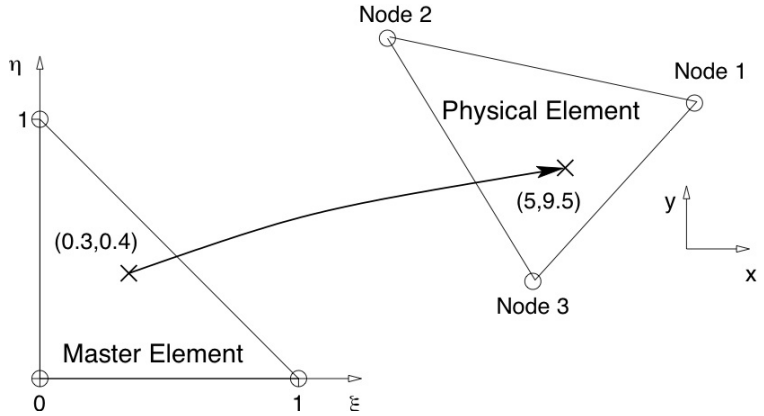


Figure 2.4: Mapping from the reference element $\hat{K}$ to the physical element K_m.

Moreover, the shape functions are given by:

$$\begin{array}{ll} \theta_1(x) = 0.25(x_1^2 - x_1)(x_2^2 - x_2) & \theta_5(x) = 0.5(1 - x_1^2)(x_2^2 - x_2) \\ \theta_2(x) = 0.25(x_1^2 + x_1)(x_2^2 - x_2) & \theta_6(x) = 0.5(x_1^2 + x_1)(1 - x_2^2) \\ \theta_3(x) = 0.25(x_1^2 + x_1)(x_2^2 + x_2) & \theta_7(x) = 0.5(1 - x_1^2)(x_2^2 + x_2) \\ \theta_4(x) = 0.25(x_1^2 - x_1)(x_2^2 + x_2) & \theta_8(x) = 0.5(x_1^2 - x_1)(1 - x_2^2) \end{array}$$

$$\theta_9(x) = (1 - x_1^2)(1 - x_2^2)$$

and satisfy $\sigma_i(\theta_j) = \theta_j(\xi_i) = \delta_{ij}$, $1 \leq i, j \leq 9$. *The finite element* $\{K, \mathbb{Q}_2, \Sigma\}$ *thus defined is therefore a Lagrange finite element.*

Other types of finite elements have been designed depending on the functions that need to be approximated. To name a few, there are the Crouzeix-Raviart finite element, the Raviart-Thomas finite element (to approximate functions in $H(\mathrm{div}, \Omega)$), the Nédélec finite element (to approximate functions in $H(\mathrm{curl}, \Omega)$), or hierarchical finite elements. For more details, we refer the reader to [2].

2.3.3 Reference finite element

In order to approximate functions on a domain Ω, the main idea is to associate with each element of the mesh a finite element. However, rather than to explicitly define all the finite elements in the mesh as shown above, it is more convenient to introduce a reference (or master) finite element $\hat{K}$ and construct the finite elements K_m through bijective mappings $T_m : \hat{K} \to K_m$, $m = 1, \ldots, N_e$ (see Fig. 2.4 in the case of a simplex in $\mathbb{R}^2$).

Let $\hat{K} \in \mathbb{R}^d$ be the unit simplex, i.e. $\hat{K} = \{\hat{x} \in \mathbb{R}^d;\ \hat{x}_i \geq 0,\ i = 1, \ldots, d;\ \hat{x}_1 + \ldots + \hat{x}_d \leq 1\}$, and let $\{\hat{K}, \hat{P}, \hat{\Sigma}\}$ be a Lagrange finite element, e.g. $\hat{P} = \mathbb{P}_1$. Affine mappings can then be defined as:

$$T_m(\hat{x}) = A_m \hat{x} + a_m$$

where A_m is a $d \times d$-matrix and a_m a vector of dimension d. The mappings T_m can in fact be defined in terms of the linear shape functions $\hat{\theta}_i$ associated with the $d + 1$ vertices of $\hat{K}$, as well as of the coordinates of the vertices a_i^m of K_m. Indeed,

$$T_m(\hat{x}) = \sum_{i=1}^{d+1} a_i^m \hat{\theta}_i(\hat{x})$$

The condition $\det A_m > 0$ would ensure that the mapping T_m be bijective; this condition holds whenever the vertices of K do not lie on the same line in 2D, nor on the same plane in 3D, and as long as the vertices are numbered counterclockwise.

Example 10 *Let $\hat{K}$ be the unit simplex in $\mathbb{R}^2$ and let K_m be the element in physical space with vertices (or nodes) $a_1^m = (4,5)$, $a_2^m = (0,6)$, and $a_3^m = (2,2)$. Then*

$$\begin{aligned} T_m(\hat{x}) &= a_1^m\hat{\theta}_1(\hat{x}) + a_2^m\hat{\theta}_2(\hat{x}) + a_3^m\hat{\theta}_3(\hat{x}) = a_1^m(1-\hat{x}_1-\hat{x}_2) + a_2^m\hat{x}_1 + a_3^m\hat{x}_2 \\ &= (a_2^m - a_1^m)\hat{x}_1 + (a_3^m - a_1^m)\hat{x}_2 + a_1^m \\ &= \begin{bmatrix} -4 \\ 1 \end{bmatrix}\hat{x}_1 + \begin{bmatrix} -2 \\ -3 \end{bmatrix}\hat{x}_2 + \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} -4 & -2 \\ 1 & -3 \end{bmatrix}\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned}$$

The mapping is bijective since $\det A_m = 12 + 2 = 14 > 0$.

Using the reference element, we construct a set of $\mathcal{T}_h$-based finite elements defined as the triplets $\{K_m, P_m, \Sigma_m\}$, $m = 1, \ldots, N_e$, such that

$$\begin{cases} K_m = T_m(\hat{K}) \\ P_m = \left\{ p = \hat{p} \circ T_m^{-1},\ \hat{p} \in \hat{P} \right\} \\ \Sigma_m = \{\sigma_{m,i},\ i = 1, \ldots, n;\ \sigma_{m,i}(p) = \hat{\sigma}_i(p \circ T_m),\ \forall p \in P_m\} \end{cases}$$

and shape functions on each element K_m are determined as

$$\theta_i^m = \hat{\theta}_i \circ T_m^{-1}, \quad i = 1, \ldots, n$$

Remark 5 *Note that $p \in P_m$, in particular $p(x) = \theta_i^m(x)$, is a polynomial function if and only if T_m is affine. For example, if $T_m(\hat{x}) = A_m\hat{x} + a_m$, then $T_m^{-1}(x) = A_m^{-1}x - A_m^{-1}a_m$, so that:*

$$\theta_i^m(x) = \hat{\theta}_i(A_m^{-1}x - A_m^{-1}a_m), \quad i = 1, \ldots, n$$

Because the mapping T_m^{-1} is also affine, the shape functions are necessarily polynomial functions.

2.3.4 Finite element space

Since we are usually interested in approximated functions in $H^1(\Omega)$, we introduce the following finite element spaces $U_h \subset H^1(\Omega)$ and $V_h \subset H^1(\Omega)$ as:

$$\begin{aligned} U_h &= \left\{ u_h \in C^0(\Omega_h);\ u_h|_{K_m} \in P,\ \forall m = 1, \ldots, N_e,\ u_h = u_d,\ \text{on } \Gamma_d \right\} \\ V_h &= \left\{ v_h \in C^0(\Omega_h);\ v_h|_{K_m} \in P,\ \forall m = 1, \ldots, N_e,\ v_h = 0,\ \text{on } \Gamma_d \right\} \end{aligned}$$

We assumed here that the function u_d can be exactly represented by functions in U_h on Γ_d. If this assumption does not hold, we may use, instead of u_d, a projection or interpolant of u_d onto U_h on the boundary Γ_d.

2.4 Galerkin approximation

Using the Galerkin approach, the finite element problem reads:

$$\boxed{\text{Find } u_h \in U_h \text{ such that} \quad B(u_h, v_h) = F(v_h), \quad \forall v \in V_h} \tag{2.7}$$

or, in the case of the model problem,

$$\text{Find } u_h \in U_h \text{ such that} \quad \int_\Omega k\nabla u_h \cdot \nabla v_h + cu_h v_h dx = \int_\Omega f v_h dx + \int_{\Gamma_n} g v_h ds, \quad \forall v \in V_h$$

Similarly to the 1D case, the above system of equations can be recast in matrix-vector form $KU = F$, in which the stiffness matrix K and loading vector F are obtained by an element-by-element assembly approach.

2.4.1 Integration of elemental matrices and load vectors

Let K_m be an element in the mesh $\mathcal{T}_h$ and let θ_i^m, $i = 1, \ldots, n$ be the shape functions associated with element K_m. Assume that k and c are constant. The elemental stiffness matrix reads:

$$K_{ij}^m = \int_{K_m} k\nabla\theta_j^m \cdot \nabla\theta_i^m + c\theta_j^m\theta_i^m \, dx = \int_{K_m} k\sum_{l=1}^{d} \frac{\partial\theta_j^m}{\partial x_l}\frac{\partial\theta_i^m}{\partial x_l} + c\theta_j^m\theta_i^m \, dx, \quad 1 \le i,j \le n$$

We consider here the first terms of the integral. Since shape functions are built with respect to the reference element, using the change of variables $x = T_m(\hat{x})$, $dx = |\det A_m| d\hat{x}$ (A_m is actually the Jacobian matrix of the transformation T_m) and the chain rule, we get

$$\int_{K_m} k\sum_{l=1}^{d} \frac{\partial\theta_j^m}{\partial x_l}\frac{\partial\theta_i^m}{\partial x_l} \, dx = \int_{\hat{K}} k\sum_{l=1}^{d} \left(\sum_{s=1}^{d} \frac{\partial\hat{\theta}_j}{\partial\hat{x}_s}\frac{\partial\hat{x}_s}{\partial x_l}\right)\left(\sum_{s=1}^{d} \frac{\partial\hat{\theta}_i}{\partial\hat{x}_s}\frac{\partial\hat{x}_s}{\partial x_l}\right) |\det A_m| \, d\hat{x}$$

From the definition of T_m, we also have:

$$\frac{\partial\hat{x}_s}{\partial x_l} = (A_m^{-1})_{s,l}$$

where the subscripts s, l indicate the s^{th} row and l^{th} column of matrix A_m^{-1}. Note that the partial derivatives $\partial\hat{\theta}_i/\partial\hat{x}_s$ can simply be computed from the definition of the shape functions on the reference element. In the same manner, the second integral of K_{ij}^m is given by:

$$\int_{K_m} c\theta_j^m\theta_i^m \, dx = \int_{\hat{K}} c\hat{\theta}_j\hat{\theta}_i |\det A_m| \, d\hat{x}$$

The elemental load vector reads:

$$F_i^m = \int_{K_m} f\theta_i^m dx + \int_{\partial K_m \cap \Gamma_n} g\theta_i^m ds, \quad 1 \le i \le n$$

For example, the first integral of F_i^m is computed on the reference element as:

$$\int_{K_m} f\theta_i^m dx = \int_{\hat{K}} f\hat{\theta}_i |\det A_m| \, d\hat{x}$$

Calculation of the second integral is left to the reader.

The integrals are never computed exactly. They are in fact approximated using numerical integration methods such as Gaussian quadratures.

	Abscissas	Weights	Truncation Error
n	$x_{n,k}$	$\omega_{n,k}$	$\mathcal{E}_n(f)$
2	± 0.5773502692	1.0000000000	$\frac{f^{(4)}(c)}{135}$
3	± 0.7745966692	0.5555555556	$\frac{f^{(6)}(c)}{15,750}$
	0.0000000000	0.8888888888	
4	± 0.8611363116	0.3478548451	$\frac{f^{(8)}(c)}{3,472,875}$
	± 0.3399810436	0.6521451549	

Table 2.1: Gauss-Legendre points and weights

2.4.2 Numerical integration by Gaussian quadratures

Suppose that we need to integrate a function $f = f(x)$ in the interval $[-1, 1]$. We know we can approximate that integral by:

$$\int_{-1}^{+1} f(x)dx \approx 2f(0) \tag{2.8}$$

and that, if the function f is linear, this approximation is exact. It turns out that the same kind of approximation can be generalized to higher-order polynomials, with also a better ratio number of points where f is evaluated versus order of the polynomials that are exactly integrated. These formulas are called Gauss-Legendre rules and are such as:

$$\int_1^1 f(x)dx = \sum_{k=1}^{n} \omega_{n,k} f(x_{n,k}) + \mathcal{E}_n(f) \approx \sum_{k=1}^{n} \omega_{n,k} f(x_{n,k})$$

where $x_{n,k}$ are the so-called Gauss-Legendre points with associated weights $\omega_{n,k}$ and $\mathcal{E}_n(f)$ is the truncation error defined in terms of the $(2n)^{th}$-derivative of f. Note that since the $(2n)^{th}$-derivative of polynomial functions of degree $(2n-1)$ vanishes, the integration of these polynomials by Gauss-Legendre n-point rules is exact (i.e. $\mathcal{E}_n(f) = 0$). We list below the first few Gauss points in Table 2.1.

