Notes on Numerical Methods for Continuous Systems

Federico Piazzon & Mario Putti
Department of Mathematics – University of Padua

September 10, 2016

Content

1 Generalities on partial differential equations

1.1 Definition and classification

In this notes we will look at the numerical solution for partial differential equations. We will be mainly concerned with differential models stemming from conservation laws, such as those arising from force conservations i.e., second Newton's law $F = ma$, such as de Saint-Venant equations, governing the equilibrium of a solid, or the Navier-Stokes equations, governing the dynamics of fluid flow. These equations are also called "equations in divergence form", to identify the fact that the divergence of a vector translates in mathematical terms the conservation of the flux represented by that vector field. As an example, let us consider the advection-diffusion equation (ADE), that governs the conservation of mass of a solute moving within the flow of the containing solvent. A typical application is the transport of a contaminant by a water body moving with laminar flow. The flow of the solvent is given by the vector (velocity) field β, and the solute is undergoing chemical (Fickian) diffusion with a diffusion field given by $D(x)$. We remark that if the density of the solvent is constant, then mass conservation is equivalent to volume concentration, and thus density does not appear in the equations. The mathematical model is then given by:

$$\frac{\partial u}{\partial t} = \text{div}\,(D\nabla u) - \text{div}(\beta u) + f \qquad \text{in } \Omega \in \mathbb{R}^d \tag{1.1}$$

where the equation is defined on a subspace of the d-dimensional Euclidean space $\mathbb{R}^d$ (generally, $d = 1$, 2, or 3), the function $u(x,t) : \Omega \times [0:T] \longrightarrow \mathbb{R}$ represents the concentration of the solute (mass/volume of solute per unit mass/volume of solvent), t is time, $\text{div} = \sum_i \partial/\partial x_i$ is the divergence operator, D is the diffusion coefficient, possibly a second order tensor, and $\nabla = \{\partial/\partial x_i, i = 1, d\}$ is the gradient operator. The conservation property mentioned above can be pointed out by a simple application of the divergence theorem. To this aim, we denote with $q = -D\nabla u - \vec{v}u$ the vector representing the flux of a quantity (e.g., mass, energy, momentum, etc), integrate equation (1.1) and use the divergence theorem to obtain:

$$\frac{\partial}{\partial t}\int_\Omega u\, dx = \int_{\partial\Omega} q \cdot \nu\, ds + \int_\Omega f\, dx$$

where $\partial\Omega$ is the boundary of Ω, assumed sufficiently regular, and ν is the unit outward normal to $\partial\Omega$. The boundary integral on the right hand side can be interpreted as the balance of the total flux, i.e., the balance between ingoing and outgoing fluxes across the domain boundary. In other words, the PDE tells us that the total flux balance of the quantity (in this case the quantity is the mass of the solute) must be equilibrated by the temporal accumulation (the time derivative) and the total source/sink terms. We note that the equation:

$$q = -D\nabla u + \vec{v}u$$

can be derived from Newton's first law $\vec{F} = m\vec{a}$, and it represents the momentum balance of the quantity of interest.
The definition of a well posed problem requires auxiliary conditions, in this case given by initial and boundary conditions. So we let the domain boundary $\Gamma = \partial\Omega$ be the union of three non overlapping sub-boundaries such that $\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_C$, so that we:

$$\begin{aligned}
u(x,0) &= u_o(x) && x \in \Omega, && t = 0 && \text{initial conditions} \\
u(x,t) &= g_o(x) && x \in \Gamma_D, && t > 0 && \text{Dirichlet BCs} \\
D\nabla u(x,t) \cdot \nu &= q_N(x) && x \in \Gamma_N, && t > 0 && \text{Neumann BCs} \\
(\vec{v}u + D\nabla u(x,t)) \cdot \nu &= q_c(x) && x \in \Gamma_C, && t > 0 && Cauchy(or mixed or Robin) BCs
\end{aligned}$$

where ν is the outward unit normal defined on Γ. Formally, under some regularity assumption and the assumption that D never vanishes, this is called a "parabolic" equation.
The term "parabolic" is used to classify partial differential equations (PDEs) on the basis of certain qualitative properties of the solution. This can be done relatively easily with linear PDEs, and becomes more complicated for nonlinear PDEs. We start this discussion by giving a general definition of a PDE:

Problem 1.1 (PDE)**.** Find a function $u(x,y,z) : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that:

$$F(x,y,z,u,u_x,u_y,u_z,u_{xx},u_{xy},u_{yy},u_{xz},u_{zz},u_{yz}) = 0. \tag{1.2}$$

where u_x e u_{xx} are the first and second partial derivatives of u with respect to x.

If F is a linear function of u and its derivatives, then the equation is called linear, and, assuming $d = 2$, it can be written as:

$$a(x,y) + b(x,y)u + c(x,y)u_x + d(x,y)u_y + e(x,y)u_{xx} + f(x,y)u_{yy} + = 0. \tag{1.3}$$

The order of a PDE is the order of the derivative of maximum degree that appears in the equation. Thus, in the previous case the order is 2. Typical examples are:

$$\begin{aligned}
\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 && \quad 2^o \text{ grade (Laplace equation)} \\
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 && \quad 1^o \text{ grade (transport or convection equation)} \\
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 && \quad 2^o \text{ grade (diffusion equation).}
\end{aligned}$$

To start in our task of classification, assume for simplicity a 2-dimensional domain $d = 2$, and a constant coefficient second order PDE:

$$au_{xx} + bu_{xy} + cu_{yy} + e = 0. \tag{1.4}$$

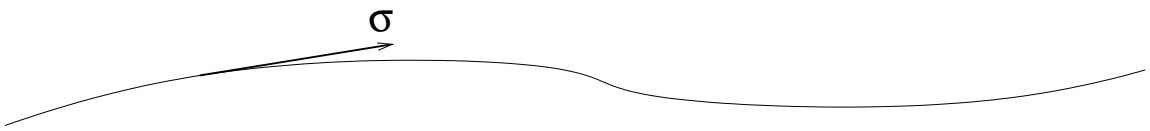


FIGURE 1.1: *Curve γ and local reference system*

We look for a curve $\gamma : \mathbb{R}^2 \longrightarrow \mathbb{R}$ that is sufficiently regular and such that when we write the PDE along this curve it turns into and Ordinary Differential Equation (ODE). We write this curve in parametric form as $\gamma(\sigma)$ (Figure 1.1) as follows:

$$\gamma = \begin{cases} x &= x(\sigma) \\ y &= y(\sigma) \end{cases}$$

Writing the above equations on a local reference system, we obtain:

$$\begin{aligned} \frac{du_x}{d\sigma} &= \frac{\partial u_x}{\partial x}\frac{dx}{d\sigma} + \frac{\partial u_x}{\partial y}\frac{dy}{d\sigma} = u_{xx}\frac{dx}{d\sigma} + u_{xy}\frac{dy}{d\sigma} \\ \frac{du_y}{d\sigma} &= \frac{\partial u_y}{\partial x}\frac{dx}{d\sigma} + \frac{\partial u_y}{\partial y}\frac{dy}{d\sigma} = u_{xy}\frac{dx}{d\sigma} + u_{yy}\frac{dy}{d\sigma}. \end{aligned}$$

Writing u_{xx} from the previous system and substituting it in (1.4), we have:

$$u_{xy}\left[a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c\right] - \left(a\frac{du_x}{dx}\frac{dy}{dx} + c\frac{du_y}{dx} + e\frac{dy}{dx}\right) = 0.$$

This equation is a re-definition of the PDE on the curve $\gamma(\sigma)$, or, in other words, the equation is satisfied on γ. Now we can choose γ so that the first term in square brackets is zero, obtaining an equation for u_x and u_y where only ordinary derivatives appear:

$$a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c = 0.$$

We note that dy/dx is the slope of γ, which can then be obtained by solving the ODE:

$$\frac{dy}{dx} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{1.5}$$

The solution of this ODE yields families of curves, that are called characteristic curves. Different families arise depending on the sign of the discriminant $\Delta = b^2 - 4ac$. We then call the equations depending on this sign, obtaining the following classification:

- $b^2 - 4ac < 0$: two complex solutions: the equation is "elliptic";
- $b^2 - 4ac = 0$: one real solution: the equation is "parabolic";

- $b^2 - 4ac > 0$: two real solutions: the equation is "hyperbolic".

Thus we have easily the following examples:

- Laplace equation:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

  $a = c = 1$ $b = 0 \longrightarrow b^2 - 4ac < 0$ is an elliptic equation;

- wave equation:

$$\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 u}{\partial x^2} = 0 \tag{1.6}$$

  $a = 1$ $b = 0$ $c = -1 \longrightarrow b^2 - 4ac > 0$ is a hyperbolic equation;

- diffusion equation:

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

  $a = 1$ $b = c = 0 \longrightarrow b^2 - 4ac = 0$ is a parabolic equation.

It is important now to understand what is the typical behavior of each of this class of equations.

1.2 Simple examples and solutions

We show in this paragraph some simple but clarifying examples of PDEs and their exact analytical solution. From these solutions we will extrapolate some typical characteristics of the solutions of PDEs.

Example 1.2. Find $u : [0, 1] \longrightarrow \mathbb{R}$ such that:

$$\begin{aligned} &-u'' = 0 && x \in [0, 1] \\ &u(0) = 1; \\ &u(1) = 0. \end{aligned}$$

This is a elliptic equation. In this simple case the solution is obtained directly by integration between $x = 0$ and $x = 1$. We have:

$$u(x) = 1 - x.$$

Example 1.3. Find $u : [0,1] \longrightarrow \mathbb{R}$ such that:

$$\begin{aligned} -(a(x)u')' &= 0 \qquad\qquad x \in [0,1] \\ u(0) &= 1; \\ u(1) &= 0; \end{aligned} \tag{1.7}$$

where the diffusion coefficient $a(x)$ assumes the values:

$$a(x) = \begin{cases} a_1 & \text{if } 0 \leq x < 0.5 \\ a_2 & \text{if } 0.5 < x \leq 1 \end{cases}$$

Since $a(x) > 0$ for each $x \in [0,1]$ is is an elliptic equation. In this case the solution can be obtained by first subdividing the domain interval in two halves and integrating the equation in each subinterval:

$$u(x) = \begin{cases} u_1(x) = c_1^1 x + c_2^1 & x \in [0, 0.5) \\ u_2(x) = c_1^2 x + c_2^2 & x \in (0.5, 1]. \end{cases}$$

We can see that the solutions are defined in terms of four constants. We need thus four equations. Two are given by the boundary conditions, but the other two are still missing. One natural condition is the request that $u(x)$ be continuous (at least $\mathcal{C}^0([0,1])$) in the domain $[0,1]$. The second condition can be determined by looking at the left-hand-side of equation (1.7) and looking for existence requirement of this term. Before we discuss this requirement we note that we can define the "flux" of $u(x)$ as $q(x) = -a(x)u'(x)$. Hence, the requirement for the existence of the left-hand-side (as long as we do not use the product rule for the derivative of the flux) is that $q(x)$ must be continuous for all $x \in [0,1]$ (again the requirement here is $q(x) \in \mathcal{C}^0([0,1])$. This observation suggests the sought condition, that yield the following system of equations for the constants c_i:

$$\begin{aligned} u_1(0) &= 0 & u_2(1) &= 0; \\ u_1(0.5) &= u_2(0.5) & q_1(0.5) &= q_2(0.5), \\ & & -a_1(0.5)u_1'(0.5) &= -a_2(0.5)u_2'(0.5). \end{aligned}$$

We note that the last condition physically means that the flux of the quantity $u(x)$ that exits from the subdomain on the left of $x = 0.5$ enters the subdomain on the right of $x = 0.5$. It is a conservation statement. Solving the system, the solution becomes:

$$u(x) = \begin{cases} 1 - \frac{2a_2}{a_1+a_2}x & x \in [0, 0.5] \\ \frac{2a_1}{a_1+a_2} - \frac{2a_1}{a_1+a_2}x & x \in [0.5, 1], \end{cases}$$

shown in Figure 1.2 in the case $a_1 = 1$ and $a_2 = 10$.

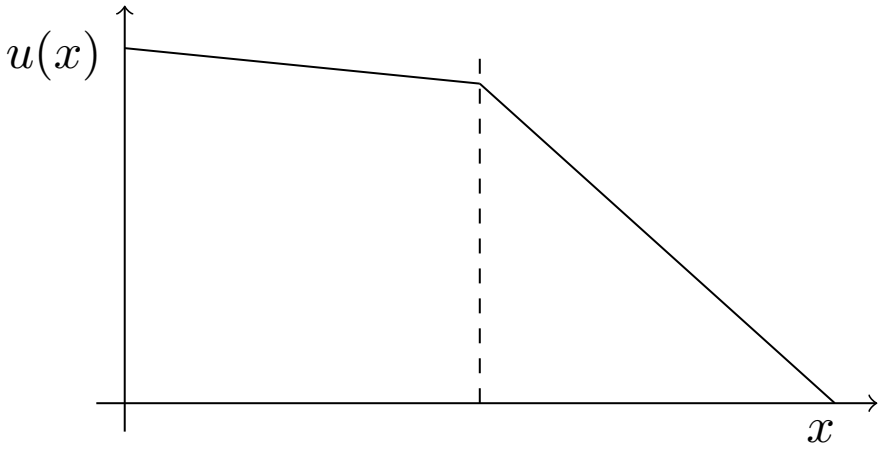


FIGURE 1.2: *Solution of problem 1.3 for* $a_1 = 1$ *and* $a_2 = 10$.

Remark 1.4. *The previous example shows that the differential equation with discontinuous coefficients has a solution that is continuous but not differentiable: the gradient is discontinuous. On the other hand the flux is continuous, and thus more regular. We will use this fact in to properly define our numerical solution. This property, that can be also shown theoretically, is very important in applications, and characterizes "conservation laws". In other words, the partial differential equation* (1.7) *represents the balance of the quantity* $u(x)$*. This quantity can be thought of as mass, then the equation is a mass-balance equation, a temperature, in which case the equation is an energy conservation equation, a fluid velocity, and then the equation is a force balance equation (first Newton law), etcetera. The determination of the conservation properties of numerical discretization schemes is an active and important field of research in the case of highly variable diffusion coefficients.*

We would like to remark that in the case of jumps in the diffusion coefficient we cannot use the product rule to expand the left-hand-side of equation (1.7)*. In fact we cannot write the following:*

$$-a(x)u''(x) - a'(x)u'(x) = 0$$

because both $u''(x)$ *and* $a'(x)$ *do not exist for* $x = 0.5$*. However, the solution* $u(x)$ *exists and is intuitively sound, i.e., without any singularity, although it does not possess a second derivative. Hence, the equation must be written exclusively as in* (1.7)*. In general, using the chain rule for derivative is numerically counterproductive even if the regularity of the mathematical objects allows it.*

Example 1.5 (Poisson equation)**.** Find $u : [0,1] \longrightarrow \mathbb{R}$ such that:

$$-u'' = f(x) \qquad x \in [0,1] \tag{1.8}$$
$$u(0) = u(1) = 0 \tag{1.9}$$

with

$$f(x) = \begin{cases} 1 & \text{if } x = 0.5, \\ 0 & \text{otherwise..} \end{cases}$$

This is an elliptic equation. The solution if this problem can be found by means of Green's functions (see Section 2.1) and is given by:

$$u(x) = \begin{cases} \frac{1}{4}(1-x) & \text{if } 0 \le x \le 0.5, \\ \frac{1}{4}(x-1) & \text{if } 0.5 \le x \le 1. \end{cases} \tag{1.10}$$

This solution is continuous but it has a piecewise constant first derivative with a jump in $x = 0.5$. Hence the second derivative $u''(x)$ does not exists in the midpoint. This seems a contradiction as in this case the left-hand-side of equation (1.8) does not exists for all $x \in [0,1]$. However, the solution $u(x)$ given in (1.10) in terms of the integral of the Green's function is mathematically sound. Thus we need to define a more "forgiving" formulation, whose solution can have discontinuous first derivatives. This is the role of the so called "weak" formulation to be seen in the next sections.

Example 1.6. Transport equation.
Given a vector field of constant velocity $\beta > 0$, find the function $u = u(x,t)$ such that:

$$u_t + \beta u_x = 0, \tag{1.11a}$$
$$u(x,0) = f(x). \tag{1.11b}$$

The characteristic curve is a line in the plane (x,t) given by:

$$x - \beta t = \text{const} = \xi. \tag{1.12}$$

Along this line the original equation (1.11a) becomes:

$$\frac{du}{dt} = \frac{\partial}{\partial t} u(\xi + \beta t, t) = \beta u_x + u_t = 0.$$

Hence, the solution u is constant along a characteristic curve and this constant is determined by the initial conditions (1.11b):

$$u(x,t) = f(\xi) = f(x - \beta t). \tag{1.13}$$

At a fixed time t_1 the solution is given by the rigid translation of the initial condition ($f(x)$) by a quantity βt_1, as shown in Figure 1.3 (right panel).

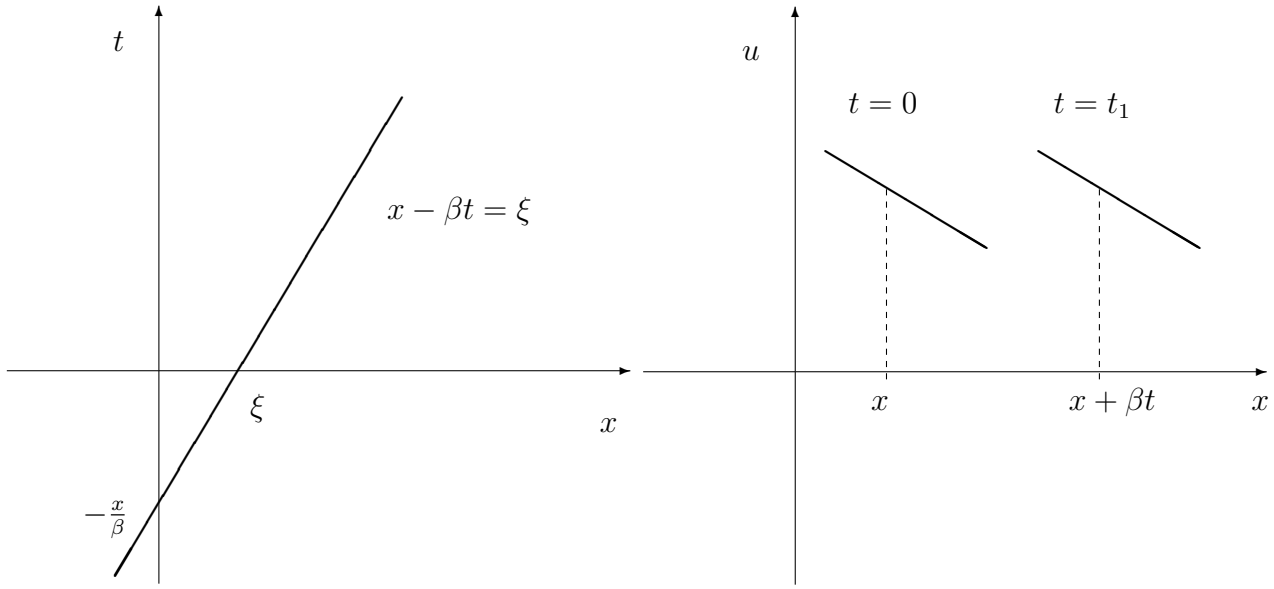


FIGURE 1.3: *Left panel: characteristic lines for equation* (1.11a) *in the* (x,t) *plane. Right panel: graph of the solution* $u(x,t)$ *at* $t=0$ *and* $t=t_1>0$ *in the* (u,x) *plane. The solution is a wave with shape given by* $f(x)$ *(a line in this case) that propagates towards the right with speed* β*.*

Example 1.7. Advection (or convection) and diffusion equation (ADE).
Find the function $u(x,t):[0,T]\times\mathbb{R}\mapsto\mathbb{R}$ such that:

$$\frac{\partial u}{\partial t}=D\frac{\partial u^2}{\partial x^2}-v\frac{\partial u}{\partial c}, \tag{1.14a}$$

$$u(x,t)=1 \qquad \text{for } x=0, \tag{1.14b}$$

$$u(x,t)=0 \qquad \text{for } x\longrightarrow\infty, \tag{1.14c}$$

$$u(x,0)=0 \qquad \text{for } t=0 \text{ and } x>0, \tag{1.14d}$$

$$u(x,0)=1 \qquad \text{for } t=0 \text{ and } x=0. \tag{1.14e}$$

The solution is given by [6]:

$$u(x,t)=\frac{1}{2}\left[\operatorname{erfc}\left(\frac{x-vt}{2\sqrt{Dt}}\right)+\exp\left(\frac{vx}{Dt}\right)\operatorname{erfc}\left(\frac{x+vt}{2\sqrt{Dt}}\right)\right],$$

where the function erfc is the complementary error function.

1.3 Conservation laws

From the physical point of view, the problems that we are facing are related to the principle of conservation. For example the equilibrium of an elastic string fixed at the end points

and subject to a distributed load is governed by an equation that determines the vertical displacement $u(x)$ of the points ox of the string and its tension stresses $\sigma(x)$, once the load $g(x)$ and the elastic characteristics of the string E (Young's modulus) are specified. The problem (D) is written as:

$$\begin{aligned} \sigma(x) &= Eu'(x) && \text{Hook's law;} \\ -\sigma'(x) &= g(x) && \text{Elastic equilibrium;} \\ u(0) &= u(1) = 0 && \text{Boundary conditions.} \end{aligned}$$

Another interpretation of the same problem can be thought of as $u(x)$ being the temperature of a rod subjected to a heat source $g(x)$. In this case the symbol k is generally used in place of E to identify the thermal conductivity of the rod material and $q(x)$ is the heat flux. The model thus is written as:

$$q(x) = -ku'(x) \qquad \text{Fourier's law;} \tag{1.15a}$$

$$q'(x) = g(x) \qquad \text{Energy conservation;} \tag{1.15b}$$

$$u(0) = u(1) = 0 \qquad \text{Boundary conditions.} \tag{1.15c}$$

The same equation can be thought as governing the diffusion of a substance dissolved in a fluid. In this case we talk about Fick's law, concentration $u(x)$. diffusion coefficient k, solute mass flux $q(x)$. Yet another interpretation of the same equation is the flow of water in a porous material. We talk then about Darcy's law. More in general, we can say that all these equations represent a conservation principle. In fact, equation (1.15a) represents the conservation of momentum deriving from Newton second law ($F = ma$), while (1.15b) states the conservation of the energy of the system.
All these problems are equations written in "divergence form" or in conservative form. For example, consider the advection-diffusion equation (1.1). From the physical point of view, our solution function u represents the density of the conserved quantity. Thus we can introduced the density flux of this quantity as:

$$\vec{q} = -D\nabla u + \vec{v}u,$$

where the first term on the right-hand-side represents the diffusive flux and the second term represents the advective flux (the quantity u is transported by the velocity $\vec{v}$ and at the same time is diffused). Equation (1.1) can then be re-written as:

$$\frac{\partial u}{\partial t} + \operatorname{div} \vec{q} = f(x).$$

The first term represents the variation in time of the mass of this quantity. The second term represents the variation in space. Integrating the above equation in a subset $U \subset \Omega$ of the domain we have:

$$\int_U \frac{\partial u}{\partial t} + \operatorname{div} \vec{q}\, dx = \int_U f(x),$$

and assuming the boundary of U to be smooth, we can apply the divergence theorem:

$$\int_U \frac{\partial u}{\partial t}\, dx + \int_{\partial U} \vec{q} \cdot \nu \, ds = \int_U f(x).$$

We recognize the classical mass conservation principle:

$$\text{rate of change} = \text{inflow-outflow}$$

Note that from a purely mathematical point of view, writing the equation in divergence form has no formal advantage with respect to any other alternative formulation. However, this is not true for the numerical formulation, in which the divergence form is always to be preferred.

1.4 Well posedness and continuous dependence of solutions on the data

The question of finding a solution to a PDE rests on the definition of solution. We can state that a solution is a function that satisfies the equation and has certain regularity properties [1]. However, the answer to the question what is a solution can be tricky. For a clear account and several interesting examples see [7]. We report here a few remarks that are useful for the developments and analysis of numerical methods.
We talk about a "classical" solution of a k-th order PDE to indicate a function that satisfies the PDE and the auxiliary conditions and that is k times differentiable. This is an intuitive requirement so that the derivatives that appear in the expression of the PDE can be formally calculated without worrying about singularities. However, this notion is often too restrictive, and there may be functions that are less regular that indeed satisfy the PDE and the auxiliary conditions. Moreover, by this strong regularity requirements, we may restrict the search of solutions only to cases that have enough regularity of the auxiliary conditions and of the data of the problem (e.g. the coefficients of the PDE). Thus we usually resort to a less restrictive definition of a solution, which is called a "weak" solution. Thus we need to change the formulation of the PDE to accommodate this lower regularity requirement, maintaining at the same time the physical notion of the process that lead to the PDE.

Remark 1.8. *Example 1.3 gives an instance of the application of this concept: the global solution, i.e., the solution over the entire domain $I = [0, 1]$, is continuous but its derivative is not. Thus a classical solution to the problem does not exist, but a "weak" solution can be defined by appropriately relaxing the continuity conditions of the search space (the space of functions that are candidate solutions).*

In any case, it is intuitive to look for solutions that are unique. There is certainly no hope to be able to find numerically a solution that is not unique. In fact, any computational algorithm

[1] This last request has to be made to avoid trivial and non interesting solutions.

in this case would never converge and would oscillate continuously among the several solutions of the problem. But uniqueness is not sufficient. We also require the notion of "continuous dependence of the solution form the data of the problem". In essence, we require that, if for example some coefficients are changed slightly, then the solution changes slightly. This notion is useful for two important reasons. First, in a computer implementation of any algorithm there is no hope to be able to specify a coefficient (which is a real number) with infinite accuracy. Next, and probably more importantly, uncertainties in physical constants or functions are intrinsically present in any model of a physical process. This uncertainty results in values of, e.g., boundary conditions or forcing functions that are not known precisely. But it is highly desirable that our mathematical model governing the physics be relatively insensitive to these uncertainties. This is reflected within the concept of well-posedness that we can make a little bit more formal by stating the following:

Definition 1.1 (Well posedness)**.** Given a problem governed by a k-th order PDE:

$$F(x, u, \partial u, \ldots, \partial^k u, \Sigma) = 0$$

where Σ denotes the set of the data defining the problem, we say that this problem is *well posed* if:

1. the solution u exists;

2. the solution u is unique;

3. the solution u depends continuously on the data, i.e., if one element $\sigma \in \Sigma$ is perturbed by a quantity δ, the corresponding solution $\tilde{u}$ to the perturbed problem is such that $\|\tilde{u} - u\| \leq L \|\delta\|$.

Note that this is not a very precise statement, as we need to specify what we mean with the symbol $\|\cdot\|$. But this definition depends on the functions with which we are dealing, and thus it must be analyzed and specified for each problem.

1.4.1 Ill-conditioning and instability

Two more concepts that are related to well-posedneess need to be clearly stated when we move from the continuous setting to the discrete (numerical) setting. The first we would like to discuss is "ill-conditioning". The "condition" of a problem is a property of the mathematical problem (not of the numerical scheme used to solve it) and can be stated intuitively as follows:

Definition 1.2. A mathematical **p**roblem is said to be *ill-conditioned* if small perturbations on the data cause large variations of the solution.

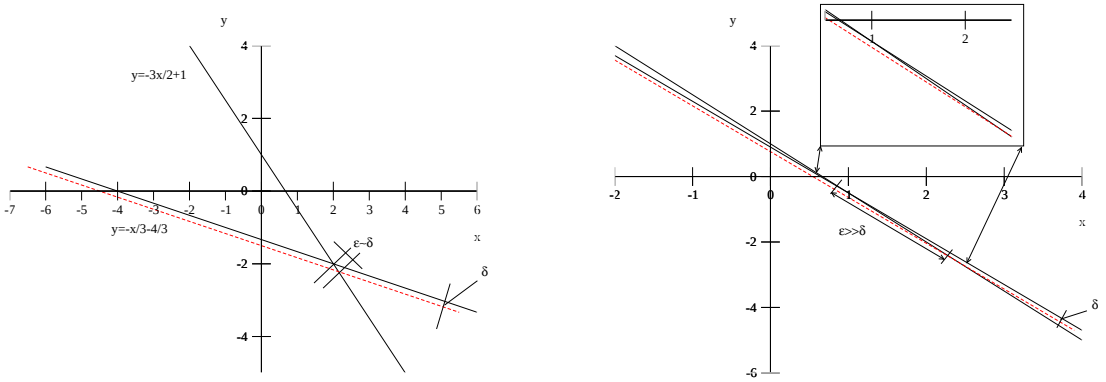


FIGURE 1.4: *Geometric interpretation of a "well-conditioned" (left) and an "ill-conditioned" linear system (right).*

The definition is problem specific, but a simple example related to linear algebra can be illuminating.

Example 1.9. Consider the following 2×2 system of linear equations:

$$\begin{aligned} 3x + 2y &= 2 && (1.16)\\ 2x + 6y &= -8. && (1.17) \end{aligned}$$

The mathematical problem can be stated as follows:

Problem 1.10. find the pair of real values (x, y) such that equations (1.16) and (1.17) are satisfied simultaneiusly.

The solution to this problem is evidently $P = (x, y) = (2, -2)$. We can rewrite the lineaer system as:

$$\begin{aligned} y &= -\frac{3}{2}x + 1 && (1.18)\\ y &= -\frac{1}{3}x - \frac{4}{3}. && (1.19) \end{aligned}$$

This reformulation, allows to change the problem into an equivalent formulation:

Problem 1.11. find the point $P = (x, y) \in \mathbb{R}^2$ that irepresents the point of intersection between the two lines identified by equations (1.18) and (1.19) (see Figure 1.4).

Now we want to analyze the conditioning of this problem. To do this we specify a small perturbation to the data of our problem and look at how its solution changes. In our case we can, for example, change the right hand side of the second equation by a quantity δ, yielding a downward translation of the line (Figure 1.4, left). The point of intersection between the two lines has now moved by a quantity $\epsilon \approx \delta$. This problem is well-condition and the ratio ϵ/δ measures somehow the conditioning of our problem.
Now, if the two lines have almos equal slopes, the situation is different (Figure 1.4, right). A small perturbation δ to one if the right hand side values yield a large movement of the solution (the point of intersection), by a quantity $\epsilon \gg \delta$. The conditioning is measured again by the quantity ϵ/δ which is now much larger than one. The problem is thus "ill-conditioned".
We note that both problems are actually "well-posed" as they admit a unique solution which is continuously depedent upon the data. But the numerical solution may loose accuracy.

The second concept is called *stability*. Unlike conditioning, stability is a property of the numerical scheme used to solve a mathematical problem. We say that a scheme is stable if errors in initial data remain bounded as the algorithm progresses. As an example, consider the following numerical algorithm given by the linear recursion:

$$u^{(k)} = Au^{(k-1)}, \qquad k = 1, 2, \ldots$$

where $u^{(k)} \in \mathbb{R}^n$, A is a constant $n \times n$ matrix, and the recursion is initiated with a given (possibly arbitrary) initial guess $u^{(0)}$. The representation $u_h^{(0)}$ of the values of $u^{(0)}$ in the computer is not exact, so the actual algorithm involves the numerical approximation $u_h^{(k)}$:

$$u_h^{(k)} = Au_h^{(k-1)}, \qquad k = 1, 2, \ldots \tag{1.20}$$

Stability of the algorithm requires that the errors with which we represent $u_h^{(0)}$ are not magnified by the algorithm process. More formally, we define the error as $e^{(k)} = u^{(k)} - u_h^{(k)}$, $k = 1, 2, \ldots$. From this last equation we have that $u^{(k)} = u_h^{(k)} + e^{(k)}$, and after substitution in eq. (1.20) we obtain the error propagation equation:

$$e^{(k)} = Ae^{(k-1)}.$$

Stability of the scheme is achieved if the norm of the error remains bounded as k increases, i.e. (using compatible norms):

$$\left\|e^{(k)}\right\| \leq \left\|Ae^{(k-1)}\right\| \leq \|A\| \left\|e^{(k-1)}\right\| \leq \|A\|^k \left\|e^{(0)}\right\|$$

which implies $\|A\| \leq 1$.

2 Galerkin Finite elements for elliptic equations

2.1 One spatial dimension

We start with some examples of simple elliptic problems. The 1-dimensional Poisson equation can be written as the following boundary value problem:

Problem 2.1 (Differential)**.**
Find the function $u : [0,1] \longrightarrow \mathbb{R}$ that satisfies:

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

where $u' = du/dx$ and $u'' = d^2u/dx^2$.

We assume the forcing function $f(x) : [0,1] \longrightarrow \mathbb{R}$ to be sufficiently regular so that the solution exists and is unique. In fact, we assume that the above problem is "well-posed" in the sense that there exists a unique solution $u(x)$ that is continuously dependent on the data of the problem, i.e., on the boundary conditions and the forcing function $f(x)$. By repeated integration we obtain:

$$\begin{aligned} -\int_0^x u''(t)\,dt &= \int_0^x f(t)\,dt; \\ -u'(x) &= -u'(0) + \int_0^x f(t)\,dt; \\ \int_0^x u'(t)\,dt &= u'(0)x - \int_0^x \left(\int_0^s f(t)\,dt \right) dx; \\ u(x) &= u(0) + u'(0)x - \int_0^x F(s)\,ds, \end{aligned}$$

where we have defined the linear functional (function of an integral function) as:

$$F(s) = \int_0^s f(t)\,dt. \tag{2.1}$$

From the boundary conditions we obtain immediately:

$$u(1) = 0 \qquad \Rightarrow \qquad u'(0) = \int_0^1 F(s)\,ds$$

from which we can write the solution to (D) as:

$$u(x) = x \left(\int_0^1 F(s)\,ds \right) - \int_0^x F(s)\,ds,$$

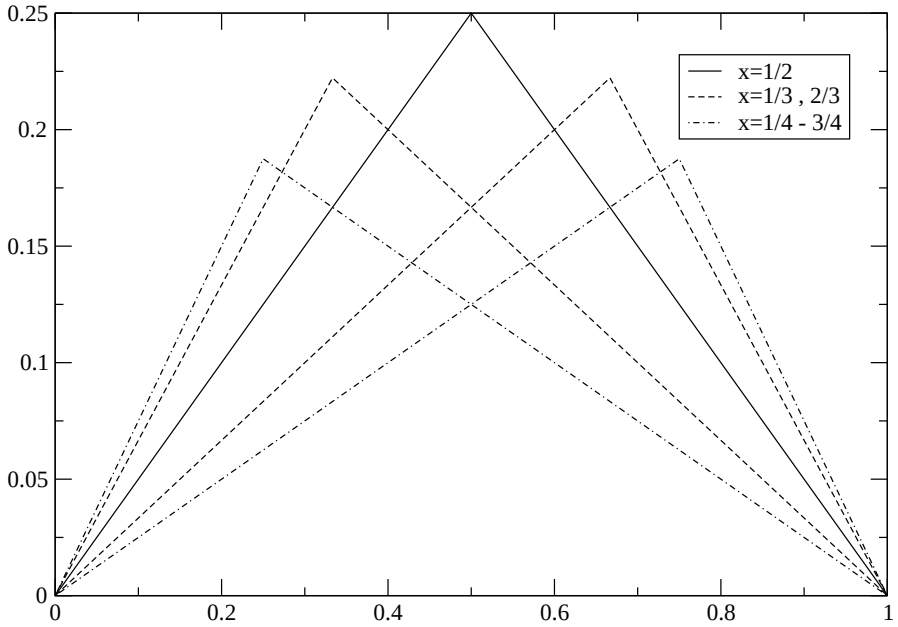


FIGURE 2.1: *Green's function for different values of x.*

which is obviously unique.
Integrating by parts equation (2.1) we obtain:

$$\int_0^x F(s)\,dt = [sF(s)]_0^x - \int_0^x sF'(s)\,ds = \int_0^x (x-t)f(t)\,dt,$$

from which the solution to problem (D) can be written as:

$$u(x) = x\int_0^1 (1-t)f(t)\,dt - \int_0^x (x-t)f(t)\,dt.$$

Define the Green's function $G(x,t)$ as:

$$G(x,t) = \begin{cases} t(1-x) & \text{if } 0 \le t \le x; \\ x(1-t) & \text{if } x \le t \le 1, \end{cases}$$

then the solution can be written in the more compact form:

$$u(x) = \int_0^1 G(x,t)f(t)\,dt.$$

The Green's function has the following properties:

- is linear for fixed t;

- is symmetric, i.e., $G(x,t) = G(t,x)$;
- is continuous;
- is non negative, assuming the value zero only at the boundary of the interval $[0,1]$;
- $\int_0^1 G(x,t)\,dt = \frac{1}{2}x(1-x)$.