Example 11 *Let $\hat{K} = [0,1] \times [0,1]$ and suppose that the Jacobian of T_m is affine (true if the physical element K_m is a parallelogram), i.e. A_m is constant. Choose $\hat{\theta}_i$ and $\hat{\theta}_j$ in $\mathbb{Q}_1$, e.g. $\hat{\theta}_i(\hat{x}) = \hat{\theta}_j(\hat{x}) = (1-\hat{x}_1)(1-\hat{x}_2)$. Then*

$$\begin{aligned}
\int_{\hat{K}} c\hat{\theta}_i\hat{\theta}_j |\det A_m| \, d\hat{x} &= c|\det A_m| \int_{-1}^{1}\int_{-1}^{1} (1-\hat{x}_1)^2(1-\hat{x}_2)^2 \, d\hat{x}_1 d\hat{x}_2 \\
&= c|\det A_m| \int_{-1}^{1} (1-\hat{x}_1)^2 \, d\hat{x}_1 \int_{-1}^{1} (1-\hat{x}_2)^2 \, d\hat{x}_2 \\
&= c|\det A_m| \sum_{k=1}^{2} \omega_{2,k}(1-x_{2,k})^2 \sum_{k=1}^{2} \omega_{2,k}(1-x_{2,k})^2
\end{aligned}$$

The integral is estimated exactly since the integrands are polynomials of degree strictly less than four.

2.4.3 Matrix and vector assembly

As in the 1D case, the global matrix K and global vector F are easily assembled from the elemental matrices $\{K_{ij}^m\}$ and elemental vectors $\{F_i^m\}$ using the connectivity array C. The matrix K and vector F are then modified to take into account for possible Dirichlet boundary conditions.

2.5 Conclusions

In this section, we have generalized the concepts that were discussed in the first section. We have also introduced formal definitions for the finite element, and discussed more numerical issues with the questions of integration (reference element and gauss quadrature). The next big step in our understanding of Finite Element Methods will come with a better understanding of error estimation. One of the important reasons for the success of finite element methods is in the error estimation analysis that they allow. We will present those in the next section. Afterwards, we will start addressing more particular cases, with different types of equations, starting with non-stationary problems.

2.6 Problems

2.6.1 Exercise 1

Let Ω be the domain shown in Fig. 2.5 with boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$. We consider the following boundary-value problem: Find $u = u(x)$ such that

$$\begin{aligned} -5\frac{\partial^2 u}{\partial x^2} - 4\frac{\partial^2 u}{\partial y^2} - 6\frac{\partial^2 u}{\partial x \partial y} + 2u &= f, \quad \text{in } \Omega \\ u &= g, \quad \text{on } \Gamma_d = \Gamma_1 \cup \Gamma_4 \\ 5\frac{\partial u}{\partial x} + 3\frac{\partial u}{\partial y} &= 2, \quad \text{on } \Gamma_2 \\ 3\frac{\partial u}{\partial x} + 4\frac{\partial u}{\partial y} &= 4, \quad \text{on } \Gamma_3 \end{aligned} \tag{2.9}$$

where f and g are given functions.

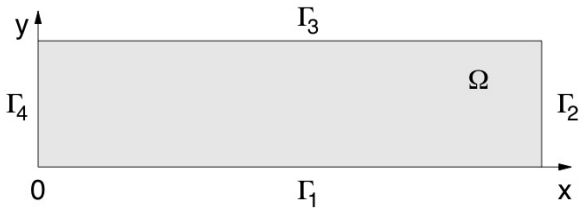


Figure 2.5: Domain Ω and boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ considered in Problem (2.9).

1. Rewrite the differential equation in the form

$$-\nabla \cdot a\nabla u + b \cdot \nabla u + cu = f, \quad \text{in } \Omega$$

   i.e., determine the tensor a, the vector b, and the coefficient c.

2. Derive the weak formulation of above problem and explicitly write the bilinear form $B(u, v)$ and linear form $F(v)$ appearing in the weak form.

   Note: you may want to use the following formula: $\nabla \cdot (qv) = (\nabla q) \cdot v + q\nabla \cdot v$, where q is a scalar in $\mathbb{R}$ and v a vector in $\mathbb{R}^2$.

3. We may define an adjoint problem, associated with above problem, which reads:

   $$\text{Find } p \in H^1(\Omega),\ p = 0 \text{ on } \Gamma_d, \text{ such that} \quad B(v, p) = Q(v), \quad \forall v \in H^1(\Omega),\ v = 0 \text{ on } \Gamma_d$$

   where $B(\cdot, \cdot)$ is the bilinear form determined in Question 2 and $Q(\cdot)$ is the linear functional defined as:

   $$Q(v) = \int_{\Gamma_2} v \, ds$$

   Derive the strong form of the dual problem. Note that $Q(v)$ is related to the average value of v along the boundary Γ_2.

4. Repeat Questions 2 and 3 in the case where b is now defined as the vector $b = [1, 1]^T$.

2.6.2 Exercise 2

Suppose that the domain Ω of Problem (2.9) is the rectangle $[0, 4] \times [0, 2]$ and that it is partitioned into eight elements as shown in Fig. 2.6. Suppose now that we want to compute the finite element solution to Problem (2.9) using elements with uniform polynomial degree $p = 2$. Compute the entries K_{14} and K_{41} of the element stiffness matrix for the element K shown in Fig. 2.6. Assume that the reference element (master element) is the square $[-1, 1] \times [-1, 1]$.

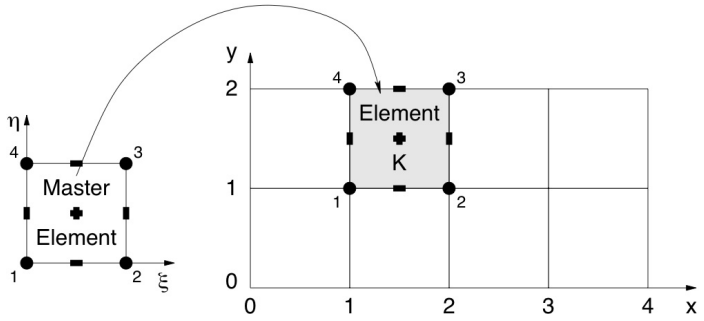


Figure 2.6: Finite element discretization of domain Ω considered in Problem (2.9).

The objective of this project is to simulate the temperature field in a one-bedroom apartment whose blueprint is shown below (all dimensions are in meters). It is composed of one living-room and one bedroom with a window in each room, a fireplace and entrance door in the living room. Let $\theta = \theta(x, y)$ denote the temperature field and let Ω be the computational domain with boundary $\partial\Omega = \overline{\Gamma \cup \Gamma_f \cup \Gamma_{1,w} \cup \Gamma_{2,w}}$ as shown below.

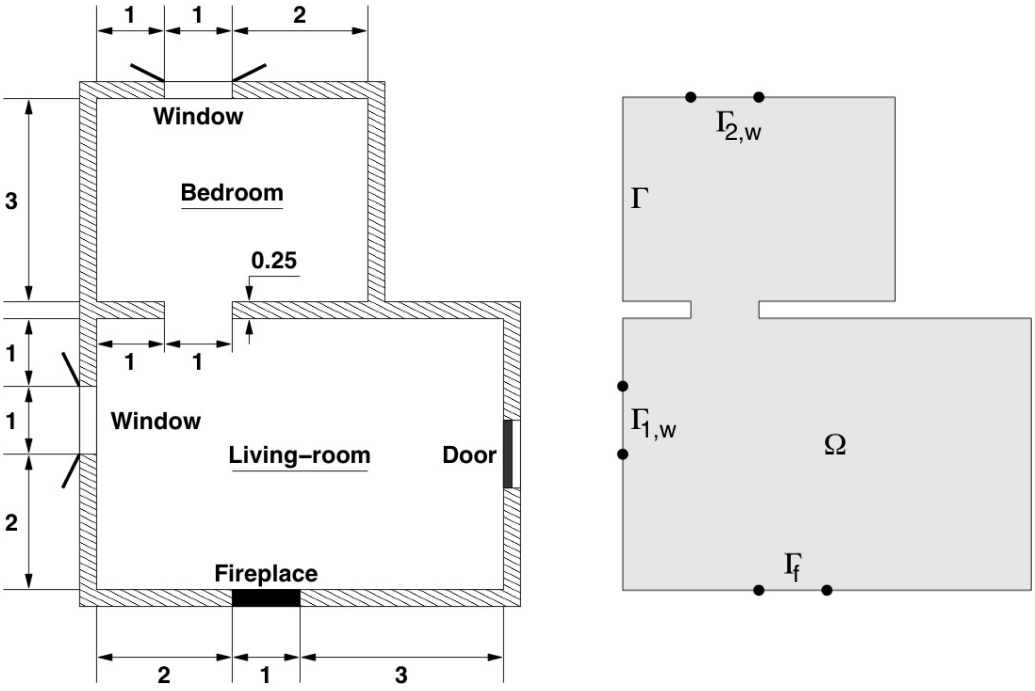
Problem 1. We shall first assume that θ is governed by the steady-state heat equation:

$$-\nabla \cdot k\nabla\theta = 0, \quad \forall x \in \Omega$$

where k is the thermal conductivity of air. In this model, there is no other heat source than heat from the fireplace whose temperature is known to be at θ_f. We shall first suppose that the walls and door are adiabatic (perfectly insulated) and that the windows are closed but that heat is exchanged with the outside air at temperature θ_o. The temperature in the apartment is thus subjected to the following boundary conditions:

$$\begin{cases} \theta = \theta_f, & \text{on } \Gamma_f \\ n \cdot (k\nabla\theta) = 0, & \text{on } \Gamma \\ n \cdot (k\nabla\theta) = -h_g(\theta - \theta_o), & \text{on } \Gamma_{1,w} \cup \Gamma_{2,w} \end{cases}$$

where h_g is the convection coefficient of the window. Take, for the experiments below, $k = 0.025$ W/(mK), $h_g = 20$ W/(m^{2}K), $\theta_f = 120°$C, and $\theta_o = 10°$C.



1. Write the weak formulation of above problem.

2. Develop an application in Comsol Multiphysics to compute a Finite Element approximation of above problem and design a mesh that should provide a reliable solution everywhere in the computational domain.

3. Suppose that you are interested in the temperature at the center of the bedroom. Design an "optimal" mesh that provides an accurate value of the temperature at that point and show the convergence of that temperature for a sequence of refined meshes.

4. Suppose now that the exterior walls and door of the apartment are not perfectly insulated. In other words, the heat flux along the door and walls is now given as:

$$n \cdot (k\nabla\theta) = -h_w(\theta - \theta_o)$$

where h_w is the convection coefficient of the walls and door (take for instance $h_w = 3$ W/(m^{2}K)). Predict the temperature at the center of the bedroom.

Problem 2. Suppose now that both windows are open and that air is blowing in through $\Gamma_{1,w}$ with velocity $u_d = (u_{1,d}, u_{2,d})$, while air is free to leave the apartment through the second window. As an approximation, the velocity and pressure fields are governed by the time-independent incompressible Stokes equations:

$$\begin{aligned}
-\nu\Delta u + \nabla p = 0, &\quad \text{in } \Omega & \\
\nabla \cdot u = 0, &\quad \text{in } \Omega & \\
u = u_d, &\quad \text{on } \Gamma_{1,w} & \text{(inflow BC)} \\
u = 0, &\quad \text{on } \Gamma & \text{(no slip BC)} \\
u \cdot n = 0 \text{ and } t \cdot (-pI + \nu(\nabla u + \nabla u^T))n = 0, &\quad \text{on } \Gamma_f & \text{(slip BC)} \\
pn - n \cdot \nu\nabla u = 0, &\quad \text{on } \Gamma_{2,w} & \text{(outflow BC)}
\end{aligned}$$

where ν is the kinematic viscosity of air ($\nu = 16 \times 10^{-6}$ m^{2}s^{-1}). t is the tangential vector to the boundary, and I the unit tensor. Consider $u_d = (u_{1,d}, u_{2,d}) = (1, 0)$ and $u_d = (u_{1,d}, u_{2,d}) = (0.8, 0.6)$ where the velocities are expressed in ms^{-1}.

The temperature is now governed by the equation and boundary conditions:

$$\begin{aligned}
-\nabla \cdot k\nabla\theta + \rho C u \cdot \nabla\theta = 0, &\quad \forall x \in \Omega \\
\theta = \theta_f, &\quad \text{on } \Gamma_f \\
n \cdot (k\nabla\theta) = 0, &\quad \text{on } \Gamma \\
\theta = \theta_o, &\quad \text{on } \Gamma_{1,w} \cup \Gamma_{2,w}
\end{aligned}$$

where $\rho = 1.2$ kg/m^3 and $C = 1000$ J/(kgK) are the density and the heat capacity of the air. Develop an application in Comsol Multiphysics to solve this coupled problem and design a mesh that provides a reliable approximation of the temperature at the center of the bedroom.

Problem 3. Create a 3D Finite Element model of the apartment and simulate the temperature for the conditions of Problem 2. Assume that the floor and ceiling are perfectly insulated.