The Green's function is shown in Figure 2.1 for different values of x.

Remark 2.2. *Problem* (D) *is* well posed*, i.e., the solution exists and is unique and it depends continuously on the data of the problem (boundary conditions and forcing function $f(x)$). We will come back in later sections on the concept of "well-posedness" and how this concept translates in numerical analysis. We observe that this is a fundamental notion that is needed to have any hope to find a numerical solution to any problem.*

As practical example, we cite the model of the equilibrium configuration of an elastic rope fixed at the two end-points subjected to a distributed load. In this case, we can indicate with $u : [0,1] \longrightarrow \mathbb{R}$ the vertical displacement of the rope points, $\sigma : [0,1] \longrightarrow \mathbb{R}$ the rope stress, and E is the Young modulus, and $g(x) : [0,1] \longrightarrow \mathbb{R}$ is the distributed load acting on the rope. Then problem (D) can be written as:

$$\begin{aligned} \sigma(x) &= Eu'(x) && \text{Hook's law;} \\ -\sigma'(x) &= g(x) && \text{elastic equilibrium;} \\ u(0) &= u(1) = 0 && \text{boundary conditions.} \end{aligned} \tag{2.2}$$

Another typical model is the energy balance of a bar subjected to a thermal load. In this case $u(x)$ represents the temperature of the bar, $g(x)$ the thermal load, $q(x)$ is used to replace $\sigma(x)$ and represents the heat flux through the bar, and k is used in place of E and it represents the thermal conductivity of the bar material. Then we can write:

$$\begin{aligned} q(x) &= -ku'(x) && \text{Fourier's law;} \\ q'(x) &= g(x) && \text{energy conservation;} \\ u(0) &= u(1) = 0 && \text{boundary conditions.} \end{aligned}$$

Again, we can write in the same way what is known as Fick's Law, which states that the mass flux of a solute is proportional to the opposite of the concentration gradient. Another typical model that gives rise to an elliptic equation is Darcy's law governing the flow of a fluid in a porous medium. In more general terms, all these equations represent the model of a flow of some quantity given a "potential" field.

2.1.1 Variational formulation

In this paragraph, we will briefly discuss the variational formulation for the solution of (D), which is at the basis of the Finite Element Method (FEM). To do this, we introduce the notion of linear normed functional spaces, i.e., spaces whose elements are functions and with operations that are defined mainly in terms of the (Lebesgue) integral operator. Note that, intuitively, there is a strong analogy between the vector spaces of linear algebra and the function spaces of functional analysis in terms of possibility of having a set of basis functions to express every element of the space. Obviously, since function spaces are infinite-dimensional (and they are uncountably so) there are many additional complications that need to be considered in these developments. In these sections we will briefly and very superficially recall some of the properties of these functional spaces only if needed, and we will describe some of the terminology typically used in the literature. Also the fact that the classical Riemann integration, typically taught in engineering calculus, has to be replaced by Lebesgue integration is a technical need that has no influence in the considerations that follow. Later chapters will deal with the more theoretical material, and thus more formal statements and more formal definitions will be adopted. For a better understanding of this material we recall the relevant literature.
Let $\mathcal{V}$ be a function space defined as:

$$\begin{aligned}\mathcal{V}([0,1]) = \{\, v(x) : &\text{ where } v(x) \text{ is a bounded and continuous function on the interval } I = [0,1],\\ &v'(x) \text{ is a piecewise continuous and bounded function in } I,\\ &\text{and } v(0) = v(1) = 0 \,\}\,.\end{aligned}$$

This function space is often called the "trial" space or the space of trial functions, i.e., the space of candidate solutions. In other words we are searching for our solution only among all functions that belong to the space $\mathcal{V}([0,1])$.
We can define an operation between the elements of this function space called "inner product" or "scalar product":

$$(v, w) = \int_0^1 v(x)w(x)\, dx,$$

from which a "functional" (function or map of functions) can be defined: $F : \mathcal{V} \longrightarrow \mathbb{R}$:

$$F(v) = \frac{1}{2}\,(v', v') - (f, v) + c.$$

We can then define the following "minimization" (M) and "variational" (V) problems, that, under hypothesis that will be verified later, are equivalent to the initial differential problem (D), equivalent in the sense that they have the same solution.

Problem 2.3 (Minimization)**.**
Find $u \in \mathcal{V}$ such that:

$$F(u) \leq F(v) \qquad \forall v \in \mathcal{V}. \tag{M}$$

Problem 2.4 (Variational)**.**
Find $u \in \mathcal{V}$ such that:

$$(u', v') = (f, v) \qquad \forall v \in \mathcal{V}. \tag{V}$$

Remark 2.5. *For the elastic problem written in* (2.2)*, the functional $F(v)$ is the total "potential energy" of the system given the admissible displacement $v(x)$. The mathematical statement equivalent to the "admissibility" requirement of the previous sentence is that the function $v(x)$ must be an element of the space $\mathcal{V}$, i.e., $v(x) \in \mathcal{V}$. Thus the term $\frac{1}{2}(v', v')$ is the elastic energy of the system and (f, v) the potential of the external forces. From this observation we can deduce that problem* (M) *is the formulation known as "principle of minimization of the potential energy", while problem* (V) *is the formulation of the "principle of virtual works".*

Equivalence between formulations (D)**,** (M)**, and** (V)**.**

(D)⇒(V)

Proof. We need to show that the solution of (D) is also solution of (V). To do so, we can multiply the PDE by an arbitrary function $v \in \mathcal{V}$ and integrate over the domain:

$$-\int_0^1 u''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx,$$

or, using the scalar product in $\mathcal{V}$:

$$-(u'', v) = (f, v)\,.$$

The left hand side can be integrated by parts, yielding:

$$-(u'', v) = -u'(1)v(1) + u'(0)v(0) + (u', v') = (u', v')\,.$$

Finally, noting that $v(0) = v(1) = 0$ we can write:

$$(u', v') = (f, v) \qquad \forall v \in \mathcal{V}. \tag{2.3}$$

□

(V)⇔(M)

Proof. We want to show that (V) and (M) have the same solution. Let $u(x)$ be a solution of (V). Take a function $v(x) \in \mathcal{V}$ and define the function $w(x) = v(x) - u(x) \in \mathcal{V}$. Then:

$$\begin{aligned} F(v) = F(u+w) &= \frac{1}{2}(u'+w', u'+w') - (f, u+w) + c \\ &= \frac{1}{2}(u', u') - (f, u) + c + (u', w') - (f, w) + \frac{1}{2}(w', w') = F(u) + \frac{1}{2}(w', w') \geq F(u), \end{aligned}$$

since equation (2.3) tells us that $(u', w') - (f, w) = 0$ and $(w', w') \geq 0$. Since w is an arbitrary function, than u is a minimizer of $F(u)$, and thus u is solution of (M).
The opposite direction can be shown as follows. Let u be solution of (M). Then for each $v \in \mathcal{V}$ and $\epsilon \in \mathbb{R}$, we have:

$$F(u) \leq F(u + \epsilon v),$$

since $u + \epsilon v \in \mathcal{V}$. Define the differentiable function $g(\epsilon)$ as:

$$g(\epsilon) := F(u + \epsilon v) = \frac{1}{2}(u', u') + \epsilon (u', v') + \frac{\epsilon^2}{2}(v', v') - (f, u) - \epsilon (f, v) + c.$$

This function has a minimum for $\epsilon = 0$, hence it is necessary that $g'(0) = 0$. This implies:

$$g'(0) = (u', v') - (f, v),$$

that shows that u is solution of (V).
By linearity it is easy to see that the solution to (V), and thus of (M), is unique. In fact, let $u_1 \in \mathcal{V}$ e $u_2 \in \mathcal{V}$ be two solutions of (V). Then:

$$\begin{aligned} (u_1', v') &= (f, v) \qquad \forall v \in \mathcal{V}; \\ (u_2', v') &= (f, v) \qquad \forall v \in \mathcal{V}. \end{aligned}$$

Subtracting the two equations and choosing $v = u_1 - u_2$, we obtain:

$$\int_0^1 (u_1' - u_2')^2 \, dx = 0,$$

from which we have $(u_1 - u_2)(x) = \text{const}$, and since $u(0) = u(1) = 0$, the constant is zero. □

(V)⇒(D). To show the thesis we need the following fundamental lemma of the calculus of variations. We indicate with $C_0^1([a, b])$ the space of $C^1((a, b))$ functions that are zero at the boundary. Then we have:

Lemma 2.1. *Let $g \in C([a, b])$ and*

$$\int_a^b g(x) \cdot \phi(x) \, dx = 0 \qquad \forall \phi(x) \in C_0^1([a, b]),$$

then $g(x) = 0$ for all $x \in [a, b]$.

Proof. We can proceed by contradiction. Assume that there is $x_0 \in [a,b]$ where $g(x_0) > 0$ (the negative case it is obviously analogous). Continuity of g guarantees that there exist a neighborhood of x_0 where $g(x) > 0$. More precisely, there exist $\delta > 0$ such that $(x_0-\delta, x_0+\delta) \subset (a,b)$ and for all $x \in (x_0-\delta, x_0+\delta)$ we have that $g(x) \geq g(x_0)/2$. We can then build the function $\phi \in \phi \in C_0^1([a,b])$:

$$\phi(x) = \begin{cases} \delta^2 - \mid x - x_0 \mid^2)^2 & \text{if } x \in (x_0-\delta, x_0+\delta) \\ 0 & \text{otherwise.} \end{cases}$$

Then:

$$\int_a^b g(x)\phi(x)\, dx \geq \frac{g(x_0)}{2} \int_{x_o-\delta}^{x_0+\delta} \phi(x)\, dx > 0,$$

which contradicts the hypothesis of the lemma. □

Proof. We need to establish that (V)$\longrightarrow$ (D). Let $u \in \mathcal{V}$ be solution of problem (V). Then:

$$\int_0^1 u'v'\, dx - \int_0^1 fv\, dx = 0 \qquad \forall v \in \mathcal{V}.$$

Assuming that u'' exists and is continuous, we can integrate by parts to get:

$$\int_0^1 u'v'\, dx - \int_0^1 fv\, dx = [u'v]_0^1 - \int_0^1 u''v\, dx - \int_0^1 fv\, dx = 0,$$

from which, using the homogeneous boundary conditions, we obtain:

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

Assuming $(u'' + f)$ continuous, we can then apply (piecewise) Lemma 2.1 to conclude:

$$-u'' + f = 0.$$

□

We have proved the equivalence between the variational and the differential problems. We would like to remark once again that this is true only under the hypothesis that the second derivative of u is continuous. But this assumption is not needed in the variational formulation. Using integration by parts we have decreased the regularity requirements of our solution. In summary, we can state that solutions of the differential problem are always also solutions of the variational problem. On the other hand, solutions of the variational problem are also solution of the differential problem only if we assume sufficient regularity.

2.1.2 Euler-Lagrange Equations

We can extend these issues to a more general context to arrive at what are called the Euler-Lagrange Equations of the calculus of variations. Remaining in a one-dimensional setting, we want to find a function $u : [0,1] \mapsto \mathbb{R}$ that satisfies the homogeneous boundary conditions $u(0) = u(1) = 0$ and that minimizes the functional:

$$F(u) = \int_0^1 L(x, u(x), u'(x))\ dx.$$

Assuming L sufficiently continuous so that its partial derivatives with respect to x, u e u' exist, the minimum is achieved at a point u characterized by the fact for that every perturbation of u, $F(u)$ assumes a greater value, i.e.:

$$F(u) \leq F(u + \epsilon v) \qquad \forall \epsilon \in \mathbb{R} \text{ and } \forall\ v \in \mathcal{V}.$$

Let $w = u + \epsilon v$. Note that v needs to satisfy $v(0) = v(1) = 0$. Then:

$$F(w) = F(\epsilon) = \int_0^1 L(\epsilon, x, w, w')\ dx,$$

which now can be considered a function of ϵ. The first variation of $F(\epsilon)$ is:

$$\frac{dF}{d\epsilon} = \frac{d}{d\epsilon}\int_0^1 L(\epsilon, x, w(x), w'(x))\ dx = \int_0^1 \frac{d}{d\epsilon} L(\epsilon, x, w(x), w'(x))\ dx.$$

Using the chain rule of differentiation we obtain:

$$\begin{aligned}\frac{dL(\epsilon)}{d\epsilon} =& \frac{\partial L}{\partial x}\frac{dx}{d\epsilon} + \frac{\partial L}{\partial w}\frac{\partial w}{\partial \epsilon} + \frac{\partial L}{\partial w'}\frac{\partial w'}{\partial \epsilon} \\ =& \frac{\partial L}{\partial w} v + \frac{\partial L}{\partial w'} v'.\end{aligned}$$

and hence:

$$\frac{dF}{d\epsilon} = \int_0^1 \left(\frac{\partial L}{\partial w} v + \frac{\partial L}{\partial w'} v'\right)\ dx.$$

For $\epsilon = 0$ we have $w = u$ and thus $F(w)|_{\epsilon=0}$ must attain its minimum, and must be stationary:

$$\frac{dF}{d\epsilon}|_{\epsilon=0} = \int_0^1 \left(\frac{\partial L}{\partial w} v + \frac{\partial L}{\partial w'} v'\right)\ dx = 0.$$

Integrating by parts we obtain:

$$\int_0^1 \frac{\partial L}{\partial w} v\ dx + v\frac{\partial L}{\partial w'}|_0^1 - \int_0^1 \frac{d}{dx}\frac{\partial L}{\partial w'} v\ dx = 0.$$

Noting that $v(0) = v(1) = 0$,we can apply the fundamental Lemma 2.1 to obtain the Euler-Lagrange equation:

$$\frac{\partial L}{\partial w} - \frac{d}{dx}\left[\frac{\partial L}{\partial w'}\right] = 0.$$

This equation determines the necessary condition (not sufficient) for the existence of the minimum of the functional $F(u) = \int_0^1 L(x, u, u')\, dx$. If $L(x, u, u')$ a convex function of u e u', then the Euler-Lagrange equation is also a sufficient condition.

Example 2.6. Consider the so called Dirichlet integral:

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

We look for the minimum of $D(u)$ within the class of continuous functions with continuous first derivatives ($u \in C^1([0, 1])$). The Euler-Lagrange equation can be written by evaluating the derivatives of $L(x, u, u') = (u')^2/2$:

$$\frac{\partial L}{\partial u} = 0; \qquad \frac{d}{dx}\left[\frac{\partial L}{\partial u'}\right] = \frac{d}{dx}\frac{1}{2}(2u') = u''(x),$$

from which we obtain:

$$u''(x) = 0,$$

i.e., La[lace equation in one dimension. Hence the solution of Laplace equation is also the minimizer of the convex functional $D(u)$.

Example 2.7. We modify the Dirichlet functional as follows:

$$D(u) = D(x, u, u') = \int_0^1 \left[\frac{1}{2}(u')^2 - fu\right]\, dx.$$

The Euler-Lagrange equations becomes:

$$\frac{\partial L}{\partial u} = f(x); \qquad \frac{d}{dx}\left[\frac{\partial L}{\partial u'}\right] = \frac{d}{dx}(2u') = u''(x),$$

i.e., the one-dimensional Poisson equation:

$$-u''(x) = f(x).$$

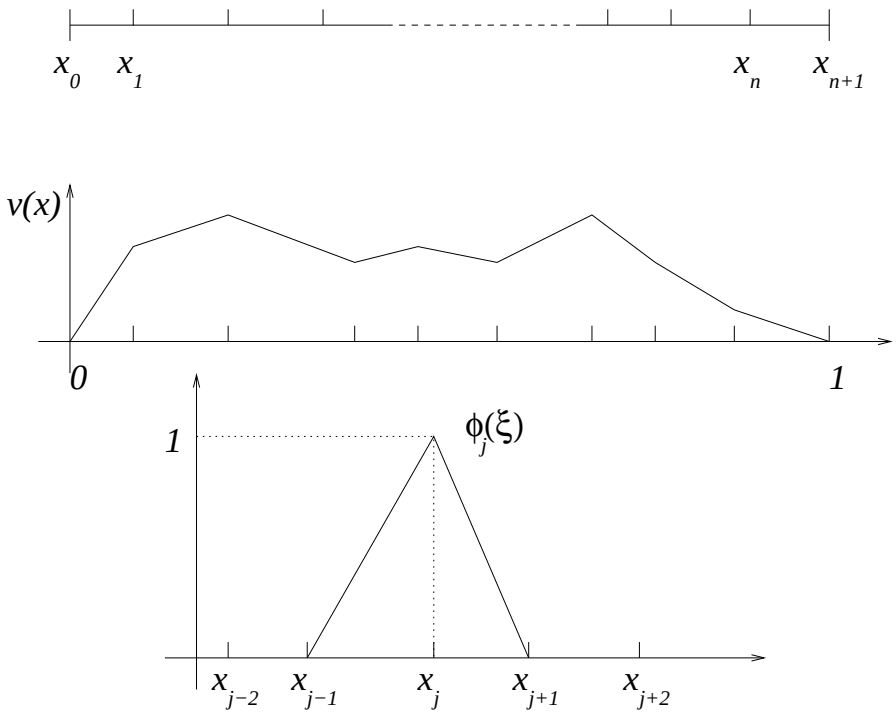


FIGURE 2.2: *Above: computational mesh defined on the interval $I = [0, 1]$; (middle) example of a function $v \in \mathcal{V}_h([0, 1])$; (bottom) example of a basis function for $\mathcal{V}_h([0, 1])$.*

2.1.3 Finite Element Formulation

The main idea of the Finite Element Method (FEM) for the solution of (V) is to discretize the functional space $\mathcal{V}$, i.e., to find an appropriate subspace $\mathcal{V}_h \subset \mathcal{V}$ of finite dimension. For example, we can choose $\mathcal{V}_h$ as the space of piecewise linear functions that interpolate the solution u. To this aim, we can define a computational mesh (or grid), i.e., a (nonuniform) partition of the interval $I = [0, 1]$ into $n + 1$ subintervals whose endpoints are given by x_i, $i = 0, 1, \ldots, n + 1$. The i-th subinterval (or element) is then $I_i = [x_i, x_{i-1}]$ and its length is $h_i = x_i - x_{i-1}$. We denote with $h = \max_i h_i$ the characteristic dimension of the mesh (Fig. 2.2). We build the subspace $\mathcal{V}_h$ as the space of piecewise linear functions v such that $v(0) = v(1) = 0$. It is obvious that $\mathcal{V}_h \subset \mathcal{V}$. We can use the Lagrange interpolation formula [23] on each I_i to construct a set of basis functions $\phi_j \in \mathcal{V}_h$ such that $\mathcal{V}_h = \text{Span}\,(\phi_1, \ldots, \phi_n)$. Following the Lagrangian approach, these functions are defined through the interpolating property:

$$\phi_j(x_i) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \quad , i, j = 1, \ldots, n. \tag{2.4}$$

Then, every function $v \in \mathcal{V}_h$ can be expressed as a linear combination of the basis functions:

$$v(x) = \sum_{j=1}^{n} v_j \phi_j(x), \tag{2.5}$$

where, because of (2.4), $v_j = v(x_j)$ is the value assumed by v in every node of the mesh. Observe that using a mesh with $n + 2$ nodes (including the endpoints $x = 0$ and $x = 1$) we have that $\text{Dim}\,(\mathcal{V}_h) = n$ and is a linear vector space.
We can now write our first FE formulation:

Problem 2.8 (Ritz method)**.**
find $u_h \in \mathcal{V}_h$ such that:

$$F(u_h) \leq F(v) \qquad \forall v \in \mathcal{V}_h. \tag{Mh}$$

Problem 2.9 (Galerkin method)**.**
Find $u_h \in \mathcal{V}_h$ such:

$$(u_h', v') = (f, v) \qquad \forall v \in \mathcal{V}_h. \tag{Vh}$$

Using (2.5), we can write immediately:

$$(u_h', \phi_i') = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{2.6}$$

and if we assume that also u_h belong to $\mathcal{V}_h$, we have that:

$$u_h(x) = \sum_{j=1}^{n} u_j \phi_j(x) \qquad u_j = u_h(x_j), \qquad u'_h(x) = \sum_{j=1}^{n} u_j \phi'_j(x) \tag{2.7}$$

and hence:

$$\sum_{j=1}^{n} \left(\phi'_i, \phi'_j\right) u_j = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{2.8}$$

which is a linear system of dimension $n \times n$. In matrix form this linear system can be written as:

$$Au = b \tag{2.9}$$

where matrix $A_{[n\times n]} = \{a_{ij}\} = \{\left(\phi'_i, \phi'_j\right)\}$ is called the *stiffness* matrix, the unknown vector is $u_{[n\times 1]} = \{u_i\}$ and the right-hand side is given by $b_{[n\times 1]} = \{b_i\} = \{(f, \phi_i)\}$.
The values of a_{ij} and b_i are easily found as follows. We first note that $a_{ij} = 0$ for $\mid i - j \mid > 1$, since in this case the supports of ϕ_i and ϕ_j have empty intersection, so that both $\phi_i(x)\phi_j(x) = 0$ and $\phi'_i(x)\phi'_j(x) = 0$. Thus, for $i = 1, \ldots, n$ we have:

$$a_{ii} = (\phi'_i, \phi'_i) = \int_{x_{i-1}}^{x_i} \frac{1}{h_i^2}\, dx + \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2}\, dx = \frac{1}{h_i} + \frac{1}{h_{i+1}},$$

and for $i = 2, \ldots, n$:

$$a_{i,i-1} = a_{i-1,i} = \left(\phi'_i, \phi'_{i-1}\right) = \left(\phi'_{i-1}, \phi'_i\right) = -\int_{x_{i-1}}^{x_i} \frac{1}{h_i^2}\, dx = -\frac{1}{h_i}.$$

Matrix A is symmetric and tridiagonal. It is also positive definite. In fact, for each $v(x) = \sum_{j=1}^{n} c_j \phi_j(x)$, we can write:

$$\sum_{i,j=1}^{n} c_i \left(\phi'_i, \phi'_j\right) c_j = \left(\sum_{i=1}^{n} c_i \phi'_i, \sum_{j=1}^{n} c_j \phi'_j\right) = (v', v') \geq 0.$$

Equality is verified only for $v'(x) \equiv 0$, or $v(x) = \text{const}$. From the boundary conditions $v(0) = v(1) = 0$ we have immediately that this constant must be zero. Thus we have:

$$\left(\sum_{i=1}^{n} c_i \phi'_i, \sum_{j=1}^{n} c_j \phi'_j\right) = \langle c, Ac\rangle > 0 \qquad \forall c \in \mathbb{R}^n,\ c \neq 0,$$

that shows that A is symmetric and positive definite, and thus invertible, and system (2.9) admits a unique solution.

Another very important property of the stiffness matrix A is that it is sparse, i.e., it has a large number of null elements. In this case, in fact, each row (or column by symmetry) has at most three nonzero elements independently of the dimension n of A. This allows the use of special solvers that are adapted to sparse linear systems and that enable the solution of very large problems.
In the case of a uniform mesh, $h_i = h = 1/(n+1)$, and constant source $f(x) = \text{const}$, the linear system takes on the form:

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \dots & \dots & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & \dots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix},$$

with $b_i = f$.

2.1.4 Analysis of one-dimensional FEM

Consistency, stability, and convergence The convergence of FEM [2] starts from the more general concepts of consistency and stability. We say that a scheme is "consistent" if the error resulting from the substitution of the real solution into the scheme tends to zero as the discretization step goes to zero. A scheme is "stable" is small variation of the data of the scheme/problem result into small variations of the numerical results.
Let us make these concepts a little bit more precise. Let $L(u, f) = 0$ our mathematical problem that needs to be solved. We can think of L as the differential operators, f the (possibly infinite) set of data of the problem, and u its real solution. We denote by $L_h(u_h, f_h) = 0$ the numerical solver for $L(u, f) = 0$ and u_h and f_h the numerical solution and the numerical data of the problem. We say that the scheme "converges" to the real solution if:

$$\|u - u_h\| \longrightarrow 0 \qquad h \longrightarrow 0,$$

where $\|\cdot\|$ is an appropriate norm. A numerical discretization scheme is "consistent" if

$$L_h(u, f) \longrightarrow 0 \qquad h \longrightarrow 0,$$

it is "strongly consistent" if:

$$L_h(u, f) = 0 \qquad \forall h.$$

[2] The study of convergence of any numerical scheme is of fundamental importance not only from the theoretical point of view but also to understand the differences between different schemes and thus be able to choose the best available method for the problem at hand. Another important point that heavily uses the theory of convergence of FEM is that a comparison between theoretical or experimental convergence rates allow a strong quality control on the correctness of the computational code and the of the input/output data.

Often it is arduous if not impossible to find the desired theoretical result showing directly the convergence of the scheme. The typical approach is then to use a fundamental theorem known as "equivalence theorem" that states that a consistent scheme is convergent if and only if it is stable [23]. On the other hand, the known the theoretical convergence speed of a scheme is useful also to understand the delicate equilibrium between the acceptable error threshold with which one solves a problem and the computational cost required to achieve that error level.
In the following we determine the convergence error of the FEM scheme using a simple one-dimensional linear elliptic model problem. More complicated, and sometime intractable, situations arise in the multidimensional case that will be treated in subsequent sections.

Error estimates for the FEM in one dimension Let $u \in \mathcal{V}$ solution of problem (D) and let $u_h \in \mathcal{V}_h$ be solution of problem (Vh). Since (V) is valid for all $v \in \mathcal{V}$ and $\mathcal{V}_h \subset \mathcal{V}$, then (V) is valid also for all functions $v \in \mathcal{V}_h$. Thus, use $v \in \mathcal{V}_h$ in (V) and in (Vh) and subtract. We obtain:

$$\begin{aligned} (u', v') &= (f, v) && \forall v \in \mathcal{V}_h \\ (u_h', v') &= (f, v) && \forall v \in \mathcal{V}_h \\ ((u' - u_h'), v') &= 0 && \forall v \in \mathcal{V}_h, \end{aligned} \tag{2.10}$$

that shows directly the the FEM scheme is strongly consistent.
We define the $\mathcal{L}^2$ norm of a function the following:

$$\|w\| = (w, w)^{\frac{1}{2}} = \left(\int_0^1 w^2 \, dx \right)^{\frac{1}{2}}.$$

We can easily see that $(v, w) = \int_0^1 vw \, dx$ satisfies the defining properties of a scalar product of two functions v and w. In particular we will often use the Cauchy inequality:

$$| \, (v, w) \, | \leq \|v\| \, \|w\|$$

It is easy to show then that u_h is the best approximation of u among all candidate functions $v \in \mathcal{V}_h$.

Theorem 2.2. *Let $u \in \mathcal{V}$ solve* (V) *and $u_h \in \mathcal{V}_h$ ($\mathcal{V}_h \subset \mathcal{V}$) be a solution of* (Vh). *Then*

$$\|(u - u_h)'\| \leq \|(u - v)'\| \qquad \forall v \in \mathcal{V}_h \tag{2.11}$$

Proof. Assume $\|(u - u_h)'\| \neq 0$, as in the case it is zero then the result is obvious.
Take $v \in \mathcal{V}_h$ form the arbitrary function $w = u_h - v$ that belongs to $\mathcal{V}_h$. Using (2.10) we have that $((u - u_h)', w') = 0$. Hence:

$$\begin{aligned} \|(u - u_h)'\|^2 &= ((u - u_h)', (u - u_h)') + ((u - u_h)', w') \\ &= ((u - u_h)', (u - u_h + w)') = ((u - u_h)', (u - v)') \\ &\leq \|(u - u_h)'\| \, \|(u - v)'\| \, . \end{aligned}$$

The results follows dividing by $\|(u - u_h)'\|$, nonzero by hypothesis. □

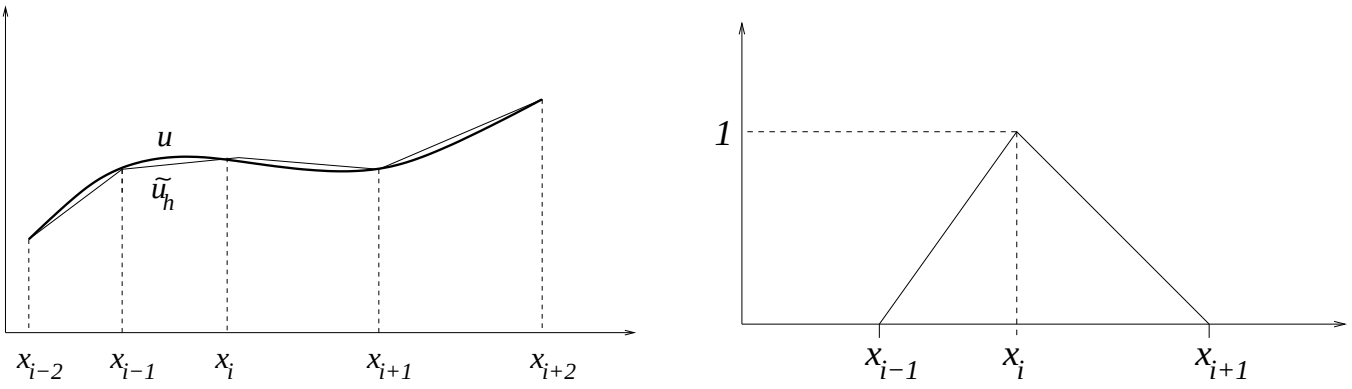


FIGURE 2.3: *Interpolant $\tilde{u}_h$ (left), basis function ϕ_i (right).*

In our case, we can also prove that $\|v\| \leq \|v'\|$ for all $v \in \mathcal{V}_h$ (a sort of Poincarè inequality):

$$\int_0^1 v^2 \, dx \leq \int_0^1 (v')^2 \, dx \qquad \forall v \in \mathcal{V}_h.$$

Note that this is true because $\mathcal{V}_h$ contains functions that vanish at the endpoints of the interval I. In fact:

$$v(x) = v(0) + \int_0^x v'(t) \, dt = \int_0^x v'(t) \, dt,$$

from which, using Cauchy inequality:

$$| \, v(x) \, | \leq \int_0^1 | \, v' \, | \, dx \leq \left(\int_0^1 1^2 \, dx \right)^{\frac{1}{2}} \left(\int_0^1 | \, v' \, |^2 \, dx \right)^{\frac{1}{2}} \leq \left(\int_0^1 | \, v' \, |^2 \, dx \right)^{\frac{1}{2}}.$$

Integrating between 0 and 1 we have finally:

$$\int_0^1 | \, v(x) \, |^2 \, dx \leq \int_0^1 \left(\int_0^1 | \, v'(x) \, |^2 \, dx \right) \, dy = \int_0^1 | \, v'(x) \, |^2 \, dx.$$

Applying the previous result to the function $v = u - u_h$ we have immediately:

$$\|u - u_h\| \leq \|(u - u_h)'\| \leq \|(u - v)'\| \qquad \forall v \in \mathcal{V}_h \tag{2.12}$$

which shows that u_h' is the best approximation (approximation of minimum norm) of u' in $\mathcal{V}_h$, i.e., it is the result of a projection.
We use this result to our advantage by trying to find an estimate of the difference between u and a particular function $v \in \mathcal{V}_h$, which we choose so that this estimate is easy to find. For convenience we then choose v as the piecewise linear interpolant $\tilde{u}_h \in \mathcal{V}_h$ of the solution u on the mesh nodes. We say that a function $\tilde{u}_h$ is an interpolant of u, or in other words that $\tilde{u_h}$

interpolates u on the mesh nodes $x_i,\ i = 0, \ldots, n+1$, if the following interpolation equations hold (see Figure 2.3):

$$\tilde{u}_h(x_i) = u(x_i) \qquad\qquad i = 0, \ldots, n+1.$$

For this purpose is convenient to use Lagrangian polynomials [23], but we will not use the definition of these polynomials but only their properties. A piecewise linear polynomial can be written as:

$$P_1(x) = \sum_{i=1}^{n} a_i \phi_i(x).$$

where the basis function on the i-th node is given by (see Figure 2.3, right)):

$$\phi_i(x) = \begin{cases} \frac{x - x_i}{x_i - x_{i-1}}, & \text{if } x_{i-1} \leq x \leq x_i, \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & \text{if } x_i \leq x \leq x_{i+1}. \end{cases}$$

The following properties are easily verified:

$$\begin{aligned} \phi_i(x) &= \begin{cases} 1, & \text{if } x = x_i, \\ 0, & \text{if } x = x_j, \quad i \neq j. \end{cases} \\ P_1(x_i) &= a_i = v(x_i) \\ P_1'(x_i) &= v'(x_i) \end{aligned}$$

Let $e(x) = v(x) - P_1(x)$ be the interpolation error. Since $P_1(x)$ is piecewise linear, its second derivative vanishes in $I = [0, 1]$, $P_1''(x) = 0$. Moreover, from the interpolation property, $e(x_i) = 0$ in all grid points $x_i,\ \ i = 0, \ldots, n+1$. Rolle's theorem states that there exist n points $\eta_i,\ i = 1, \ldots, n$ with $\eta_i \in [x_i, x_{i+1}]$ where $e'(\eta_i) = 0$. Thus, for $x_i \leq x \leq x_{i+1}$, we can write:

$$e'(x) = \int_{\eta_i}^{x} e''(t)\, dt = \int_{\eta_i}^{x} v''(t)\, dt,$$

from which:

$$\begin{aligned} \mid e'(x) \mid &\leq \int_{x_i}^{x_{i+1}} \mid v''(t) \mid dt = \int_{x_i}^{x_{i+1}} 1 \cdot \mid v''(t) \mid dt \leq \text{(using Cauchy inequality)} \\ &\leq \left(\int_{x_i}^{x_{i+1}} 1^2\, dt \right)^{\frac{1}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}} = h^{\frac{1}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}}, \end{aligned} \tag{2.13}$$

and:

$$\mid e'(x) \mid^2 \leq h \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right).$$

Integration between x_i and x_{i+1} yields:

$$\int_{x_i}^{x_{i+1}} \mid e'(x) \mid^2 dx \leq h^2 \int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt.$$

To evaluate $e(x)$, we first note that $e(x) = \int_{x_i}^{x} e'(t)\ dt$. Then, using (2.13) and again after integration, we obtain:

$$\mid e(x) \mid \leq h^{\frac{3}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}},$$

from which we have:

$$\int_{x_i}^{x_{i+1}} \mid e(x) \mid^2 dx \leq h^4 \int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt.$$

Summing over all mesh elements (intervals) we have the following interpolation error:

$$\begin{aligned} \left(\int_0^1 \mid e(x) \mid^2 \right)^{\frac{1}{2}} &\leq h^2 \left(\int_0^1 \mid v''(x) \mid^2 dx \right)^{\frac{1}{2}} \\ \left(\int_0^1 \mid e'(x) \mid^2 \right)^{\frac{1}{2}} &\leq h \left(\int_0^1 \mid v''(x) \mid^2 dx \right)^{\frac{1}{2}} \end{aligned}$$

or in terms of norms:

$$\begin{aligned} \|v - P_1(x)\| &\leq h^2 \|v''(x)\| \\ \|v' - P_1'(x)\| &\leq h \|v''(x)\| \end{aligned}$$

Using (2.11) and (2.12) we have the following error estimates:

$$\|u - u_h\| \leq h \|u''\| \tag{2.14}$$

$$\|(u - u_h)'\| \leq h \|u''\| \tag{2.15}$$

that show that if the second derivative of the solution is bounded, then FEM converges with an error that tends to zero proportionally to h as the mesh size parameters $h \longrightarrow 0$ ($\mathcal{O}(() h)$). We can actually prove that the error converges to zero quadratically:

$$\|u - u_h\| \leq h^2 \|u''\|. \tag{2.16}$$

if $\|u''\|$ is bounded, but this proof requires some extra work that will be done once and for all in the general multidimensional case.