Chapter3

Error Estimation in FEM

3.1 Introduction

The purpose of this chapter is to give a brief introduction to a priori and a posteriori error estimation for the finite element method. However, it is important to show beforehand that boundary-value problems and corresponding finite element problems derived by the Galerkin method are well-posed, in the sense that the problems admit a unique solution and that the solution continuously depends on the data (stability). There exist several existence theorems for linear boundary-value problems, in particular, the Lax-Milgram and Generalized Lax-Milgram theorems that we shall present below. Once the problems are shown to be well-posed, it is then possible to introduce without ambiguity the notion of error, defined as the difference between the solution u of the exact problem and the finite element solution u_h of the approximate problem, i.e. $e = u - u_h$. An important consideration in finite element methods is to make sure that the solution u_h actually converges to the solution u as the mesh is refined, or equivalently, that measures of the error e tend to zero as the mesh size h goes to zero. This is the primary objective in the derivation of a priori error estimates with, as byproduct, the determination of rates of convergence with respect to h. Unfortunately, a priori error estimates are useless to quantify the error in a given solution u_h as they involve constants as well as norms of the unknown solution u. The derivation of computable error estimates in a finite element solution u_h is the subject of a posteriori error estimation and at the basis of mesh adaptation.

3.2 Existence and uniqueness of solutions of BVP

Let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$, $d = 1$, 2, or 3, with boundary $\partial\Omega$, composed of two parts, Γ_d and Γ_n, such that $\partial\Omega = \overline{\Gamma_d \cup \Gamma_n}$. We consider the weak formulation of the abstract boundary-value problem:

$$\boxed{\text{Find } u \in U \text{ such that} \quad B(u,v) = F(v), \quad \forall v \in V} \tag{3.1}$$

where U denotes the space of admissible solutions, V the space of test functions, $B(\cdot,\cdot)$ is a bilinear form defined on $U \times V$, and $F(\cdot)$ a linear form defined on V.

There exist several existence theorems for boundary-value problems; the most important ones are certainly the Lax-Milgram and generalized Lax-Milgram theorems that we state below without proof.

3.2.1 Lax-Milgram Theorem

This theorem was first established by Peter Lax (1926–), Professor Emeritus at the Courant Institute of Mathematical Sciences, New York University, and Arthur Norton Milgram (1912–1961).

Theorem 1 (Lax-Milgram) *Let V be a Hilbert space with inner product $(\cdot,\cdot)_V$ and associated norm $\|\cdot\|_V = (\cdot,\cdot)_V$ and let $U = V$. Moreover, the following conditions hold:*

1) $\exists\ M > 0$ such that $\forall u, v \in V$, $|B(u,v)| \leq M\|u\|_V\|v\|_V$ (Continuity of $B(\cdot,\cdot)$).

2) $\exists\ C > 0$ such that $\forall v \in V$, $|F(v)| \leq C\|v\|_V$ (Continuity of $F(\cdot,\cdot)$).

3) $\exists\ \alpha > 0$ such that $\forall u \in V$, $B(u,u) \geq \alpha\|u\|_V^2$ (Coercivity of $B(\cdot,\cdot)$).

Then, Problem (3.1) is well-posed (in Hadamard's sense), i.e. there exists a solution, the solution is unique and it continuously depends on the data according to the following estimate:

$$\|u\|_V \leq \frac{C}{\alpha}$$

Example 12 *We consider the following problem: Find u such that*

$$\begin{cases} -\nabla\cdot(k\nabla u) + cu = f, & \text{in } \Omega \subset \mathbb{R}^d \\ u = 0, & \text{on } \Gamma_d \\ n\cdot(k\nabla u) = g, & \text{on } \Gamma_n \end{cases} \tag{3.2}$$

where n is the unit outward normal vector to the boundary, $f = f(x)$, $c = c(x)$, $k = k(x)$, $x \in \Omega$, and $g = g(x)$, $x \in \Gamma_n$, are given scalar functions (the data). We assume that $f \in L^2(\Omega)$, $g \in L^2(\Gamma_n)$, and, for simplicity, that $k, c \in C^0(\bar{\Omega})$ and that there exist k_{min}, $k_{max} \in \mathbb{R}$ and c_{min}, $c_{max} \in \mathbb{R}$ such that:

$$\begin{aligned} 0 < k_{min} \leq k(x) \leq k_{max}, \qquad & \forall x \in \Omega \\ 0 < c_{min} \leq c(x) \leq c_{max}, \qquad & \forall x \in \Omega \end{aligned}$$

Let V be given such as:

$$V = \{v \in H^1(\Omega);\ \gamma v = 0 \text{ on } \Gamma_d\}$$

where γ is the trace operator, which maps a function defined on Ω into a function on Γ_d. Introducing the bilinear form and linear form as:

$$\begin{aligned} B(u,v) &= \int_\Omega k\nabla u\cdot\nabla v + cuv dx \\ F(v) &= \int_\Omega fv dx + \int_{\Gamma_n} gv ds \end{aligned}$$

the above classical boundary-value problem can be recast in weak form as in (3.1) with $U = V$.

We now show that the conditions of the Lax-Milgram Theorem hold for the above problem. The norm in V is chosen as the H^1 norm (we could also consider the H^1 seminorm as we know that it is equivalent in V to the H^1 norm due to the Poincaré inequality, i.e. $\forall v \in V$, $\|v\|_0 \leq C_p|v|_1$):

$$\|v\|_1 = \sqrt{\int_\Omega |\nabla v|^2 + |v|^2 dx}$$

First, since k and c are bounded above, we have, using Cauchy-Schwarz,

$$|B(u,v)| \leq k_{max}\int_\Omega |\nabla u \cdot \nabla v| dx + c_{max}\int_\Omega |uv| dx \leq M\int_\Omega |\nabla u \cdot \nabla v| + |uv| dx \leq M\|u\|_1\|v\|_1$$

where we have introduced the continuity constant $M = \max(k_{max}, c_{max})$.

Next, we observe that for $u \in V$:

$$B(u,u) \geq k_{min}\int_\Omega (\nabla u)^2 dx + c_{min}\int_\Omega u^2 dx \geq \alpha\int_\Omega (\nabla u)^2 + u^2 dx = \alpha\|u\|_1^2$$

which shows that $B(\cdot,\cdot)$ is coercive with coercivity constant $\alpha = \min(k_{min}, c_{min})$.

Finally, F is continuous since:

$$\begin{aligned} |F(v)| &\leq \int_\Omega |fv| dx + \int_{\Gamma_n} |gv| ds \leq \|f\|_0\|v\|_0 + \|g\|_{0,\Gamma_n}\|v\|_{0,\Gamma_n} \\ &\leq \|f\|_0\|v\|_1 + C_n\|g\|_{0,\Gamma_n}\|v\|_1 \leq C\|v\|_1 \end{aligned}$$

with continuity constant $C = (\|f\|_0 + C_n\|g\|_{0,\Gamma_n})$. We conclude that the problem is well-posed and that the solution is bounded by C/α.

Example 13 *Let V be the finite-dimensional space $\mathbb{R}^n$ with norm the standard Euclidean norm $\|u\| = \sqrt{u_1^2 + \ldots + u_n^2} = \sqrt{u^T u}$. In this case, the bilinear form can be represented in terms of a matrix A in $\mathbb{R}^{n\times n}$ such that:*

$$B(u,u) = u^T A u$$

Let λ be an eigenvalue of A with associated eigenvector u, so that $Au = \lambda u$. It follows that

$$\frac{u^T A u}{u^T u} = \frac{u^T \lambda u}{u^T u} = \lambda$$

If B is coercive, then the eigenvalues are all strictly positive, i.e. A is positive definite, that is, $\det A > 0$, which means the matrix is invertible. Moreover, the coercivity constant is simply the smallest eigenvalue of A.

More general theorems for well-posedness of boundary-value problems are available. In particular, we have the so-called generalized Lax-Milgram theorem which weakens the coercivity condition.

3.2.2 Generalized Lax-Milgram Theorem

The generalized Lax-Milgram theorem extends the existence and uniqueness result of the Lax-Milgram theorem to the case of non-Hilbert spaces. This theorem has been widely popularized by Babuška (1926–), Professor at the Institute for Computational Engineering and Sciences at The University of Texas at Austin, in the case of the finite element method.

Theorem 2 (Generalized Lax-Milgram) *Let U be a Banach space and V be a reflexive Banach space with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively. The following conditions hold:*

1) B is continuous on $U \times V$.

2) F is continuous on V.

3) Inf-sup condition: $\exists \alpha > 0$ such that

$$\inf_{u \in U} \sup_{v \in V} \frac{|B(u,v)|}{\|u\|_U \|v\|_V} \geq \alpha \tag{3.3}$$

4) B satisfies the property:

$$\forall v \in V, \qquad (B(u,v) = 0, \quad \forall u \in U) \Rightarrow v = 0 \tag{3.4}$$

Problem (3.1) is well-posed and the solution continuously depends on the data, i.e. $\|u\|_U \leq C/\alpha$.

Remark 6 *Note that the coercivity of B implies that B satisfies (3.3) and (3.4). The contrary is not necessarily true. Moreover, the inf-sup condition (3.3) is often restated as:*

$$\exists \alpha > 0, \quad \forall u \in U, \qquad \sup_{v \in V} \frac{|B(u,v)|}{\|v\|_V} \geq \alpha$$

and referred to as the LBB condition (after Ladyzenskaya, Babuška, and Brezzi).

Example 14 *Let $U = \mathbb{R}^m$ and $V = \mathbb{R}^n$. Then given a matrix $A \in \mathbb{R}^{n \times m}$, the inf-sup condition reads (since the infimum and supremum are attained in finite dimensional spaces):*

$$\alpha = \min_{u \in U} \max_{v \in V} \frac{v^T A u}{\|u\|_U \|v\|_V}$$

However, we know that the maximum is attained for $v = Au$. Therefore:

$$\alpha = \min_{u \in U} \frac{u^T A^T A u}{\|u\|_U \|Au\|_V} = \min_{u \in U} \frac{\|Au\|_V^2}{\|u\|_U \|Au\|_V} = \min_{u \in U} \frac{\|Au\|_V}{\|u\|_U} = \min_{u \in U} \sqrt{\frac{u^T A^T A u}{u^T u}}$$

As before, we can conclude that the inf-sup constant is given by:

$$\alpha = \sqrt{\lambda_{\min}(A^T A)}$$

where $\lambda_{\min}(A^T A)$ is the smallest eigenvalue of $A^T A$, i.e. α is the smallest singular value of A.

3.3 Finite element problem and approximation error

Let $U_h \subset U$ and $V_h \subset V$ be two conforming finite element subspaces such that $\dim U_h = \dim V_h$. A finite element approximation of Problem 3.1 can be given as follows:

$$\boxed{\text{Find } u_h \in U_h \text{ such that} \quad B(u_h, v_h) = F(v_h), \quad \forall v_h \in V_h} \tag{3.5}$$

The main issue is to analyze the error $e = u - u_h$, which, in the case of conforming spaces, belongs to U, and to study the convergence of u_h to u as the mesh size h goes to zero, with p fixed, or, as the polynomial degree p is increased, keeping h fixed, or by varying both h and p.

3.3.1 Error equation and Galerkin orthogonality

Substracting $B(u_h, v)$ from both sides of (3.1) and recalling that $B(\cdot, \cdot)$ is bilinear, the approximation error $e = u - u_h$ satisfies:

$$B(u, v) - B(u_h, v) = B(u - u_h, v) = B(e, v) = F(v) - B(u_h, v) \equiv \mathcal{R}(v), \qquad \forall v \in V$$

where $\mathcal{R}(v)$ is the residual. The residual is a linear functional defined on V and represents the sources of error due to the finite element discretization. The problem for the error can simply be written as:

$$\boxed{\text{Find } e \in U \text{ such that} \quad B(e, v) = \mathcal{R}(v), \qquad \forall v \in V} \tag{3.6}$$

Moreover,

$$B(e, v_h) = \mathcal{R}(v_h) = F(v_h) - B(u_h, v_h) = F(v_h) - F(v_h) = 0, \qquad \forall v_h \in V_h$$

This is the so-called "Galerking orthogonality property" which plays a fundamental role in error estimation.

3.3.2 Error estimate for the coercive case

We first assume that $U = V$, $U_h = V_h$, such that $V_h \subset V$, and that B and F satisfy the conditions of the Lax-Milgram Theorem. It immediately follows that the finite element problem (3.5) is well-posed since $V_h \subset V$.