Remark 2.10. *From the error estimate we can derive an estimate on the condition number of the stiffness matrix A. In fact:*

$$\kappa(A) = \frac{\lambda_1}{\lambda_N} = Ch^{-2}$$

where λ_1 and λ_N are the maximum and minimum (positive) eigenvalues of A and the constant C does not depend on h. If we use the conjugate gradient (CG) scheme to solve the linear system, this estimates tells us that the number of iterations needed to achieve e prescribed tolerance in the residual of the linear system is $\mathcal{O}\left(\sqrt{\kappa(A)}\right) = \mathcal{O}(h)$.

Some simple examples Consider the problem:

$$\begin{aligned} -u''(x) &= q \qquad\qquad x \in [0,1], \\ u(0) &= u(1) = 0. \end{aligned}$$

let $F(u)$ be the functional given by:

$$F(u) = \int_0^1 \left[\frac{1}{2}(u')^2 - qu\right] \, dx,$$

and let the numerical solution be expressed as:

$$u_n(x) = \sum_{j=1}^{n} a_j \phi_j(x).$$

Minimization of the functional $F(u)$ (Ritz method) requires that u be a stationary point for F. This yields a linear system of equations whose i-th row is given by:

$$\frac{\partial F}{\partial a_i} = \int_0^1 \left[\left(\sum_{j=1}^{n} a_j \phi_j'(x)\right)\phi_i'(x) - q\phi_i(x)\right] \, dx = 0.$$

We now need to choose the basis functions $\phi_i(x) \in \mathcal{V}_h$.

Example 2.11. We can choose the canonical basis of the vector space of polynomials of degree n:

$$\phi_i(x) = x^i \qquad\qquad i = 0, 1, \ldots, n-1.$$

Then our solution can be written as:

$$u_n(x) = x(x-1)\sum_{i=1}^{n} a_i x^{i-1}$$

where the first two terms (x and $(x-1)$ were added so that $u_n \in \mathcal{V}_h$ (recall, it must satisfy homogeneous boundary conditions). The space $\mathcal{V}_h$ is then formed by the following functions:

$$\begin{aligned}
\phi_1(x) &= x(x-1) \\
\phi_1'(x) &= 2x-1 \\
\dots &\quad \dots \\
\phi_i(x) &= x(x-1)x^{i-1} = x^{i-1} - x^i \\
\phi_i'(x) &= (i+1)x^i - ix^{i-1} \\
\dots &\quad \dots
\end{aligned}$$

For $n = 1$ we have $i = 1$ and:

$$u_n(x) = x(x-1)a_1$$

$$u_n'(x) = 1(x-1)a_1$$

$$\begin{aligned}
\frac{\partial F}{\partial a_1} &= \int_0^1 \left[a_1(2x-1)^2 - qx(x-1)\right]\, dx \\
&= \int_0^1 \left[a_1(4x^2+1-4x) - qx^2 + qx\right]\, dx = 0,
\end{aligned}$$

from which immediately we have $a_1 = -q/2$, and the numerical solution takes on the expression:

$$u_n(x) = -x(x-1)\frac{q}{2}.$$

Differentiating twice the above equation, it is immediate to see that u_n satisfies the original PDE, leading to the conclusion that $a_2 = a_3 = \ldots = a_n = 0$.

Example 2.12. Let

$$u_n(x) = \sum_{i=1}^{n} a_i \sin(i\pi x).$$

The basis functions are identified by:

$$\phi_i(x) = \sin(i\pi x) \qquad\qquad \phi_i'(x) = i\pi\cos(i\pi x).$$

The linear system (by Ritz method) becomes:

$$\frac{\partial F}{\partial a_i} = \int_0^1 \left[\left(\sum_{j=1}^{n} a_j\phi_j'(x)\right)\phi_i'(x) - q\phi_i(x)\right]\, dx = 0,$$

from which, solving for a_1 in the case $n = 1$, we have:

$$\frac{\partial F}{\partial a_1} = \int_0^1 \left[a_1 \pi^2 \cos^2(\pi x) - -q \sin(\pi x)\right] \, dx = 0,$$

or:

$$a_1 = \frac{\int_0^1 q \sin(\pi x) \, dx}{\int_0^1 \pi^2 \cos(\pi x) \, dx} = \frac{4}{\pi^3} q.$$

The numerical solution is thus:

$$u_n(x) = \frac{4}{\pi^3} q \sin(\pi x)$$

The following table reports a comparison between the numerical and the explicit (closed form) solution of the problem.

x	u/q	u_n/q
0.00	0.00	0.00
0.25	0.09375	0.09122
0.50	0.125	0.12901
0.75	0.09375	0.09122
1.00	0.00	0.00

Example 2.13. Let

$$u_n(x) = \sum_{i=1}^{n} a_i \sin(2\pi i x)$$

In this case we have $a_1 = a_2 = \ldots = a_n = 0$. What is happening is that in this case the space $\mathcal{V}_h$ spanned by the basis functions $\phi_i(x) = \sin(2\pi i x)$ does not contain the solution of our problem, and thus the FE scheme evaluates a solution that is identically zero.

2.2 Multidimensional extension

2.2.1 Differential operators.

Let $\Omega \subset \mathbb{R}^d$, and u a function $u : \Omega \longrightarrow \mathbb{R}$.

The gradient. The gradient of u is a d-dimensional vector formed by the first derivatives if u:

$$\nabla u = (\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d}).$$

Divergence. Given a vector field $q(x) \in \mathbb{R}^d$, the divergence of the vector q is formally given by the scalar product between the operator ∇ and the vector q:

$$\operatorname{div} q = \langle \nabla, q \rangle = \nabla \cdot q = \frac{\partial q_1}{\partial x_1} + \ldots + \frac{\partial q_n}{\partial x_n}.$$

Laplacian. The Laplacian of u is the function:

$$\Delta u = \operatorname{div} \nabla u = \langle \nabla, \nabla u \rangle = \nabla \cdot \nabla u.$$

Curl. The curl of the vector field q is given by the vector (external) product between the gradient vector and q. For $d = 3$ we have:

$$\operatorname{curl} q = \nabla \times q = (\frac{\partial q_3}{\partial x_2} - \frac{\partial q_3}{\partial x_3}, \frac{\partial q_2}{\partial x_3} - \frac{\partial q_3}{\partial x_1}, \frac{\partial q_2}{\partial x_1} - \frac{\partial q_1}{\partial x_2}).$$

Higher order derivatives. We will often use the "multi-index" notation for derivatives. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$ be a multi-index of order $k \mid \alpha \mid = k = \sum_{i=1}^{d} \alpha_i$. Then:

$$\partial^\alpha u = \frac{\partial^{\mid \alpha \mid} u}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}.$$

Given an integer $k \geq 0$, the symbol ∂^k is the set of all derivatives of u of order k: $\partial^k u = \{\partial^\alpha, \mid \alpha \mid = k\}$.

Weak derivative. The weak derivative (or derivative in the sense of distributions or generalized derivative) can be defined by means of formula of integration by parts.

Definition 2.3. given two functions $u, v : \Omega \longrightarrow \mathbb{R}$ and a multi-index α. Then $v = \partial^\alpha u$ is a weak derivative of u if for all smooth (infinitely continuous) functions $\phi \in C^\infty(\Omega)$ with compact support (they are zero in $\partial\Omega$) we have:

$$\int_\Omega v\phi \, dx = (-1)^{|\alpha|} \int_\Omega u \partial^\alpha \phi \, dx.$$

It is intuitive, and indeed it can be proved, that the weak derivative of a function u coincides with its standard derivative if the latter exists.

Remark 2.14. *To better appreciate the theoretical developments we should mention that all these results are valid "almost everywhere" or "everywhere except subsets of zero measure", according to the theory of Lebesgue integration and the theory of function [9, 3]. Readers who are not interested in these theoretical aspects can simply think of continuous functions that*

have enough smooth derivatives so that have finite squared integral, i.e., belong to $\mathcal{L}^2(\Omega)$ and, moreover, the scalar products and norms used in the formulas are well-defined. More detailed information can be found on any book of functional analysis [3]. Simple compendia of the results needed in these notes are given in [22, 21], and are summarized in the chapters on the mathematical theory of finite elements reported in the second part of these notes.

Gauss or divergence theorem. The principal tool that we will be using in this chapter is integration by parts and the divergence (or Gauss') theorems. Green's formula, also known as Green's first identity or Green's lemma, is the multidimensional equivalent of the well-known integration by parts. Let us start by stating the divergence theorem: Let $\Omega \subset \mathbb{R}^d$ be a compact subset of $\mathbb{R}^d$ with boundary denoted by $\Gamma = \partial\Omega$ which is sufficiently smooth. Let $\vec{F} \in \Omega$ be a vector field defined in Ω. Then

$$\int_\Omega \operatorname{div} \vec{F}\, dx = \int_\Gamma \vec{F} \cdot \nu\, ds, \tag{2.17}$$

where ν is the outward unit normal to Γ, dx is the volume measure on Ω (in $\mathbb{R}^d$) and ds is surface measure on Γ (in $\mathbb{R}^{d-1}$), and $\vec{F} \cdot \nu$ is the standard scalar product between two vectors in $\mathbb{R}^d$. Let $\vec{F} = v\vec{q}$, i.e. the vector field $F(x)$ is given by the product of a real-valued function $v(x)$ times a vector field $\vec{q}(x)$ Using the product rule of differentiation for each component of the vector scalar product, we obtain:

$$\int_\Omega \nabla v \cdot \vec{q}\, dx = \int_\Gamma v\, \vec{q} \cdot \nu\, ds - \int_\Omega v\, \operatorname{div} \vec{q}\, dx.$$

In the case that $\vec{q} = \nabla w$ we have the first Green identity or Green's Lemma:

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v\, \nabla w \cdot \nu\, ds - \int_\Omega v\, \Delta w\, dx, \tag{2.18}$$

that can be thought of a the multidimensional extension of the theorem of integration by parts by interpreting v as the primitive of ∇v and $\Delta w - \operatorname{div} \nabla w$ the derivative of ∇w.

Linear and bilinear forms. A linear form $F(v)$ is a mapping from a set of functions to the real space, $F : \mathcal{V} \mapsto \mathbb{R}$. In other words, it is a function (in more abstract sense a map) that takes functions as its only argument and returns a real value. It is linear when $F(\alpha v + \beta w) = \alpha F(v) + \beta F(w)$.
A bilinear form is a mapping from a set of pairs of functions (v, w) and the real space, $a : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$, that is linear separately in each of its two arguments, i.e., $a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$ and $a(w, \alpha u + \beta v) = \alpha a(w, u) + \beta a(w, v)$. A bilinear form is symmetric if $a(v, w) = a(w, v)$.

2.2.2 Weak formulation and FEM

Consider Poisson equation in d dimensions with $d = 2$ or $d = 3$:

Problem 2.15 (differential)**.**
Find $u : \Omega \mapsto \mathbb{R}$ such that:

$$\begin{aligned} -\Delta u &= f(x), && x \in \Omega \subset \mathbb{R}^d \\ u(x) &= 0 && x \in \Gamma, \end{aligned} \tag{2.19}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain of $\mathbb{R}^d = \{x = [x_1, x_2, \ldots, x_d], x_i \in \mathbb{R}\}$ having boundary $\Gamma = \partial\Omega$, which is assumed sufficiently regular, and Δ is the Laplacian operator:

$$\Delta = \operatorname{div} \nabla = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}.$$

The weak formulation for (2.19) is given by:

Problem 2.16 (variational)**.**
find $u \in \mathcal{V}$ such that:

$$a(u, v) = (f, v)_\Omega \qquad \forall v \in \mathcal{V}, \tag{2.20}$$

where:

$$\begin{aligned} a(u, v) &= \int_\Omega \nabla u \cdot \nabla v \, dx \\ (f, v)_\Omega &= \int_\Omega f \, v \, dx \\ \mathcal{V} &= \{v(x) : v \text{ is continuous in } \Omega, \nabla v \text{ is piecewise continuous in } \Omega \text{ and } v(x) = 0 \text{ for } x \in \Gamma\}. \end{aligned}$$

We note here that in the sequel we will drop the subscript Ω when referring to Ω in $(\cdot, \cdot)_\Omega$ and no confusion should arise.
This formulation can be derived from (2.19) as follows. Multiply be a test function$v(x) \in \mathcal{V}$ and integrate over Ω. Using Greens' Lemma we obtain:

$$(f, v) = -\int_\Omega (\Delta u) v \, dx = -\int_\Gamma v \nabla u \cdot \nu \, ds + \int_\Omega \nabla u \cdot \nabla v \, dx = a(u, v),$$

where the boundary integral is zero because $v(x) = 0$ for $x \in \Gamma$. Analogously to the one-dimensional case we can see that:

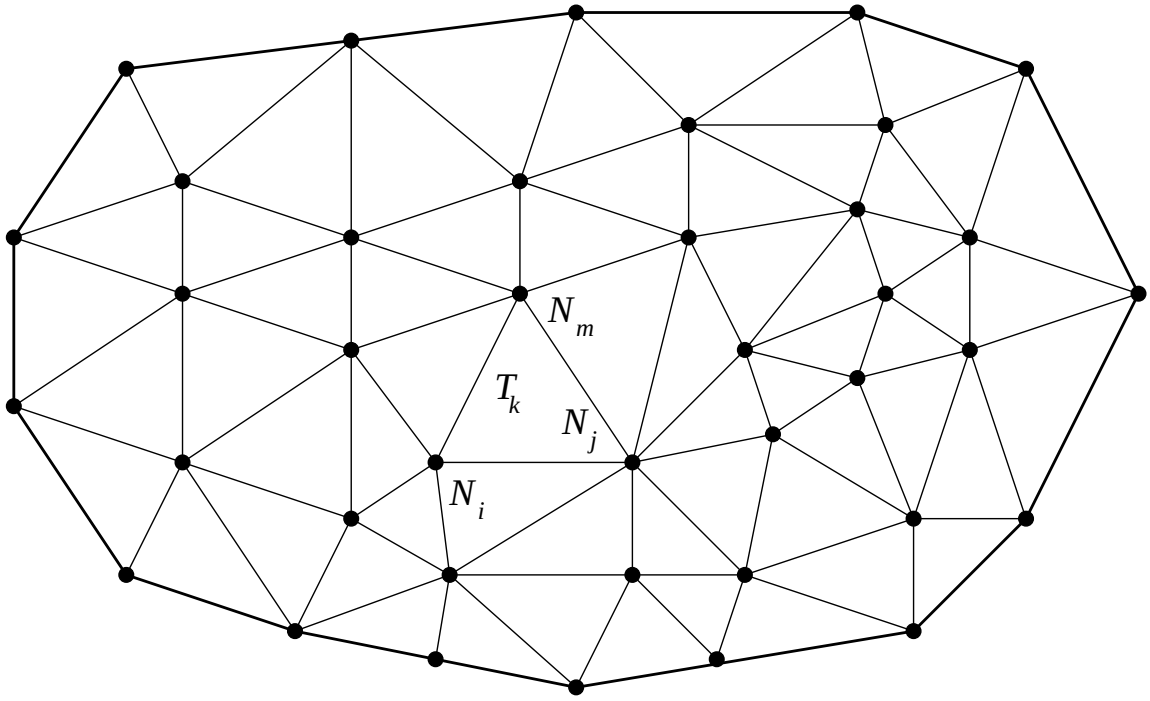


FIGURE 2.4: *Example of an admissible triangulation of* Ω. *The boundary* $\Gamma = \partial\Omega$ *is drawn with the thicker line.*

- the solution of the variational problem is solution of the differential problem if u is regular;
- the variational problem is equivalent to the following minimization problem:

  Problem 2.17 (minimization)**.**
  Find $u \in \mathcal{V}$ such that:

  $$F(u) \leq F(v) \qquad \forall v \in \mathcal{V}, \tag{2.21}$$

  where:

  $$F(v) = \frac{1}{2} a(u, v) - (f, v).$$

We need to give an appropriate definition of the basis functions. As done for $d = 1$, we need to build a computational mesh or grid, i.e., a partition of the domain Ω. We do this for $d = 2$ for simplicity. We can define a triangulation $\mathcal{T}_h(\Omega)$ of Ω formed by the union of triangles T_k such that:

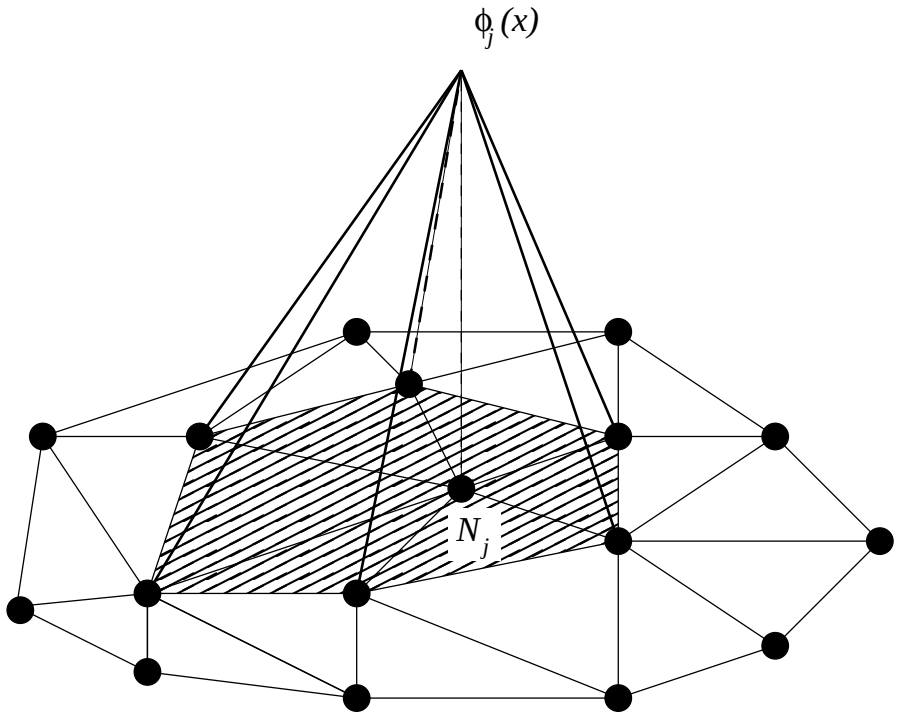


FIGURE 2.5: *Linear (pyramidal) basis function* $\phi_j(x) \in \mathcal{V}_h$.

- $\mathcal{T}_h(\Omega)$ is formed by n nodes (triangle vertices named N_i, $i = 1, \ldots, n$, with given coordinates), and m triangles (identified by T_k, $k = 1, \ldots, m$);
- $\Omega = \bigcup_{T_k \in \mathcal{T}_h} T_k = T_1 \cup T_2 \ldots \cup T_m$;
- $T_i \cap T_j = e_{ij}$, $i \neq j$, where e_{ij} denotes the edge shared by triangles T_i and T_j;
- no vertex N_i lies in the interior of a triangle edge;
- the boundary triangles have at least one vertex on the boundary.

An example of an admissible triangulation is given in Figure 2.4. Note that to derive convergence estimates we need to require that the domain boundary does not change as the mesh varies. For this reason our domain boundary is formed by piecewise linear segments.
We now introduce the mesh parameter h defined as:

$$h = \max_{T_i \in \mathcal{T}_h} \text{diam } (T_i), \tag{2.22}$$

where the triangle diameter diam (T_i) is defined as the longest edge of T_i. We can define the finite dimensional space $\mathcal{V}_h$ as:

$$\mathcal{V}_h = \{v(x) : v \text{ is continuous in } \Omega, v|_{T_i} \text{ is linear in each } T_i \in \mathcal{T}_h, v(x) = 0 \text{ for } x \in \Gamma\}.$$

where $v|_{T_i}$ is the restriction of the test function v to triangle T_i. Obviously $\mathcal{V}_h \subset \mathcal{V}$. Now we want to be able to define Lagrangian interpolation of functions. We use the internal nodes $\tilde{N}_i$ of the triangulation and build the basis functions $\phi_i(x)$, $i = 1, \ldots, n$ as:

$$\phi_i(x_j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \qquad i, j = 1, \ldots, n$$

These are piecewise linear functions of pyramidal shape, as shown in Figure 2.5, and with support given by the unions of all triangles sharing node N_j. Thus we can express a generic function $v \in \mathcal{V}_h$ as:

$$v(x) = \sum_{j=1}^{n} \eta_j \phi_j(x), \qquad \eta_j = v(x_j),$$

and the Galerkin FEM method becomes:

Problem 2.18 (Galerkin FEM)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a(u_h, v) = (f, v) \qquad \forall v \in \mathcal{V}_h. \tag{2.23}$$

This yields the following

$$\sum_{j=1}^{n} a(\phi_i, \phi_j)\, u_j = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{2.24}$$

which is the FEM linear systems that in matrix form can be written as:

$$Au = b$$

where the stiffness matrix A and the load vector b are given by:

$$A_{[n\times n]} = \{a_{ij}\} \qquad a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, dx \tag{2.25}$$

$$u_{[n\times 1]} = \{u_i\}, \qquad b_{[n\times 1]} = \{b_i\} \qquad b_i = (f, \phi_i) = \int_\Omega f\phi_i \, dx. \tag{2.26}$$

Note that now the matrix coefficients are evaluated via a d-dimensional scalar product on Ω. Analogously to what done for $d = 1$, we can prove that A is SPD.

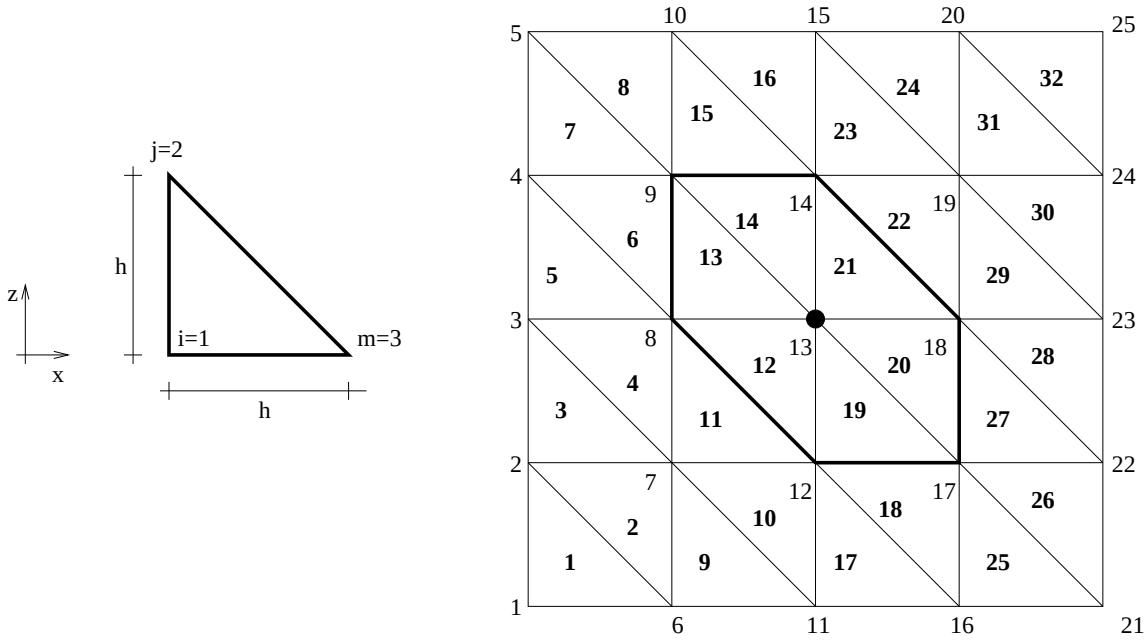


FIGURE 2.6: *Regular triangulation on a square domain.*

In the case of a square domain and regular discretization with equal triangles of edge length h (Figure 2.6) the stiffness matrix is penta-diagonal and of the form:

$$\begin{bmatrix} 4 & -1 & 0 & 0 & 0 & -1 & \dots & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & 0 & 0 & -1 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & -1 & \dots & -1 & 4 & -1 & \dots & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & -1 & \dots & 0 & -1 & 4 & -1 \\ 0 & \dots & \dots & \dots & -1 & 0 & 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix},$$

where in the case $f = cost$ we calculate:

$$b_i = fh^2$$

It coincides with the "5=point stencil" of the second order finite difference discretization of the Laplacian operator [23].

Remark 2.19. *The practical evaluation of the stiffness matrix proceeds via the so called "assembly" process, an element-by-element procedure typical of FE methods. This procedure evaluates local elemental matrices and then sums their contributions to build the global system matrix. This procedure introduces flexibility in handling complicated geometries and heterogeneities in the coefficients of the PDE, which can then be described element-wise.*

Another advantage deriving essentially from the organization of this procedure is the fact that it is possible to assemble local elements into a "super element". This is called "static condensation" and is useful in particular in parallel applications.
Finally, again the element-by-element assembly procedure allows the efficient handling of local mesh refinements. We can think of building a mesh with a characteristic size h_j that can be dynamically adapted to the size of the error. For these, we need what are called "a-posteriori" error estimations that will be treated in subsequent chapters.

2.2.3 Convergence of FEM in the multidimensional case

The results derived for the one-dimensional case in section 2.1.4 can be extended to the multivariate case with only technical difficulties. We assume that we have a homogeneous Dirichlet problem, so that $u \in \mathcal{V} = \mathcal{H}_0^1$. Moreover, we assume that the bilinear form $a(\cdot, \cdot)$ is continuous and coercive, i.e.:

- *continuity* there exists a constant $\gamma > 0$ such that:

$$| a(u, v) | \leq \gamma \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \qquad \forall u, v \in \mathcal{V}; \tag{2.27}$$

- *coerciveness* there exists a constant $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|_{\mathcal{V}}^2 \qquad \forall v \in \mathcal{V}. \tag{2.28}$$

Continuity implies that small changes in the arguments of the bilinear form do to not lead to large changes in the values attained by the bilinear form. Coerciveness implies that the bilinear form is always bounded away from zero, and can thus be inverted in some sense.
Under these assumptions, we can show that $u_h \in \mathcal{V}_h$ is the best approximation in the sense that there exist a constant C independent of h such that:

$$\|\nabla u - \nabla u_h\| \leq C \|\nabla u - \nabla v\| \qquad \forall v \in \mathcal{V}_h,$$

where now the norm is defined as:

$$\|\nabla v\| = a(v, v)^{\frac{1}{2}} = \left(\int_\Omega |\nabla v|^2 \, dx \right)^{\frac{1}{2}},$$

showing the optimality of Galerkin solution with respect to the $\mathcal{L}^2$ norm. We note here that, for a homogeneous Dirichlet problem where v vanishes at the boundary, the above norm is equivalent to the norm of the function v in the sense that there exist two constants C_1 and C_2 such that

$$C_1 \|\nabla v\| \leq \|v\| \leq C_2 \|\nabla v\|$$

where

$$\|v\| = \left(\int_\Omega v^2 + |\nabla v|^2 \, dx \right)^{\frac{1}{2}},$$

We can then use Lagrangian interpolation to show immediately that:

$$\|\nabla u - \nabla \tilde{u}_h\| \leq Ch.$$

Finally, also in the multidimensional case it is possible to show that:

$$\|u - u_h\| = \left(\int_\Omega (u - u_h)^2 \, dx \right)^{\frac{1}{2}} \leq Ch^2,$$

which is the equivalent of (2.16) in $\mathbb{R}^d$. To arrive at this result we need to assume some regularity properties of the triangulation, namely that when the diameter given in (2.22) $h \longrightarrow 0$ the triangles do not degenerate, i.e., the vertices of a triangle never align on a line.

Optimality of the solution of the variational problem. We look here at the simple Dirichlet boundary value problem:

$$\begin{aligned} -\Delta u + u &= f \qquad x \in \Omega \\ u &= 0 \qquad x \in \Gamma = \partial\Omega \end{aligned}$$

The variational formulation becomes:

Problem 2.20 (Variational Formulation)**.** Find $u \in \mathcal{H}_0^1(\Omega)$ such that:

$$a(u, v) = (f, v) \qquad \forall v \in \mathcal{H}_0^1(\Omega) \tag{2.29}$$

where:

$$a(u, v) = \int_\Omega \left[\nabla u \cdot \nabla v + uv \right] \, dx$$

The related FEM problem becomes:

Problem 2.21 (FEM Problem)**.** Find $u_h \in \mathcal{V}_h(\Omega) \subset \mathcal{H}_0^1(\Omega)$ such that:

$$a(u_h, v) = (f, v) \qquad \forall v \in \mathcal{V}_h(\Omega). \tag{2.30}$$

Subtracting (2.30) from (2.29), we can appreciate the strong consistency of the FE scheme:

$$a(u - u_h, v) = 0 \qquad \forall v \in \mathcal{V}_h(\Omega)$$

which also says that the error function $e = u - u_h$ is orthogonal to the basis functions of $\mathcal{V}_h(\Omega)$ with respect to the scalar product $a(\cdot,\cdot)_{\mathcal{V}}$. This statement is equivalent to saying that u_h is the orthogonal projection of u onto $\mathcal{V}_h(\Omega)$ with respect to the scalar product $a(\cdot,\cdot)$. In other words, noting that this scalar product the the $\mathcal{H}^1(\Omega)$ scalar product , u_h is characterized by an $\mathcal{H}^1$ norm of the error $u - u_h$ that is smaller than any other function $v \in \mathcal{V}_h(\Omega)$:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq \|u - v\|_{\mathcal{H}^1(\Omega)} \qquad \forall v \in \mathcal{V}_h(\Omega)$$

stating the optimality of u_h in $\mathcal{V}_h(\Omega)$. These theoretical results will be derived with more details and rigor in section 2.6

2.3 Non-homogeneous boundary conditions

2.3.1 Neumann problem: natural and essential boundary conditions

Consider the following pure-Neumann problem:

$$\begin{aligned} -\Delta u + u &= f && \text{in } \Omega, \\ \nabla u \cdot \nu &= g && \text{in } \Gamma = \partial\Omega. \end{aligned} \tag{2.31}$$

Multiplying the first equation by the test function $v \in \mathcal{V}$ and integrating over Ω we obtain:

$$-\int_\Omega (\Delta u v - uv)\, dx = \int_\Omega fv\, dx. \tag{2.32}$$

Applying Green's lemma:

$$\int_\Omega uv\, dx - \int_\Gamma \nabla u \cdot \nu v\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega fv\, dx,$$

where $\Gamma = \partial\Omega$ is the (sufficiently smooth) boundary of Ω. The second integral of the left-hand-side of this equation contains exactly the flux boundary term $\nabla u \cdot \nu$. In the cases previously encountered of homogeneous Dirichlet conditions, we required that the test and basis functions where zero on the boundary, leading to the nullification of the boundary integral deriving from Green's lemma. In this case, requiring that the test functions are nonzero at the boundary automatically implies that the Neumann boundary conditions are satisfied once we have substituted q in place of $\nabla u \cdot \nu$. Thus, $\mathcal{V} =$ we can write the following:

Problem 2.22 (Variational formulation)**.**
Find $u \in \mathcal{V}$ such that:

$$a(u,v) = (f,v) + (g,v)_\Gamma \qquad \forall v \in \mathcal{V}, \tag{2.33}$$

where:

$$\begin{aligned} a(u,v) &= \int_\Omega (\nabla u \cdot \nabla v + uv)\, dx \\ (f,v) &= \int_\Omega f\, v\, dx \\ (g,v)_\Gamma &= \int_\Gamma g\, v\, ds \\ \mathcal{V} &= \{v(x) : v \text{ is continuous in } \Omega, \nabla v \text{ is piecewise continuous in } \Omega\}, \end{aligned}$$

equivalent to the minimization problem:

Problem 2.23 (Minimization problem)**.**
Find $u \in \mathcal{V}$ such that:

$$F(u) \leq F(v) \qquad \forall v \in \mathcal{V} \tag{2.34}$$

where:

$$F(v) = \frac{1}{2} a(v,v) - (f,v) - (g,v)_\Gamma .$$

As done before, assuming u sufficiently regular and applying Green's lemma "backward" to (2.33) we have:

$$\int_\Omega (-\Delta u + u - f)\, v\, dx + \int_\Gamma (\nabla u \cdot \nu - g)\, v\, ds = 0 \qquad \forall v \in \mathcal{V}. \tag{2.35}$$

Since all the functions $v \in \mathcal{V}$ are non zero at the boundary we derive the following two conditions:

$$\int_\Omega (-\Delta u + u - f)\, v\, dx = 0 \qquad \forall v \in \mathcal{V},$$

and

$$\int_\Gamma (\nabla u \cdot \nu - g)\, v\, ds = 0 \qquad \forall v \in \mathcal{V}.$$

Varying $v \in \mathcal{V}$ (again v is nonzero in Γ), we can apply Lemma 2.1 to obtain:

$$-\Delta u + u - f = 0 \qquad \text{in } \Omega,$$

and

$$\nabla u \cdot \nu - g = 0 \qquad \text{in } \Gamma;$$

which states that the boundary conditions are satisfied.