Using the coercivity B, introducing an arbitrary function $w_h \in V_h$, using the orthogonality property, and finally the continuity of B, we get:

$$\begin{aligned} \alpha \|e\|_V^2 \leq B(e, e) &= B(e, u - u_h) \\ &= B(e, u - w_h + w_h - u_h) = B(e, u - w_h) \\ &\leq M \|e\|_V \|u - w_h\|_V \end{aligned}$$

so it immediately follows that:

$$\|e\|_V \leq \frac{M}{\alpha} \|u - w_h\|_V$$

and since w_h is arbitrary, we can write:

$$\|e\|_V \leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|_V$$

Remark 7 *This estimate can be further sharpened if the bilinear form $B(\cdot, \cdot)$ is also symmetric, i.e. $B(u, v) = B(v, u)$, $\forall u,\ v \in V$. Indeed, being both symmetric and positive definite, it defines an inner product with associated norm $\| \cdot \|_e$ (the so-called energy norm):*

$$\|v\|_e = \sqrt{B(v, v)}, \qquad \forall v \in V$$

From the definition of the energy norm and upon using Cauchy-Schwarz, note that B is continuous with continuity constant $M_e = 1$ and coercive with coercivity constant $\alpha_e = 1$, i.e.

$$\begin{aligned} &B(u, v) \leq \|u\|_e \|v\|_e, \quad \forall u,\ \forall v \in V \\ &B(u, u) = \|u\|_e^2,\ \forall u \in V \end{aligned}$$

Since $e \in V$, we then have, for an arbitrary $w_h \in V_h$:

$$\begin{aligned}
\|e\|_e^2 = B(e,e) &= B(u - w_h + w_h - u_h, u - w_h + w_h - u_h) \\
&= B(u - w_h, u - w_h) + B(w_h - u_h, w_h - u_h) + 2B(u - w_h, w_h - u_h) \\
&= B(u - w_h, u - w_h) + B(w_h - u_h, w_h - u_h) - 2B(w_h - u_h, w_h - u_h) \\
&= B(u - w_h, u - w_h) - B(w_h - u_h, w_h - u_h) \\
&= \|u - w_h\|_e^2 - \|w_h - u_h\|_e^2 \\
&\leq \|u - w_h\|_e^2
\end{aligned}$$

where we have used the fact that $B(u_h, v_h) = F(v_h) = B(u, v_h)$, $\forall v_h \in V_h$. It follows that an error estimate with respect to the energy norm is given as:

$$\|e\|_e \leq \inf_{w_h \in V_h} \|u - w_h\|_e$$

The energy norm is equivalent to the V-norm since:

$$\alpha \|v\|_V^2 \leq B(v,v) = \|v\|_e^2 \leq M \|v\|_V^2, \quad \forall v \in V$$

Combining above inequalities yields:

$$\|e\|_V \leq \frac{1}{\sqrt{\alpha}} \|e\|_e \leq \frac{1}{\sqrt{\alpha}} \inf_{w_h \in V_h} \|u - w_h\|_e \leq \sqrt{\frac{M}{\alpha}} \inf_{w_h \in V_h} \|u - w_h\|_V$$

Example 15 *For the problem of Example 12, we observe that the bilinear form is symmetric. Indeed,*

$$B(u,v) = \int_\Omega k \nabla u \cdot \nabla v + cuv dx = \int_\Omega k \nabla v \cdot \nabla u + cvu dx = B(v,u), \quad \forall u, v \in V$$

and since the form satisfies the coercivity condition, we know that B defines an inner product with associated norm:

$$\|u\|_e = \sqrt{\int_\Omega k(\nabla u)^2 + cu^2 dx}$$

The error in a Galerkin approximation u_h of u can be estimated in the energy norm as:

$$\|e\|_e \leq \inf_{w_h \in V_h} \|u - w_h\|_e$$

or, in the H^1 norm as:

$$\|e\|_1 \leq \sqrt{\frac{\max(k_{max}, c_{max})}{\min(k_{min}, c_{min})}} \inf_{w_h \in V_h} \|u - w_h\|_1$$

We conclude that, whenever possible, it is usually better to use the energy norm rather than other norms.

3.3.3 Error estimate for the non-coercive case

Since $U_h \subset U$ and $V_h \subset V$, we know that the bilinear form B and linear form F are continuous on $U_h \times V_h$ and on V_h, respectively. To ensure that the finite element problem (3.5) has a unique solution, the conditions (3.3) and (3.4) need to be satisfied with respect to the finite element spaces, i.e.:

$$\exists \alpha_h > 0, \qquad \inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{|B(u_h, v_h)|}{\|u_h\|_U \|v_h\|_V} \geq \alpha_h \tag{3.7}$$

$$\forall v_h \in V_h, \qquad (B(u_h, v_h) = 0, \quad \forall u_h \in U_h) \Rightarrow v_h = 0 \tag{3.8}$$

However, if dim U_h = dim V_h, these two conditions are equivalent and the main condition that needs to hold for existence and uniqueness of finite element solutions is the so-called discrete inf-sup condition (3.7). Note that α_h is in general different from α.

Theorem 3 (Céa's Lemma) *Let the discrete inf-sup condition hold with $U_h \subset U$, $V_h \subset V$, and dim U_h = dim V_h. Let u and u_h be the solutions of the boundary-value problem (3.1) and of the finite element problem (3.5), respectively. Then the error $e = u - u_h$ satisfies the following estimate:*

$$\|e\|_U \leq \left(1 + \frac{M}{\alpha_h}\right) \inf_{w_h \in U_h} \|u - w_h\|_U$$

Proof. Let $v_h \in V_h$. The triangle inequality yields

$$\|e\|_U \leq \|u - w_h\|_U + \|u_h - w_h\|_U$$

From the orthogonality condition, we have:

$$B(u_h - w_h, v_h) = B(u - w_h, v_h) - B(u - u_h, v_h) = B(u - w_h, v_h), \qquad \forall v_h \in V_h$$

Then, using the discrete inf-sup condition and continuity of B, we obtain:

$$\alpha_h \|u_h - w_h\|_U \leq \sup_{v_h \in V_h} \frac{|B(u_h - w_h, v_h)|}{\|v_h\|_V} \leq \sup_{v_h \in V_h} \frac{|B(u - w_h, v_h)|}{\|v_h\|_V} \leq M \|u - w_h\|_U$$

which completes the proof.

3.4 A priori error estimation and rate of convergence

The objective of a priori error estimation for the finite element method is to provide an estimate of the term:

$$\inf_{w_h \in U_h} \|u - w_h\|_U$$

A particular choice for w_h is given by the interpolant of u in V_h, i.e. $w_h = \mathcal{I}_h u$, we would be able to have: where we define the interpolation operator as:

$$\mathcal{I}_h : U \longrightarrow U_h, \quad \text{such that} \quad \mathcal{I}_h u = \sum_{i=1}^{N} \sigma_i(u) \phi_i$$

where N is the number of degrees of freedom, i.e. $N = \dim U_h$, σ_i are the degrees of freedom, and ϕ_i are the basis functions in U_h. Note that other choices for w_h could be considered, such as the projection of u on U_h. However, the advantage of using the interpolation operator is the fact that estimates of the interpolation error have been derived with respect to finite element spaces.

For instance, in the case of Lagrange finite elements (those used in Comsol Multiphysics), the interpolation error can be estimated as, given a function $u \in H^{r+1}(\Omega)$, $r \geq 0$, when using elements of order $p \geq r$,

$$\boxed{\begin{aligned} \|u - \mathcal{I}_h^p u\|_0 &\leq C_I h^{r+1} |u|_{r+1} \\ \|u - \mathcal{I}_h^p u\|_1 &\leq C_I h^{r} |u|_{r+1} \end{aligned}}$$

or simply:

$$\boxed{\begin{aligned} \|u - \mathcal{I}_h^p u\|_0 &\leq C_I h^{\min(p,r)+1} |u|_{r+1} \\ \|u - \mathcal{I}_h^p u\|_1 &\leq C_I h^{\min(p,r)} |u|_{r+1} \end{aligned}}$$

where $C_I > 0$ is a constant independent of h. The exponent of h will define the rate of convergence of the method.

Example 16 *Returning to the problem of Example 12, if the data f and g are such that we know that the solution u belongs to $H^2(\Omega)$, and if we use linear finite elements, i.e. $p = 1$, we thus obtain the following interpolation estimate:*

$$\|u - \mathcal{I}_h^1 u\|_1 \leq C_I h |u|_2$$

so that an a priori estimate of the finite element error is:

$$\|e\|_1 \leq \sqrt{\frac{\max(k_{max}, c_{max})}{\min(k_{min}, c_{min})}} \|u - \mathcal{I}_h^1 u\|_1 \leq \sqrt{\frac{\max(k_{max}, c_{max})}{\min(k_{min}, c_{min})}} C_I h |u|_2$$

The rate of convergence is $k = 1$. If we know that the solution u belongs to $H^2(\Omega)$, and use quadratic finite elements, i.e. $p = 2$, we would obtain the estimate:

$$\|e\|_1 \leq \sqrt{\frac{\max(k_{max}, c_{max})}{\min(k_{min}, c_{min})}} C_I h^2 |u|_3$$

with rate of convergence $k = 2$.

In practice, the exact solution u, the interpolation constant C_I, and the continuity and coercivity constants are usually unknown; a priori error estimates cannot be employed to quantify the error, but they are nevertheless useful to compute rates of convergence, which measure how fast (or how slow) the finite element solution u_h converges to the exact solution u of a given problem. Suppose, for example, that we would like to know the rate of convergence k of a given method. We can pose:

$$\|e\|_U \leq C_e h^k$$

with C_e an unknown positive constant. Then

$$\log \|e\|_U \leq \log C_e + k \log h$$

Convergence studies are frequently used to verify the correct implementation of finite element codes, an activity usually referred to as Code Verification. In this case, a problem with known exact solution u is considered such that the error $u - u_h$ can be exactly computed. However, because the constant C_e is unknown, we need to introduce a sequence of meshes $\mathcal{T}_{h_i}$, so that the rate of convergence can be estimated as:

$$k_{i,\text{est}} \approx \frac{\log \|e_{h_i}\|_U - \log \|e_{h_{i+1}}\|_U}{\log h_i/h_{i+1}}$$

The study can be performed on uniform meshes such that $h_i/h_{i+1} = 2$. For example, in 1D, with $\Omega = (0, 1)$, we would consider meshes with 2, 4, 8, ... elements so that $h_1 = 1/2$, $h_2 = 1/4$, $h = 1/8$, ... The rate of convergence k would then be obtained as the limit of $k_{i,\text{est}}$ as i goes to infinity, i.e.:

$$k = \lim_{i \to \infty} k_{i,\text{est}}$$

For better comparison between methods, it is usually more convenient to represent the relation $\log \|e_h\|$ versus $\log N$ (rather than $\log h$), where N is the number of degrees of freedom. In 1D, with $\Omega = (0, 1)$, the relationship between h and N is given by $h \approx p/N$, so that:

$$\log \|e\|_U \leq \log C_e + k \log(p/N) = (\log C_e + k \log p) - k \log N$$

and

$$k_{i,\text{est}} \approx \frac{\log \|e_{h_i}\|_U - \log \|e_{h_{i+1}}\|_U}{\log N_{i+1}/N_i}$$

Finally, when the exact solution of the problem is unknown, we need to estimate the error in order to derive the rate of convergence. This can be accomplished by computing $e_{h_i} \approx u_{h_{i+1}} - u_{h_i}$. The rate of convergence is then calculated as:

$$k_{i,\text{est}} \approx \frac{\log \|u_{h_{i+1}} - u_{h_i}\|_U - \log \|u_{h_{i+2}} - u_{h_{i+1}}\|_U}{\log N_{i+1}/N_i}$$

An alternative approach is to solve for a reference solution $\tilde{u} \approx u$, if possible, on a very fine mesh and to compute:

$$k_{i,\text{est}} \approx \frac{\log \|\tilde{u} - u_{h_i}\|_U - \log \|\tilde{u} - u_{h_{i+1}}\|_U}{\log N_{i+1}/N_i}$$

3.5 A brief introduction to a posteriori error estimation

At this point, we understand that we cannot use a priori error estimates to evaluate the error in a given solution u_h of the finite element problem. This is indeed the subject of a posteriori error estimation, where "a posteriori" stands here for the fact that such approaches require the knowledge of u_h.

There are essentially two types of a posteriori error estimates: those defined with respect to norms, usually the energy norm (e.g. recovery-type estimators, explicit and implicit residual estimators), and those defined with respect to quantities of interest, referred to as goal-oriented error estimators (which of course involve the use of the adjoint problem). Interest in the development of a posteriori error estimators dates back to the mid-seventies with the work of Ladevèze (ENS Cachan) and

that of Babuška (ICES, UT Austin) and Rheinboldt. Since that time, several methods have been proposed and it would be too long to describe all of them, even a few, in this sequel.