Remark 2.24. *Neumann boundary conditions are not applied explicitly in the variational formulation but appear as a natural term that does not vanish because the test functions do not vanish at the boundary. We can then say that while Dirichlet conditions must be imposed directly on the functional space where the solution is sought, i.e., must be imposed explicitly on the solution candidate function, Neumann boundary conditions are satisfied naturally by the formulation. For these reasons Neumann boundary conditions are called "natural" while Dirichlet boundary conditions are called "essential". Note that, if $g = 0$ (no flow at the boundary) the term $(g, v)_\Gamma$ would disappear from the formulation (from this the word "natural").*

The Galerkin FE formulation can be written as:

Problem 2.25 (Galerkin FEM)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a(u_h, v) = (f, v) + (g, v)_\Gamma \qquad \forall v \in \mathcal{V}_h, \tag{2.36}$$

where:

$$\begin{aligned} a(u_h, v) &= \int_\Omega (\nabla u_h \cdot \nabla v + u_h v)\, dx \\ \mathcal{V}_h &= \{v(x) : v \text{ is continuous in } \Omega, v|_{T_k} \text{ is linear } \ \forall T_k \in \mathcal{T}_h\}. \end{aligned}$$

Remark 2.26. *We observe that problem 2.31 admits a unique solution. In fact, if $\tilde{u} \neq u$ solves the problem, then the function $w = \tilde{u} - u$ solves the same problem with $f = 0$ and $g = 0$. Using, e.g., the maximum principle or energy methods, we see immediately that this problem admits the only solution $w = 0$, contradicting the hypothesis. This is a consequence of the presence of the term u in the right hand side of the equation. Had we tackled Poisson equation, the same reasoning show that the solution of the pure Neumann problem is defined up to an additive constant. We can think intuitively that the role of the presence of the term u in the equation is to "fix" the constant. This is similar to the role played by Dirichlet boundary conditions. Thus we conclude that to obtain a well-posed problem we need "at least" one point where Dirichlet conditions are specified.*

2.3.2 Cauchy (or Robin) problem

Consider the following problem:

$$-\Delta u = f \qquad \text{in } \Omega, \tag{2.37a}$$
$$\nabla u \cdot \nu + \gamma u = g \qquad \text{in } \Gamma = \partial\Omega, \tag{2.37b}$$

where $\gamma > 0$. Multiplying the first equation by the test function $v \in \mathcal{V}$ and integrating over Ω we obtain:

$$-\int_\Omega (\Delta u v)\, dx = \int_\Omega f v\, dx.$$

Application of Green's lemma yields:

$$-\int_\Gamma \nabla u \cdot \nu v\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx,$$

where $\Gamma = \partial\Omega$ is the (sufficiently smooth) boundary of Ω. Using (2.37b) we have:

$$-\int_\Gamma (g - \gamma u) v\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx,$$

or:

$$\int_\Omega \nabla u \cdot \nabla v\, dx + \int_\Gamma \gamma u v\, ds = \int_\Omega f v\, dx + \int_\Gamma g v\, ds$$

giving rise to the following variational formulation:

Problem 2.27 (Variational formulation)**.**
Find $u \in \mathcal{V}$ such that:

$$a_\gamma(u, v) = (f, v) + \gamma (g, v)_\Gamma \qquad \forall v \in \mathcal{V}, \tag{2.38}$$

where:

$$\begin{aligned} a_\gamma(u, v) &= a(u, v) + \gamma (u, v)_\Gamma \\ &= \int_\Omega \nabla u \cdot \nabla v + \gamma \int_\Gamma u\, v\, dx \end{aligned}$$

The Galerkin FE formulation can be written as:

Problem 2.28 (Galerkin FEM)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a_\gamma(u_h, v) = (f, v) + \gamma (g, v)_\Gamma \qquad \forall v \in \mathcal{V}_h. \tag{2.39}$$

Note that as in the pure Neumann case $v \in \mathcal{V}_h$ is nonzero at the boundary Γ. We would like to note that the bilinear form $a_\gamma(\cdot, \cdot)$ may not be coercive for $\gamma < 0$.

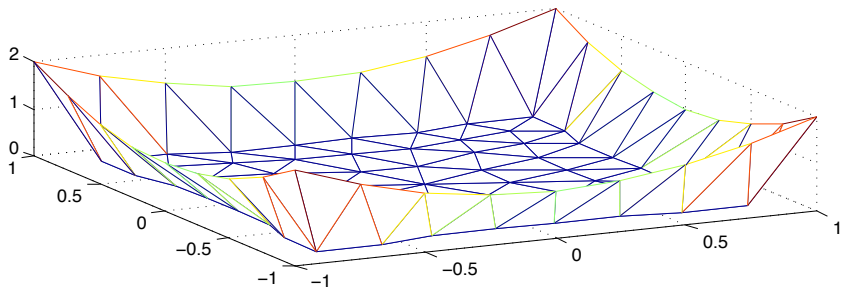


FIGURE 2.7: *Projection onto* $\mathcal{V}_h \subset \mathcal{H}^1$ *of* u_g *(eq.* (2.44)*).*

2.3.3 Non homogeneous Dirichlet problem

Consider the following problem:

$$-\Delta u = f \qquad \text{in } \Omega, \tag{2.40a}$$

$$u = g \qquad \text{in } \Gamma. \tag{2.40b}$$

To incorporate non-homogeneous Dirichlet conditions we can make use of the remark 2.24 and impose the boundary conditions on all the solution candidates. In other words we look for functions $u \in \mathcal{H}^1_\Gamma$, where $\mathcal{H}^1_\Gamma$ is the set of functions in $\mathcal{H}^1$ that coincide with g in Γ. We obtain the following variational formulation: find $u \in \mathcal{H}^1_\Gamma$ such that:

$$a(u, v) = (f, v) \qquad \forall v \in \mathcal{H}^1_0.$$

However, $\mathcal{H}^1_\Gamma$ is not an affine space, as the sum of two such functions does not belong to this space (it is equal to $2g$ on the boundary). Hence we cannot use linear combinations of this space to approximate the solution and set up a FE formulation. The solution is to define a smooth enough "lifting" function u_g that satisfies (2.40b) on Γ and we let:

$$u = u_g + u_0 \tag{2.41}$$

where $u_g \in \mathcal{H}^1_\Gamma$ and $u_0 \in \mathcal{H}^1_0$. We can write the following variational formulation:

Problem 2.29 (Variational formulation for non-homogeneous Dirichlet BCs)**.**
Find $u_0 \in \mathcal{H}^1_0$ such that:

$$a(u_0, v) = (f, v) - a(u_g, v) \qquad \forall v \in \mathcal{H}^1_0. \tag{2.42}$$

This problem has now the sought form. It is easy to see that everything is well defined and we will show in later chapters that all the important well-posedness, stability theorems, and convergence theorems hold. The corresponding Galerkin FE formulation reads:

Problem 2.30 (Galerkin FEM for non-homogeneous Dirichlet BCs)**.**
Find $u_{0,h} \in \mathcal{V}_h \subset \mathcal{H}_0^1$ such that:

$$a\left(u_{0,h}, v\right) = \left(f, v\right) - a\left(u_{g,h}, v\right) \qquad \forall v \in \mathcal{V}_h. \tag{2.43}$$

The problem is indeed well defined but the question remains on how should we choose the "lifting" function u_g. There are general regularity theorems that guarantees the existence of such function in the appropriate spaces that are called "trace" theorems. We are not discussing these theorems in these notes for the main reason that in practical applications one is not interested in reproducing the exact function g but rather a numerical approximation of g. The easiest procedure is then to define $u_{h,g}$ on the space $\mathcal{V}_h$ as the projection of g onto $\mathcal{V}_h$. Thus, noting that g is defined only on Γ, we can define u_g as:

$$u_g(x) = \begin{cases} g(x), & \text{if } x \in \Gamma, \\ 0, & \text{if otherwise.} \end{cases} . \tag{2.44}$$

and we use the approximation $u_{g,h}$ defined by the projection of this function onto the subspace $\mathcal{V}_h{}'$ of $\mathcal{H}^1$ (and not of $\mathcal{H}_0^1$) generated by FEM basis functions that are nonzero on the boundary, exactly as in the case of a pure Neumann problem. An example of $u_{g,h}$ if given in Figure 2.7.

Remark 2.31. *The Dirichlet boundary conditions are imposed explicitly in "strong form", while the Neumann BCs are imposed in "weak" form. In practical applications this correspond that local errors on Dirichlet nodes are proportional to the residual of the system solution, while in Neumann or Cauchy nodes errors are governed by the FEM approximation, and thus go to zero quadratically with the mesh parameter h (see* (2.16)*).*

The penalty method. A different approach to impose non-homogeneous Dirichlet conditions, sometimes called the "penalty" method, is via Cauchy boundary treatment. Using this idea, the Dirichlet problem (2.40) is transformed into the following:

$$-\Delta u = f \qquad \text{in } \Omega, \tag{2.45a}$$

$$\alpha \nabla u \cdot n + \lambda(u - g) = 0 \qquad \text{in } \Gamma. \tag{2.45b}$$

Using (2.38), we obtain the variational formulation:

Problem 2.32 (Variational formulation with penalty)**.**
Find $u \in \mathcal{V}$ such that:

$$a_\lambda\left(u, v\right) = \left(f, v\right) + \lambda\left(g, v\right)_\Gamma \qquad \forall v \in \mathcal{V}, \tag{2.46}$$

The corresponding Galerkin FE formulation can be written as:

Problem 2.33 (Galerkin FEM)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a_\lambda(u_h, v) = (f, v) + \lambda (g, v)_\Gamma \qquad \forall v \in \mathcal{V}, \tag{2.47}$$

In practice, the i-th equation becomes:

$$\sum_{k=1}^{i-1} a_{ik} u_k + \lambda (\phi_i, \phi_i)_\Gamma u_i + \sum_{k=i+1}^{n} a_{ik} u_k = \lambda (g, \phi_i)_\Gamma$$

If λ is large enough, all the extra-diagonal terms (the terms with the summations) of the left-hand-side be negligible with respect to the diagonal term. Hence, the previous equation practically corresponds to:

$$\lambda (\phi_i, \phi_i)_\Gamma u_i = \lambda (g, \phi_i)_\Gamma, \qquad \text{or} \qquad \lambda u_i = \lambda g_i,$$

i.e., the direct enforcement of the Dirichlet boundary function, projected on the discretized boundary. A typical value for λ is $\lambda = 10^{30}$, but care must be taken that the extra-diagonal coefficient matrix are far away from the value of the penalty, so that no ill-conditioning is introduced in the system matrix.

2.3.4 Implementation notes

It is intuitive that a problem can have varying boundary conditions, i.e., the boundary may be formed by the union of three non-overlapping subsets $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C$ where Dirichlet, Neumann, or Cauchy boundary conditions are specified. The problem now reads:

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= g_D && \text{in } \Gamma_D, \\ \nabla u \cdot \nu &= g_N && \text{in } \Gamma_N, \\ \nabla u \cdot \nu + \gamma u &= g_C && \text{in } \Gamma_C \end{aligned}$$

In light of Remark 2.26, we will require $\Gamma_D \neq \emptyset$. To better understand the practical implementation for a general boundary condition, we work with a triangulation $\mathcal{T}_h$ having n nodes. The boundary is discretized with $N_D + N_N + N_C$ nodes, and we require $N_D \neq 0$. In the case $N_N = N_C = 0$ then Γ_N and Γ_C are empty and the corresponding integrals vanish. A general FEM implementation works with basis functions that are nonzero at the boundary, so that Neumann and Cauchy boundary conditions can be accommodated easily. We renumber the mesh nodes so that the first N_D the Dirichlet boundary nodes while the other nodes (internal

plus Neumann and Cauchy) are labeled from N_D+1 to $n = N+N_D$[3]. The numerical solution is then expressed as a linear combination of these basis functions:

$$u_h(x) = u_{g_D,h} + u_{0,h} = \sum_{j=1}^{N_D} u_{g_D,j}\phi_j(x) + \sum_{j=N_D+1}^{n} u_j\phi_j(x).$$

Introducing this expression, the FEM system can be rewritten as:

$$\sum_{j=N_D+1}^{n} a_\gamma(\phi_i,\phi_j)\, u_j = (f,\phi_i) - \sum_{j=1}^{N_D} a(\phi_i,\phi_j)\, u_{g_D,j} + (g_N,\phi_i)_{\Gamma_N} + \gamma(g_C,\phi_i)_{\Gamma_C} \quad i = N_D+1,\ldots,n.$$

As usual, the elements of the symmetric stiffness matrix (of dimension $n-N_D$) can be expressed as $a_{ij} = a_\gamma(\phi_i,\phi_j)$. This system can then be completed with the Dirichlet boundary conditions. Thus, we can write the global FEM system using block matrices as follows:

$$\begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where the vectors are given by:

$$u^{(1)} = \left\{u_i\right\} \qquad b_1 = \left\{u_{g_D,i}\right\} \qquad i = 1,\ldots,N_D$$

$$u^{(2)} = \left\{u_i\right\} \qquad b_2 = \Big\{ (f,\phi_i) - \sum_{j=1}^{N_D} a(\phi_i,\phi_j)\, u_{g_D,j} + (g_N,\phi_i)_{\Gamma_N} + \gamma(g_C,\phi_i)_{\Gamma_C} \Big\} \qquad i = N_D+1,\ldots,n$$

The upper left block of this system (here I denotes the N_D-dimensional identity matrix) imposes the Dirichlet conditions. In essence, the term $-\sum_{j=1}^{N_D} a(\phi_i,\phi_j)\, u_{g_D,j}$ is the consequence of the fact that $u_{h,j} = u_{g_D,j}$ on the Dirichlet node j, and are moved on the right hand side of all equations (i.e., for all i) of the linear system.

2.4 Types of Finite Elements

We have seen so far the use of linear basis functions to define the discrete space $\mathcal{V}_h$. It is intuitive that we can use for this purposes interpolating polynomials of any degree defined on the elements. For example, quadratic functions can easily be introduced in both one-dimensional intervals and two-dimensional triangles. In this case, we will need to specify 1D elements with three nodes and 2D triangles with 6 nodes (see Figure 2.8, all of them at the triangle boundary. This approach guarantees continuity of the representation of the solution, a

[3]This node renumbering is not done in practice, but it helps the exposition.

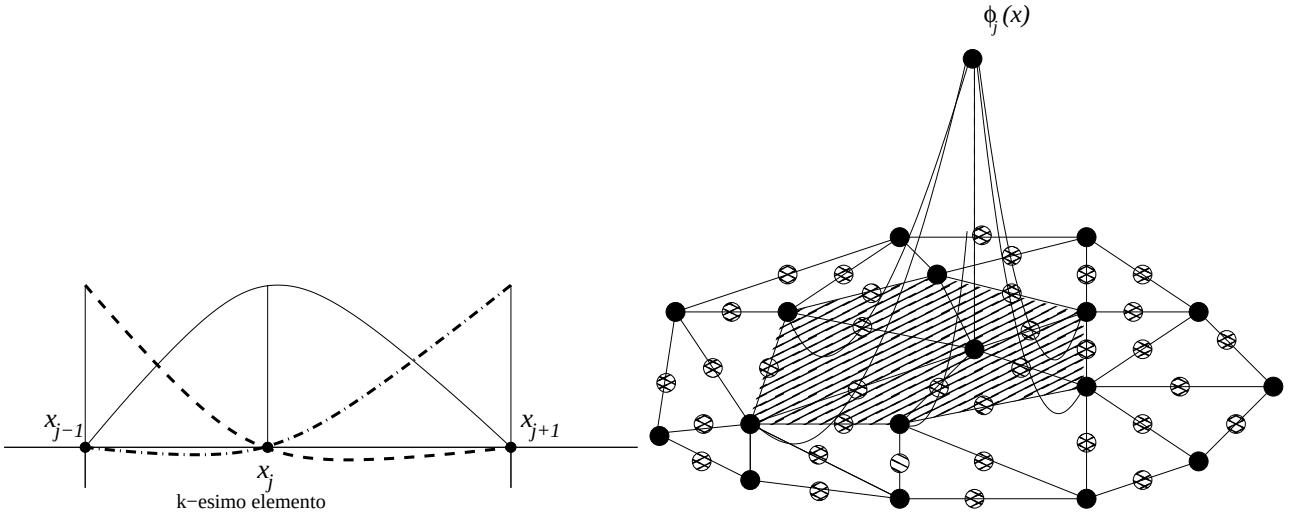


FIGURE 2.8: *Basis functions for quadratic elements in* $1D$ *(left) and 2D (right).*

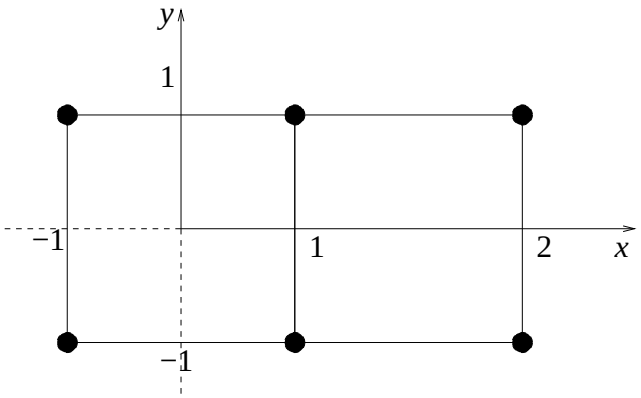


FIGURE 2.9: *Bilinear basis function on square elements.*

requirement for functions in $\mathcal{H}^1$. To see this, we need to look at element boundaries and verify that three nodes in each triangle side define uniquely the same quadratic function along the side. In other words, the two quadratic polynomials defined on the two neighboring elements have the same trace on the common side. The same can be done with three-dimensional tetrahedra. If we want to introduce different shaped elements, e.g., quadrilaterals in 2D, maintaining the continuity at inter-element edges becomes more complicated. This is done using the so-called "isoparametric" elements via appropriate transformations that allow the simple definition of basis functions to be used essentially in the evaluations of the integrals in (2.25).

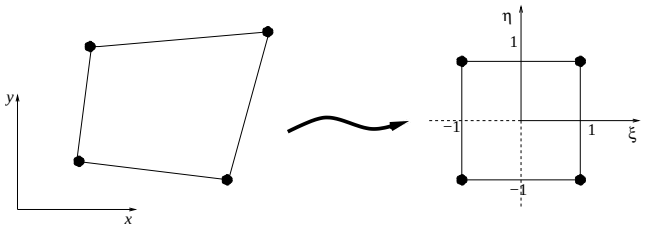


FIGURE 2.10: *Generic quadrilateral element and the transformed reference square element.*

2.4.1 Isoparametric elements

We do not want to go into details of approximation theory and work out the most general case. We content ourself on the description of the technique for general quadrilateral elements in two dimensions. We start from the simple example of square elements. In Figure 2.9 we have shown two adjacent square elements. Continuity of u_h is ensured is we use bilinear interpolation functions, i.e., polynomials that are separately linear in x and y:

$$\phi_i(x, y) = (a_i + b_i x)(c_i + d_i y).$$

It is easy to see that at the element edge located at $x = 1$ the representation of the basis function is linear (function of y) and is determined uniquely by the two nodes defining the endpoints of the edge that are common to both squares. The expression for ϕ_i there is then independent on the location of the other nodes of the two adjacent elements. In our case we have four coefficients that are used to determine the basis function elementwise. Since these are Lagrangian interpolating polynomials, we have exactly four independent conditions and the elementwise evaluation of the coefficient is well-posed. Extension to higher order (always x and y separate as before) is straight forward.
In the case of general quadrilateral elements, where the edges are not aligned with the coordinate axes, we need to resort to a coordinate transformation for each element. In practice, the idea is to transform the reference system for the element of interest so that its edges are aligned with the local reference axis. In this transformed reference system we can define the basis functions as done above to guarantee continuity of the representation. The inverse transformation is then used to evaluate the needed integrals. Figure 2.10 shows an example of such a transformation, typically a conformal mapping.
We exemplify this approach using an example involving bilinear basis functions. With reference to Figure2.10, it is easy to see that the mapping $(x, y) \mapsto (\eta, \xi)$ is given by:

$$\begin{aligned} x &= \tfrac{1}{4}\left[(1-\xi)(1-\eta)x_i + (1+\xi)(1-\eta)x_j + (1+\xi)(1+\eta)x_m + (1-\xi)(1+\eta)x_k\right] \\ y &= \tfrac{1}{4}\left[(1-\xi)(1-\eta)y_i + (1+\xi)(1-\eta)y_j + (1+\xi)(1+\eta)y_m + (1-\xi)(1+\eta)y_k\right]. \end{aligned}$$

We recall that we want to evaluate integrals of the type (2.25). Thus we need to evaluate the

Jacobian of the transformation:

$$J = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix},$$

so that the integral of the generic function f can be evaluated as:

$$\int_{\Omega^e} f(x,y)\mathrm{det}J\ dxdy = \int_{-1}^{1}\int_{-1}^{1} f(\eta,\xi)\ d\eta d\xi.$$

The actual value is generally calculated using quadrature formulas, e.g. a Gaussian formula with 4 points, recalling for vector quantities the (covariant) transformation formula:

$$\nabla_{(\xi,\eta)} f = J\nabla f.$$

2.5 Convection diffusion equation

We address here the more complicated elliptic problem given by the following convection-diffusion (or advection-diffusion) equation:

$$\begin{aligned} -\operatorname{div}(D\nabla u) + \operatorname{div}(\beta u) &= f && \text{in } \Omega \\ u &= 0 && \text{in } \Gamma_D \\ D\nabla u \cdot \nu &= g && \text{in } \Gamma_N, \end{aligned} \tag{2.48}$$

where D is the diffusion coefficient (for now a positive scalar) and $\beta(x)$ is a vector field. From the physical point of view, this equation may represent the transport of a solute dissolved in a fluid that moves with the velocity field $\beta(x)$.
The typical variational formulation can be obtained as done before by multiplying by a test function v and integrating on the domain Ω:

$$-\int_\Omega \operatorname{div} D\nabla u\, v\ dx + \int_\Omega \operatorname{div}(\beta u) v\ dx = \int_\Omega f\, v\ dx.$$

Application of Green's Lemma only to the first term yields:

$$-\int_{\Gamma_N} g\, v\ ds + \int_\Omega D\nabla u \cdot \nabla v\ dx + \int_\Omega \operatorname{div}(\beta u)\, v\ dx = \int_\Omega f\, v\ dx,$$

from which we can deduce the following (Galerkin) finite element formulation:

Problem 2.34 (Galerkin)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a(u_h, v) = (f, v) + (g, v)_\Gamma \qquad \forall v \in \mathcal{V}_h, \tag{2.49}$$

where:

$$\begin{aligned} a(u_h, v) &= \int_\Omega (D\nabla u_h \cdot \nabla v + \operatorname{div}(\beta u_h) v)\ dx \\ (f, v) &= \int_\Omega f\, v\, dx \\ (g, v)_\Gamma &= \int_\Gamma g\, v\, ds \\ \mathcal{V}_h &= \{v(x) : v \text{ is continuous in } \Omega, v(x) = 0 \text{ in } \Gamma_D, v|_{T_k} \text{ is linear } \ \forall T_k \in \mathcal{T}_h; \}. \end{aligned}$$

Remark 2.35. *Note that now the bilinear form is not symmetric anymore, $a(u, v) \neq a(v, u)$, and hence there is no associated minimization (Ritz) problem. Moreover, the FE linear system is not symmetric, although it remains obviously sparse.*

The corresponding linear system becomes:

$$(A + B)u = c,$$

where A the classical symmetric stiffness matrix seen before and B represents the non-symmetric transport (or convection) matrix:

$$\begin{aligned} A = \{a_{ij}\} \qquad & a_{ij} = \int_\Omega D\nabla\phi_j \cdot \nabla\phi_i\, dx \\ B = \{b_{ij}\} \qquad & b_{ij} = \int_\Omega \operatorname{div}(\beta\phi_j)\, \phi_i\, dx \\ c = \{c_i\} \qquad & c_i = \int_\Omega f\, \phi_i\, dx + \int_{\Gamma_n} g\, \phi_i\, ds_N. \end{aligned}$$

2.5.1 One dimensional case

Consider the following one-dimensional problem:

$$\begin{aligned} -Du'' + bu' = 0, \qquad & 0 < x < 1, \\ u(0) = 0; \qquad & u(1) = 1. \end{aligned} \tag{2.50}$$

It is an ordinary differential equation and the corresponding boundary-value problem can be solved easily. The characteristic equation is given by:

$$-D\lambda^2 + b\lambda = 0,$$

whose roots are $\lambda_1 = 0$ e $\lambda_2 = b/D$. The general solution is then given by:

$$u(x) = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x} = c_1 + c_2 e^{bx/D}.$$

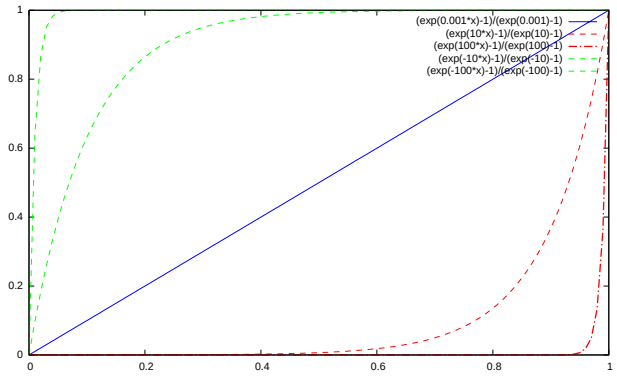


FIGURE 2.11: *Solution of the one-dimensional convection-diffusion equation for different values of the ratio b/D.*

Imposing the two boundary conditions we find:

$$u(x) = \frac{e^{\frac{b}{D}x} - 1}{e^{\frac{b}{D}} - 1},$$

whose behavior at different values of b/D is shown in Figure 2.11. Looking at this plot, we see that for small values of b/D the solution tends to be linear, while for large values the solution shows a strong exponential behavior characterized by local areas of the domain where large gradients are found.
The FE formulation that uses linear basis functions can be readily derived and shown to be equivalent to the standard finite different method (see 2.1.3). The i-th equation is given by:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0. \tag{2.51}$$

This equation corresponds to a second-order center discretization of the first and second derivatives (see Appendix A).
We introduce now the mesh Péclet number as the non-dimensionalized ratio between the local convective and diffusive fluxes:

$$\mathbb{P}\mathrm{e} = \frac{|b|h}{D}.$$

Assuming $b > 0$, the following difference equation is derived:

$$(\mathbb{P}\mathrm{e} - 2)u_{i+1} + 4u_i - (\mathbb{P}\mathrm{e} + 2)u_{i-1} = 0 \qquad i = 1, \ldots, n-1. \tag{2.52}$$

In analogy to the procedure followed for the ODE above, we can find the solution of this finite-difference equation by assuming the solution is a linear combination of the solutions of the type $u_i = \lambda^i$. Substituting we obtain:

$$(\mathbb{P}\mathrm{e} - 2)\lambda^2 + 4\lambda - (\mathbb{P}\mathrm{e} + 2) = 0,$$

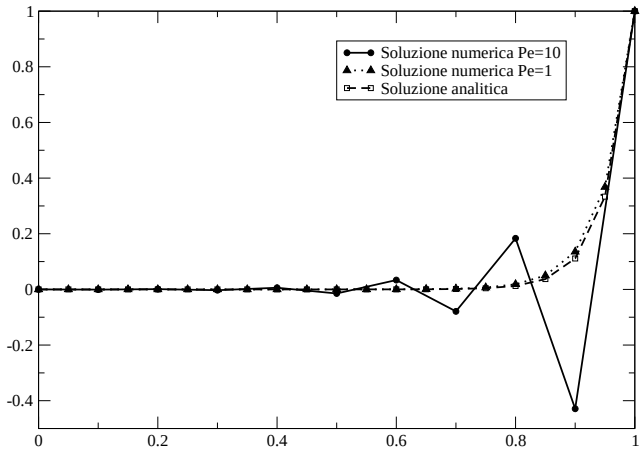


FIGURE 2.12: *Solution of the FEM difference equation for the solution of the one-dimensional convection-diffusion equation compared with the real solution of the differential equation in the case of* $\mathbb{P}e = 0.5$ *e* $\mathbb{P}e = 2$.

from which:

$$\lambda_{1,2} = \frac{-2 \pm \sqrt{4 + (\mathbb{P}\mathrm{e} - 2)(\mathbb{P}\mathrm{e} + 2)}}{\mathbb{P}\mathrm{e} - 2} = \begin{cases} (2 + \mathbb{P}\mathrm{e})/(2 - \mathbb{P}\mathrm{e}), \\ 1. \end{cases}$$

The general solution of (2.52) is then given by:

$$u_i = c_1 \lambda_1^i + c_2 \lambda_2^i$$

Using the boundary conditions we finally obtain:

$$u_i = \frac{1 - \left(\frac{2+\mathbb{P}\mathrm{e}}{2-\mathbb{P}\mathrm{e}}\right)^i}{1 - \left(\frac{2+\mathbb{P}\mathrm{e}}{2-\mathbb{P}\mathrm{e}}\right)^n} \qquad i = 0, 1, \ldots, n,$$

that gives the solution of the FEM (or FDM) problem on each grid node.
It is easy to see now that we have a problem. In fact, in the case in which $\mathbb{P}\mathrm{e} > 2$ the solution of the FEM scheme oscillates from node to node. In fact, the denominator of the term raised to the power i becomes in this case negative, exposing the solution to changes of signs corresponding to odd or even exponents. This behavior is shown in Figure 2.12.
We can try to correct the situation by resorting to different finite difference approximations of the first derivative, using a lower-accuracy non-centered discretization. For example, we could use a forward difference approximation of the convective term yielding:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{h}(u_{i+1} - u_i) = 0,$$

The difference equation is now:

$$(\mathbb{P}\mathrm{e}-1)u_{i+1}-(\mathbb{P}\mathrm{e}-2)u_i-u_{i-1}=0 \qquad i=1,\ldots,n-1.$$

and the roots of the characteristic equation are:

$$\lambda_{1,2}=\frac{\mathbb{P}\mathrm{e}-2\pm\sqrt{(\mathbb{P}\mathrm{e}-2)^2+4(\mathbb{P}\mathrm{e}-1)}}{2(\mathbb{P}\mathrm{e}-1)}=\begin{cases}1/(1-\mathbb{P}\mathrm{e}),\\ 1,\end{cases} \quad .$$

Using the boundary conditions the solution becomes:

$$u_i=\frac{1-\left(\frac{1}{1-\mathbb{P}\mathrm{e}}\right)^i}{1-\left(\frac{1}{1-\mathbb{P}\mathrm{e}}\right)^n} \qquad i=0,1,\ldots,n,$$

showing instabilities for $\mathbb{P}\mathrm{e}<1$, thus worsening the situation. Using instead a backward discretization (upwind) we obtain:

$$\frac{D}{h^2}(-u_{i-1}+2u_i-u_{i+1})+\frac{b}{h}(u_i-u_{i-1})=0,$$

whose solution results stable for any value of $\mathbb{P}\mathrm{e}$. Simple algebraic manipulations show that the previous difference equation can be written as:

$$\frac{D}{h^2}(-u_{i-1}+2u_i-u_{i+1})+\frac{b}{2h}(u_{i+1}-u_{i-1})+\frac{bh}{2}(\frac{-u_{i-1}+2u_i-u_{i+1}}{h^2})=0,$$

which shows that the "upwind" stable formulation is equivalent to (or can be interpreted as) a center formulation with an added diffusion term corresponding to an increased diffusion coefficient equal to $D+bh/2$. The term $bh/2$ is called "numerical diffusion" or "numerical viscosity". The new Péclet number becomes then:

$$\mathbb{P}\mathrm{e}=\frac{bh}{D+bh/2},$$

always less or equal than 2 for every value of D and $b(>0)$, and thus always stable.
This exercise shows that stabilization is obtained by adding numerical diffusion to the scheme. In other words, we are solving a problem that is different from the original problem, hence we may ask the question if this is procedure procedure is legitimate. Actually, the strategy of adding a term to stabilize the numerical scheme is often used in practice and is called a "variational crime". The idea is that the convergence of the scheme should not be hampered by the additional term, or, equivalently, the added term should tend to zero as $h\longrightarrow 0$. In the present case, the additional term is proportional to $bh/2$, and thus we can expect first order convergence ($\mathcal{O}\left(()\,h\right)$) of our numerical scheme, decreased by one with respect to the optimal second order that we found for $b=0$ (see eq. (2.16)).

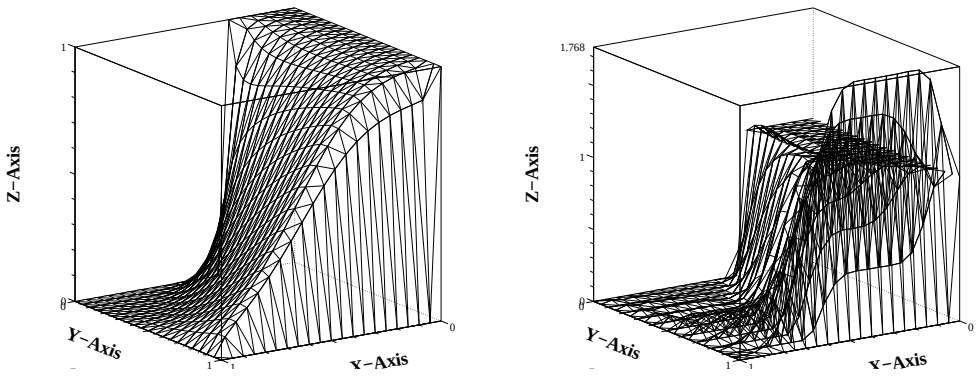


FIGURE 2.13: *Convection-diffusion equation solved with linear (P1) Galerkin in the case of constant coefficients and* $\beta = (1,3)^T$ *and* $D = 0.1$, $\mathbb{P}e_h = 1$ *(left panel) and* $D = 0.01$, $\mathbb{P}e_h = 10$*(right panel). Note the strong oscillations appearing in the case of larger* $\mathbb{P}e_h$*.*

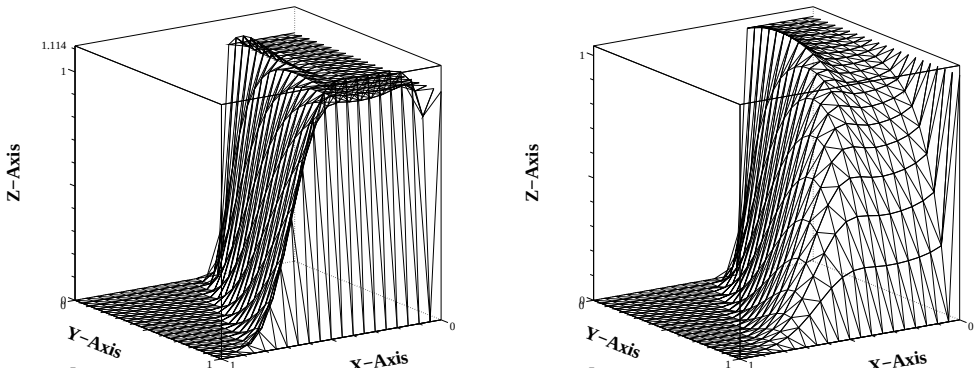


FIGURE 2.14: *Convection-diffusion equation solved with linear (P1) Galerkin with Streamline Diffusion stabilization in the case* $D = 0.01$ *and* $\beta = (1,3)^T$ *(*$\mathbb{P}e_h = 10$*). The left panel shows the case* $\tau = 0.01$ *and the panel on the right shows the corresponding solution for* $\tau = 1$*. Note that the oscillations in the latter case are much dumped, but, correspondingly, the effect of the numerical diffusion is clearly noticeable.*

2.5.2 Multidimensional extension and FEM

A simple and naive approach is to replicate the above sketched procedure to the multidimensional weak form of the convection-diffusion equation. To this aim, we add to our bilinear form a bilinear term of the type:

$$\int_\Omega (\beta \cdot \nabla u)(\beta \cdot \nabla v)\ dx,$$

which in practice correspond to a numerical diffusion somehow proportional to the velocity β. Actually, a more clever idea is to add numerical diffusion only along the streamlines, leading to the definition of the so called Streamline-Diffusion (SD) finite element:

Problem 2.36 (Streamline Diffusion)**.**
Find $u_h \in \mathcal{V}_h$ such that:

$$a_{_h}(u_h, v) = (f, v) \qquad \forall v \in \mathcal{V}_h, \tag{2.53}$$

dove:

$$a_{_h}(u_h, v) \;=\; \int_\Omega \left[D\nabla u_h \cdot \nabla v + \mathrm{div}(\beta u_h)v + \tau \frac{\mathbb{P}\mathrm{e}_h}{|\ \beta\ |^2} (\beta \cdot \nabla u_h)(\beta \cdot \nabla v) \right] dx$$

where $\mathbb{P}\mathrm{e}_h$ is the mesh Péclet defined element by element by:

$$\mathbb{P}\mathrm{e}_h = \frac{|\beta_k| h_k}{D^{(k)}}$$

with $D^{(k)}$ and β_k the diffusion coefficient and the velocity vector considered constant on element T_k but that can vary from element to element. We recognize immediately that the presence of h_k in the definition of the Péclet number force $a_{_h}(\cdot,\cdot)$ to converge to $a(\cdot,\cdot)$ when $h \longrightarrow 0$. Hence we are adding a term that resembles a diffusion but that is projected along velocity vector (from this the name Streamline Diffusion). The coefficient τ is is an empirical parameter introduced to tune the amount of numerical diffusion introduced to control oscillations on a case-by-case scenario. In fact, there is no multidimensional theory that determines the exact value of $\mathbb{P}\mathrm{e}_h$ that guarantees convergence. Figures 2.13 and 2.14 show some exemplifying example of solutions obtained in the stable and unstable regime.

Remark 2.37. *We would like to stress that much improved methods exist for the solution of the convection diffusion equations with respect to the SD approach. All these methods are based essentially to the introduction of minimal numerical diffusion only when necessary. They rely thus on algorithms that intercept potential oscillations and as such are much more complicated. These topics will be dealt with, although not exhaustively, in the context of finite volume methods that will be discussed in subsequent chapters.*

Remark 2.38. *The practice of introducing an additional term into the variational formulation maintaining consistency of the overall scheme is sometimes referred to a "variational crime". It is typical when addressing solutions to ill-posed or degenerate systems. The numerical symptoms in these cases is the appearance of oscillations depending on the data of the problem. This was the case in the convection-diffusion equation for large Péclet numbers. The crucial idea in these cases is to maintain consistency of the overall scheme and to introduce the smalles amount of the extra stabilizing terms. As seen above, this is generally obtained by multiplication by an appropriate power of the mesh parameter h. Moreover, wheneveer possible, the added terms are what are called "residuals", i.e., terms that are zero when the real solution of the original problem is substituted in place of the approximate (numerical) solution. These stratagems ensure the consistency (weak or strong) of the ensuing scheme and thus its convergence.*

2.6 Mathematical theory of Galerkin Finite Elements

2.6.1 Preliminaries

A measurable space (Ω, Σ, μ), with nonnegative measure μ, will be denoted simply with Ω. In general, Ω is an open, bounded, and connected subset of $\mathbb{R}^d$ with $d = 1, 2$ or 3. Its closure is assumed to be sufficiently smooth (e.g., Lipschitz) and will be denoted with $\Gamma = \partial\Omega$. A measurable space gives us the ability of measuring the global size of a function and thus perform comparisons between different functions. Thus we need the notion of functional norm of a function that can be thought of roughly as an appropriate infinite dimensional extension of the norm of a vector space. Thus given two functions u and v in $\mathcal{V}(\Omega)$, $u, v : \Omega \longrightarrow \mathbb{R}$, we can define the scalar product of these to functions in $\mathcal{V}$ as follows:

Definition 2.4 (scalar product (in the real field))**.** A scalar product between two functions u and v defined in a space $\mathcal{V}(\Omega)$ is a bilinear form $(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$ that satisfies the following properties:

1. symmetry: $(u, v) = (v, u)$;
2. linearity (in the first argument): $(\alpha u, v) = \alpha\,(u, v)$, $(u + v, w) = (u, w) + (v, w)$;
3. positiveness: $(u, u) =\geq 0$, $(u, u) = 0 \Leftrightarrow u = 0$.

Definition 2.5 (Norm of a function)**.** Given a function u defined in a domain $\Omega \subset \mathbb{R}$ the norm of u is a linear form (functional) $\|\cdot\| : \mathcal{V} \longrightarrow \mathbb{R}$ that satisfies the following properties:

1. $\|u\| > 0$, $\|u\| = 0$ if and only if $u = 0$;
2. given $\alpha \in \mathbb{R}$, $\|\alpha u\| = \mid \alpha \mid \|u\|$;

3. triangular inequality: $\|u+v\| \leq \|u\| + \|v\|$;

We will be talking of "semi-norm", denoted by $|\cdot|$, when the first property is substituted with the requirement that $\|u\| \geq 0$.
We will be using often the Cauchy-Schwartz inequality:

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

The proof of this property can be obtained as follows.

Proof. If $v = 0$ the result is obvious. Assume then $v \neq 0$. Let $\lambda \in \mathbb{R}$, with $\lambda = (v,u) / (v,v)$. The function $z = u - \lambda v$ is the orthogonal projection of u along v, so that $(v, u - \lambda v) = (u,v) - \lambda (u,u) = 0$. The symmetry and positive definiteness of the scalar product imply:

$$0 \leq (u - \lambda v, u - \lambda v) = (u, u - \lambda v) - \lambda (v, u - \lambda v) = (u, u - \lambda v) = (u,u) - \lambda (u,v),$$

summing $\lambda (u,v)$ on both sides we have:

$$\lambda (u,v) \leq (u,u)$$

and multiplying by (u,u) we obtain:

$$(u,v)^2 \leq (u,u) \langle v,v \rangle .$$

□

We will work with:

Continuous functions; The space of continuous functions $\mathcal{C}^0(\Omega)$ is given by:

$$\mathcal{C}^0(\Omega) = \{ u : \Omega \longrightarrow \mathbb{R} : u \text{ is continuous and bounded } \},$$

with norm expressed by:

$$\|u\|_\infty = \sup_{x \in \Omega} | \, u(x) \, |; \tag{2.54}$$

bounded function; the space of bounded functions is characterized by:

$$\mathcal{L}^\infty(\Omega) = \{ u : \Omega \longrightarrow \mathbb{R} : u \text{ is measurable and } \mu\text{-a.e. bounded} \},$$

and the same norm as in (2.54) can be defined;

integrable functions; given $0 < p < \infty$, the space of integrable (measurable) functions is given by:

$$\mathcal{L}^p(\Omega) = \left\{ u : \Omega \longrightarrow \mathbb{R} : u \text{ is measurable } \int_\Omega \mid u \mid^p d\mu < +\infty \right\};$$

given $1 \leq p < \infty$, the norm can be written as:

$$\|u\|_{\mathcal{L}^p(\Omega)} = \|u\|^p = \left(\int_\Omega \mid u(x) \mid^p d\mu \right)^{\frac{1}{p}}; \tag{2.55}$$

in the case $0 < p < 1$, we cannot define the norm as above, and we use the distance metric:

$$d_p(u, v) = \int_\Omega \mid u(x) - v(x) \mid^p d\mu;$$

differentiable functions; $\mathcal{C}^k(\Omega)$ is the space of functions that are k times continuously differentiable:

$$\mathcal{C}^k(\Omega) = \left\{ u : \Omega \longrightarrow \mathbb{R} : \forall \alpha, \mid \alpha \mid \leq k : \partial^\alpha u \text{ is continuous in } \overline{\Omega} \right\};$$

and admissible norm is given by:

$$\|u\| = \sum_{0 \leq |\alpha| \leq k} \|\partial^\alpha u\|_\infty$$

Sobolev Spaces. We denote by $W^{k,p}(\Omega)$ the Sobolev space of functions whose derivative up to order k belong to $\mathcal{L}^p(\Omega)$:

$$W^{k,p}(\Omega) = \{ u : \Omega \longrightarrow \mathbb{R} : u \in \mathcal{L}^p(\Omega); \partial^\alpha u \in \mathcal{L}^p(\Omega) \forall \alpha : \mid \alpha \mid \leq k \}$$

with norm defined by:

$$\|u\|_{k,p} = \left(\sum_{0 \leq |\alpha| \leq k} \int_\Omega \mid \partial^\alpha u \mid^p \right)^{\frac{1}{p}},$$

if $p < \infty$, and for $p = \infty$:

$$\|u\|_{k,\infty} = \max_{0 \leq |\alpha| \leq k} \|\partial^\alpha u\|_\infty .$$

The $\mathcal{L}^2(\Omega)$ and $\mathcal{H}^1(\Omega)$ spaces. In this section we recall the results about Hilbert spaces that are used in the FEM method. The space $\mathcal{L}^2(\Omega)$ is the space of square integrable functions with respect to the Lebesgue measure, i.e., given $f : \Omega \longrightarrow \mathbb{R}$, we have:

$$\int_\Omega \mid f(x) \mid^2 dx < \infty.$$

If we associate the following scalar product:

$$(u, v)_{\mathcal{L}^2(\Omega)} = \int_\Omega u(x)v(x)\, dx,$$

and the induced norm:

$$\|u\|_{\mathcal{L}^2(\Omega)} = \left(\int_\Omega u(x)v(x)\, dx\right)^{1/2},$$

then $\mathcal{L}^2(\Omega)$ is a Hilbert space.
Given a subspace $\mathcal{V}$ (e.g. of $\mathcal{L}^2(\Omega)$), we will work with linear functionals and forms (roughly functions of functions) $F : \mathcal{V} \longrightarrow \mathbb{R}$, and $b : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$. Given a linear subspace $\mathcal{V}(\Omega) \subset \mathcal{L}^2(\Omega)$, an operator $a : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ defines a bilinear form in $\mathcal{V} \times \mathcal{V}$ if:

$$\begin{aligned} a : \mathcal{V} \times \mathcal{V} &\mapsto \mathbb{R}, \\ a(u, v) &= a(v, u), \\ a(\alpha u + \beta v, w) &= \alpha a(u, w) + \beta a(v, w), \\ a(u, \alpha v + \beta w) &= \alpha a(u, v) + \beta a(u, w). \end{aligned}$$

The bilinear form $a(\cdot, \cdot)$ defines a scalar product in $\mathcal{V}(\Omega)$ if it is symmetric and:

$$a(v, v) > 0 \qquad \forall v \in \mathcal{V},\ \ v \neq 0.$$

The induced norm is given by:

$$\|v\|_{\mathcal{V}} = (a(v, v))^{\frac{1}{2}}.$$

The scalar product satisfies the Cauchy-Schwartz inequality:

$$|a(u, v)| \leq \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}.$$

A linear subspace $\mathcal{V}$ endowed with a scalar product and the corresponding induced norm is a Hilbert space if it is complete, i.e., every Cauchy sequence[4] converges with respect to the norm $\|\cdot\|_{\mathcal{V}}$.

[4] A sequence of functions $v_i \in \mathcal{V}$ is a Cauchy sequence if there exist $\epsilon > 0$ such that $\|v_i = v_j\|_{\mathcal{V}} < \epsilon$ for sufficiently large i and j. We say that v_i converges to v if $\|v - v_i\|_{\mathcal{V}} \longrightarrow 0$ $i \longrightarrow \infty$.

For example, the space of square integrable functions in the interval $\Omega = [a, b]$:

$$\mathcal{L}^2(I) = \left\{ v(x) : I \longrightarrow \mathbb{R} \text{ such that } \int_a^b v^2 \, dx < \infty \right\}$$

is a Hilbert space with scalar product:

$$(u, v) = \int_a^b u(x)v(x) \, dx$$

and norm:

$$\|u\|_{\mathcal{L}^2(I)} = \|u\|_2 = \left(\int_a^b [u(x)]^2 \, dx \right)^{\frac{1}{2}}.$$

Remark 2.39. *Note that the space $\mathcal{L}^2$ is exactly the space of functions that were mentioned in Remark 2.14.*

Example 2.40. The function $v(x) = x^{-\alpha}$, $x \in I = [0, 1]$, belongs to $\mathcal{L}^2(I)$ only for $\alpha < 1/2$.

The natural space of admissible functions that are candidate solutions of our elliptic equations is the Hilbert space $\mathcal{H}^1(I) = \{v : v \text{ and } v' \text{ belong to } \mathcal{L}^2(I)\}$ equipped with the scalar product:

$$(u, v)_{\mathcal{H}^1(I)} = \int_a^b \left[u(x)v(x) + u'(x)v'(x) \right] \, dx$$

and the norm:

$$\|u\|_{\mathcal{H}^1(I)} = \left(\int_a^b \left[u(x)^2 + u'(x)^2 \right] \, dx \right)^{\frac{1}{2}}$$

Remark 2.41. *We note here that the derivatives above are always to be intended in the sense of distributions (weak derivatives). Thus we may write:*

$$\mathcal{H}^1(I) = \{ v \in \mathcal{L}^2(I) : \text{ there exists } g \in \mathcal{L}^2(I) \text{ such that } \int_I v\phi' = -\int_I g\phi \qquad \forall \phi \in \mathcal{C}_c^1(I) \}$$

and we will always denote $v' = g$. The function ϕ is called the test function and can be chosen to belong to $\mathcal{C}_c^\infty(I)$ as well (see [3])

Remark 2.42. *Note that the space $\mathcal{V}$ defined in paragraph 2.1.1 is also a Hilbert space and is denoted by:*

$$\mathcal{V}(I) = \mathcal{H}_0^1(I) = \left\{ v(x) : \mathbb{R} \longrightarrow \mathbb{R} \text{ such that } v(x) \in \mathcal{L}^2(I), v'(x) \in \mathcal{L}^2(I) \text{ and } v(0) = v(1) = 0 \right\}$$

The subscript 0 denotes the fact that the functions are zero on the boundary of Ω.

All these notions can be easily extended to the multidimensional case. Given an open and bounded domain $\Omega \in \mathbb{R}^d$ with a smooth boundary $\Gamma = \partial\Omega$:

$$\mathcal{L}^2(\Omega) = \left\{ v(x) : \Omega \longrightarrow \mathbb{R} \text{ such that } \int_\Omega v(x)^2 < \infty \right\}$$
$$\mathcal{H}^1(\Omega) = \left\{ v(x) : \Omega \longrightarrow \mathbb{R} \text{ such that } v(x) \in \mathcal{L}^2(\Omega) \text{ and } \frac{\partial v(x)}{\partial x_i} \in \mathcal{L}^2(\Omega) \text{ for } i = 1, \ldots, d \right\}$$
$$\mathcal{H}^k(\Omega) = \left\{ v(x) : \Omega \longrightarrow \mathbb{R} \text{ such that } v(x) \in \mathcal{L}^2(\Omega) \text{ and } \partial^\alpha v \in \mathcal{L}^2(\Omega) \text{ for all } \mid \alpha \mid \leq k \right\}$$

with the following scalar products:

$$(u, v)_{\mathcal{L}^2(\Omega)} = \int_\Omega uv \, dx \qquad \|u\|_{\mathcal{L}^2(\Omega)} = \left(\int_\Omega u^2 \, dx \right)^{\frac{1}{2}}$$
$$(u, v)_{\mathcal{H}^1(\Omega)} = \int_\Omega \left[uv + \nabla u \cdot \nabla v \right] \, dx \qquad \|u\|_{\mathcal{H}^1(\Omega)} = \left(\int_\Omega \left[u^2 + |\nabla u|^2 \right] \, dx \right)^{\frac{1}{2}}$$

The seminorm is given by:

$$\mid v \mid_{H^k(\Omega)} = \left(\int_\Omega \mid \nabla u \mid^2 dx \right)^{\frac{1}{2}} = \left(\int_\Omega \mid \partial u \mid^2 dx \right)^{\frac{1}{2}} = \|\partial v\|_{\mathcal{L}^2(\Omega)} .$$

Note that the above is not a norm (thus the name "seminorm") because it vanishes for all nonzero constant functions.

More on the space $\mathcal{H}_0^1(\Omega)$ and Poincaré inequality. The subscript "0" is used to denote the space of functions that vanish on the boundary of Ω, i.e., that satisfy homogeneous Dirichlet conditions on the boundary:

$$\mathcal{H}_0^1(\Omega) = \{ v(x) \in \mathcal{H}^1(\Omega) \text{ such that } v(x) = 0 \text{ for } x \in \Gamma \}.$$

and is equipped with the same scalar product of $\mathcal{H}^1(\Omega)$. In the case $u \in \mathcal{H}_0^1(\Omega)$, the homogeneous Dirichlet boundary conditions transform the seminorm into an equivalent norm:

Lemma 2.6 (Poincarè)**.** *Let $\Omega \subset \mathbb{R}^d$ a bounded open subset. Then there exist a constant C depending only on Ω such that*

$$\|u\|_{\mathcal{L}^2(\Omega)} \leq C \, \|\nabla u\|_{\mathcal{L}^2(\Omega)}$$

for all $u \in \mathcal{H}_0^1(\Omega)$.

Proof. We report the proof for $\Omega = I \subset \mathbb{R}$, pointing to more specialized functional analysis books for a more general proof.
The hypothesis is then to have a bounded open interval $I = (a, b)$ and functions u that vanish in a and b ($u \in \mathcal{H}_0^1(I)$). Then we have:

$$| u(x) | = | u(x) - u(a) | = | \int_a^x u'(\tau)\, d\tau | \leq \|u'\|_{\mathcal{L}^1} .$$

Hence $\|u\|_{\mathcal{L}^\infty} \leq \|u'\|_{\mathcal{L}^1}$ from which the result follows by Hölder inequality (see also the discussion on page 29). □

Corollary 2.7. *The gradient norm:*

$$\|\nabla u\| = \left(\int_\Omega \nabla u \cdot \nabla u \, dx \right)^{\frac{1}{2}}$$

is equivalent (with regards to the induced topology and thus for the notion of convergence) to the usual norm $\|u\|_{\mathcal{H}^1(\Omega)}$.

In fact:

$$\|\nabla u\|_2^2 \leq \|u\|_{H^1}^2 = \|u\|_2^2 + \|\nabla u\|_2^2 \leq (1 + C^2)\|\nabla u\|_2^2$$

and hence, given u e v in $\mathcal{H}_0^1(\Omega)$, we can define a scalar product:

$$(u, v)_{\mathcal{H}_0^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, dx.$$

In general the trace of a function g in $\partial\Omega$ (essentially the value that g takes on the boundary) is not always well defined, a simple example being $\sin(1/x)$ for $x = 0$. But we can define a trace operator γ so that γg is appropriately extended on $\partial\Omega$ and we can properly use the trace of any function of $\mathcal{H}^1(\Omega)$ to define $\mathcal{H}_0^1(\Omega)$.

2.6.2 Lax-Milgram Theorem

Definition 2.8. Let $\mathcal{V}$ a Hilbert space and $a(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$ a bilinear form. We say that the bilinear form is:

- *continuous* if there is a constant $\gamma > 0$ such that:

$$| a(u, v) | \leq \gamma \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \qquad \forall u, v \in \mathcal{V}; \tag{2.56}$$

- *V-elliptic or coercive* if there exists a constant $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|_{\mathcal{V}}^2 \qquad \forall v \in \mathcal{V}. \tag{2.57}$$

Analogously, a linear form $F(\cdot) : \mathcal{V} \longrightarrow \mathbb{R}$ is said to be *continuous* if there exists a constant $\Lambda > 0$ such that

$$| F(v) | \leq \Lambda \|v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}.$$

Remark 2.43. *The coercivity of the continuous operator is generally inherited by the discrete FEM operator. This property is fundamental to obtain convergence estimate. But it is not only a theoretical property. In fact it is the property that guarantees that the system FEM matrix can be inverted (is nonsingular). In other words, in the discrete case $\mathcal{V} \equiv \mathbb{R}^n$, the coercivity of $a(\cdot,\cdot)$ implies that the stiffness matrix (i.e., the linear operator associated to $a(\cdot,\cdot)$) $A = \{a_{ij}\}$, $a_{ij} = a(\phi_i, \phi_j)$, is positive definite. However, sometimes it is too restrictive, and the coercivity of the differential operator is not inherited by the discrete operator. In these cases experimental convergence is observed but theoretical convergence cannot be proved. This is the case some finite volume schemes that we will see in future chapters.*

For linear forms in Hilbert spaces there is the following fundamental theorem:

Theorem 2.9 (Riesz representation)**.** *For all continuous linear forms $\phi_u(\cdot)$ in a Hilbert space $\mathcal{V}$ there exists a unique $u \in \mathcal{V}$ such that $\phi_u(v) = a(u, v)$ for all $v \in \mathcal{V}$. Moreover, $\|u\| = \|\phi_u\|$.*

We denote with $\mathcal{V}^*$ (or $\mathcal{V}'$) the space (dual of $\mathcal{V}$) formed by all linear forms from $\mathcal{V}$ to $\mathbb{R}$. The Riesz theorem ensures that every element of $\mathcal{V}^*$ can be uniquely written as $\phi_u(v) = a(u, v)$. In other words, the map $\Phi : \mathcal{V} \longrightarrow \mathcal{V}^*$ defined by $\Phi(u) = \phi_u(v)$ is an isomorphism. A consequence of the Riesz representation theorem is the Lax-Milgram theorem for continuous and coercive bilinear forms:

Theorem 2.10 (Lax-Milgram)**.** *Let $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$ be a continuous and coercive bilinear form. For all linear forms $F(v) : \mathcal{V} \longrightarrow \mathbb{R}$, there exist a unique function $u \in \mathcal{V}$ such that:*

$$a(u, v) = F(v) \qquad \forall v \in \mathcal{V}.$$

Proof. (Sketch) From the Riesz representation theorem we can define a linear and continuous map $A : \mathcal{V} \mapsto E$ as:

$$a(u, v) = (A(u), v) \qquad \text{con} \qquad \|A(u)\| \leq C \|u\|_{\mathcal{V}}.$$

We can associate to a linear form $F(\cdot) \in \mathcal{V}^*$ a function of $\mathcal{V}$ such that $F(v) = (f, v)$. Thus we need to show that the solution of the problem $A(u) = f$ in $\mathcal{V}$ is unique. For this, we can use the Banach-Cacciopoli fixed point theorem to show that the map $T : E \mapsto E$:

$$T_\epsilon(u) = u - \epsilon A(u) + \epsilon f$$

is a contraction for sufficiently small ϵ. □

2.7 Abstract formulation of the FEM method for elliptic equations

2.7.1 Weak formulation

Let $\mathcal{V}$ be a Hilbert space with scalar product $(\cdot,\cdot)_{\mathcal{V}}$ and norm $\|\cdot\|_{\mathcal{V}}$. Let $a(\cdot,\cdot) : \mathcal{V}\times\mathcal{V} \longrightarrow \mathbb{R}$ a continuous and coercive bilinear form and $F(\cdot) : \mathcal{V} \longrightarrow \mathbb{R}$ a continuous linear form. We have the following:

Problem 2.44 (M)**.** Find $u \in \mathcal{V}$ such that

$$F(u) = \min_{v\in\mathcal{V}} F(v),$$

where

$$F(v) = \frac{1}{2}a(v,v) - F(v). \tag{2.58}$$

Problem 2.45 (V)**.** Find $u \in \mathcal{V}$ such that

$$a(u,v) = F(v) \qquad \forall v \in \mathcal{V}. \tag{2.59}$$

We have the following

Theorem 2.11. *If the bilinear form is symmetric, i.e., $a(u,v) = a(v,u)$, the two Problems 2.44 and 2.45 are equivalent in the sense that $u \in \mathcal{V}$ is solution of Problem 2.44 if and only if it is solution of Problem 2.45. Moreover, there exist a unique solution that satisfies the stability estimate:*

$$\|u\|_{\mathcal{V}} \leq \frac{\Lambda}{\alpha} \tag{2.60}$$

Proof. Existence comes from Lax-Milgram theorem.
The equivalence is a trivial extension of what done in $\mathbb{R}^1$. Thus, we first show that if $u \in \mathcal{V}$ is solution of 2.44 then it is solution of 2.45. Let then $v \in \mathcal{V}$ and $\epsilon \in \mathbb{R}$ be arbitrary. The condition that $u \in \mathcal{V}$ is a point of minimum for F can be written as:

$$F(u) \leq F(u+\epsilon v) \qquad \forall \epsilon \in \mathbb{R}.$$

Let $g(\epsilon) = F(u+\epsilon v)$. The function g has a minimum for $\epsilon = 0$ which is characterized by $g'(0) = 0$. Then, using the symmetry of $a(\cdot,\cdot)$ we have:

$$\begin{aligned} g(\epsilon) &= \frac{1}{2}a(u+\epsilon v, u+\epsilon v) - F(u+\epsilon v) \\ &= \frac{1}{2}a(u,u) - F(u) + \epsilon a(u,v) - \epsilon F(v) + \frac{\epsilon^2}{2}a(v,v); \end{aligned}$$

from which the results follows by noting that:

$$g'(0) = a(u, v) - F(v).$$

Let now $u \in \mathcal{V}$ be solution of 2.45. We need to show that for such u we have $F(u) \leq F(u+v)$ for all $v \in \mathcal{V}$. Observe that:

$$F(u+v) = \frac{1}{2}a(u, u) - F(u) + a(u, v) - F(v) + \frac{1}{2}a(v, v),$$

from which the results follows because of the coercivity of $a(\cdot, \cdot)$.
The stability estimate can be deduced by setting $v = u$ in (2.59) and invoking the coercivity of $a(\cdot, \cdot)$ and the continuity of $F(\cdot)$. Then we have:

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

Also uniqueness follows from this last inequality, since, if u_1 and u_2 are two functions satisfying Problem 2.45, then:

$$a(u_1 - u_2, v) = 0 \qquad \forall v \in \mathcal{V}.$$

The stability estimate with $F(\cdot) = 0$ and $\Lambda = 0$ implies that $\|u_1 - u_2\| = 0$, from which $u_1 = u_2$. □

We observe that the most important result coming out of the Lax-Milgram theorem, for what we are concerned, is that the continuity and coercivity of the bilinear form, together with continuity of the linear form stemming from the source term are the key ingredient to guarantee existence and uniqueness, and thus we will see in the discrete setting also convergence of the FEM scheme if the FEM spaces are chosen appropriately.
However, there are problems where the bilinear form is not coercive, and we need some weaker statements. This condition, which we will specialize for systems of equation where it finds its typical application, is called the "inf-sup" or "LBB" (Ladyzhenskaya-Babuska-Brezzi) condition [4], and again guarantees well-posedness. We have the following:

Definition 2.12. A bilinear form $a(\cdot, \cdot)$ satisfies the *inf-sup* condition in $\mathcal{V}$ if there is $\alpha > 0$ such that

$$\sup_{v \in \mathcal{V}} \frac{a(u, v)}{\|v\|_{\mathcal{V}}} \geq \alpha \|u\|_{\mathcal{V}} \qquad \forall u \in \mathcal{V}; \tag{2.61}$$

and at the same time:

$$\sup_{u \in \mathcal{V}} \frac{a(u, v)}{\|u\|_{\mathcal{V}}} \geq \alpha \|v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}; \tag{2.62}$$

It is obvious that if $a(\cdot,\cdot)$ is symmetric then the two conditions above are equivalent. Moreover, eq. (2.61) (and at the same time eq. (2.62)) can be re-written as:

$$\inf_{u\in\mathcal{V}} \sup_{v\in\mathcal{V}} \frac{a(u,v)}{\|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \geq \alpha > 0 \tag{2.63}$$

from which the name "inf-sup". We have the following:

Lemma 2.13. *If $a(\cdot,\cdot)$ is coercive then it satisfies the "inf-sup" condition 2.63.*

Proof. Coercivity of $a(\cdot,\cdot)$ implies

$$a(u,u) \geq \alpha \|u\|_{\mathcal{V}}^2, \qquad \forall u \in \mathcal{V}.$$

Then we can write:

$$\sup_{v\in\mathcal{V}} \frac{a(u,v)}{\|v\|_{\mathcal{V}}} \geq \frac{a(u,u)}{\|u\|_{\mathcal{V}}} \geq \alpha \|u\|_{\mathcal{V}}.$$

□

Remark 2.46. *In the discrete setting (FEM) we can write the "inf-sup" condition for the linear operator A associated to $a(\cdot,\cdot)$ and its adjunct A^*. In fact we have:*

$$A : \mathcal{V}' \longrightarrow \mathcal{V} \qquad\qquad A^* : \mathcal{V} \longrightarrow \mathcal{V}',$$

with $\mathcal{V}'$ the dual space of $\mathcal{V}$ (with respect to the linear form F) and where A and A^* are defined by:

$$(Au,v)_{\mathcal{V}'\times\mathcal{V}} = a(u,v) \qquad (u,Av)_{\mathcal{V}\times\mathcal{V}'} = a(u,v),$$

then the condition (2.63) is equivalent to:

$$\|Au\|_{\mathcal{V}'} \geq \alpha \|u\|_{\mathcal{V}} \qquad \forall u \in \mathcal{V}; \tag{2.64}$$

$$\|A^*v\|_{\mathcal{V}'} \geq \alpha \|v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}. \tag{2.65}$$

In the case $\mathcal{V} = \mathbb{R}^n$, coercivity of the bilinear form, as seen before, implies that matrix A is positive definite, while the "inf-sup" condition implies only that A is invertible.

We have the following:

Theorem 2.14. *The continuous bilinear form $a(\cdot,\cdot)$ satisfies the "inf-sup" condition if and only if the operator A is a bijection.*

Note that if A is a bijection then problem (2.45) has a unique solution for each $F(\cdot)$, i.e., A has continuous inverse and $\|u\|_{\mathcal{V}} \leq C \|F\|_{\mathcal{V}'}$.

Proof. If $a(\cdot,\cdot)$ satisfies the "inf-sup" condition, then eq. (2.64) and (2.65) show that the operators A e A^* are injections. Thus it is enough to show that the image $R(A)$ of A is closed. Let $Au_n \longrightarrow w$, then

$$\|A(u_n - u_m)\|_{\mathcal{V}'} \geq \alpha \|u_n - u_m\|_{\mathcal{V}},$$

that shows that $\{u_n\}$ is a Cauchy sequence and thus it converges to $u \in \mathcal{V}$. From the continuity A we have $w = Au \in R(A)$.
On the other hand, if A is a bijection, then also A^* is and thus the operator has continuous inverse. □

2.7.2 FEM formulation

The FEM Formulation is obtained directly by approximating the appropriate functional space with a finite-dimensional subset $\mathcal{V}_h \subset \mathcal{V}$ generated by a finite number of basis functions $\{\phi_1, \ldots, \phi_n\}$. Then, every function $v \in \mathcal{V}_h$ can be written as:

$$v = \sum_{j=1}^{n} \xi_i \phi_i(x). \tag{2.66}$$

We have then:

Problem 2.47 (FEM, Ritz method)**.** Find $u_h \in \mathcal{V}_h$ such that

$$F(u_h) \leq F(v) \qquad \forall v \in \mathcal{V}_h. \tag{2.67}$$

or equivalently:

Problem 2.48 (FEM, Galerkin method)**.** find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = (f, v) \qquad \forall v \in \mathcal{V}_h. \tag{2.68}$$

Using the representation of u_h as linear combination of ϕ_i:

$$u_h = \sum_{j=1}^{n} u_j \phi_j, \qquad u_j \in \mathbb{R},$$

we obtain:

$$\sum_{j=1}^{n} u_j a(\phi_j, \phi_i) = F(\phi_i), \qquad i = 1, \ldots, n,$$

or in matrix form:

$$Au = b,$$

where:

$$u = \{u_i\}, \qquad A = \{a_{ij}\},\ a_{ij} = a\left(\phi_j, \phi_i\right), \qquad b = b_i,\ b_i = F\left(\phi_i\right). \tag{2.69}$$

Matrix A is called the *stiffness* matrix. We have the following:

Theorem 2.15. *The stiffness matrix A is symmetric and positive definite.*

Proof. The symmetry of A is inherited by the symmetry of the bilinear form.
Using (2.66) we have:

$$a\left(v, v\right) = a\left(\sum_{i=1}^{n} \xi_i \phi_i, \sum_{i=1}^{n} \xi_i \phi_i\right) = \sum_{i,j=1}^{n} \xi_i a\left(\phi_i, \phi_j\right) \xi_j = \xi \cdot A\xi,$$

where $\xi = \{\xi_i\}$ is a vector of $\mathbb{R}^n$ and the $\cdot$ denotes the vector scalar product. From the coercivity (eq. (2.57)) of the bilinear form we obtain immediately:

$$\xi \cdot A\xi = a\left(v, v\right) \geq \alpha \left\|v\right\|_{\mathcal{V}}^2 > 0$$

if $v \neq 0$, i.e., if $\xi \neq 0$. □

Theorem 2.16. *The two Problems* (2.47) *and* (2.48) *are equivalent and have a unique solution $u_h \in \mathcal{V}_h$. Moreover, we have the following stability estimate:*

$$\left\|u_h\right\|_{\mathcal{V}} \leq \frac{\Lambda}{\alpha}.$$

Proof. Existence and uniqueness comes from Theorem 2.15. Now, choose $v = u_h$ in (2.68) and using the properties of $a\left(\cdot, \cdot\right)$ and $F\left(\cdot\right)$ we have:

$$\alpha \left\|u_h\right\|_{\mathcal{V}}^2 \leq a\left(u_h, u_h\right) \leq \Lambda \left\|u_h\right\|_{\mathcal{V}}.$$

Dividing by $\left\|u_h\right\|_{\mathcal{V}} \neq 0$ we obtain the desired result. □

Next we show Céa Lemma, that states that the error for u_h is optimal in $\mathcal{V}_h$.