We propose here to briefly present the *explicit residual method* to show how it is possible to obtain computable estimates and to explain how these can be used as refinement indicators for mesh adaptation. We describe the approach on the problem of Example 12. From Remark 7, using the equation for the error, as well as the orthogonality property, we have:

$$\alpha\|e\|_1^2 \leq \|e\|_e^2 = B(e,e) = R(e)$$

Now, for all $v \in V$, we can compute:

$$\begin{aligned} R(v) &= F(v) - B(u_h, v) \\ &= \int_\Omega fv\,dx + \int_{\Gamma_n} gv\,ds - \int_\Omega k\nabla u_h \cdot \nabla v + cu_h v\,dx \\ &= \sum_{K\in\mathcal{T}_h} \left\{ \int_K fv\,dx + \int_{K\cap\Gamma_n} gv\,ds - \int_K k\nabla u_h \cdot \nabla v + cu_h v\,dx \right\} \\ &= \sum_{K\in\mathcal{T}_h} \left\{ \int_K (f + \nabla\cdot k\nabla u_h - cu_h)v\,dx + \int_{K\cap\Gamma_n} gv\,ds - \int_{\partial K} n\cdot(k\nabla u_h)v\,ds \right\} \end{aligned}$$

where we have decomposed the global integral into a sum of elemental integrals and integrated by parts these integrals. We introduce the interior residuals r_K over each element K and flux jumps j_γ over each in element interface γ such as:

$$r_K = f + \nabla\cdot k\nabla u_h - cu_h$$

$$j_\gamma = \begin{cases} g - n\cdot(k\nabla u_h) & \text{on } \gamma = \partial K \cap \Gamma_n \\ \langle n\cdot(k\nabla u_h)\rangle & \text{on } \gamma = \partial K \cap \partial L \\ 0 & \text{on } \gamma = \partial K \cap \Gamma_d \end{cases}$$

where $\langle n\cdot(k\nabla u_h)\rangle$ denotes the averaged flux at the interface of two elements, i.e.:

$$\langle n\cdot(k\nabla u_h)\rangle = \frac{1}{2}\left(n_K\cdot(k\nabla u_h)_K + n_L\cdot(k\nabla u_h)_L\right), \qquad \forall K, L \in \mathcal{T}_h,\ K \neq L$$

We then have

$$R(v) = \sum_{K\in\mathcal{T}_h} \left[\int_K r_K v\,dx + \sum_{\gamma\subset\partial K} \int_\gamma j_\gamma v\,ds \right]$$

Using the orthogonality property, Cauchy-Schwarz inequality, and making use of the relation $ab + cd \leq \sqrt{a^2+c^2}\sqrt{b^2+d^2}$, we deduce that:

$$\begin{aligned} |R(v)| = |R(v-v_h)| &\leq \sum_{K\in\mathcal{T}_h} \left[\|r_K\|_{0,K}\|v-v_h\|_{0,K} + \sum_{\gamma\subset\partial K} \|j_\gamma\|_{0,\gamma}\|v-v_h\|_{0,\gamma} \right], \qquad \forall v_h \in V_h \\ &\leq \sum_{K\in\mathcal{T}_h} \left[\|r_K\|_{0,K}\|v-v_h\|_{0,K} + \|j_\gamma\|_{0,\partial K}\|v-v_h\|_{0,\partial K} \right], \qquad \forall v_h \in V_h \end{aligned}$$

Using special interpolation functions of v in V_h, we can show, skipping many steps in the demonstration, that:

$$|R(v)| \leq C_e\|v\|_1 \sqrt{\sum_{K\in\mathcal{T}_h} \left[h_K^2\|r_K\|_{0,K}^2 + h_K\|j_\gamma\|_{0,\partial K}^2 \right]}$$

with C_e a positive constant independent of h. Substituting v for e, we then get:

$$\alpha\|e\|_1^2 \leq R(e) \leq C_e\|e\|_1\sqrt{\sum_{K\in\mathcal{T}_h}\left[h_K^2\|r_K\|_{0,K}^2 + h_K\|j_\gamma\|_{0,\partial K}^2\right]}$$

that is:

$$\|e\|_1 \leq \frac{C_e}{\alpha}\sqrt{\sum_{K\in\mathcal{T}_h}\left[h_K^2\|r_K\|_{0,K}^2 + h_K\|j_\gamma\|_{0,\partial K}^2\right]}$$

We see that the quantity η, such as:

$$\eta = \sqrt{\sum_{K\in\mathcal{T}_h}\left[h_K^2\|r_K\|_{0,K}^2 + h_K\|j_\gamma\|_{0,\partial K}^2\right]}$$

defines an upper bound on the error $\|e\|_1$, up to the constant C_e/α and as such constitutes an explicit error estimator (error indicator). Moreover, the global quantity η can be decomposed into the elemental contributions:

$$\eta_K = \sqrt{h_K^2\|r_K\|_{0,K}^2 + h_K\|j_\gamma\|_{0,\partial K}^2}$$

These local quantities can serve as refinement indicators in an adaptive strategy, following, for example, the rule:

$$\text{if } \frac{\eta_K}{\max_K \eta_K} \geq C_{\text{ad}}, \text{ then refine element } K$$

i.e. the elements with the largest contributions η_K are selected for refinement. Here, C_{ad} is a user-prescribed tolerance usually set to 0.5.

3.6 Problems

3.6.1 Exercise 1

Let u be the solution of

$$\begin{cases} -\nabla\cdot\nabla u + u = f, & \text{in } \Omega \subset \mathbb{R}^d \\ u = u_d, & \text{on } \Gamma_d \\ n\cdot\nabla u = g, & \text{on } \Gamma_n \end{cases}$$

where n is the unit outward normal vector to the boundary, $f = f(x)$, $x \in \Omega$, $g = g(x)$, $x \in \Gamma_n$, $u_d = u_d(x)$, $x \in \Gamma_d$, are given scalar functions (the data). Assuming that $f \in L^2(\Omega)$, $g \in L^2(\Gamma_n)$, and $u_d \in L^2(\Gamma_d)$, show that the problem is well-posed and derive an a priori error estimate for piecewise linear finite element approximations obtained using the standard Galerkin method.

3.6.2 Exercise 2

Let $\Omega = (0,1)$, $f \in L^2(\Omega)$, and $\beta \in \mathbb{R}$. Consider the following problem: Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega u'v'dx + \int_\Omega \beta u'vdx + \int_\Omega uvdx = \int_\Omega fvdx, \qquad \forall v \in H_0^1(\Omega) \tag{3.9}$$

1. Show that the problem is well-posed.
2. Derive the corresponding PDE and boundary conditions.
3. The problem is discretized using Lagrange finite elements of degree 2 and uniform size h, and such that the finite element space is H_0^1 conformal. Show that the discrete problem, defined using the standard Galerkin method, is well-posed.
4. Knowing that the following interpolation estimate holds for $l = 1, 2$ and all $v \in H^{l+1}(\Omega)$

$$\|v - \mathcal{I}_h v\|_{0,\Omega} + h|v - \mathcal{I}_h v|_{1,\Omega} \leq ch^{l+1}|v|_{l+1,\Omega}$$

derive a priori error estimates for the finite element approximation u_h of u in the $L^2(\Omega)$ and $H^1(\Omega)$ norms and in the $H^1(\Omega)$ "seminorm". Be as explicit as you can about the constants.

3.6.3 Exercise 3

Let u be the solution of the steady-state convective-diffusion problem:

$$\begin{aligned} -\Delta u + \beta \cdot \nabla u &= f, \quad \text{in } \Omega \subset \mathbb{R}^d \\ u &= 0, \quad \text{on } \partial\Omega \end{aligned}$$

where $d = 2$ or 3 and the velocity field β is given such that $\nabla \cdot \beta = 0$ in Ω and $\beta \cdot n = 0$ on $\partial\Omega$.

1. Derive the weak formulation of the problem.
2. Show that the problem is well-posed (specify requirements on β and f, if necessary, for well-posedness).
3. Let $u_h \in U_h$ be the solution of the corresponding finite element problem, obtained by the Galerkin method, where e.g. $U_h = \{u_h \in C^0(\overline{\Omega});\ u_h \circ T_K \in \mathbb{P}_1,\ \forall K \in \mathcal{T}_h;\ u_h = 0 \text{ on } \partial\Omega\}$. Derive an a posteriori error estimator based on the explicit residual method.

3.6.4 Exercise 4

The objective here is to study the convergence of uniform h-adaptive refinement for the following problem:

$$\begin{aligned} -5\frac{\partial^2 u}{\partial x^2} - 4\frac{\partial^2 u}{\partial y^2} - 6\frac{\partial^2 u}{\partial x \partial y} + 2u &= f, \quad \text{in } \Omega \\ u &= g, \quad \text{on } \partial\Omega \end{aligned} \tag{3.10}$$

where f and g are given functions.

The function is approximated with Lagrange finite elements with $P = \mathbb{Q}_2$, i.e. the degree of the polynomial functions is given by $p = 2$. We start with a mesh (shown in Fig. 3.1) of 4×2 elements. At each refinement iteration, the mesh is refined by divided the elements into four elements. The relative error E_r:

$$\mathcal{E}_r = \frac{\|u - u_h\|_1}{\|u\|_1}$$

is computed exactly (the exact solution u is supposed to be known) with respect to the H^1 norm. In this experiment, the number of elements N_e, number of degrees of freedom N, and relative error E_r are reported in Table 3.1.

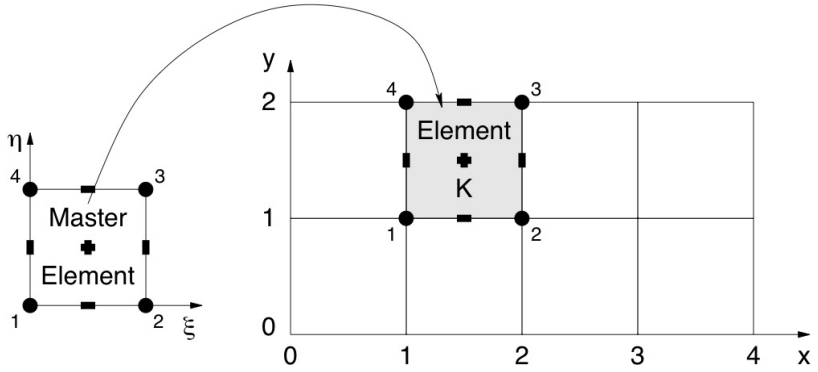


Figure 3.1: Finite element mesh of domain Ω.

N_e	N	$\mathcal{E}_r$
4×2	45	0.24E+00
8×4	153	3.27E-02
16×8	561	1.62E-02
32×16	2113	4.17E-03
64×32	8385	1.05E-03

Table 3.1: Convergence analysis of the finite element method.

1. Compute the rate of convergence of the method with respect to mesh size h.

2. Is the rate of convergence confirmed by the following error estimate:

$$\|u - u_h\|_1 \leq Ch^{\min(p,r)}\|u\|_{r+1}$$

where C is a constant independent of h and $\|u\|_{r+1}$ denotes the H^{r+1} norm of u. What can you say about the regularity of the solution u? Do you think the exact solution u is a very smooth function?

The objective of this project is to investigate numerical errors in finite element approximations of boundary-value problems. In particular, rates of convergence are evaluated for simple problems in 1D and 2D.

Problem 1 (Boundary layer problem). We first consider the convection dominated diffusion problem:

$$\begin{aligned} -(\varepsilon u')' + \alpha u' = 0, &\quad \forall x \in \Omega = (0,1) \\ u = 1, &\quad \text{at } x = 0 \\ u = 0, &\quad \text{at } x = 1 \end{aligned}$$

where α represents the velocity field and ε the kinematic viscosity of the fluid. The first term of the differential equation is the diffusion term while the second is the convection term. We shall choose $\alpha = 1$ and vary ε only as the important parameter is in fact the ratio between the two terms.

1. Derive the weak formulation of the problem and show that the problem is well-posed.
2. Calculate the exact solution.
3. Estimate the rates of convergence with respect to the H^1 and L^2 norms for h-uniform meshes and h-adaptive meshes. Repeat the calculations for different values of the polynomial degrees, e.g. $p = 1$, 2, 3, and different values of the viscosity, e.g. $\varepsilon = 1$, 0.1, and 0.01.
4. Provide convergence plots (i.e. log(error) versus N, where N is the number of degrees of freedom) and comment your results.

Problem 2 (Problem with singular solution). We now consider the following problem:

$$\begin{aligned} -u'' + u = f, &\quad \forall x \in \Omega = (0,1) \\ u = a, &\quad \text{at } x = 0 \\ u' + u = b, &\quad \text{at } x = 1 \end{aligned}$$

where f is a scalar-valued function defined on Ω, and a, $b \in \mathbb{R}$. We suppose that the data f, a, and b are defined such that the exact solution of this problem is given by:

$$u(x) = x^{3/5}$$

1. Compute f, a, and b.
2. Derive the weak formulation of the problem, show that the problem is well-posed, and compute the regularity of the solution (degree of smoothness).
3. Estimate the rate of convergence with respect to the L^2 and H^1 norms for different values of h and p.
4. Provide convergence plots (i.e. log(error) versus N, where N is the number of degrees of freedom) and comment your results.

Problem 3 (A 2D elliptic problem). We consider in this example the boundary-value problem as presented in Problem 1 of Project 3. The problem consisted in finding the temperature in the

"apartment" such that:

$$\begin{aligned} -\nabla \cdot k\nabla\theta = 0, &\quad \forall x \in \Omega \\ \theta = \theta_f, &\quad \text{on } \Gamma_f \\ \theta = \theta_o, &\quad \text{on } \Gamma_{1,w} \\ \theta = \theta_r, &\quad \text{on } \Gamma_{1,r} \cup \Gamma_{2,r} \\ n \cdot (k\nabla\theta) = -h_g(\theta - \theta_o), &\quad \text{on } \Gamma_{2,w} \\ n \cdot (k\nabla\theta) = -h_w(\theta - \theta_o), &\quad \text{on } \Gamma \end{aligned}$$

The data is given as before, namely $k = 0.025$ W/(mK), $h_g = 20$ W/(m^{2}K), $h_w = 2$ W/(m^{2}K), $\theta_f = 120°$C, $\theta_o = 10°$C, and $\theta_r = 30°$C.

1. Show that the problem is well-posed.
2. Estimate the rate of convergence with respect to the L^2 and H^1 norms using h-uniform triangular and quadrangular meshes for different values of p.
3. Derive the adjoint problem associated with the quantity of interest

   $$Q(\theta) = \frac{1}{|\omega|} \int_\omega \theta(x, T)\, dx$$

   where ω is the unit square located at the center of the bedroom.
4. Estimate the rates of convergence with respect to the quantity of interest and the rates of convergence of the adjoint solution in the H^1 norm. Comments?

Chapter4

Finite Element Method for Time-Dependent Problems

4.1 Introduction

The objectives of these lecture notes is to provide a brief introduction to finite element approximations of time-dependent problems. We first show how to derive weak formulations for a model problem given in strong form and derive space and time discretization of the weak formulation. Finally, we introduce the notion of adjoint problems for the prediction of specific quantities of interest and provide a few examples.

4.2 Model problem: strong and weak formulation

4.2.1 Strong formulation

Let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$, $d = 1$, 2, or 3, with boundary $\partial\Omega$. For the sake of simplicity, we will consider homogeneous Dirichlet boundary conditions in the exposition below. Let $\Omega_T = (0, T)$ be an interval in $\mathbb{R}$. We are interested in solving for the scalar function $u = u(x, t)$, $x \in \overline{\Omega}$, $t \in \overline{\Omega}_T$, that satisfies the time-dependent convection-diffusion-reaction equation:

$$\partial_t u + \alpha \cdot \nabla u - \nu \Delta u + cu = f, \quad \text{in } \Omega \times \Omega_T \tag{4.1}$$

subjected to the boundary condition and initial condition:

$$\begin{aligned} u(x,t) &= 0, \qquad \forall x \in \partial\Omega,\ \forall t \in \Omega_T \\ u(x,0) &= u_0, \qquad \forall x \in \Omega, \end{aligned} \tag{4.2}$$

where $f = f(x, t)$, $\forall (x, t) \in \Omega \times \Omega_T$, is a scalar function, and ν and c are strictly positive constants. In addition, the velocity field $\alpha = (\alpha_1, \ldots, \alpha_d) = \alpha(x, t)$ is assumed to be known and solenoidal, i.e. $\nabla \cdot \alpha = 0$ almost everywhere in $\Omega \times \Omega_T$.

4.2.2 Weak formulations

As usual, a weak formulation can be obtained by multiplying the equation by an arbitrary "smooth" test function $v = v(x)$, $v = 0$ on $\partial\Omega$, and integrating over the domain Ω:

$$\int_\Omega [\partial_t u(t) + \alpha \cdot \nabla u(t) - \nu \Delta u(t) + cu(t)] v \, dx = \int_\Omega f v \, dx, \qquad \forall t \in \Omega_T \tag{4.3}$$

Note that we write $u = u(t)$ in above equation to emphasize the dependence on time. Integrating by parts and making use of the divergence theorem (or Green-Ostrogradski), we arrive at:

$$\int_\Omega \partial_t u(t)v + \alpha \cdot \nabla u(t)\, v + \nu \nabla u(t) \cdot \nabla v + cu(t)v \, dx = \int_\Omega fv \, dx, \qquad \forall t \in \Omega_T \tag{4.4}$$

At this stage, it is important to ask ourselves to which spaces the functions u, v, a and f should belong in order for the above integrals to be well-defined. We introduce some notation in the section below and recall preliminary results about function spaces.

Notation and function spaces

Let $g = g(x,t)$ be a function of space and time and V a space of functions defined on Ω. Assume $g \in L^2(0,T;V)$, meaning that the V-norm of the function g: $(0,T) \to V$, such that at a given time t, $g(t) \equiv g(\cdot,t)$, is square integrable with respect to time (the symbol $\equiv$ is used here to mean "by definition"). The norm of the function g is therefore given by:

$$\|g\|_{L^2(0,T;V)} = \sqrt{\int_0^T \|g(t)\|_V^2 dt} \; < \; \infty$$

where $\|g(t)\|_V$ is the V-norm of g at t.

Dual space and duality pairing. The dual space of a space V is denoted by V'. Symbolically, the product of a function $w \in V'$ with a function $v \in V$ is denoted by the duality pairing:

$$\langle w, v \rangle_{V',V} \equiv \int_\Omega wv \, dx, \qquad w \in V', \; v \in V$$

and can be read as: "the product of a non-smooth function with a sufficiently smooth function is well-defined". For example, the dual space of $L^2(\Omega)$ is simply $L^2(\Omega)$ and the duality pairing corresponds in that case to the L^2 inner product. Now, if $V = H_0^1(\Omega)$, then the dual space V' is defined as $H^{-1}(\Omega)$. Since functions in $H_0^1(\Omega)$ are smoother than those in $L^2(\Omega)$, we can infer that the functions in $H^{-1}(\Omega)$ are less smooth than those in $L^2(\Omega)$. We indeed have the following embedding:

$$H^{-1}(\Omega) \subset L^2(\Omega) \subset H_0^1(\Omega)$$

Example 17 *As an example, it is well-known that the Heaviside (or step) function $H(x)$ is in $L^2(-1,1)$ but that the Dirac function $\delta(x)$, $x \in (-1,1)$ is not. The Dirac function actually lives in $H^{-1}(-1,1)$ and can be defined as the "distributional derivative" of $H(x)$. Indeed, since $H \in L^2(-1,1)$ and $v \in H_0^1(-1,1)$, then the inner product of H and v' is well-defined, so that:*

$$\int_{-1}^1 Hv' \, dx + \int_{-1}^1 H'v \, dx = \int_{-1}^1 (Hv)' \, dx = H(1)v(1) - H(-1)v(-1) = 0$$

meaning that the integral

$$\int_{-1}^1 H'v \, dx$$

is well-defined as well (although H' cannot be defined analytically). By definition, the Dirac delta is the distributional derivative H' of H and we can write:

$$\langle \delta, v \rangle_{H^{-1},H_0^1} = \int_{-1}^{1} \delta(x) v(x)\, dx = v(0) \equiv \int_{-1}^{1} H'(x) v(x)\, dx$$

Finally, if $g \in L^2(0,T;V')$ and $v \in V$, then $\langle g, v \rangle_{V',V}$ is well-defined "almost everywhere" in $(0,T)$ (a.e. except for a few points in the interval).

Formulations

We suppose here that $f \in L^2(0,T;H^{-1}(\Omega))$, that $\alpha \in L^\infty(0,T;H(\mathrm{div}) \cap (L^\infty(\Omega))^d)$, and that $u_0 \in L^2(\Omega)$. Let us introduce the bilinear form $a(\cdot,\cdot)$ defined on $H_0^1(\Omega) \times H_0^1(\Omega)$ as:

$$a(u,v) = \int_\Omega \nu \nabla u \cdot \nabla v + \alpha \cdot \nabla u\, v + cuv\, dx$$

Therefore, (4.4) can be rewritten in the form:

$$\int_\Omega \partial_t u(t) v\, dx + a(u(t),v) = \langle f, v \rangle_{H^{-1},H_0^1}, \quad \forall v \in H_0^1(\Omega), \text{ a.e. in } \Omega_T$$

From now on, and for the sake of clarity in the notation, we will simply write $\langle \cdot,\cdot \rangle$ instead of $\langle \cdot,\cdot \rangle_{H^{-1},H_0^1}$. We observe that the integral is well-defined if and only if $\partial_t u \in L^2(0,T;H^{-1}(\Omega))$. In that case only, we are allowed to write:

$$\langle \partial_t u(t), v \rangle + a(u(t),v) = \langle f, v \rangle, \quad \forall v \in H_0^1(\Omega), \text{ a.e. in } \Omega_T$$

Let us define the space U of admissible solutions as:

$$U = \{u : (0,T) \to H_0^1(\Omega);\ u \in L^2(0,T;H_0^1(\Omega));\ \partial_t u \in L^2(0,T;H^{-1}(\Omega))\}$$

Formulation 1: A weak formulation of Problem (4.1)-(4.2) may read:

$$\boxed{\begin{array}{l}\text{Find } u \in U \text{ such that } u(0) = u_0 \text{ and} \\ \qquad \langle \partial_t u(t), v \rangle + a(u(t),v) = \langle f, v \rangle, \quad \forall v \in H_0^1(\Omega), \text{ a.e. in } \Omega_T\end{array}} \tag{4.5}$$

Formulation 2: The above formulation can be written in an alternative form by considering test functions in the space V:

$$V = \{v : (0,T) \to H_0^1(\Omega);\ v \in L^2(0,T;H_0^1(\Omega))\}$$

Therefore, it is equivalent to write the weak formulation as:

$$\boxed{\begin{array}{l}\text{Find } u \in U \text{ such that } u(0) = u_0 \text{ and} \\ \qquad \displaystyle\int_0^T \langle \partial_t u(t), v \rangle + a(u(t),v)\, dt = \int_0^T \langle f, v \rangle\, dt, \quad \forall v \in V\end{array}} \tag{4.6}$$

Formulation 3: The initial condition can be weakly imposed so that a third formulation of the problem reads:

$$\boxed{\begin{array}{l}\text{Find } u \in U \text{ such that} \\ \displaystyle\int_0^T \langle \partial_t u(t), v\rangle + a(u(t), v)\, dt + \int_\Omega u(0)v(0)\, dx = \int_0^T \langle f, v\rangle\, dt + \int_\Omega u_0 v(0)\, dx, \quad \forall v \in V\end{array}} \tag{4.7}$$

Remark 8 *Thanks to a theorem by J.-L. Lions, above problems are known to be well-posed in the sense that there exists a unique solution, which is continuously dependent on the data.*

Remark 9 *If it is clear that the integrals are understood in terms of duality pairings, Equation (4.7) can also be written as:*

$$\begin{aligned}\int_0^T\int_\Omega \partial_t u\, v\, dxdt + \int_0^T\int_\Omega \nu\nabla u\cdot\nabla v + \alpha\cdot\nabla u\, v + cuv\, dxdt + \int_\Omega u(0)v(0)\, dx \\ = \int_0^T\int_\Omega fv\, dxdt + \int_\Omega u_0 v(0)\, dx\end{aligned} \tag{4.8}$$

4.3 Time and space discretization

A fully finite element approximation of time-dependent problems is rarely considered as it would become cumbersome to deal with finite elements in four dimensions. The following approach is often (generally) considered instead: (i) first, approximate the solution to (4.5) in space only by the finite element method so as to obtain a system of coupled ordinary differential equations, where time is then the only independent variable, (ii) construct an approximation in time by making use of the vast collection of methods for the solution to ordinary differential equations. This approach is often referred to as the method of lines. Discretization in time is usually obtained using finite difference methods.

Attention: *In this section, we assume that $\alpha = 0$, i.e.*

$$a(u, v) = \int_\Omega \nu\nabla u\cdot\nabla v + cuv\, dx$$

Indeed, discretization of the convective term $\alpha\cdot\nabla u$ by classical Galerkin methods yields unstable schemes. The reason is that information is propagated in the domain in the direction of the convective velocity. In order to stabilize the discretization schemes, it is necessary to either add artificial viscosity (not too much, not too little), or use special test functions (Petrov-Galerkin methods, such as SUPG), or least-squares methods, etc. This topic is out of the scope of these lecture notes and will not be treated here.

4.3.1 Space discretization

Let $\mathcal{T}_h$ be a mesh of Ω (it is assumed here that the domain $\Omega_h = \cup_{K\in\mathcal{T}_h} K$ exactly coincides with $\bar{\Omega}$) and let V_h be a finite element subspace of $H_0^1(\Omega)$, i.e. $V_h \subset H_0^1(\Omega)$. Moreover, for the sake of simplicity, we suppose that $f \in C^0(0, T; L^2(\Omega))$ (f is continuous in time and at each time, $f(t)$ is

in $L^2(\Omega)$). In this case, the duality pairing can simply be written in terms of the L^2-inner product $(\cdot,\cdot)$:

$$\langle f, v\rangle = (f, v) = \int_\Omega fv \; dx$$

Problem (4.5) can be approximated in space as follows:

$$\boxed{\begin{array}{l}\text{Find } u_h \in C^1(0,T;V_h) \text{ such that } u_h(0) = u_{h,0} \text{ and} \\ \qquad\qquad (\partial_t u_h, v_h) + a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \ \forall t \in [0,T]\end{array}} \tag{4.9}$$

where $u_{h,0}$ is an approximation of u_0 in V_h, which may either be obtained by projection or interpolation.

Let ϕ_i, $i = 1, \ldots, m$, denote the basis functions of V_h. The function u_h is to be understood as:

$$u_h(x,t) = \sum_{i=1}^{m} u_i(t)\phi_i(x)$$

where the coefficients $u_i(t)$ are scalar functions of time only. However, the test functions $v_h \in V_h$ are independent of time and can be expressed as:

$$v_h(x) = \sum_{i=1}^{m} v_i\phi_i(x)$$

where the coefficients v_i are simply real numbers. Note that the first derivative of u_h is given by:

$$\partial_t u_h(x,t) = \sum_{i=1}^{m} d_t u_i(t)\phi_i(x) \equiv \sum_{i=1}^{m} \dot{u}_i(t)\phi_i(x)$$

The systems of ordinary differential equations (4.9) can be recast in matrix form. Let $U(t) = [u_1(t), u_2(t), \ldots, u_m(t)]^T$ denote the vector of unknowns at time t, $\dot{U}$ the first derivative of U, F the loading vector, M the mass matrix, and K the stiffness matrix associated with the bilinear form $a(\cdot,\cdot)$. The objective is then to find $U = U(t)$ such that

$$M\dot{U} + KU = F, \quad \forall t \in [0,T], \qquad U(0) = U_0$$

where U_0 is the vector of degrees of freedom associated with $u_{h,0}$. This system of equations is sometimes referred to as the semi-discrete equations. The solution u_h (or equivalently U) can be shown to converge to the exact solution of the problem as h goes to zero. Finite difference techniques can be used to approximate the semi-discrete equations. They are based on Taylor expansions.