Theorem 2.17 (Céa)**.** *Let $u \in \mathcal{V}$ be a solution of Problem 2.45 and let $u_h \in \mathcal{V}_h \subset \mathcal{V}$ a solution of Problem 2.48. Then:*

$$\left\|u - u_h\right\|_v \leq \frac{\gamma}{\alpha} \left\|u - v\right\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}_h.$$

Proof. Since $\mathcal{V}_h \subset \mathcal{V}$, subtracting eqs. (2.59) and (2.68), we have that the Galerkin FEM scheme, and as a consequence also the Ritz scheme, are strongly consistent:

$$a\left(u - u_h, v\right) = 0 \qquad \forall v \in \mathcal{V}_h. \tag{2.70}$$

Take $w = u_h - v$ ($w \in \mathcal{V}_h$) so that $v = u_h - w$ and invoke consistency and coerciveness to obtain:

$$\begin{aligned}\alpha \left\|u - u_h\right\|_{\mathcal{V}}^2 \leq a\left(u - u_h, u - u_h\right) &= a\left(u - u_h, u - u_h\right) + a\left(u - u_h, v\right) \\ &= a\left(u - u_h, u - v\right) \leq \gamma \left\|u - u_h\right\|_{\mathcal{V}} \left\|u - v\right\|_{\mathcal{V}}.\end{aligned}$$

The result follows after division by $\left\|u - u_h\right\|_{\mathcal{V}} > 0$. □

Remark 2.49. *Céa Lemma suggests that we can obtain a quantitative estimate of the error by choosing any $v \in \mathcal{V}_h$ for which we can derive a quantitative estimate for $\|u - v\|$. A typical way of doing this for FEM is to use a Lagrangian interpolation function of degree r, for example for $r = 1$, we can use a piecewise linear interpolator $v = \pi_{h,1} u$.*

The energy norm. If the bilinear form $a\left(\cdot, \cdot\right)$ is symmetric, we can define a new norm, called the energy norm, in $\mathcal{V}$:

$$\left\|v\right\|_a^2 = a\left(v, v\right), \qquad v \in \mathcal{V}.$$

This norm is equivalent to the classical norm in $\mathcal{V}$:

$$\sqrt{\alpha} \left\|v\right\|_{\mathcal{V}} \leq \left\|v\right\|_a \leq \sqrt{\gamma} \left\|v\right\|_{\mathcal{V}},$$

with scalar product given by:

$$\left(u, v\right)_a = a\left(u, v\right).$$

Using the energy norm, we can say that u_h is the orthogonal projection of u on $\mathcal{V}_h$ with respect to the scalar product $\left(\cdot, \cdot\right)_a$, and u_h is the best approximation of u in the energy norm.

Non coercive operators. Not that in the proof of Céa's lemma 2.17 we did not use the symmetry of $a\left(\cdot, \cdot\right)$. In fact, this lemma can be extended to non symmetric and non coercive operators. For the latter, we need to use the "inf-sup" condition, so that we need to have a $\beta > 0$ such that:

$$\sup_{v \in \mathcal{V}_h} \frac{a\left(u, v\right)}{\left\|v\right\|_{\mathcal{V}}} \geq \alpha \left\|u\right\|_{\mathcal{V}} \qquad \forall u \in \mathcal{V}_h.$$

The second "inf-sup"condition comes from the previous one since $\mathcal{V}_h$ has finite dimension. If the constant α is independent on h, then we have:

Theorem 2.18 (Céa Lemma)**.** *Let $u \in \mathcal{V}$ be solution of 2.45 and $u_h \in \mathcal{V}_h \subset \mathcal{V}$ be solution of 2.48. Then:*

1. *if $a(\cdot,\cdot)$ is not coercive:*

$$\|u - u_h\|_v \leq \left(1 + \frac{\|a(\cdot,\cdot)\|}{\alpha}\right) \inf_{v \in \mathcal{V}_h} \|u - v\|_{\mathcal{V}},$$

where

$$\|a(\cdot,\cdot)\| = \sup_{v \in \mathcal{V}_h, v \neq 0} \frac{a(v,v)}{\|v\|_{\mathcal{V}}^2};$$

2. *if $a(\cdot,\cdot)$ is continuous and coercive:*

$$\|u - u_h\|_v \leq \frac{\gamma}{\alpha} \inf_{v \in \mathcal{V}_h} \|u - v\|_{\mathcal{V}};$$

3. *if $a(\cdot,\cdot)$ is also symmetric:*

$$\|u - u_h\|_v \leq \sqrt{\frac{\gamma}{\alpha}} \inf_{v \in \mathcal{V}_h} \|u - v\|_{\mathcal{V}}.$$

Proof. We show only the first point, since the other two are immediate. Let $v \in \mathcal{V}_h$. using the "inf-sup" condition and the consistency of the scheme, we have, analogously to the coercive case:

$$\alpha \|v - u_h\|_{\mathcal{V}} \leq \sup_{w \in \mathcal{V}_h} \frac{a(v - u_h, w)}{\|w\|_{\mathcal{V}}} = \sup_{w \in \mathcal{V}_h} \frac{a(v - u, w)}{\|w\|_{\mathcal{V}}} \leq M \|v - u\|_v,$$

where

$$M = \|a(\cdot,\cdot)\| = \sup_{v \in \mathcal{V}_h, v \neq 0} \frac{a(v,v)}{\|v\|_{\mathcal{V}}^2};$$

the proofs concludes using the triangular inequality. □

Corollary 2.19. *Let $\{\mathcal{V}_h\}$ a sequence of finite-dimensional subspaces of $\mathcal{V}$ indexed by the parameter h. If for $h \longrightarrow 0$ we have that:*

$$\inf_{v_h \in \mathcal{V}_h} \|v - v_h\|_{\mathcal{V}} \longrightarrow 0,$$

then, for $\alpha = \inf_h \alpha_h > 0$, u_h converges to u in $\mathcal{V}$.

Example 2.50. Let $\mathcal{V} = H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^2$. Consider

$$a(v,w) = \int_\Omega \nabla v \cdot \nabla w \, dx; \qquad F(v) = \int_\Omega fv \, dx,$$

with $f \in L^s(\Omega)$. The bilinear form is symmetric and continuous. Coercivity derives from application of Poincarè theorem 2.6. Thus we have:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch,$$

if u is sufficiently smooth.

Example 2.51. Consider the convection-diffusion equation in $\mathbb{R}^2$:

$$\begin{aligned} -\mu\Delta u + \operatorname{div}(\beta u) + u &= f \qquad in \Omega \\ u &= 0 \qquad in \partial\Omega \end{aligned}$$

where $\beta = (\beta_1, \beta_2)$ is a vector field of $\mathbb{R}^2$. Multiplying by $v \in \mathcal{V} = H_0^1(\Omega)$, integrating over Ω. and applying Green's lemma we have:

$$a(u,v) = F(v) \qquad \forall v \in \mathcal{V},$$

where:

$$a(v,w) = \int_\Omega \left(\nabla v \cdot \nabla w + \operatorname{div}(\beta v) w\right) \, dx, \qquad F(v) = \int_\Omega fv \, dx.$$

Assume $\mu = 1$ and $|\beta|/\mu$ small. The problem is coercive. In fact, application of Green's lemma to the second term yields:

$$\begin{aligned} \int_\Omega \operatorname{div}(\beta u) v \, dx = \\ \int_\Gamma (\beta \cdot n) u \, v \, ds - \int_\Omega \operatorname{div}(\beta v) u \, dx \\ = -\int_\Omega \operatorname{div}(\beta v) u \, dx. \end{aligned}$$

This is valid for all $v \in \mathcal{V}$, and thus also for $v = u$, which yields:

$$\int_\Omega \operatorname{div}(\beta u) u \, dx = 0.$$

and

$$a(v,v) = \int_\Omega \left(|\nabla v|^2 + v^2\right) \, dx = \|v\|^2_{\mathcal{H}^1(\Omega)}.$$

Remark 2.52. *We note that coerciveness of the bilinear form is regulated by the diffusive term if* $\operatorname{div}\beta = 0$*, i.e., the advective field that transports the quantity* u *must be divergence free, or in other terms, must be "conservative".*

We can formulate a convergent FEM problem: find $u_h \in \mathcal{V}_h$ such that:

$$a(u_h, v) = F(v) \qquad \forall v \in \mathcal{V}_h.$$

The stiffness matrix is not symmetric any longer but coercivity guarantees that it is invertible. Moreover ($\alpha = 1$):

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq \gamma \|u - v\|_{\mathcal{H}^1(\Omega)} \qquad \forall v \in \mathcal{V}_h.$$

Example 2.53. Let u be the temperature of a solid whose shape is given by $\Omega \in \mathbb{R}^3$. The thermal flux is given by Fourier's law:

$$q_i(x) = -k_i(x)\frac{\partial u}{\partial x_i} \qquad x \in \Omega; i = 1, 2, 3, ;$$

Conservation of energy states that:

$$\operatorname{div} q = \sum_{i=1}^{3} \frac{\partial}{\partial x_i}\left(k_i(x)\frac{\partial u}{\partial x_i}\right) = f \qquad x \in \Omega;$$

This is an example of a PDE with variable coefficient. The variational formulation requires boundary conditions:

$$\begin{aligned} u &= 0 && in\Gamma_D \\ -q \cdot n &= g && in\Gamma_N \end{aligned}$$

with $\partial\Omega = \Gamma = \Gamma_D \cup \Gamma_N$.
Let $\mathcal{V} = \{v \in \mathcal{H}^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. With a standard approach we obtain:

$$\int_\Omega fv\, dx = \int_\omega v \operatorname{div} q\, dx = \int_\Gamma vq \cdot n\, ds - \int_\Omega q \cdot \nabla v\, dx = \sum_{i=1}^{3} \int_\Omega k_i(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\, dx - \int_{\Gamma_N} gv\, ds,$$

and thus the following: Find $u \in \mathcal{V}$ such that:

$$a(u, v) = F(v) \qquad \forall v \in \mathcal{V},$$

where:

$$a(v,w) = \sum_{i=1}^{3} \int_{\Omega} k_i(x) \frac{\partial u}{\partial x_i}$$
$$F(v) = \int_{\Omega} fv \, dx + \int_{\Gamma_N} gv \, ds.$$

The bilinear form is symmetric, continuous and coercive and the linear form is continuous if there exist constants c and C such that:

$$c \leq k_i(x) \leq C, \qquad \forall x \in \Omega; i = 1,2,3,$$

and moreover $f \in \mathcal{L}^2(\Omega)$ e $g \in \mathcal{L}^2(\Gamma_N)$ the measure of Γ_D is positive (non-zero).

2.8 Finite element spaces

The final definition of the finite element method requires now a proper definition of $\mathcal{V}_h$. The choice made by FEM is to use piecewise continuous polynomials defined on appropriate subdivisions of the domain $\Omega \in \mathbb{R}^d$, called generically triangulations. A triangulation $\mathcal{T}_h = \{K\}$ is thus formed by the union of elements T (or subdivisions) that cover Ω without superposition. The spaces we are looking for are finite dimensional subspaces of $\mathcal{H}^1(\Omega)$ (or of $H^2(\Omega)$ for PDE of fourth order, for example). To properly define the piecewise continuous polynomials we need to require:

$$\mathcal{V}_h \subset \mathcal{H}^1(\Omega) \Leftrightarrow \mathcal{V}_h \subset \mathcal{C}^0(\bar{\Omega})$$
$$\mathcal{V}_h \subset \mathcal{H}^2(\Omega) \Leftrightarrow \mathcal{V}_h \subset \mathcal{C}^1(\bar{\Omega})$$

where $\bar{\Omega} = \Omega \cup \Gamma$.

2.8.1 Two-dimensional case ($d = 2$)

Let the domain $\Omega \in \mathbb{R}^2$ be characterized by a polygonal boundary Γ. Let $\mathcal{T}_h = \{T\}$ be a triangulation formed by triangular elements T, and let $\pi_r(T)$ the polynomial of degree r in T:

$$\mathcal{P}_r(T) = \{v : v \text{ polynomial of degree } \leq r \text{ in } T\}.$$

A linear polynomial $\pi_1(T)$ can be written as

$$v(x) = a_{00} + a_{10}x_1 + a_{01}x_2, \qquad x \in T, \tag{2.71}$$

with $a_{ij} \in \mathbb{R}$. We note immediately that $\phi_1(x) = 1$, $\phi_2(x) = x_1$, $\phi_3(x) = x_2$ are a basis for $\mathcal{P}_1(T)$.

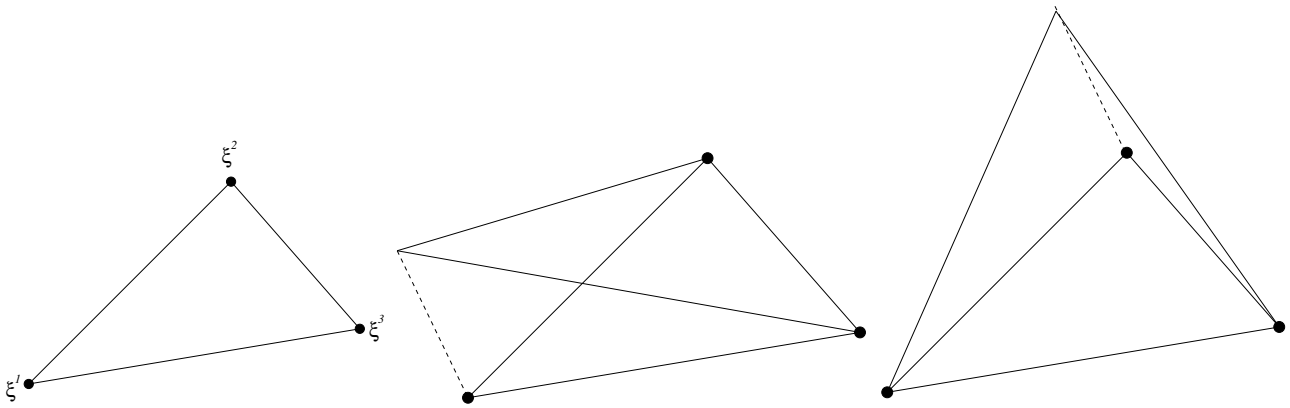


FIGURE 2.15: *Left: triangle with degrees of freedom defining a linear function. Center and right: examples of linear basis functions.*

In the quadratic case we may write:

$$v(x) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{11}x_1x_2 + a_{02}x_2^2, \qquad x \in T,$$

with $a_{ij} \in \mathbb{R}$ and basis $\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$. In general we can write:

$$\mathcal{P}_r(T) = \left\{ v : v(x) = \sum_{0 \le i+j \le r} a_{ij} x_1^i x_2^j \text{ for } x \in T \right\} \qquad \dim \mathcal{P}_r(T) = \frac{(r+1)(r+2)}{2}.$$

Example 2.54. Affine polynomials on triangles (Fig.2.15:

$$\mathcal{V}_h = \left\{ v \in C^0(\bar{\Omega}) : v|_T \in \mathcal{P}_1(T), \forall T \in \mathcal{T}_h \right\}.$$

The space $\mathcal{V}_h$ is then formed by functions that are piecewise continuous and with derivatives that are piecewise constant. To describe these functions we use the "degrees of freedom", in this case the nodes of $\mathcal{T}_h$. Every function $v \in \mathcal{V}_h(T)$ is uniquely determined by its values at the nodes of T. Let $\xi^{(i)}$, $i = 1, 2, 3$ be the coordinates of these nodes. Then, for $\alpha_i \in \mathbb{R}$ we have:

Theorem 2.20. *Let $T \in \mathcal{T}_h$ a triangle with vertices having coordinates $\xi^{(i)}$, $i = 1, 2, 3$. A function $v(x) \in \mathcal{P}_1(T)$ is uniquely determined by its values at the vertices. In other words, given the values $\alpha_i \in \mathbb{R}$, $i = 1, 2, 3$, $v(x) \in \mathcal{P}_1(T)$ is uniquely determined by:*

$$v(\xi^{(i)}) = \alpha_i \qquad i = 1, 2, 3 \tag{2.72}$$

Proof. The generic function $v(x)$ can be written as in (2.71). This the linear system (2.72) has a unique solution if matrix

$$A = \begin{bmatrix} \xi_1^{(1)} & \xi_2^{(1)} & 1 \\ \xi_1^{(2)} & \xi_2^{(2)} & 1 \\ \xi_1^{(3)} & \xi_2^{(3)} & 1 \end{bmatrix}$$

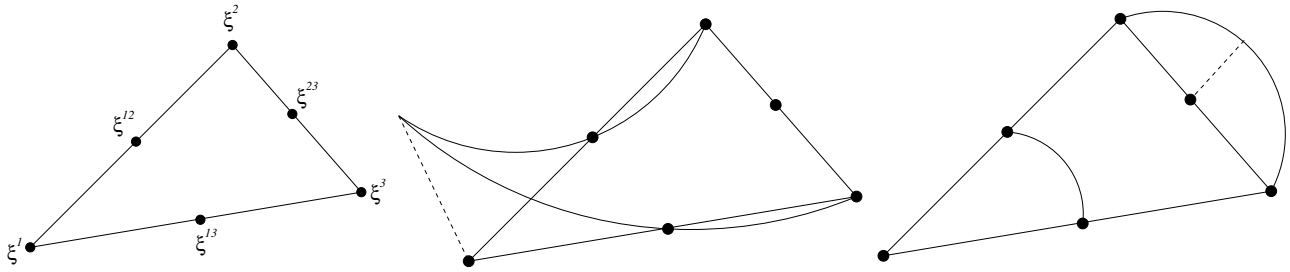


FIGURE 2.16: *Left: degrees of freedom for a triangle defining a quadratic function. Center and right: example of quadratic basis functions.*

is nonsingular. This is true since $\mathrm{Ker}(A) = \emptyset$. In fact, it was not empty, there would be a nonzero vector $a = (a_1, a_2, a_3)$ such that $Aa = 0$. Then we would have a polynomial of degree 1 in $\mathbb{R}^2$ with three roots. □

To determine the basis functions it is sufficient to choose α_i appropriately. Then, following the Lagrangian interpolation idea, we can chose α_i equal to $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. We can then define functions $\phi_i(x)$ that are continuous on $\bar{\Omega}$ with piecewise constant gradient. For each element T we have:

$$v(x)|_T = \sum_{i=1}^{3} v(\xi^{(i)})\phi_i(x) \qquad \nabla v(x)|_T = \sum_{i=1}^{3} v(\xi^{(i)})\nabla\phi_i(x).$$

Example 2.55. Quadratic basis functions. The space $\mathcal{V}_h$ is given by:

$$\mathcal{V}_h = \left\{ v \in C^0(\bar{\Omega}) : v|_T \in \mathcal{P}_2(T), \forall T \in \mathcal{T}_h \right\}.$$

To describe these functions we need six degrees of freedom $T \in \mathcal{T}_h$. We choose the vertices of T and the midpoints of each edge (Fig. 2.16). We have then:

Theorem 2.21. *Let $T \in \mathcal{T}_h$ a triangle whose vertices have coordinates $\xi^{(i)}$, $i = 1, 2, 3$. Let $\xi^{(ij)}$ be the coordinates of the midpoints of the edge between nodes i and j. A function $v(x) \in \mathcal{P}_2(T)$ is uniquely determined by:*

$$v(\xi^{(i)}) = \alpha_i \qquad i = 1, 2, 3 \qquad\qquad v(\xi^{(ij)}) = \alpha_{ij} \qquad i < j, \;\; i, j = 1, 2, 3.$$

Proof. Again it suffices to determine as before that the conditions $v(\xi^{(i)}) = 0$ and $v(\xi^{(ij)}) = 0$ $i < j, i, j = 1, 2, 3$ imply $v = 0$ on the whole T. Consider an edge, e.g, between nodes 2 and 3 (Fig. 2.16). The quadratic function restricted on this edge is uniquely determined by the three points $\xi^{(2)}$, $\xi^{(23)}$, and $\xi^{(3)}$. if v is zero on these nodes, then v is identically zero on edge 2-3. Thus we can factor a function $\phi_1(x)$ (the polynomial of degree 1 of the previous example):

$$v(x) = \phi_1(x)w_1(x).$$

Repeating this argument on the edge between nodes 1 and 3, we have:

$$v(x) = \phi_1(x)\phi_2(x)w_0,$$

where now w_0 is constant. Now take $v(\xi^{(12)}) = 0$. We find:

$$0 = v(\xi^{(12)}) = \phi_1(\xi^{(12)})\phi_2(\xi^{(12)})w_0 = \frac{1}{2}\frac{1}{2}w_0,$$

which yields $w_0 = 0$. □

Quadratic basis functions can be determined from the linear basis functions as:

$$v(x)|_T = \sum_{i=1}^{3} v(\xi^{(i)})\phi_i(x)(2\phi_i(x) - 1) + \sum_{\substack{i,j=1 \\ i<j}}^{3} v(\xi^{(ij)})4\phi_i(x)\phi_j(x).$$

2.9 Error estimates for elliptic problems

For elliptic equation whose bilinear form is coercive with constant α and continuous with constant γ, Céa's lemma ensures that:

$$\|u - u_h\|_{\mathcal{V}} \leq \frac{\gamma}{\alpha} \|u - v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}_h.$$

The idea is to take for v the interpolating polynomial $\pi_{h,r}u$ of u, so that we are left with the problem of estimating the interpolation error $\|u - \pi_{h,r}u\|_{\mathcal{V}}$. It is evident that it is sufficient to estimate this error on each element $T \in \mathcal{T}_h$ and then sum over all elements, as done in the one-dimensional case.

Interpolation error. We consider here a polygonal domain $\Omega \in \mathbb{R}^2$ (the extension to $\mathbb{R}^3$ is immediate) and a triangulation $\mathcal{T}_h(\Omega)$. We identify with T_j the j-th triangle of $\mathcal{T}_h(\Omega)$. Then:

$$\mathcal{T}_h(\Omega) = \bigcup_{j=1}^{M} T_j$$

$$T_j \cap T_i = \begin{cases} \emptyset \\ \sigma_{ij} \end{cases}$$

where σ_{ij} is the edge between nodes i and j. For each $T \in \mathcal{T}_h$ we denote;

$h_T =$ diameter of $T =$ edge of maximum length of T;
$\rho_T =$ diameter of the circle inscribed in T;

The triangulation $\mathcal{T}_h$ will be characterized by a single grid parameter h, defined as;

$$h = \max_{T \in \mathcal{T}_h} h_T.$$

Consider a family of triangulations $\{\mathcal{T}_h(\Omega)\}$ indexed by h and a corresponding family of functional spaces $\mathcal{V}_h = \{v \in \mathcal{H}^1(\Omega) : v|_T \in \mathcal{P}_1(T)\}$. We have:

Definition 2.22 (Regular triangulation)**.** A triangulation $\{\mathcal{T}_h(\Omega)\}$ is "regular" if there exists a constant $\beta > 0$ independent of h and of the member $\mathcal{T}_h$ of the family $\{\mathcal{T}_h\}$ such that

$$\frac{\rho_T}{h_T} \geq \beta \qquad \forall T \in \mathcal{T}_h.$$

The constant β estimates the measure of the smallest angle among all triangles T. The regularity of the triangulation ensures that during the limit process $h \longrightarrow 0$ no angle of the triangulation tends to zero. We report here the classical interpolation result, whose proof can be found for example in [22]:

Theorem 2.23. *Let $T \in \mathcal{T}_h$ be a triangle with vertices $\xi^{(i)}$, $i = 1, 2, 3$. Let $v(x) \in \mathcal{H}^{r+1}(T)$ and $\pi_{h,r} v \in \mathcal{P}_r(T)$ be its Lagrangian interpolating polynomial of degree r. Then for each triangle T:*

$$\|v - \pi_{h,r} v\|_{\mathcal{L}^2(T)} \leq C h_T^{r+1} \left\|\partial^{r+1} v\right\|_{\mathcal{L}^2(T)},$$
$$|\, v - \pi_{h,r} v \,|_{\mathcal{H}^1(T)} \leq C \frac{h_T^{r+1}}{\rho_T} \left\|\partial^{r+1} v\right\|_{\mathcal{L}^2(T)} \leq \frac{C}{\beta} h_T^r \left\|\partial^{r+1} v\right\|_{\mathcal{L}^2(T)}.$$

Remark 2.56. *We note that the second inequality contains the grid parameter ρ_T, which comes in play when we estimate the gradients of v and $\pi_{h,r} v$ on triangle T. In fact the norm of the gradient of functions $v \in \mathcal{H}^{r+1}(T)$ is bounded by $1/\rho_T$.*

Remark 2.57. *The basis functions $\phi_j \in \mathcal{P}_1(T)$ two-dimensional triangle T have the following useful properties:*

Lemma 2.24. *For $j = 1, 2$ and $x \in T$ the following properties hold:*

$$\sum_{i=1}^{3} \phi_i(x) = 1 \qquad\qquad \sum_{i=1}^{3} \frac{\partial \phi_i}{\partial x_j}(x) = 0$$

The following corollary holds for the family of triangulations:

Corollary 2.25. *If the members of the family $\{\mathcal{T}_h\}$ are regular triangulations, then there exist two constants C_1 and C_2 independent of h and of $v \in H^{r+1}(\Omega)$ such that*

$$\|v - \pi_{h,r} v\|_{\mathcal{L}^2(\Omega)} \leq C_1 h^{r+1} \left\|\partial^{r+1} v\right\|_{\mathcal{L}^2(\Omega)}, \tag{2.73}$$

$$|\, v - \pi_{h,r} v \,|_{\mathcal{H}^1(\Omega)} \leq C_2 h^{r} \left\|\partial^{r+1} v\right\|_{\mathcal{L}^2(\Omega)}. \tag{2.74}$$

Proof. We show the corollary for $r = 1$ (linear interpolation). In this case theorem (2.23) specializes in:

$$\|v - \pi_{h,1} v\|_{\mathcal{L}^2(T)} \leq C h_T^2 \left\|\partial^2 v\right\|_{\mathcal{L}^2(T)},$$

$$|\, v - \pi_{h,1} v \,|_{\mathcal{H}^1(T)} \leq C \frac{h_T^2}{\rho_T} \left\|\partial^2 v\right\|_{\mathcal{L}^2(T)}.$$

Summing over all $T \in \mathcal{T}_h$ we have:

$$\|v - \pi_{h,k} v\|^2_{\mathcal{L}^2(\Omega)} = \sum_{T \in \mathcal{T}_h} \|v - \pi_{h,k} v\|^2_{\mathcal{L}^2(T)} \leq \sum_{T \in \mathcal{T}_h} C^2 h_T^4 \left\|\partial^2 v\right\|^2_{\mathcal{L}^2(T)} \leq C^2 h^4 \left\|\partial^2 v\right\|^2_{\mathcal{L}^2(\Omega)}.$$

For the second inequality, recall that $h_T/\rho_T \leq 1/beta$. Then:

$$|\, v - \pi_{h,k} v \,|^2_{\mathcal{H}^1(\Omega)} = \sum_{T \in \mathcal{T}_h} |\, v - \pi_{h,k} v \,|^2_{\mathcal{H}^1(T)} \leq \sum_{T \in \mathcal{T}_h} C^2 \frac{h_T^4}{\rho_T^2} \left\|\partial^2 v\right\|^2_{\mathcal{L}^2(T)} \leq \frac{C^2}{\beta} h^2 \left\|\partial^2 v\right\|^2_{\mathcal{L}^2(\Omega)}.$$

□

As typical of Lagrangian interpolation, accuracy is determined by the order of the interpolating polynomial and by the smoothness of the interpolated function. In general, we have for $1 \leq s \leq r + 1$:

$$\|v - \pi_{h,s} v\|_{\mathcal{L}^2(T)} \leq C h_T^s \|\partial^s v\|_{\mathcal{L}^2(T)},$$

$$|\, v - \pi_{h,s} v \,|_{\mathcal{H}^1(T)} \leq C h_T^{s-1} \|\partial^s v\|_{\mathcal{L}^2(T)}.$$

FEM error and regularity of the solution From Céa Lemma we have immediately an estimate of the FEM error:

$$\|u - u_h\|_{\mathcal{V}} \leq \frac{\gamma}{\alpha} \|u - \pi_{h,k} u\|_{\mathcal{V}},$$

and using the interpolation error estimates above we can estimate the FEM error for different problems.
For example, consider the following Poisson problem:

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{in } \Gamma = \partial\Omega \end{aligned}$$

Let $\mathcal{V} = H_0^1(\Omega)$ and $\mathcal{V}_h = \{v \in \mathcal{V} : v|_T \in \mathcal{P}_r(T) \forall T \in \mathcal{T}_h\}$. Then we have:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch^r | \, u \, |_{\mathcal{H}^{r+1}(\Omega)}.$$

The regularity theory for elliptic equations ensures that if Γ is sufficiently regular, then we have:

$$\|u\|_{H^{s+2}(\Omega)} \leq C \, \|f\|_{H^s(\Omega)} \, , \tag{2.75}$$

or, intuitively, the solution gains two orders of derivatives with respect to the forcing function f.
If Γ is non smooth, the estimate may not be true (even for $s = 0$). For example, if Ω is not convex with an angular vertex pointing towards the interior of Ω, the solution will have a singularity even though f is smooth. To give some intuition, we could approximate the solution u in such a point as (we use polar coordinates centered in the angular point):

$$u(r, \theta) = r^\gamma \alpha(\theta) + \beta(r, \theta) \qquad \gamma = \frac{\pi}{\omega}, \tag{2.76}$$

where ω is the measure of the angle at the boundary. It is possible to show that (2.75) is valid with $s = 0$ if $\omega < \pi$, i.e., convex domain with polygonal boundary. If $\omega > \pi$, a function of the form (2.76) does not belong to $H^2(\Omega)$ if $\alpha \neq 0$. It is easy to verify that:

$$\int_\Omega | \, \partial^s u \, |^2 \, dx \approx C \int_0^R [r^{\gamma - s}]^2 r \, dr.$$

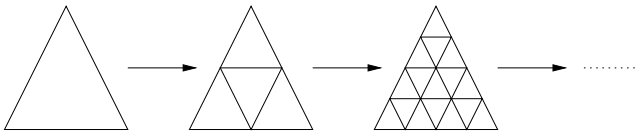
This integral exists and is finite, thus $u \in H^s(\Omega)$, if $s < \gamma + 1$. The error of the FEM formulation can be described for every $\epsilon > 0$ as:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch^{\gamma - \epsilon} \, \|u\|_{H^{\gamma - \epsilon + 1}(\Omega)} = Ch^{\gamma - \epsilon},$$

where $\gamma = \pi/\omega$, and ω is the measure of the largest angle of the angular point of Γ. For example, if $\gamma = 2/3$, i.e., an angle with $\omega = 3\pi/2$, we loose the $(O(h))$ convergence of the method:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch^{\frac{2}{3} - \epsilon}.$$

We can however conceive adaptive methods that try to decrease h_T locally to prevent this loss of accuracy. In principle these methods work very well, although the complication of dynamically adjusting the triangulation renders problematic their application to large scale problems problems. We will not pursue this approach further and refer the reader to the specialized literature.



ℓ	$\|u-u_{h,\ell}\|_{\mathcal{L}^2}$	$\frac{\|u-u_{h,\ell-1}\|_{\mathcal{L}^2}}{\|u-u_{h,\ell}\|_{\mathcal{L}^2}}$	p
0	0.2241	–	
1	0.0483	4.64	2.32
2	0.0117	4.13	2.06
3	0.0029	4.03	2.02

FIGURE 2.17: *Left: Uniform refinements of initial triangle ($\ell = 0$, $\ell = 1$, $\ell = 2$). Right: $\mathcal{L}^2$ error norms and convergence rates p for $\ell = 0, \ldots, 3$.*

$\mathcal{L}^2$ error for the Poisson equation. The Nitsche-Aubin- trick. From the preceding analysis we realize that there is a discrepancy from the order of convergence of the FE scheme as derived from Céa's lemma and the order of convergence of the interpolation scheme. In fact, we can only state the following FEM convergence estimate:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch|\, u\,|_{H^2(\Omega)},$$

and the following interpolation convergence estimate:

$$\|u - \pi_{h,1}u\|_{\mathcal{L}^2(\Omega)} \leq Ch^2|\, u\,|_{H^2(\Omega)}.$$

Example 2.58 (FEM for Poisson equation with smooth solution)**.** We solve the following problem:

$$\begin{aligned} -\Delta u &= f(x,y) && (x,y) \in \Omega \subset \mathbb{R}^2 \\ u &= 0 && (x,y) \in \partial\Omega \end{aligned}$$

where $\Omega = \{(x,y) : (x,y) : -1 \leq x \leq 1, -1 \leq y \leq 1\}$ and $f(x,y) = -5/4\pi^2 \sin(\pi\, x)\cos(\pi\, y/2)$. The problem admits the solution $u(x,y) = \sin(\pi\, x)\cos(\pi\, y/2)$.
We solve this problem with linear Galerkin FEM and use a sequence of four $\ell \in \{0,1,2,3\}$ unstructured triangulations $\mathcal{T}_{h,\ell}(\Omega)$ constructed by uniform refinements of an initial mesh $\mathcal{T}_{h,0}(\Omega)$ obtained by halving h at each refinement level (Fig. 2.17, left panel). The table on Figure 2.17 (right panel) shows the experimental $\mathcal{L}^2$ error norms and their ratio at the different mesh levels. The errors decrease by a factor almost 4 at each refinement, showing that convergence of the FEM method is optimal, in the sense that the $\mathcal{L}^2$ norm of the error is proportional to h^2, i.e., the same convergence rate of the theoretical interpolation error.

Backed up by numerical convergence, we see that we should find theoretical evidence for the above calculations. This is obtained using by duality arguments using the so called "Aubin-Nitsche trick". We have the following:

Theorem 2.26. *Let Ω a convex polygonal domain and u_h the FEM solution of the Poisson equation with piecewise linear basis functions. Then there exists a constant C independent of h and u such that:*

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq Ch^2 |\, u \,|_{H^2(\Omega)}.$$

Proof. Let $e = u - u_h$ be the error function. Strong consistency of the scheme implies:

$$a\,(e, v) = 0 \qquad \forall v \in \mathcal{V}_h, \tag{2.77}$$

We want to estimate the $\mathcal{L}^2$ error norm, which we note can be equivalently defined as:

$$\|e\|_{\mathcal{L}^2(\Omega)} = (e, e)^{1/2} = \sup_{v \in \mathcal{L}^2} \frac{\int_\Omega e\, v\, dx}{\|v\|_{\mathcal{L}^2}}.$$

The integral in the numerator reminds us of the linear form of the right-hand-side of a variational formulation for a Poisson equation with e as source function. Then we can let φ be solution of the following dual (adjoint) problem:

$$\begin{aligned} -\Delta\varphi &= e \qquad \text{in } \Omega \\ \varphi &= 0 \text{ in } \Gamma. \end{aligned}$$

Since Ω is convex, the elliptic estimate (2.75) holds with $s = 0$. Then:

$$\|\varphi\|_{H^2(\Omega)} \leq C\, \|e\|_{\mathcal{L}^2(\Omega)}\,. \tag{2.78}$$

Now we can use Green's Lemma, the fact that $e = 0$ in Γ and the consistency of the FEM scheme (2.77) stating that $a\,(e, \pi_{h,1}\varphi) = 0$, to obtain:

$$(e, e) = (e, -\Delta\varphi) = a\,(e, \varphi) = a\,(e, \varphi - \pi_{h,1}\varphi)\,.$$

Using again Green's Lemma and noting that $\varphi = 0$ at the boundary, we have:

$$\begin{aligned} \|e\|^2_{\mathcal{L}^2(\Omega)} &= \int_\Omega \nabla e \nabla(\varphi - \pi_{h,1}\varphi) \leq \|\nabla e\|_{\mathcal{L}^2(\Omega)}\, \|\nabla(\varphi - \pi_{h,1}\varphi)\|_{\mathcal{L}^2(\Omega)} \\ &\leq \|e\|_{\mathcal{H}^1(\Omega)}\, \|\varphi - \pi_{h,1}\varphi\|_{\mathcal{H}^1(\Omega)}\,. \end{aligned}$$

We now use the interpolation error estimate (2.74) with $r = 1$:

$$\|e\|^2_{\mathcal{L}^2(\Omega)} \leq C\, \|e\|_{\mathcal{H}^1(\Omega)}\, h |\, \varphi \,|_{H^2(\Omega)};$$

finally, using (2.78) on the auxiliary (adjoint) Poisson problem forced by the error:

$$\|e\|^2_{\mathcal{L}^2(\Omega)} \leq Ch\, \|e\|_{\mathcal{H}^1(\Omega)}\, h\, \|e\|_{\mathcal{L}^2(\Omega)}\,.$$

Division by $\|e\|_{\mathcal{L}^2(\Omega)}$ yields:

$$\|e\|_{\mathcal{L}^2(\Omega)} \leq Ch\, \|e\|_{\mathcal{H}^1(\Omega)}\,,$$

which using Céa Lemma gives immediately the desired result:

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq ch^2 |\, u \,|_{H^2(\Omega)}.$$

□

2.10 Estimate of the condition number of the stiffness matrix

We have now almost all the tools that allow to show the result stated in Remark 2.10. We take as an example the Poisson equation with homogeneous Dirichlet conditions discretized via linear finite elements on a regular triangulation $\mathcal{T}_h$ (cfr. paragraph 2.9). The stiffness matrix is given by::

$$A = \{a_{ij}\} \qquad a_{ij} = a(\phi_i, \phi_j) \qquad a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j,$$

with $\phi_j \in \mathcal{P}_1(\mathcal{T}_h)$. We have the following:

Theorem 2.27. *The condition number of the stiffness matrix A can be estimated by:*

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

In particular, the largest and smallest eigenvalues are $\lambda_1(A) = \mathcal{O}(1)$ and $\lambda_n(A) = \mathcal{O}(h^2)$, respectively.

Before proceeding with the proof we show the following result, known as "inverse estimate" (inverse of the Poincaré inequality), used to estimate the gradient of the solution with the solution itself. Note that this is obtained at the cost of the appearance of a factor $1/h$ in the estimate.

Lemma 2.28 (Inverse estimate)**.** *There exist two constants c and C depending only on the regularity constants of the triangulation $\mathcal{T}_h$ such that for all $v = \sum_{i=1}^N \alpha_i\phi_i \in \mathcal{V}_h$ we have:*

$$ch^2 |\, \alpha \,|^2 \leq \|v\|^2_{\mathcal{L}^2(\Omega)} \leq Ch^2 |\, \alpha \,|^2; \tag{2.79}$$

$$a(v, v) = \int_\Omega |\, \nabla v \,|^2 \, dx \leq Ch^{-2} \|v\|^2_{\mathcal{L}^2(\Omega)}. \tag{2.80}$$

Proof. We need to show that for each triangle $T \in \mathcal{T}_h$ with vertex coordinates given by $\xi^{(i)}$, $i = 1, 2, 3$ and for all $v \in \mathcal{P}_1(T)$, we have:

$$ch_T^2 \sum_{i=1}^3 |\, v(\xi^{(i)}) \,|^2 \leq \|v\|^2_{L^2(T)} \leq Ch_T^2 \sum_{i=1}^3 |\, v(\xi^{(i)}) \,|^2, \tag{2.81}$$

$$\int_T |\, \nabla v \,|^2 \, dx \leq Ch_T^{-2} \int_T |\, v \,|^2 \, dx. \tag{2.82}$$

Then summing up over all $T \in \mathcal{T}_h$ we obtain the result.
The strategy is to show the inequalities for a reference triangle $\hat{T}$ with nodal coordinates given by $\hat{\xi}^{(1)} = (0, 0)$, $\hat{\xi}^{(2)} = (1, 0)$, and $\hat{\xi}^{(3)} = (0, 1)$ and then use an affine transformation to translate the inequalities for a general triangle T in the reference system (x_1, x_2)

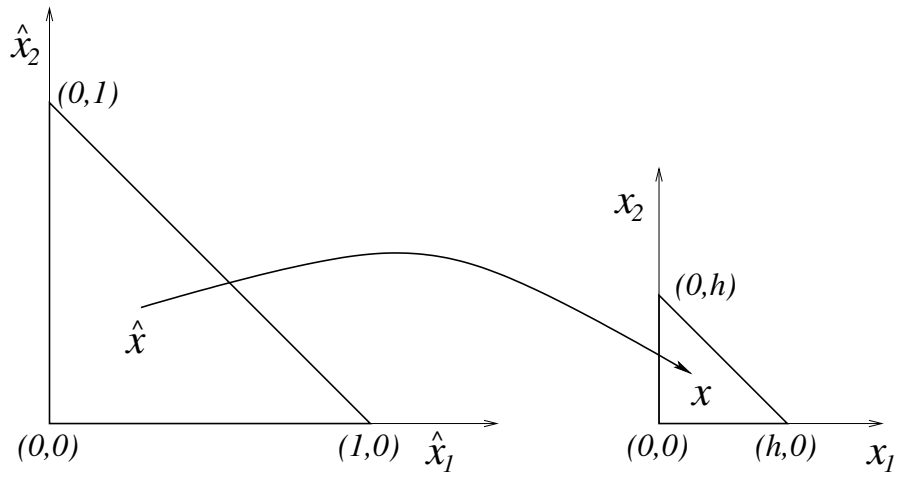


FIGURE 2.18: *Coordinate transformation from a reference triangle $\hat{T}$ to the scaled triangle T.*

Let $\hat{\phi}_i(\hat{x})$ be the classical basis function for $\mathcal{P}_1(\hat{T})$ and let:

$$\hat{v}(\hat{x}) = \sum_{i=1}^{3} \gamma_i \hat{\phi}_i(\hat{x}), \qquad \forall \hat{x} \in \hat{T}.$$

Let $\gamma = (\gamma_1, \gamma_2, \gamma_3)$. We want to show that the function $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ defined as

$$f(\gamma) = \frac{\int_{\hat{T}} |\, \nabla \hat{v} \,|^2 \, d\hat{x}}{\int_{\hat{T}} \hat{v}^2 \, d\hat{x}}, \gamma \neq 0,$$

satisfies:

$$f(\gamma) \leq C \qquad \forall \gamma \in \mathbb{R}^3, \gamma \neq 0. \tag{2.83}$$

From this we obtain (2.82) with $T = \hat{T}$ e $h_T = \sqrt{2}$. Note that $f(\gamma)$ is a homogeneous function of degree zero ($f(\alpha\gamma) = f(\gamma)\ \forall \alpha \in \mathbb{R}$, $\alpha \neq 0$). To show (2.83) we show that $f(\gamma)$ is continuous and bounded in a ball $B = \{\gamma \in \mathbb{R}^3 : \|\gamma\| = 1\}$. In fact, $f(\gamma) \neq 0$ for $\gamma \in B$ and is continuous since γ_i are the baricentric coordinates of $\hat{v}$ in $\hat{T}$. Since B is closed and bounded in $\mathbb{R}^3$, then f reaches its maximum value in B contained in $\hat{T}$.
Now we work on a simplified generic triangle T, similar to $\hat{T}$ but with edges of length h and hypotenuse $h_T = \sqrt{h}$ (see Fig. 2.18). The map $F : \hat{T} \longrightarrow T$ is:

$$x = F(\hat{x}) = (h\hat{x}_1, h\hat{x}_2)\,, \qquad \hat{x} \in \hat{T}.$$

For every function $v \in \mathcal{P}_1(T)$ we have:

$$\hat{v}(\hat{x}) = v\,(F(\hat{x}))\,, \qquad \hat{x} \in \hat{T},$$

The Jacobian of the transformation is given by:

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = \frac{\partial v}{\partial x_1}\frac{\partial x_1}{\partial \hat{x}_i} + \frac{\partial v}{\partial x_2}\frac{\partial x_2}{\partial \hat{x}_i} = \frac{\partial v}{\partial x_i} h.$$

Hence we have that $\nabla \hat{v} = h \nabla v$ and obviously $dx = h^2 d\hat{x}$, $d\hat{x} = dx/h^2$, yielding:

$$\int_{\hat{T}} |\, \nabla \hat{v} \,|^2 \, d\hat{x} = \int_T h^{-2} |\, \nabla v \,|^2 h^2 \, dx \leq C \int_{\hat{T}} \hat{v}^2 \, d\hat{x} = C h^{-2} \int_T v^2 \, dx.$$

In analogy, to go from $\hat{T}$ to a general triangle, we can form the coordinate transformation:

$$x = F(\hat{x}) = \xi^{(1)} + \left(\xi^{(2)} - \xi^{(1)}\right) \hat{x}_1 + \left(\xi^{(3)} - \xi^{(1)}\right) \hat{x}_2 = B_T \hat{x} + p.$$

Using the fact that $|\, \xi^{(i)} - \xi^{(1)} \,| \leq C h_T$, $i = 1, 2, 3$ and $dx = C h_t^2 d\hat{x}$, which are true because of the regularity property of $\mathcal{T}_h$, we obtain the sought result. □

Proof of theorem 2.27. A generic function $v \in \mathcal{V}_h$ can be written as a linear combination of the basis functions:

$$v(x) = \sum_{i=1}^{N} \beta_i \phi_i(x),$$

hence:

$$a(v, v) = \beta \cdot A\beta,$$

with $\beta = \{\beta_i\}$. Using the inverse inequality (2.79) and (2.80) of Lemma 2.28, we have:

$$\frac{\beta \cdot A\beta}{\|\beta\|^2} = \frac{a(v, v)}{\|\beta\|^2} \leq C h^{-2} \frac{\|v\|^2_{\mathcal{L}^2(\Omega)}}{\|\beta\|^2} \leq C^2 \qquad \forall \beta \in \mathbb{R}^N.$$

The coercivity of the bilinear form $a(\cdot, \cdot)$ together with eq. (2.79), yields ($\|v\|_{\mathcal{H}^1(\Omega} \geq \|v\|_{\mathcal{L}^2(\Omega)}$):

$$\frac{\beta \cdot A\beta}{\|\beta\|^2} = \frac{a(v, v)}{\|\beta\|^2} \geq \alpha \frac{\|v\|^2_{\mathcal{L}^2(\Omega)}}{\|\beta\|^2} \geq C\alpha h^2 \qquad \forall \beta \in \mathbb{R}^N.$$

Thus there exist two constants c and C independent of h such that:

$$\lambda_{\max} \leq C, \qquad \lambda_{\min} \geq c h^2,$$

and thus $\kappa(A) = \lambda_{\max}/\lambda_{\min} \leq C h^{-2}$. □

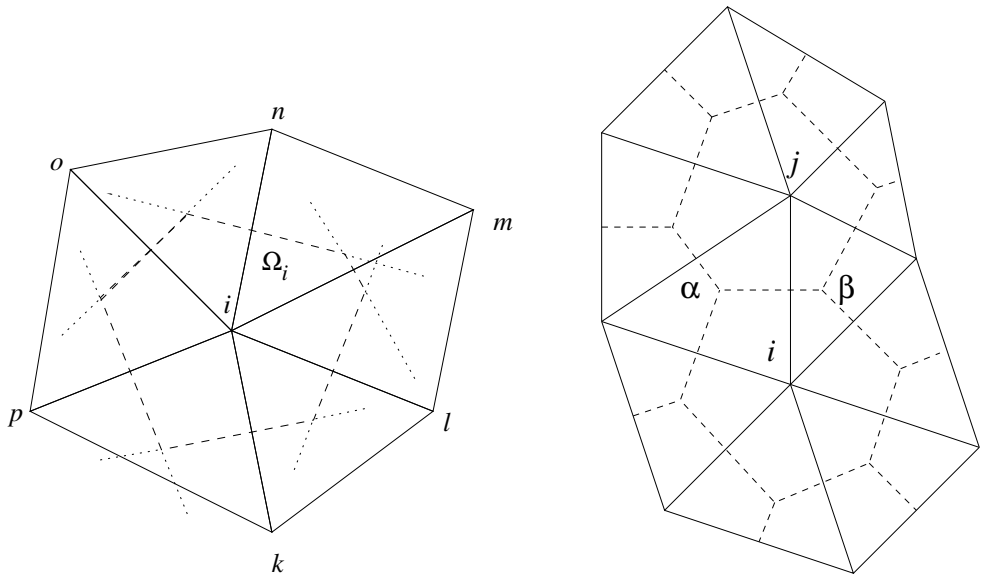


FIGURE 2.19: *Nodal patch and corresponding Voronoi control volume in Delaunay triangulation (left). A subset of a Delaunay triangulation and Voronoi cells (right).*

Remark 2.59. *The stiffness matrix A is often scaled with a constant of the order $\mathcal{O}(h^2)$, e.g., every row i is multiplied by the inverse of the area of the patch relative to node i. In this case the matrix has eigenvalues that are $\lambda_{min} = \mathcal{O}(1)$ and $\lambda_{max} = \mathcal{O}(h^{-2})$. It can be shown that these eigenvalues tend to the eigenvalues of the Laplace operator as $h \longrightarrow 0$. Note that the eigenvalues of the Laplace operator are all localized in the unbounded interval $[\Lambda, \infty)$, $\Lambda > 0$.*

Remark 2.60. *Recall that the conjugate gradient (CG) algorithm for the solution of a linear system converges with a number of iterations that is proportional to the square root of the spectral condition number of the system matrix. Thus, the number of iterations for CG when solving a diffusion equation on a sequence of refinements increases linearly with h. For example, halving h at each refinement step means that the number of iterations for convergence of CG (or PCG) doubles every time. This is a typical phenomenon common to all "elliptic" problems.*

2.11 Galerkin $\mathcal{P}_1$ Finite Elements and Finite Volumes

In two dimensional triangulations the Galerkin $\mathcal{P}_1$ can be interpreted as finite volumes. To show this, we first note the following properties of the FEM basis functions and consequently of the FEM stiffness matrix. The Galerkin FEM basis function form a partition of unity. In

fact, since the basis functions satisfy the interpolation property (2.4) we have immediately:

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

From the definition of the stiffness matrix (2.69), we have:

$$a_{ij} = a\left(\phi_i, \phi_j\right), \qquad \sum_{j=1}^{M} a_{i,j} = a\left(\phi, \sum_{j=1}^{N} \phi_j\right) = 0, \qquad a_{ii} = -\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}. \tag{2.84}$$

From the above considerations, it follows immediately that the kernel of the Galerkin stiffness matrix (before Dirichlet boundary conditions have been imposed) has dimension 1 and is generated by the constant vector. Now, the i-th Galerkin equation is written as:

$$\sum_{j=1}^{N} a_{ij} u_j = b_i \qquad a_{ii} u_i + \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} u_j = b_i.$$

Using (2.84), we have immediately:

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}(u_j - u_i) = b_i, \qquad \text{or} \qquad \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} \mid \sigma_{ij} \mid \frac{u_j - u_i}{\mid \sigma_{ij} \mid} = b_i,$$

where $\mid \sigma_{ij} \mid$ is the length of edge σ_i with endpoints given by the nodes i and j. We can now interpret the equation on the right as the sum of the fluxes entering/exiting a nodal control volume Ω_i centered in i. In fact, the term $(u_j - u_i)/\mid \sigma_{ij} \mid$ is a first order finite difference approximation of the gradient along σ_{ij}. Conservation, or the divergence theorem, calls for the sum of the flux projected along the normal to the boundary of Ω_i. Thus we can identify Ω_i as the region surrounding node i that is bounded by the polyline formed by the normals passing through the midpoint of each edge σ_{ij}, the so called axes of the triangular elements (see Fig. 2.19). This region is called the Voronoi cell, the dual of a Delaunay triangulation, and is always convex as long as the triangulation is of Delaunay type [24]. Using this identification and assuming a constant diffusion coefficient, the normal flux on σ_{ij} can be approximated by the following:

$$q_{ij} = -D \mid \sigma_{\alpha\beta} \mid \frac{u_j - u_i}{\mid \sigma_{ij} \mid},$$

which implies that the stiffness matrix coefficient a_{ij} must be equal to:

$$a_{ij} = -D \frac{\mid \sigma_{\alpha\beta} \mid}{\mid \sigma_{ij} \mid}.$$

Indeed this is true for Galerkin $\mathcal{P}_1$ in two dimensional Delaunay triangulations and constant coefficients and can be proved using the properties of the classical scalar product [20]. Thus, the Galerkin $\mathcal{P}_1$ finite element approach can be interpreted as a finite volume method where the interpolation of the unknown is performed on the triangles, while Gauss' theorem is applied on the dual Voronoi control volumes. This approach is known as "Control Volume Finite Element" method [15], an approach often used in multiphase simulation in porous media. We will see in later chapters that classical finite volume methods rely on performing both of the above tasks on the same control volume, and are less susceptible to ill-conditioning in case of highly deformed elements linked to the difficulty in the precise evaluation of the geometrical quantities of the control volumes.

3 Mixed formulation for elliptic equations

3.1 Equations in mixed form

We start this section with two important examples. The first concerns Stokes equation, the second concerns Darcy's equation. Both are important models in the application of computational fluid dynamics.

Example 3.1 (Stokes equations)**.** The stationary Stokes equations for an incompressible Newtonian fluid are a linear approximation of the corresponding Navier-Stokes equations that is valid for small Reynolds numbers (tending to zero). They are:

$$\begin{aligned} -\mu\Delta u + \nabla p &= f && \text{in } \Omega, \\ \operatorname{div} u &= 0 && \text{in } \Omega, \\ u &= 0 && \text{in } \Gamma, \end{aligned} \tag{3.1}$$

where $\mu > 0$ is the fluid dynamic viscosity, $u \in \mathbb{R}^d$ (d =1,2, or 3) is the fluid velocity, the forcing function is a vector function $f \in \mathbb{R}^d$, and the Laplace operator Δ is applied to the vector u component by component. A variational formulation can be derived as follows. We choose to use vector test functions $v \in [\mathcal{H}_0^1(\Omega)]^d$ that satisfy the further condition of being divergence free, i.e., $\div v = 0$. After scalar multiplication by v and application of Green's Lemma, we obtain:

$$\int_\Omega f \cdot v \, dx = \mu \int_\Omega \nabla u : \nabla v \, dx - \int_\Gamma (\nabla u \cdot n) \cdot v \, ds + \int_\Gamma p v \cdot n \, ds - \int_\Omega p \operatorname{div} v \, dx,$$

where the "double dot product" (or dyadic product) ":" symbol is defined as $\nabla u : \nabla v = \operatorname{Tr}\left((\nabla u)(\nabla v)^T\right)$ where ∇, u, and v identified as column vectors. This equation can be equivalently obtained using Einstein notation whereby repeated indices are summed over the spatial dimension d, as:

$$\int_\Omega f_i v_i \, dx = \mu \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \int_\Gamma (\nabla u \cdot n)_i v_i \, ds + \int_\Gamma p v_i n_i \, ds - \int_\Omega p v_{i,i} \, dx;$$

where we have used Einstein convention whereby repeated indices denote summation over the spatial dimension d. Since $v = 0$ on Γ and $\operatorname{div} v = v_{i,i} = 0$ in Ω. we can write the equation in a more compact form:

$$\mu \int_\Omega \nabla u : \nabla v \, dx = \int_\Omega f \cdot v \, dx \qquad \forall v \in \mathcal{V}(\Omega),$$

where $\mathcal{V}(\Omega) = \left\{ v \in [\mathcal{H}_0^1(\Omega)]^d : \operatorname{div} v = 0 \right\}$. The variational formulation becomes then: find $u \in \mathcal{V}$ such that:

$$a(u, v) = F(v) \qquad \forall v \in \mathcal{V}, \tag{3.2}$$

where:

$$a(v, w) = \mu \int_\Omega \nabla u : \nabla v \, dx, \qquad F(v) = \int_\Omega f \cdot v \, dx.$$

We note that there is no variational equation for the pressure, as we work in a divergence free space, and it seems that there is something we are missing for a sound development.
The numerical formulation is developed as usual by finding a finite dimensional subset of the variational space. To do this, we analyze with more details the space $\mathcal{V}$ in a two-dimensional domain ($d = 2$). In this case this space can be written as:

$$\mathcal{V}(\Omega) = \left\{ v = (v_1, v_2) \in \left[\mathcal{H}_0^1(\Omega)\right]^2 : \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} = 0 \text{ in } \Omega \right\},$$

with $\Omega \subset \mathbb{R}^2$. If $\Omega \subset \mathbb{R}^2$ is simply connected, then $\operatorname{div} v = 0$ if and only if there exists a so called "stream" function $\varphi \in H_0^2(\Omega)$ defining the vector potential $\boldsymbol{\varphi} = (0, 0, \varphi)^T$, so that:

$$v = \operatorname{curl} \boldsymbol{\varphi} = \nabla \times \boldsymbol{\varphi} = \left(\frac{\partial \varphi}{\partial x_2}, -\frac{\partial \varphi}{\partial x_1} \right).$$

The functions φ are of class $\mathcal{C}^1(\Omega)$. We call our discrete subspace as $\mathcal{W}_h \subset \mathcal{H}_0^2(\Omega)$ and see that the basis functions need to be polynomials of degree 5 $\varphi \in \mathcal{P}_5(T)$, that are determined by the following conditions (we denote with $\xi^{(i)}$ the coordinates of the three vertices of the triangle T and with $\xi^{(ij)}$ the coordinate of the midpoint of the edge σ_{ij} having endpoints given by vertices i and j):

$$\begin{aligned} &D^\alpha \varphi(\xi^{(i)}), \qquad i = 1, 2, 3; \mid \alpha \mid \leq 2 \\ &\frac{\partial \varphi}{\partial n}(\xi^{(ij)}), \qquad i, j = 1, 2, 3; i < j. \end{aligned}$$

Our FEM space is then:

$$\mathcal{V}_h = \{ v : v = \operatorname{curl} \boldsymbol{\varphi}, \varphi \in \mathcal{W}_h \},$$

and our numerical scheme is obtained by substituting this space in place of $\mathcal{V}$ in eq. (3.2). Intuitively, assuming optimal order of convergence, we will have the following error estimate:

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq Ch^4 |\ u\ |_{H^5(\Omega)}.$$

As we can see from the previous example, looking for basis function in a divergence free space $\mathcal{V}_h$ (i.e., functions that satisfy the incompressibility condition ($\operatorname{div} v = 0$) introduces limiting constraints already for $d = 2$ and that become even more stringent in a three-dimensional setting. It is then useful to consider the "mixed" form of the equation using the explicit unknowns of the problem, i.e., fully utilizing the velocity u and the pressure p as unknowns. Note that pressure is defined up to a constant given that only the gradient of p is present in our equations, thus we add the condition on the average pressure:

$$\int_\Omega p\, dx = 0.$$

Before addressing the problem of the Stokes equation, we look at a simpler example, namely the "mixed" form of the diffusion equation:

$$-\operatorname{div}(K(x)\nabla p) = f \qquad \text{in } \Omega \tag{3.3}$$

$$p = 0 \qquad \text{in } \Gamma = \partial\Omega, \tag{3.4}$$

where the diffusion coefficient $a(x)$ is bounded from above and below by positive constants and $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. This equation, for example, governs the motion of a fluid in laminar conditions, thus we talk about fluid pressure p and fluid velocity $u = -K(x)\nabla p$. The idea of the mixed formulation is to simultaneously approximate both p and u hoping to obtain properties of (u, p) that are more easily transferred to the discrete setting. Writing $\mu(x) = (K(x))^{-1}$, the problem is transformed into:

$$\mu u + \nabla p = 0 \qquad \text{in } \Omega, \tag{3.5}$$

$$\operatorname{div} u = f \qquad \text{in } \Omega, \tag{3.6}$$

$$p = 0 \qquad \text{in } \Gamma, \tag{3.7}$$

where the similarity with the Stokes problem is self-evident. We use vector and scalar test functions for the first and second equation, respectively, and apply Green's lemma to the second term of the first equation to obtain the following variational formulation[5]:

[5]This is the so called Dual Mixed Formulation and is the most widely used in applications. The Primal Mixed Formulation is obtained by applying Green's lemma to the second equation (for more details see [4]).

Problem 3.2 (Mixed Dual formulation)**.** Find $(u, p) \in \mathcal{V}(\Omega) \times \mathcal{Q}(\Omega)$ such that:

$$\begin{aligned} \int_\Omega \mu u \cdot v \, dx - \int_\Omega p \operatorname{div} v \, dx = 0 \qquad & \forall v \in \mathcal{V}(\Omega), \\ \int_\Omega q \operatorname{div} u \, dx = \int_\Omega f q \, dx \qquad & \forall q \in \mathcal{Q}(\Omega), \end{aligned}$$

We clearly have that $\mathcal{Q} = \mathcal{L}^2(\Omega)$. The function space $\mathcal{V}$ contains vector functions that are in $\mathcal{H}^1(\Omega)$ and with divergence in $\mathcal{L}^2$. In other words, $\mathcal{V}(\Omega) = \mathcal{H}(\operatorname{div}, \Omega) = \mathcal{H}_{\operatorname{div}}(\Omega)$ is the Hilbert space given by vector functions that admit divergence in $\mathcal{L}^2$, i.e.:

$$\mathcal{H}_{\operatorname{div}}(\Omega) = \left\{ v \in \left[\mathcal{L}^2\right]^d : \operatorname{div} v \in \mathcal{L}^2(\Omega) \right\}.$$

The norm defined as:

$$\|v\|^2_{\mathcal{H}_{\operatorname{div}}\Omega)} = \|v\|^2_{\mathcal{L}^2(\Omega)} + \|\operatorname{div} v\|^2_{\mathcal{L}^2(\Omega)}.$$

Obviously, we have that $[\mathcal{H}^1(\Omega)]^d \subset \mathcal{H}_{\operatorname{div}}(\Omega)$ and $\|v\|_{\mathcal{H}_{\operatorname{div}}} \leq m \|v\|_{\mathcal{H}^1}$. In fact, since $\forall v \in \mathcal{H}^{\operatorname{div}}(\Omega)$:

$$\|v\|^2_{\mathcal{H}^1} = \int_\Omega v \cdot v + \nabla v : \nabla v \, dx, \qquad \|v\|^2_{\mathcal{H}_{\operatorname{div}}} = \int_\Omega v \cdot v + (\operatorname{div} v)^2 \, dx,$$

and since for $d = 2$:

$$(\operatorname{div} v)^2 = \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}\right)^2 \leq 2 \left[\left(\frac{\partial v_1}{\partial x_1}\right)^2 + \left(\frac{\partial v_2}{\partial x_2}\right)^2 \right] \leq 2(\nabla v : \nabla v),$$

and for $d = 3$:

$$(\operatorname{div} v)^2 = \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right)^2 \leq 3 \left[\left(\frac{\partial v_1}{\partial x_1}\right)^2 + \left(\frac{\partial v_2}{\partial x_2}\right)^2 + \left(\frac{\partial v_3}{\partial x_3}\right)^2 \right] \leq 3(\nabla v : \nabla v),$$

we have that:

$$\int_\Omega v \cdot v + (\operatorname{div} v)^2 \, dx \leq 3 \int_\Omega v \cdot v + \nabla v : \nabla v \, dx.$$

Remark 3.3. *We observe that in this Dual Mixed formulation homogeneous Dirichlet boundary conditions are "natural" boundary conditions, i.e., they are contained naturally within the weak formulation, as opposed to the standard formulation for which "natural" boundary conditions are of homogeneous Neumann type.*

We have the following theorem due to Brezzi:

Theorem 3.1 (Brezzi splitting theorem)**.** *If $a : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ is continuous and coercive and $b : \mathcal{V} \times \mathcal{Q} \mapsto \mathbb{R}$ is continuous and satisfies the inf-sup condition:*

$$\inf_{q \in \mathcal{Q}} \sup_{v \in \mathcal{V}} \frac{b(v,q)}{\|v\|_{\mathcal{V}} \|q\|_{\mathcal{Q}}} \geq \beta, \tag{3.8}$$

then the problem 3.2 has a unique solution $(u,p) \in \mathcal{V} \times \mathcal{Q}$ that satisfies the following stability property:

$$\|u\|_{\mathcal{V}} + \|p\|_{\mathcal{Q}} \leq C \|f\|_{\mathcal{Q}}.$$

Note that the coercivity assumption on $a(\cdot,\cdot)$ can be relaxed, as will be seen later in § 3.3.1. The two bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are obviously bounded, and thus continuous:

$$| a(v,w) | \leq \|K\|_{\infty} \|v\|_{\mathcal{V}} \|w\|_{\mathcal{V}} \qquad | b(v,q) | \leq \|v\|_{\mathcal{V}} \|q\|_{\mathcal{Q}},$$

where $\mathcal{V} = \mathcal{H}_{\text{div}}(\Omega)$ and $\mathcal{Q} = \mathcal{L}^2(\Omega)$. The kernel of the bilinear form $b(\cdot,\cdot)$, or better the kernel of the operator B associated with the bilinear form, is given by:

$$\operatorname{Ker} b = \left\{ v \in \mathcal{H}_{\text{div}}(\Omega) : (\operatorname{div} v, q) = 0 \text{ for all } q \in \mathcal{L}^2(\Omega) \right\},$$

and thus $\|\operatorname{div} v\|^2_{\mathcal{L}^2(\Omega)} = 0$ for all $v \in \operatorname{Ker} b \subset \mathcal{H}_{\text{div}}(\Omega)$. Hence:

$$a(v,v) = \|v\|^2_{\mathcal{L}^2(\Omega)} = \|v\|^2_{\mathcal{H}_{\text{div}}(\Omega)},$$

showing the coercivity of the $a(\cdot,\cdot)$ bilinear form. For the other form, we have that for any there exists a function $v_q \in [\mathcal{H}^1(\Omega)]^d$ such that for every $q \in \mathcal{L}^2(\Omega)$ we have that $\operatorname{div} v_q = q$ and $\|v\|_{\mathcal{H}^1} \leq C \|q\|_{\mathcal{L}^2}$. Thus we have, for all $q \in \mathcal{Q} = \mathcal{L}^2(\Omega)$ and $v \in \mathcal{V} = \mathcal{H}_{\text{div}}(\Omega)$ (taking $\operatorname{div} v_q = q$):

$$\sup_{q \in \mathcal{Q}} \frac{| b(v,q) |}{\|v\|_{\mathcal{V}}} \geq \sup_{q \in \mathcal{Q}} \frac{| b(v_q,q) |}{\|v\|_{\mathcal{V}}} \geq \sup_{q \in \mathcal{Q}} \frac{(q,q)}{C \|v\|_{\mathcal{L}^2}} = \frac{1}{C} \|q\|_{\mathcal{L}^2}.$$

Thus the problem admits a unique solution $(u,p) \in \mathcal{V} \times \mathcal{Q}$ with the stability estimate:

$$\|u\|_{\mathcal{H}_{\text{div}}} + \|p\|_{\mathcal{L}^2} \leq C \|f\|_{\mathcal{L}^2}.$$

We note here t,hat although the above estimate for p is in $\mathcal{L}^2(\Omega)$, it is easy to see, using again integration by parts, that the pressure p has a weak derivative and satisfies the homogeneous Dirichlet boundary conditions, yielding the desired result that $p \in \mathcal{H}^1_0(\Omega)$.
Alternatively, the mixed formulation can be rewritten using a symmetric bilinear form as:

$$c((u,p),(v,q)) = \int_{\Omega} \mu u \cdot v \, dx - \int_{\Omega} p \operatorname{div} v \, dx - \int_{\Omega} q \operatorname{div} u \, dx.$$

and a linear form:

$$F\left((v,q)\right) = -\int_\Omega fq\,dx.$$

Note that the original two equations can be obtained using $(v,0)$ and $(0,q)$ in the second argument of the above bilinear form. Thus we have:

$$c\left((u,p),(v,q)\right) = F\left((v,q)\right) \qquad \forall (v,q) \in \mathcal{H}_{\text{div}}(\Omega) \times \mathcal{L}^2(\Omega).$$

However, the form $c\left(\cdot,\cdot\right)$ is not coercive, but it can be shown that is satisfies the "inf-sup" condition (2.61) and thus, because of its symmetry, also (2.62).

3.2 Mixed finite elements

Let $\mathcal{T}_h(\Omega)$ be a regular triangulation of Ω with grid parameter h. We want to build the FEM spaces:

$$\mathcal{V}_h \subset \mathcal{H}_{\text{div}}(\Omega) \text{ and } \mathcal{Q}_h \subset \mathcal{L}^2(\Omega).$$

It easy to see that the "inf-sup" condition required for stability introduces a relationships between $\mathcal{V}_h$ and $\mathcal{Q}_h$, in the sense that $\mathcal{V}_h$ must be "sufficiently richer" than $\mathcal{Q}_h$ (we will see this in more details in the next section). Then the mixed finite element method can be written directly as:

Problem 3.4 (Mixed FEM)**.** Find $(u_h, p_h) \in \mathcal{V}_h \times \mathcal{Q}_h$ such that:

$$\begin{aligned} a\left(u_h, v\right) - b\left(p_h, v\right) &= 0 \qquad \forall v \in \mathcal{V}_h, \\ b\left(q, u_h\right) &= F\left(q\right) \qquad \forall q \in \mathcal{Q}_h, \end{aligned} \tag{3.9}$$

where:

$$a\left(v,w\right) = \int_\Omega \mu v \cdot w\,dx \qquad b\left(v,q\right) = \int_\Omega q \operatorname{div} v \qquad F\left(q\right) = \int_\Omega fq\,dx.$$

Remark 3.5. *We note that using density arguments*[6]*, via integration by parts, one can show that functions with continuous normal derivatives on edges of $\mathcal{T}_h(\Omega)$ are in $\mathcal{H}_{\text{div}}(\Omega)$, i.e., given $T_i, T_j \in \mathcal{T}_h(\Omega)$ two neighboring elements of the computational mesh, with $\sigma_{ij} = T_i \cap T_j$, we have that:*

$$\left\{v \in [\mathcal{L}^2(\Omega)]^d : v|_{T_i} \in [\mathcal{H}^1(T_i)]^d \text{ and } v|_{T_i} \cdot \nu = v|_{T_j} \cdot \nu \text{ for all } \sigma_{ij} \in \mathcal{T}_h(\Omega)\right\} \subset \mathcal{H}_{\text{div}}(\Omega).$$

[6]Typical density arguments rely on the fact that $\mathcal{C}^\infty(\bar{\Omega})$ is dense in $\mathcal{L}^2(\Omega)$ and on the Sobolev (or Rellich-Kondrachov) embedding theorem, and can be used, e.g., to show that weak normal derivatives on element edges exist and that integration by parts formula hold (both globally on Ω and locally on $\mathcal{T}_h$) (see for example [7, Thm. 3, §5.3.3, §5.6, §5.7]).

Thus, we will chose vector basis functions for $\mathcal{H}_{\text{div}}(\Omega)$ that have continuous normal component. We will see later on a stronger result.