4.3.2 Review of Taylor expansions

Let Δt denote the timestep such that $\Delta t = T/N$ and let $t^n = n\Delta t$, $n = 0, \ldots, N$. Then, useful expansions of a function $g = g(t)$ are:

$$g(t^n + \Delta t) = g(t^n) + \dot{g}(t^n)\Delta t + \ddot{g}(t^n)\frac{\Delta t^2}{2} + \dddot{g}(t^n)\frac{\Delta t^3}{6} + \ldots$$
$$g(t^n - \Delta t) = g(t^n) - \dot{g}(t^n)\Delta t + \ddot{g}(t^n)\frac{\Delta t^2}{2} - \dddot{g}(t^n)\frac{\Delta t^3}{6} + \ldots$$

where $\dot{g}(t^n)$, $\ddot{g}(t^n)$, and $\dddot{g}(t^n)$ denote the first, second, and third derivative of g at t^n, respectively. If we denote $g^{n+1} = g(t^n + \Delta t)$, $g^n = g(t^n)$, and $g^{n-1} = g(t^n - \Delta t)$, these expansions can be rewritten:

$$g^{n+1} = g^n + \dot{g}(t^n)\Delta t + \ddot{g}(t^n)\frac{\Delta t^2}{2} + \dddot{g}(t^n)\frac{\Delta t^3}{6} + \dots$$
$$g^{n-1} = g^n - \dot{g}(t^n)\Delta t + \ddot{g}(t^n)\frac{\Delta t^2}{2} - \dddot{g}(t^n)\frac{\Delta t^3}{6} + \dots$$

The first derivative at t^n can be approximated by:

$$\dot{g}(t^n) = \frac{g^{n+1} - g^n}{\Delta t} - \frac{\ddot{g}(t^n)}{2}\Delta t - \frac{\dddot{g}(t^n)}{6}\Delta t^2 + \dots = \frac{g^{n+1} - g^n}{\Delta t} + \mathcal{O}(\Delta t) \tag{4.10}$$

$$\dot{g}(t^n) = \frac{g^n - g^{n-1}}{\Delta t} + \frac{\ddot{g}(t^n)}{2}\Delta t - \frac{\dddot{g}(t^n)}{6}\Delta t^2 + \dots = \frac{g^n - g^{n-1}}{\Delta t} + \mathcal{O}(\Delta t) \tag{4.11}$$

or, using a combination of the two expansions:

$$\dot{g}(t^n) = \frac{g^{n+1} - g^{n-1}}{2\Delta t} - \frac{\dddot{g}(t^n)}{12}\Delta t^2 + \dots = \frac{g^{n+1} - g^{n-1}}{2\Delta t} + \mathcal{O}(\Delta t^2) \tag{4.12}$$

4.3.3 Examples of time discretization schemes

Using these approximations of the first derivative, common finite difference schemes are:

1. **Forward Euler scheme:** From (4.10), we have:

$$MU^{n+1} = MU^n - \Delta t K U^n + \Delta t F^n, \quad n = 0, \dots, N-1, \qquad U^0 = U_0$$

2. **Backward Euler scheme:** From (4.11), we can write:

$$(M + \Delta t K)U^{n+1} = MU^n + \Delta t F^{n+1}, \quad n = 0, \dots, N-1, \qquad U^0 = U_0$$

3. **Crank-Nicolson scheme:** Finally, the last expression for the derivative gives the scheme:

$$(M + \frac{\Delta t}{2}K)U^{n+1} = (M - \frac{\Delta t}{2}K)U^n + \frac{\Delta t}{2}(F^{n+1} + F^n), \quad n = 0, \dots, N-1, \qquad U^0 = U_0$$

The forward Euler scheme is conditionally stable in the sense that Δt needs to be restricted to ensure stability of the scheme. The backward Euler and Crank-Nicolson schemes are unconditionally stable; in principle, the schemes are stables for any value of Δt. The forward and backward Euler schemes are first-order accurate since the approximation of the first derivative is $\mathcal{O}(\Delta t)$. The Crank-Nicolson scheme is second-order accurate.

4.4 Adjoint problems

In general, the solution itself of a given problem is not what we are looking after; what is of interest are output functionals of the solution, or simply, quantities of interest. Examples of quantities of interest are, for instance, solutions at a specified point in the domain, local averages of the solution in given subdomains, fluxes through parts of the boundary, or any other observable quantities that could be, in principle, measured in actual experiments.

The quantity of interest, thereafter, will be denoted by $Q(u)$ where u is the solution of an initial boundary-value problem. Although such a quantity could be a nonlinear functional of u (such as the kinetic energy in the whole domain if u is a velocity field), we shall only consider bounded linear functionals $Q \in \mathcal{L}(U, \mathbb{R})$, i.e. Q: $U \longrightarrow \mathbb{R}$ and there exists a constant $C > 0$ such that $|Q(u)| \leq C\|u\|_U$.

Examples of quantities of interest: Let $\omega \in \Omega$ and let k_ω be a kernel function defined on Ω defined as $k_\omega(x) = 1$, if $x \in \omega$, and $k_\omega(x) = 0$, otherwise. Then, if u is a scalar-valued function:

$$Q(u) = u(x_0) = \int_\Omega \delta(x - x_0)u(x)dx \qquad (= \text{solution at point } x_0)$$
$$Q(u) = \frac{1}{|\omega|}\int_\omega u dx = \int_\Omega k(x)u(x)dx / \int_\Omega k(x)dx \qquad (= \text{averaged solution in } \omega)$$

If u is a vector-valued function (velocity for example), one might be interested in the solution in direction β, where β is a given unit vector:

$$Q(u) = u(x_0) \cdot \beta = \int_\Omega \delta(x - x_0)u(x) \cdot \beta dx$$

Derivation of the adjoint problem: Suppose that a boundary-value problem (or initial boundary-value problem) is given in the abstract weak form:

$$\boxed{\text{Find } u \in U \text{ such that} \quad B(u, v) = F(v), \quad \forall v \in V} \tag{4.13}$$

where $B(\cdot, \cdot)$ is a bilinear form defined on $U \times V$ and $F(\cdot)$ a linear form on V. We also suppose that the goal of the prediction, as mentioned earlier, is to estimate the quantity $Q(u)$ rather than u itself. The question is whether it is possible to bypass the calculation of u in order to evaluate $Q(u)$. The answer is yes (only for linear problems and linear quantities of interest) by introducing the function $p \in V$ (additional mathematical arguments are omitted for the sake of simplicity) such that:

$$Q(u) = F(p) \tag{4.14}$$

Substituting p for v in (4.13), we thus obtain:

$$Q(u) = B(u, p) \tag{4.15}$$

It now suffices to introduce the following problem, the so-called adjoint or dual problem,

$$\boxed{\text{Find } p \in V \text{ such that} \quad B(v, p) = Q(v), \quad \forall v \in U} \tag{4.16}$$

in order to solve for $p \in V$. Note that if p satisfies above problem, then replacing v by u in (4.16) and using (4.13) immediately yields:

$$Q(u) = B(u, p) = F(p)$$

The function p is sometimes referred to as the influence function (it indicates how the loading influences the quantity of interest) or generalized Green's function.

Green's Function: Let u be the solution of the Poisson problem $-\Delta u = f$ in Ω, $u = 0$ on $\partial\Omega$. Let $U = V = H_0^1(\Omega)$. The problem can be recast in weak form as: Find $u \in V$ such that

$$B(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega fv \, dx = F(v), \qquad \forall v \in V$$

Suppose now that the quantity of interest is the solution u at a given point x_0 in Ω, i.e.

$$Q(u) = u(x_0) = \int_\Omega \delta(x - x_0) u(x) \, dx$$

Then the adjoint problem is given by:

$$\boxed{\text{Find } G \in V \text{ such that:} \quad B(v,G) = \int_\Omega \nabla v \cdot \nabla G \, dx = \int_\Omega \delta(x - x_0) v(x) \, dx = Q(v), \quad \forall v \in V}$$

where we have used for the adjoint solution the notation G (as in Green). Integrating by part the first integral (symbolically), we have:

$$\int_\Omega v(-\Delta G - \delta(x - x_0)) \, dx = 0$$

which provides us with the adjoint problem in strong form:

$$\boxed{\begin{aligned} -\Delta G(x) &= \delta(x - x_0) \quad \text{in } \Omega \\ G &= 0 \quad \text{on } \partial\Omega \end{aligned}}$$

and we know that the solution at x_0 can simply be computed as:

$$u(x_0) = Q(u) = F(G) = \int_\Omega f(x) G(x) dx$$

where G is the Green's function associated with point x_0, sometimes denoted as $G(x, x_0)$. The adjoint problem generalized the concept of the Green's function to any problem and any quantity of interest.

Example 18 *Let $\Omega \subset \mathbb{R}^d$. We supposed that $u = u(x,t)$ is governed by the parabolic equation:*

$$\rho \frac{\partial u}{\partial t} - \nu \Delta u = f, \quad \forall x \in \Omega, \ t \in (0,T)$$

and is subjected to the following boundary and initial conditions:

$$\begin{cases} u = u_d, & \text{on } \Gamma_d \times (0,T) \\ n \cdot \nu \nabla u = g, & \text{on } \Gamma_n \times (0,T) \\ u = u_0, & \text{at } t = 0 \end{cases}$$

where ρ, $\nu \in \mathbb{R}$, and f, u_d, g, and u_0 are given data. Moreover, $\partial\Omega = \overline{\Gamma_d \cup \Gamma_n}$.

We introduce the lift $\tilde{u}$ such that $\tilde{u} = u_d$ on $\Gamma_d \times (0,T)$, $n \cdot (k \nabla \tilde{u}) = 0$ on $\Gamma_n \times (0,T)$, and $\tilde{u} = 0$ at $t = 0$. We also introduce the spaces:

$$W = \{w \in H^1(\Omega); \ w = 0, \ \text{on } \Gamma_d\}$$
$$V = \{v \in L^2(0,T;W); \ \partial_t v \in L^2(0,T;W')\}$$

We pose $u = \tilde{u} + w$. A weak formulation of above problem can be written as:

$$\text{Find } w \in V \text{ such that} \quad B(w,v) = F(v) - B(\tilde{u},v), \quad \forall v \in V$$

where

$$B(w,v) = \int_0^T\!\!\int_\Omega \rho \frac{\partial w}{\partial t} v \, dxdt + \int_0^T\!\!\int_\Omega \nu \nabla w \cdot \nabla v \, dxdt + \int_\Omega \rho w(x,0) v(x,0) \, dx$$
$$F(v) = \int_0^T\!\!\int_\Omega f v \, dxdt + \int_0^T\!\!\int_{\Gamma_n} g v \, dsdt + \int_\Omega \rho u_0 v(x,0) \, dx$$

Suppose now that we are interested in a quantity of interest $Q(u)$ of the form:

$$\boxed{Q(u) = \int_0^T\!\!\int_\Omega k_\omega(x) u(x,t) \, dxdt}$$

The adjoint problem is then given by $B(v,p) = Q(v)$ with

$$B(v,p) = \int_0^T\!\!\int_\Omega \rho \frac{\partial v}{\partial t} p \, dxdt + \int_0^T\!\!\int_\Omega \nu \nabla v \cdot \nabla p \, dxdt + \int_\Omega \rho v(x,0) p(x,0) \, dx$$

In order to derive the strong form of the adjoint problem, we first integrate by parts some of the integrals of the bilinear form as follows:

$$B(v,p) = \int_0^T\!\!\int_\Omega -v\rho \frac{\partial p}{\partial t} \, dxdt + \int_\Omega \rho v(x,T) p(x,T) \; dx - \int_\Omega \rho v(x,0) p(x,0) \, dx$$
$$- \int_0^T\!\!\int_\Omega v \, \nu \Delta p \, dxdt - \int_0^T\!\!\int_{\Gamma_n} v n \cdot (\nu \nabla p) \, dxdt + \int_\Omega \rho v(x,0) p(x,0) \, dx$$

Simplifying, we get:

$$B(v,p) = \int_0^T\!\!\int_\Omega v \left[-\rho \frac{\partial p}{\partial t} - \nu \Delta p) \right] dxdt + \int_0^T\!\!\int_{\Gamma_n} v \left[-n \cdot (\nu \nabla p) \right] dxdt + \int_\Omega \rho v(x,T) p(x,T) \, dx$$

Equating $B(v,p) = Q(v)$, we derive the strong form of the adjoint problem as:

$$\begin{cases} -\rho \dfrac{\partial p}{\partial t} - \nu \Delta p = k_\omega, & \forall x \in \Omega, \; t \in (0,T) \\ p = 0, & on \; \Gamma_d \times (0,T) \\ n \cdot \nu \nabla p = 0, & on \; \Gamma_n \times (0,T) \\ p = 0, & at \; t = T \end{cases}$$

Note that the adjoint problem for a time-dependent problem needs to be solved backward in time. Introducing the change of variable $\tau = T - t$, the adjoint problem can be recast as the forward problem:

$$\begin{cases} \rho \dfrac{\partial p}{\partial \tau} - \nu \Delta p = k_\omega, & \forall x \in \Omega, \; \tau \in (0,T) \\ p = 0, & on \; \Gamma_d \times (0,T) \\ n \cdot \nu \nabla p = 0, & on \; \Gamma_n \times (0,T) \\ p = 0, & at \; \tau = 0 \end{cases}$$

This problem has exactly the same structure as the primal problem with different loading data.

Example 19 *We consider the same problem as before but suppose here that we are interested in the local solution at the final time T, i.e.*

$$\boxed{Q(u) = \int_\Omega k_w(x) u(x,T) dx}$$

then it is straightforward to derive the adjoint problem as:

$$\begin{cases} \rho \dfrac{\partial p}{\partial \tau} - \nu \Delta p = 0, & \forall x \in \Omega,\ \tau \in (0,T) \\ p = 0, & on\ \Gamma_d \times (0,T) \\ n \cdot \nu \nabla p = 0, & on\ \Gamma_n \times (0,T) \\ p = k_\omega, & at\ \tau = 0 \end{cases}$$

Example 20 *We now consider the field $u = u(x,t)$ governed by the convection-diffusion equation:*

$$\rho \frac{\partial u}{\partial t} + \rho \alpha \cdot \nabla u - \nu \Delta u = f, \quad \forall x \in \Omega,\ t \in (0,T)$$

and subjected to the same boundary and initial conditions as in Example 18. For simplicity, we suppose that the velocity field α is independent of time and satisfies $\nabla \cdot \alpha = 0$ and $\alpha \cdot n = 0$ on Γ_n. The only difference with Example 18 is due to the convective term. We have, since p, $v \in V$ (i.e. p, $v = 0$ on Γ_d):

$$\begin{aligned} \int_\Omega \rho \alpha \cdot \nabla v\, p\, dx &= \int_\Omega \nabla \cdot (\rho p v \alpha)\, dx - \int_\Omega v \nabla \cdot (\rho p \alpha)\, dx \\ &= \int_{\Gamma_n} \rho p v \alpha \cdot n\, dx - \int_\Omega v\, \rho p \nabla \cdot \alpha\, d - \int_\Omega v\, \rho \alpha \cdot \nabla p\, dx = - \int_\Omega \rho v \alpha \cdot \nabla p\, dx \end{aligned}$$

due to the properties satisfied by α. If we are interested in the quantity $Q(u)$ of Example 18, the strong form of the adjoint problem then reads:

$$\begin{cases} -\rho \dfrac{\partial p}{\partial t} - \rho \alpha \cdot \nabla p - \nu \Delta p = k_\omega, & \forall x \in \Omega,\ t \in (0,T) \\ p = 0, & on\ \Gamma_d \times (0,T) \\ n \cdot \nu \nabla p = 0, & on\ \Gamma_n \times (0,T) \\ p = 0, & at\ t = T \end{cases}$$

Using the change of variable $\tau = T - t$, the adjoint problem can be recast as a problem forward in time:

$$\rho \frac{\partial p}{\partial \tau} - \rho \alpha \cdot \nabla p - \nu \Delta p = k_\omega, \quad \forall x \in \Omega,\ \tau \in (0,T)$$

with same boundary conditions as before but with the final condition replaced by the initial condition $p = 0$, at $\tau = 0$. We note that the structure of this problem is similar to the primal problem with negative velocity $-\alpha$ for the convective term.

The adjoint problems can then be approximated following the same approaches as for the primal problems.

4.5 Problems

4.5.1 Exercise 1

Let $u = u(x,t)$ be governed by the convection-diffusion equation:

$$\rho\frac{\partial u}{\partial t} + \rho\alpha\cdot\nabla u - \nu\Delta u = f, \quad \forall x\in\Omega,\ t\in(0,T)$$

and subjected to the boundary and initial conditions:

$$\begin{cases} u = 0, & \text{on } \Gamma_d\times(0,T) \\ n\cdot\nu\nabla u + \kappa u = g, & \text{on } \Gamma_n\times(0,T) \\ u = u_0, & \text{at } t=0 \end{cases}$$

Assume that the velocity field α is independent of time and satisfies $\nabla\cdot\alpha = 0$ and $\alpha\cdot n = 0$ on Γ_n. Derive the strong form of the adjoint problem associated with the quantity of interest:

$$Q(u) = \frac{1}{|\Gamma_n|}\int_{\Gamma_n} u(x,T)dx$$

4.5.2 Exercise 2

Consider the Stokes problem: Find (u,p) such that

$$\begin{aligned} -\nu\Delta u + \nabla p &= f, \quad \text{in } \Omega\subset\mathbb{R}^d \\ \nabla\cdot u &= 0, \quad \text{in } \Omega \\ u &= 0, \quad \text{on } \partial\Omega \end{aligned}$$

Suppose that the goal of the simulation is to estimate the averaged velocity in direction β in a subdomain $\omega\in\Omega$, where β is a unit vector. The quantity of interest reads:

$$Q(u) = \frac{1}{|\omega|}\int_\omega \alpha\cdot u\,dx$$

Derive the strong form of the corresponding adjoint problem.

4.5.3 Exercise 3

Let u be the solution of the steady-state convective-diffusion problem:

$$\begin{aligned} -\nu\Delta u + \alpha\cdot\nabla u &= f, \quad \text{in } \Omega\subset\mathbb{R}^d \\ u &= 0, \quad \text{on } \partial\Omega \end{aligned}$$

where $d = 2$ or 3 and the velocity field α is given such that $\nabla\cdot\alpha = 0$ in Ω and $\alpha\cdot n = 0$ on $\partial\Omega$. Suppose that one is interested in the averaged flux on a portion γ of the domain boundary $\partial\Omega$. This quantity of interest can be expressed as:

$$Q(u) = \frac{1}{|\gamma|}\int_\gamma n\cdot\nabla u\,ds$$

Derive the strong form of the adjoint problem.

The objective of this project is to investigate time-dependent simulations and adjoint problems. The model application is the same as in Project 2 except for the following features: two "heaters" of dimension 1.5 m $\times$ 0.2 m are installed in the apartment (which is now in Paris and no longer in Austin, Texas), one in the bedroom along the right wall, and one in the living room along the top wall. The heaters are placed 5 cm away from the walls and are perfectly centered with respect to these walls (see figures below).

Let $\theta = \theta(x, y, t)$, $t \geq 0$, denote the temperature field and let Ω be the computational domain with boundary $\partial\Omega = \overline{\Gamma \cup \Gamma_f \cup \Gamma_{1,w} \cup \Gamma_{2,w} \cup \Gamma_{1,r} \cup \Gamma_{2,r}}$, where $\Gamma_{1,r}$ and $\Gamma_{2,r}$ represent the boundaries of the "radiators".

Problem 1. We first look for the steady-state solution assuming that the air velocity is zero in the apartment; the temperature θ is therefore governed by the steady-state heat equation:

$$-\nabla \cdot k\nabla\theta = 0, \quad \forall x \in \Omega$$

We assume that the window in the living room is wide open and that the walls, door, and the window in the bedroom are not perfectly insulated. We also suppose that the radiators are kept at constant temperature θ_r. The temperature in the apartment is thus subjected to the following boundary conditions:

$$\begin{cases} \theta = \theta_f, & \text{on } \Gamma_f \\ \theta = \theta_o, & \text{on } \Gamma_{1,w} \\ \theta = \theta_r, & \text{on } \Gamma_{1,r} \cup \Gamma_{2,r} \\ n \cdot (k\nabla\theta) = -h_g(\theta - \theta_o), & \text{on } \Gamma_{2,w} \\ n \cdot (k\nabla\theta) = -h_w(\theta - \theta_o), & \text{on } \Gamma \end{cases}$$

For the following simulations, the parameters are chosen as $k = 0.025$ W/(mK), $h_g = 20$ W/(m^{2}K), $h_w = 2$ W/(m^{2}K), $\theta_f = 120°$C, $\theta_o = 10°$C, and $\theta_r = 30°$C.

Give a prediction of the temperature at the center of the bedroom using first triangular elements, then quads. You may also use different polynomial degrees. Do you obtain the same temperature in all cases?

Problem 2. We now solve for the time-dependent heat equation (assuming again zero velocity):

$$\rho C \frac{\partial\theta}{\partial t} - \nabla \cdot k\nabla\theta = 0, \quad \forall x \in \Omega,\ t > 0$$

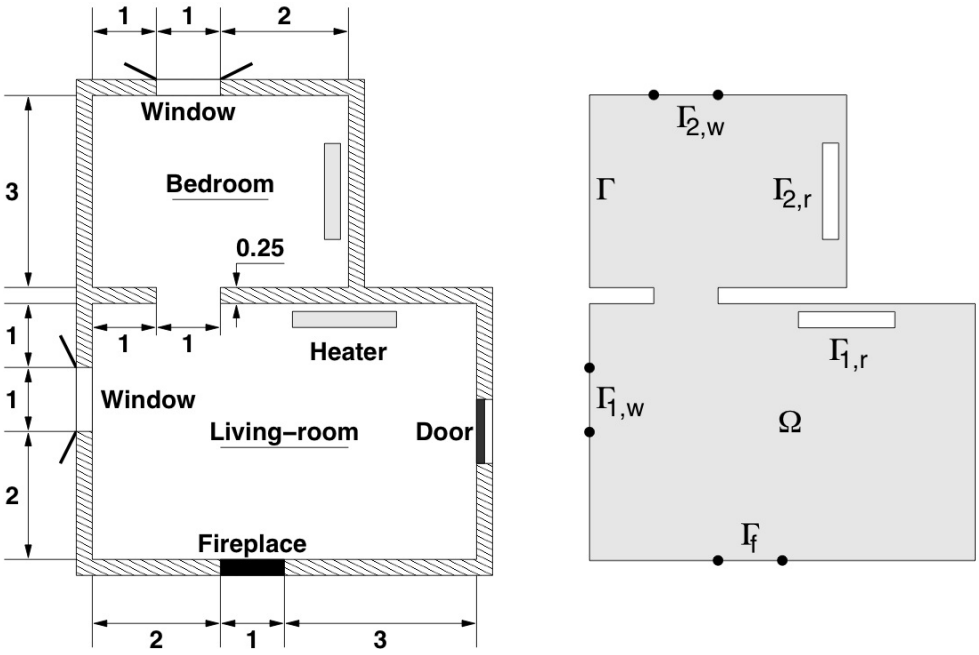
with initial condition $\theta(0) = \theta_0$. Boundary conditions are as prescribed in Problem 1. Setting $\rho = 1.2$ kg/m^3, $C = 1000$ J/(Kg·K), and $\theta_0 = 10°$C,

1. evaluate the "time" T it takes to reach the steady-state solution,

2. solve for the adjoint solution corresponding to the averaged temperature in the unit square located at the center of the bedroom, defined first as an average over the whole time interval $(0, T)$, and then defined at time T only, i.e.

$$Q_1(\theta) = \frac{1}{T}\int_0^T \left[\frac{1}{|\omega|}\int_\omega \theta(x,t)\,dx\right] dt, \qquad Q_2(\theta) = \frac{1}{|\omega|}\int_\omega \theta(x,T)\,dx$$

   where ω is the unit square.



Problem 3. Suppose now that the window of the living room is open and that air is continuously blowing along $\Gamma_{1,w}$ with velocity $u_d = (u_{1,d}, u_{2,d})$. The velocity and pressure fields are estimated by the time-independent incompressible Stokes equations (Navier-Stokes with $\rho = 0$):

$$\begin{aligned}
-\nu\Delta u + \nabla p = 0, &\quad \text{in } \Omega \\
\nabla \cdot u = 0, &\quad \text{in } \Omega \\
u = u_d, &\quad \text{on } \Gamma_{1,w} \qquad \text{(inflow BC)} \\
u = 0, &\quad \text{on } \Gamma \cup \Gamma_{1,r} \cup \Gamma_{2,r} \qquad \text{(no slip BC)} \\
n \cdot u = 0 \text{ and } t \cdot (-pI + \nu(\nabla u + \nabla u^T))n = 0, &\quad \text{on } \Gamma_f \cup \Gamma_{2,w} \qquad \text{(slip BC)}
\end{aligned}$$

where ν is the kinematic viscosity of air ($\nu = 16 \times 10^{-6}$ m^{2}s^{-1}), t the tangential vector to the boundary, and I the unit tensor. Consider the case $u_d = (0, 0.2)$ and the case $u_d = (0, -0.2)$ where the velocities are expressed in ms^{-1}.

Meanwhile, the temperature is governed by the partial differential equation:

$$\rho C \left(\frac{\partial \theta}{dt} + u \cdot \nabla \theta \right) - \nabla \cdot k \nabla \theta = 0, \quad \forall x \in \Omega, \ t > 0$$

with initial condition $\theta_0 = 10°$C. Boundary conditions are as prescribed in Problem 1. Note that $n \cdot u = 0$ on $\Gamma_{2,w}$ and that $u = 0$ on Γ.

1. Evaluate the "time" T it takes to reach the steady-state solution and provide the temperature at the center of the bedroom for the two cases of "inflow" velocity.

2. Write the adjoint problem corresponding to the quantities of interest $Q_1(\theta)$ and $Q_2(\theta)$ defined above and solve for the adjoint solution.

3. Use information from the adjoint solution and temperature field to design an "optimal" finite element mesh for estimating the temperature at the center of the bedroom.

Problem 4. Do you believe that these results are realistic? How could you improve the mathematical model to obtain more realistic results?

Bibliography

[1] COMSOL. Multiphysics modeling and simulation. http://www.comsol.com.

[2] A. Ern and J.-L. Guermond. Theory and practice of finite elements, volume 159 of Applied Mathematical Sciences. Springer, 2004.

[3] Wikipedia.org. Partial differential equation. http://en.wikipedia.org/wiki/Partial_differential_equation.