We have the following equivalent to the continuous case, which is easy to prove:

Theorem 3.2. *If the following discrete inf-sup conditions hold;*

$$\inf_{u_h \in \operatorname{Ker} b(\cdot,\cdot)} \sup_{v_h \in \operatorname{Ker} b(\cdot,\cdot)} \frac{a\left(u_h, v_h\right)}{\|u_h\|_{\mathcal{V}} \|v_h\|_{\mathcal{V}}} \geq \alpha_h,$$

and

$$\inf_{q_h \in \mathcal{Q}_h} \sup_{v_h \in \mathcal{V}_h} \frac{b\left(v_h, q_h\right)}{\|q_h\|_{\mathcal{Q}} \|v_h\|_{\mathcal{V}}} \geq \beta_h, \tag{3.10}$$

then there exists a unique solution $(u_h, p_u) \in \mathcal{V}_h \times \mathcal{Q}_h$ of Problem 3.4 and the pair satisfies the following stability estimate:

$$\|u_h\|_{\mathcal{V}} + \|p_h\|_{\mathcal{Q}} \leq C \|f\|_{\mathcal{V}^*}.$$

For a proof of this theorem see [4].
Note that although the continuous problem satisfies the inf-sup condition, the discrete inf-sup does not follows directly from the continuous counterpart. There is however a characterization that allows to identify appropriate FEM spaces $\mathcal{V}_h$ and $\mathcal{Q}_h$ that satisfy the above discrete inf-sup:

Lemma 3.3 (Fortin criterion)**.** *If the continuous inf-sup condition is satisfied (eq.* (3.8)*), then the discrete inf-sup condition* (3.10) *is satisfied is and only if there exists a linear operator $\Pi_h : \mathcal{V} \mapsto \mathcal{V}_h$ such that:*

$$b\left(\Pi_h v, q\right) = b\left(v, q\right) \qquad \forall q \in \mathcal{Q}_h,$$

and such that:

$$\|\Pi_h v\|_{\mathcal{V}} \leq \gamma_h \|v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}.$$

Proof. **Forward implication.** Assume such a Π_h exists. Then, since $\Pi_h(\mathcal{V}) \subset \mathcal{V}_h$, we have, for all $q \in \mathcal{Q}_h$:

$$\sup_{v \mathcal{V}_h} \frac{b\left(v, q\right)}{\|v\|_{\mathcal{V}}} \geq \sup_{v \mathcal{V}} \frac{b\left(\Pi_h v, q\right)}{\|\Pi_h v\|_{\mathcal{V}}} \geq \sup_{v \mathcal{V}} \frac{b\left(v, q\right)}{\gamma_h \|v\|_{\mathcal{V}}} \geq \frac{\beta}{\gamma_h} \|q\|_{\mathcal{Q}},$$

i.e., the discrete inf-sup condition (3.10).
Backward implication. Assume the discrete inf-sup condition (3.10) holds. This means that, given any $v \in \mathcal{V}$, the bilinear form $b\left(\cdot,\cdot\right)$, identified as an operator, is surjective and has continuous right inverse. Hence, there exist $v_h = \Pi_h v \in \mathcal{V}_h$ such that $b\left(\Pi_h v, q\right) = b\left(v, q\right)$ for every $q \in \mathcal{Q}_h$. The stability inequality follows directly. □

To obtain convergence estimates we assume some properties of the spaces $\mathcal{V}_h$ and $\mathcal{Q}_h$ so that the Fortin criterion is satisfied:

$$\operatorname{div} \mathcal{V}_h = \mathcal{Q}_h, \tag{3.11}$$

and that there exists a projection operator $\Pi_h : [\mathcal{H}^1(\Omega)]^d \longrightarrow \mathcal{V}_h$ such that:

$$\int_\Omega \operatorname{div}(u - \Pi_h u) q = 0 \qquad \forall u \in \left[\mathcal{H}^1(\Omega)\right]^d, \forall q \in \mathcal{Q}_h. \tag{3.12}$$

Under these hypothesis we can show uniqueness of the solution of the discrete system. Thus we prove that $f = 0$ implies $(u_h, p_h) = (0, 0)$. For $f = 0$ the linear system becomes:

$$\begin{aligned} \int_\Omega \mu u_h \cdot v \, dx - \int_\Omega p_h \operatorname{div} v \, dx = 0 \qquad & \forall v \in \mathcal{V}_h \\ \int_\Omega q \operatorname{div} u_h \, dx = 0 \qquad & \forall q \in \mathcal{Q}_h \end{aligned}$$

Since $\operatorname{div} \mathcal{V}_h \subset \mathcal{Q}_h$ and $u_h \in \mathcal{V}_h$, we can take $q = \operatorname{div} u_h \in \mathcal{Q}_h$ in the second equation to obtain $\operatorname{div} u_h = 0$. Then taking $v = u_h$ in the first equation yields immediately $u_h = 0$. But since $\operatorname{div} \mathcal{V}_h \supset \mathcal{Q}_h$ and $p_h \in \mathcal{Q}_h$, we can choose $v \in \mathcal{V}_h$ such that $\operatorname{div} v = p_h$, from which $p_h = 0$.
The error estimate is given by:

Theorem 3.4. *Let $\mathcal{V}_h$ and $\mathcal{Q}_h$ be mixed FEM spaces, satisfying* (3.11) *and let Π_h be the projector defined in* (3.12)*. Then there exist a constant C independent of h such that:*

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq C \left\{ \|u - \Pi_h u\|_{\mathcal{L}^2(\Omega)} \right\}.$$

Proof. The error equation is easily recovered by subtraction:

$$a\left((u - u_h), v\right) - b\left((p - p_h), v\right) = 0 \qquad \forall v \in \mathcal{V}_h, \tag{3.13}$$

$$b\left(q, (u - u_h)\right) = 0 \qquad \forall q \in \mathcal{Q}_h. \tag{3.14}$$

Using (3.12), we can write the last equation as:

$$b\left(q, (\Pi_h u - u_h)\right) = 0 \qquad \forall q \in \mathcal{Q}_h.$$

Take $q = \operatorname{div}(\Pi_h u - u_h)$ to write:

$$\operatorname{div}(\Pi_h u - u_h) = 0.$$

Now take $v = \Pi_h u - u_h$ in (3.13) to obtain:

$$a\left((u - u_h), (\Pi_h u - u_h)\right) = \int_\Omega \mu (u - u_h) \cdot (\Pi_h u - u_h) \, dx = 0. \tag{3.15}$$

We can write:

$$\begin{aligned}\|u-u_h\|^2_{\mathcal{L}^2(\Omega)} &= \int_\Omega (u-u_h)\cdot(u-u_h)\,dx \le \|K\|_\infty \int_\Omega \mu(u-u_h)\cdot(u-u_h)\,dx \\ &= \|K\|_\infty \left[\int_\Omega \mu(u-u_h)\cdot(u-u_h)\,dx + \int_\Omega \mu(u-u_h)\cdot(\Pi_h u-u_h)\,dx\right] \\ &\le \|K\|_\infty a\left((u-u_h),(\Pi_h u-u_h)\right) \\ &\le \|K\|_\infty \|\mu\|_\infty \|u-u_h\|_{\mathcal{L}^2(\Omega)} \|\Pi_h u-u_h\|_{\mathcal{L}^2(\Omega)}.\end{aligned}$$

The final result follows by dividing by the nonzero term $\|u-u_h\|_{\mathcal{L}^2(\Omega)}$. □

The error estimate of the mixed FEM relies upon interpolation error estimates for scalar and vector functions. Thus we assume the following (interpolation) estimates:

$$\|p\|_{\mathcal{H}^2(\Omega)} \le C\|f\|_{\mathcal{L}^2(\Omega)} \tag{3.16}$$

$$\|q-\pi_h q\|_{\mathcal{L}^2(\Omega)} \le Ch\|q\|_{\mathcal{H}^1(\Omega)} \qquad \forall q\in\mathcal{H}^1(\Omega); \tag{3.17}$$

$$\|v-\Pi_h v\|_{\mathcal{L}^2(\Omega)} \le Ch\|v\|_{\mathcal{H}^1(\Omega)} \qquad \forall v\in\left[\mathcal{H}^1(\Omega)\right]^d; \tag{3.18}$$

$$\|\Pi_h v\|_{\mathcal{L}^2(\Omega)} \le C\|v\|_{\mathcal{H}^1(\Omega)}. \tag{3.19}$$

Then we can prove the following:

Theorem 3.5. *Let $\mathcal{V}_h$ and $\mathcal{Q}_h$ be mixed FEM spaces satisfying* (3.11) *and* (3.12). *Let Π_h be a projector satisfying* (3.19). *Then there exists a constant C independent of h such that:*

$$\|\pi_h p - p_h\|_{\mathcal{L}^2(\Omega)} \le C\|u-\Pi_h u\|_{\mathcal{L}^2(\Omega)}.$$

Proof. Note that (3.12) together with (3.19) implies that for each $q\in\mathcal{Q}_h$ exists a $v\in\mathcal{V}_h$ such that $\operatorname{div} v = q$ and $\|v\|_{\mathcal{L}^2(\Omega)} \le C\|q\|_{\mathcal{L}^2(\Omega)}$. In fact, we can use the auxiliary problem:

$$\begin{aligned}\Delta\varphi &= q && \in\Omega,\\ \varphi &= 0 && \in\partial\Omega,\end{aligned}$$

and define $w=\nabla\varphi$. From (3.16) we have $\|w\|_{\mathcal{H}^1(\Omega)} \le C\|f\|_{\mathcal{L}^2(\Omega)}$. Then, using (3.11) and (3.19) we see that the function $v=\Pi_h w$ satisfies the requested conditions.
We first note that for all $v\in\mathcal{V}_h$ eq. (3.11) ensures that $(\pi_h p, q) = (p,q)$ for all $q\in\mathcal{Q}_h$ and $\operatorname{div} v\in\mathcal{Q}_h$. From the error equation we then have:

$$\int_\Omega (p-p_h)\operatorname{div} v\,dx = \int_\Omega (\pi_{h,p}-p_h)\operatorname{div} v\,dx = \int_\Omega (\pi_h p - p_h)\operatorname{div} v\,dx = \int_\Omega (u-u_h)\cdot v\,dx;$$

Take $v\in\mathcal{V}_h$ such $\operatorname{div} v = (\pi_h p - p_h)$ and that verifies:

$$\|v\|_{\mathcal{L}^2(\Omega)} \le C\|\pi_h p - p_h\|_{\mathcal{L}^2(\Omega)}.$$

We obtain:

$$\|\pi_h p - p_h\|^2_{\mathcal{L}^2(\Omega)} \leq C \|u - u_h\|_{\mathcal{L}^2(\Omega)} \|\pi_h p - p_h\|_{\mathcal{L}^2(\Omega)} .$$

The proof is concluded by invoking Theorem 3.4 and the triangular inequality. □

Theorem 3.6. *Let $\mathcal{V}_h$ and $\mathcal{Q}_h$ be mixed FEM spaces satisfying* (3.11) *and* (3.12)*. Let Π_h be a projector satisfying* (3.19)*. Then there exists a constant C independent of h such that:*

$$\|\pi_h p - p_h\|_{\mathcal{L}^2(\Omega)} \leq C \left\{ h \|u - u_h\|_{\mathcal{L}^2(\Omega)} + h^2 \|\mathrm{div}(u - u_h)\|_{\mathcal{L}^2(\Omega)} + \right\}$$

Proof. We need to use a duality argument similar to the Aubin-Nitsche trick. Let ϕ be solution of the dual problem:

$$\begin{aligned} \mathrm{div}(K\nabla\phi) &= \pi_h p - p_h && \text{in } \Omega \\ \phi &= 0 && \text{in } \partial\Omega. \end{aligned}$$

Then, assuming $K(x)$ sufficiently regular and using (3.16)-(3.19), we can write:

$$\begin{aligned} \|\pi_h p - p_h\|^2_{\mathcal{L}^2(\Omega)} &= \int_\Omega (\pi_h p - p_h)\, \mathrm{div}(K\nabla\phi)\, dx \\ &= \int_\Omega (\pi_h p - p_h)\, \mathrm{div}(\Pi_h(K\nabla\phi))\, dx \\ &= \int_\Omega (p - p_h)\, \mathrm{div}(\Pi_h(K\nabla\phi))\, dx \\ &= \int_\Omega \mu(u - u_h)(\Pi_h(K\nabla\phi) - K\nabla\phi)\, dx + \int_\Omega (u - u_h)\nabla\phi)\, dx \\ &= \int_\Omega \mu(u - u_h)(\Pi_h(K\nabla\phi) - K\nabla\phi)\, dx - \int_\Omega \mathrm{div}(u - u_h)(\phi - \pi_h\phi)\, dx \\ &\leq C \|u - u_h\|_{\mathcal{L}^2(\Omega)} h \|\phi\|_{\mathcal{H}^2(\Omega)} + C \|\mathrm{div}(u - u_h)\|_{\mathcal{L}^2(\Omega)} h^2 \|\phi\|_{\mathcal{H}^2(\Omega)} \end{aligned}$$

□

These theorems, together with (3.17) and (3.18), tells us that the Mixed Finite Element scheme converges linearly for p_h and u_h as long as the real solution is sufficiently regular. In practice, there are "super-convergence" theorems that show that the pressure p_h converges super-linearly in specific points of the triangles for all those cases in which (3.17) can be written with an exponent of h larger than one.

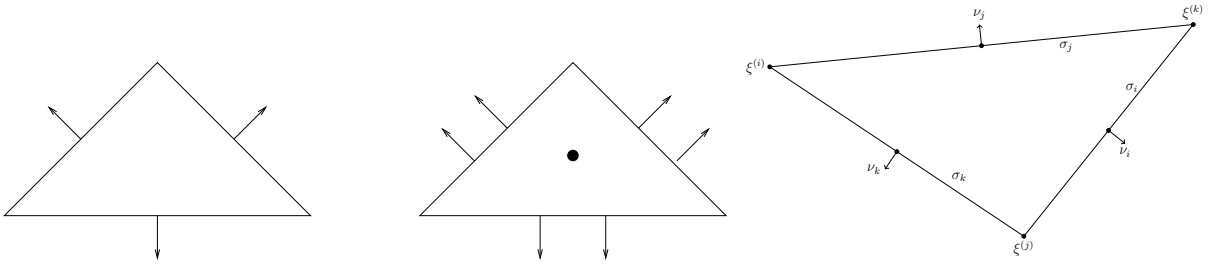


FIGURE 3.1: *Location of the "degrees of freedom" in the triangle for the spaces $\mathcal{RT}_0$ (left) and $\mathcal{RT}_1$ (right). The arrows indicate the normal component $v \cdot n$ of the vector basis functions, while the dot indicates the function v (left and middle). Triangular element with notation for nodes, edges, and exterior normals (right).*

3.2.1 Raviart-Thomas $\mathcal{RT}_k$ finite dimensional spaces

To complete the numerical formulation we need to characterize the MFEM spaces $\mathcal{V}_h$ and $\mathcal{Q}_h$ that satisfy the requested properties. We recall that $\mathcal{V}_h$ is a subspace of $\mathcal{H}_{\text{div}}(\Omega)$ and that we will need to find an appropriate projector Π_h. We consider in these notes only the case of a two-dimensional domain ($d = 2$) and a regular triangulation $\mathcal{T}_h = \{T\}$. Moreover we consider here only Raviart-Thomas spaces of degree k ($\mathcal{RT}_k$), referring the reader to, e.g., [4] for further details.
We define the following family of $\mathcal{RT}$ spaces on the triangle $T \in \mathcal{T}_h$:

$$\mathcal{RT}_k(T) = [\mathcal{P}_k]^2 \oplus x\mathcal{P}_k \tag{3.20}$$

where $x \in \mathbb{R}^d$. The finite dimensional space $\mathcal{V}_h$ can be described as:

$$\mathcal{V}_h = \{v \in \mathcal{H}_{\text{div}}(\Omega) : v|_T \in \mathcal{RT}_k(T) \ \ \forall T \in \mathcal{T}_h\}\,.$$

As done before, we denote with $\xi^{(i)}$ both the i-th node of T and its coordinate vector, with σ_i the edge (face for $d = 3$) opposite to the i-th node and with ν_i the corresponding edge outer unit normal (see Figure 3.1, right). For now, we work in a local (triangle based) enumeration $i = 1, 2, 3$. We have the following:

Lemma 3.7. *The family of mixed finite element spaces $\mathcal{RT}_k$ has the properties:*

1. $\dim \mathcal{RT}_k(T) = (k+1)(k+3)$;
2. *if $v \in \mathcal{RT}_k(T)$, then $v \cdot \nu_i \in \mathcal{P}_k(\sigma_i)$;*
3. *if $\operatorname{div} v = 0$, $v \in \mathcal{RT}_k(T)$, then $v \in [\mathcal{P}_k(T))^n$.*

For the space $\mathcal{Q}_h \subset \mathcal{L}^2(\Omega)$, there are no special regularity requirements::

$$\mathcal{Q}_h = \left\{q \in \mathcal{L}^2(\Omega) : q|_T \in \mathcal{P}_k(T) \ \ \forall T \in \mathcal{T}_h\right\}.$$

We are missing only the construction of the projection operator Π_h. We first note that a vector function, whose components are elementwise continuous polynomials and with trace (assuming enough regularity for its existence) having continuous normal projection along subdomain edges, belongs to $\mathcal{H}_{\text{div}}(\Omega)$. More precisely, we have the following:

Lemma 3.8. *Given an arbitrary partition $\mathbb{P}(\Omega)$ of the domain Ω into subdomains $\tilde{\Omega}_k$ with sufficiently regular boundary, i.e., $\mathbb{P}(\Omega) = \cup_k \tilde{\Omega}_k$, and let $\mathcal{S}(\Omega)$ be the space of piecewise $\mathcal{H}_{\text{div}}$ functions, i.e., $\mathcal{S}(\mathbb{P}) = \{v \in [\mathcal{L}^2(\Omega)]^d : v|_{\tilde{\Omega}_k} \in \mathcal{H}_{\text{div}}(\tilde{\Omega}_k)\ \forall \tilde{\Omega}_k \in \mathbb{P}(\Omega)\}$. Then, $v \in \mathcal{S}(\mathbb{P})$ belongs to $\mathcal{H}_{\text{div}}(\Omega)$ if and only if its normal traces are continuous across all subdomain interfaces.*

Proof. Let Γ_k be portion of the boundary of subdomain $\tilde{\Omega}_k$ that does not lie in $\partial\Omega$. By the divergence theorem we can write:

$$\int_\Omega v\ dx = \int_{\partial\Omega} v \cdot n\ ds = \int_{\mathbb{P}(\Omega)} v\ dx = \sum_k \int_{\tilde{\Omega}_k} v\ dx = \sum_k \int_{\Gamma_k} v \cdot n_\Gamma\ ds + \int_{\partial\Omega} v \cdot n\ ds,$$

where n_Γ is the outward unit normal on Γ_k. Thus we have:

$$\sum_k \int_{\Gamma_k} v \cdot n_\Gamma\ ds = 0.$$

The inverse implication is obvious. □

The definition of the basis function can then rely on degrees of freedom associated to $k+1$ points on each edge of T. For example, for $k = 0$ in two dimensional triangulations we take one point per edge (identified generally with the midpoint) where we impose the continuity of the normal component of the basis functions of the two neighboring elements. In addition we take the central point of the triangle for the function v (see Figure 3.1). Then we have the following:

Lemma 3.9. *Given a triangle $T \in \mathcal{T}_h$ and a vector function $v \in [\mathcal{H}^1(T)]^2$, there is a unique operator $\Pi_h v \in \mathcal{RT}_k(T)$ such that:*

$$\int_{\sigma_i} \Pi_h v \cdot \nu_i p_k\ ds = \int_{\sigma_i} v \cdot \nu_i p_k\ ds \qquad \forall p_k \in \mathcal{P}_k(T), i = 1, 2, 3,$$

and:

$$\int_T \Pi_T v \cdot p_{k-1}\ dx = \int_T v \cdot p_{k-1}\ dx \qquad \forall p_{k-1} \in [\mathcal{P}_{k-1}(T)]^2.$$

Now convergence can be shown using similar reasoning as done for Galerkin methods using appropriate transformations of the triangle into a reference triangle $\tilde{T}$ having vertex coordinates $\xi^{(1)} = (0, 0)$, $\xi^{(2)} = (1, 0)$, $\xi^{(3)} = (0, 1)$. The transformation must conserve the properties

in particular of the vector interpolation. This achieved by means of the (contravariant) Piola transform, which is defined as follows. Given the (affine) map F that transforms triangle $\tilde{T}$ into triangle T, we define $\tilde{v} \in [\mathcal{L}^2(\tilde{T})]^2$ via:

$$v(x) = \frac{1}{\mid \det J(\tilde{x}) \mid} J(\tilde{x})\tilde{v}(\tilde{x}),$$

where $x = F(\tilde{x})$ e $J(\tilde{x})$ is the Jacobian of F. Then we have:

Lemma 3.10. *There exists a constant $C > 0$ such that for all $v \in [\mathcal{H}^m(T)]$ with $1 \leq m \leq k+1$:*

$$\|v - \Pi_T v\|_{\mathcal{L}^2(T)} \leq C h_T^m \|v\|_{\mathcal{H}^m(T)}.$$

Using this lemma and the interpolation estimates, we obtain:

Theorem 3.11. *Let $\{\mathcal{T}_h\}$ be a family of regular triangulations and given the functions $u \in \left[\mathcal{H}^{k+1}(\Omega)\right]^2$ and $p \in \mathcal{H}^{k+1}(\Omega)$, then the numerical solution $(u_h, p_h) \in \mathcal{V}_h \times \mathcal{Q}_h$ obtained with the mixed finite element method satisfies:*

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq C h^{k+1} \|u\|_{\mathcal{H}^{k_1}(\Omega)},$$

and:

$$\|p - p_h\|_{\mathcal{L}^2(\Omega)} \leq C h^{k+1} \left[\|u\|_{\mathcal{H}^{k_1}(\Omega)} + \|p\|_{\mathcal{H}^{k_1}(\Omega)}\right].$$

3.2.2 Practical implementation of $\mathcal{RT}_0 - \mathcal{P}_0$ MFEM on triangles

As we have seen, mixed finite elements need information on the boundaries of the elements, e.g., edges in two dimensional triangulations. In this section we look at a system for efficiently defining this information on a regular triangulation. Consider a polygonal domain $\Omega \subset \mathbb{R}^2$ and let $\mathcal{T}_h(\Omega)$ be a regular triangulation of Ω. An example of the application of our data structure to a specific triangulation is given in Figure 3.2. We start with the evaluation of the $\mathcal{RT}_0$ basis functions and proceed next with the calculation of the elemental matrices to conclude with some information of how assembly of these local matrices into the global system matrices is performed.

$\mathcal{RT}_0$ basis functions. We first show some properties of the vector basis functions that are used in the lowest order Mixed FEM space. Recall that $\mathcal{RT}_0 \subset \mathcal{H}_{\text{div}}$. Thus, because of Lemma 3.8, we need to have basis functions with continuous normal trace. From (3.20), $\mathcal{RT}_0$ basis functions are of the form:

$$w_m(x) = \begin{bmatrix} a_m x + b_m \\ a_m y + c_m \end{bmatrix},$$

Edge connectivity		
f_k	v_1	v_2
1	2	1
2	3	2
3	1	5
4	6	3
5	5	8
6	9	6
7	8	10
8	10	9
9	1	4
10	4	2
11	4	3
12	5	4
13	4	6
14	5	7
15	7	4
16	7	6
17	8	7
18	7	9
19	10	7

Edge-cell connectivity	
f_k	T^L	T^R
1	1	0
2	2	0
3	3	0
4	4	0
5	7	0
6	8	0
7	9	0
8	10	0
9	1	3
10	1	2
11	2	4
12	3	5
13	4	6
14	5	7
15	5	6
16	6	8
17	7	9
18	8	10
19	9	10

Cell connectivity			
T_i	v_1	v_2	v_3
1	1	4	2
2	2	4	3
3	1	5	4
4	3	4	6
5	5	7	4
6	7	6	4
7	5	8	7
8	6	7	9
9	8	10	7
10	7	10	9

Cell-Cell connectivity			
T_i	f_1	f_2	f_3
1	9	10	1
2	10	11	2
3	3	12	9
4	11	13	4
5	14	15	12
6	16	13	15
7	5	17	14
8	16	18	6
9	7	19	17
10	19	8	18

Vertex coordinates		
v_k	x_1	x_2
1	1.5	0.5
2	1.0	2.0
3	3.0	3.5
4	3.0	1.5
5	4.0	0.0
6	5.0	4.0
7	6.0	2.0
8	7.0	0.0
9	7.0	3.0
10	8.0	2.0

FIGURE 3.2: *Triangulation of a two-dimensional polygonal domain with cell, node, and edge global numbering, and the (redundant) data structures that completely characterize the geometry and topology of the mesh.*

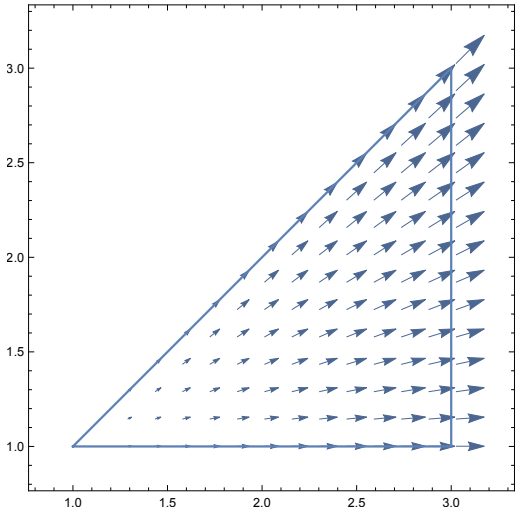


FIGURE 3.3: *Example of vector basis function for a rectangular triangle.*

where subscript m is used to index triangle edges, ($m = 1, 2, 3$ in a local numbering system, or i, j, k in a global numbering system). Let us focus on a general triangle $T \in \mathcal{T}_h(\Omega)$ with vertices labeled i, j, and k, and corresponding opposite edges σ_i, σ_j, and σ_k.
The unit outward normal is denoted by ν_k, while the edge normal is denoted by ν_{σ_k}. These two may differ by a sign depending on the triangulation node numbering. Using the data structure proposed in Figure 3.2, we use the convention that the direction of the edge normal is defined by the "Edge-cell connectivity" table according to the "Edge connectivity" node ordering. In practice, given triangle T and edge i, if the left triangle in the "Edge-cell' connectivity" table is element T itself, then the edge normal for edge i points inward, otherwise, it points outward (e.g., looking at element 1 and edge 1, the left triangle in the "Edge-cell connectivity" table is equal to 1, thus the normal points inward to element 1).
We observe that using Lemma 3.7 with $k = 0$ we have immediately that $\operatorname{div} w = const$ and $w \cdot \nu_k = const$. In fact we have the following:

Lemma 3.12. *Given the triangle T with nodes $\xi^{(i)}$, $\xi^{(j)}$, and $\xi^{(k)}$, the function $w_i(x) \in \mathcal{RT}_0(T)$ given by:*

$$w_i(x) = \langle \nu_{\sigma_i}, \nu_i \rangle \frac{|\ \sigma_i\ |}{2|\ T\ |}(x - \xi^{(i)})$$

has the following properties:

(i) $\left\langle w_i(x)|_{\sigma_i}, \nu_{\sigma_j} \right\rangle = \delta_{ij}$;

(ii) $\{w_i, w_j, w_k\} = \operatorname{Span}(\mathcal{RT}_0(T))$;

(iii) $\operatorname{div} w_i = 2a_i = \dfrac{|\ \sigma_i\ |}{|\ T\ |}$.

Proof. To prove the first property, we observe that for $i \neq j$ and $x \in \sigma_j$ we have:

$$(x - \xi^{(i)})|_{\sigma_j} \cdot \nu_j = 0,$$

since $\xi^{(i)} \in \sigma_j$. If $i = j$, then $\mid (x - \xi^{(j)}) \cdot \nu_{\sigma_j} \mid$ is the height of T passing through $\xi^{(j)}$. Thus, $\mid T \mid = \frac{1}{2}(x - \xi^{(j)}) \cdot \nu_j \left\langle \nu_j, \nu_{\sigma_j} \right\rangle \mid \sigma_j \mid$, since $(x - \xi^{(j)})$ is in the direction of ν_j. The second statement derives directly from the fact that $\mathrm{Dim}\,(\mathcal{RT}_0 T) = 3$, and property (i) tells us that the functions $\{w_i, w_j, w_k\}$ are orthogonal. Finally, the last property is a direct calculation:

$$\int_T \mathrm{div}\, w_i \; dx = 2a_i \mid T \mid = \int_{\partial T} w_i \cdot n \; ds = \mid \sigma_i \mid,$$

from which $a_i = \dfrac{\mid \sigma_i \mid}{2 \mid T \mid}$. □

An example of such a function is given in Figure 3.3.

3.3 A closer look at the "inf-sup" condition

We first rewrite system (3.9) changing sign to the definition of the bilinear and linear forms $b\,(p, q)$ and $F\,(q)$, respectively:

$$\begin{aligned} a\,(u_h, v) + b\,(p_h, v) &= 0 \qquad \forall v \in \mathcal{V}_h, \\ +b\,(q, u_h) &= F\,(q) \qquad \forall q \in \mathcal{Q}_h, \end{aligned}$$

where:

$$a\,(v, w) = \int_\Omega \mu v \cdot w \; dx \qquad b\,(v, q) = -\int_\Omega q \,\mathrm{div}\, v \qquad F\,(q) = -\int_\Omega f q \; dx.$$

For simplicity we use the lowest order spaces $\mathcal{RT}_0 - \mathcal{P}_0$ an a triangulation $\mathcal{T}_h$ formed by N_T triangles and N_σ edges. We can then express u_h and p_h as a linear combination of the basis functions $v_k \in \mathcal{V}_h$, $k = 1, \ldots, N_\sigma$ and $q_t \in \mathcal{Q}_h$, $t = 1, \ldots, N_T$:

$$u_h = \sum_{i=1}^{N_\sigma} u_i v_i, \tag{3.21}$$

$$p_h = \sum_{m=1}^{N_T} p_m q_m. \tag{3.22}$$

Substitution of the above equations into the linear system yields:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{3.23}$$

where matrices A, B have dimensions $N_\sigma \times N_\sigma$ and $N_T \times N_\sigma$, respectively, and are given by::

$$A = \{a_{ij}\} \;\; a_{ij} = a(v_i, v_j), \qquad i,j = 1,\ldots,N_\sigma$$
$$B = \{b_{lm}\} \;\; b_{lm} = b(q_l, v_m), \qquad l = 1,\ldots,N_T\,, m = 1,\ldots,N_\sigma.$$

The vectors $u \in \mathbb{R}^{N_\sigma}$ and $p \in \mathbb{R}^{N_T}$ contain the coefficients of the linear combinations (3.21) and (3.22), while vectors $f \in \mathbb{R}^{N_\sigma}$ and $g \in \mathbb{R}^{N_T}$ are the known right hand sides. In our previous case we had $f = 0$, but if we had non homogeneous Neumann boundary conditions we would have had $f \neq 0$.
We take the algebraic point of we, and will indicate with $\mathcal{A}$ the system matrix:

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

so that the full system will be denoted with $\mathcal{A}x = b$.
We first observe that A, coming from the discretization of the $\mathcal{L}^2$ scalar product, is symmetric and positive definite. The full system is symmetric and corresponds to a "saddle point" problem: the solution $x = (u,p) \in \mathbb{R}^{N_\sigma \times N_t}$ can be viewed as the solution of the following constrained minimization problem:

$$\min_{u \in \mathbb{R}^{N_\sigma}} \frac{1}{2} u^T A u - f^T v \tag{3.24}$$
$$\text{subject to } Bu = g, \tag{3.25}$$

where variable p plays now the role of a Lagrange multiplier. Every solution (u^*, p^*) is a saddle point for the Lagrangian:

$$\mathcal{L}u, p = \frac{1}{2} u^T A u - f^T u + (Bu - g)^T p,$$

as the pair (u,p) must satisfy:

$$\mathcal{L}u^*, p \leq \mathcal{L}u^*, p^* \leq \mathcal{L}u, p^*, \qquad \forall u \in \mathbb{R}^{N_\sigma} \qquad \forall p \in \mathbb{R}^{N_T},$$

or, equivalently:

$$\min_u \max_p \mathcal{L}u, p = \mathcal{L}u^*, p^* = \max_p \min_u \mathcal{L}u, p.$$

Matrix $\mathcal{A}$ can be block-factorized as follows:

$$\begin{aligned} \mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} &= \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} A & 0 \\ B & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ 0 & S \end{bmatrix} \end{aligned}$$

where the *Schur complement* S is $S = -(BA^{-1}B^T)$. It is now easy to see that, since A is nonsingular, the condition for the existence of the inverse of $\mathcal{A}$ is that B have maximum rank ($\text{rank}\,(B) = N_T$). In fact, there is the following [2]:

Theorem 3.13. *Let A be a symmetric and semi-positive definite matrix, and let B be of maximum rank. Then matrix $\mathcal{A}$ is nonsingular if and only if* $\ker(A) \cap \ker(B) = \{0\}$.

Proof. *S*ufficient condition. Let $x = (u, p)^T$ such that $\mathcal{A}x = 0$. Then we have $Au + B^T p = 0$ and $Bu = 0$, from which $u^T Au = -u^T B^T p = -(Bu)^T p = 0$. By hypothesis, A is spd, so that $u^T Au = 0$ implies $Au = 0$ and $u \in \ker(A) \cap \ker(B)$, which implies $u = 0$. Moreover, $B^T p = 0$ and the conclusion $p = 0$ is a consequence of the fact that B has maximum rank.
*N*ecessary condition. Assume now that $\ker(A) \cap \ker(B) \neq \{0\}$ and let $u \in \ker(A) \cap \ker(B)$ with $u \neq 0$. Then, for $x = (u, 0)^T$ we have $\mathcal{A}x = 0$ and thus $\mathcal{A}$ is singular and the condition is also necessary. □

It can be shown that u^* is the A-orthogonal projection (orthogonal projection with respect to the scalar product $(v, w)_A = v^T Aw$) on the space $\mathcal{C} = \{p \in \mathbb{R}^{N_T} : Bp = g\}$, interpreted as the space of the constraints. We assume B is of maximum rank and denote with $\beta^2 = \sigma_{min}(B)$ the smallest singular value of B, which we will assume always strictly positive. We assume that also α, the minimum eigenvalue of A, is strictly positive. Then the inverse of $\mathcal{A}$ can be written explicitly as:

$$\mathcal{A}^{-1} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}(I - B^T S^{-1} B A^{-1}) & A^{-1} B^T S^{-1} \\ S^{-1} B A^{-1} & S^{-1} \end{bmatrix}, \tag{3.26}$$

and we get immediately the following estimates:

$$\|u\|_A \leq \|f\|_{A^{-1}} \leq \frac{1}{\alpha} \|f\| \qquad \|p\| \leq \frac{1}{\beta} \|f\|_{A^{-1}} \leq \frac{1}{\alpha\beta} \|f\| .$$

The problem is well posed if there exist constants β^* and α^* independent of h such that $\beta > \beta^* > 0$ and $\alpha > \alpha^* > 0$. We have the following:

Lemma 3.14. *The condition $\beta = \sigma_{min}(B) > 0$ is equivalent to the "inf-sup" condition for the following saddle point problem:*

$$\inf_{q \in \mathbb{R}^{N_T}} \sup_{v \in \mathbb{R}^{N_\sigma}} \frac{q^T B v}{\|q\| \, \|v\|} > \beta^2 > 0 \qquad \forall q \neq 0, \;\; e \; \forall v \neq 0,$$

or, equivalently:

$$\max_{v \in \mathbb{R}^{N_\sigma}} \frac{q^T B v}{\|v\|} > \beta^2 \|q\| \qquad \forall q \in \mathbb{R}^{N_T}, q \neq 0.$$

Proof. Let $B = U\Sigma V^T$ be the singular value decomposition of B and let:

$$q = u_i, i = 1, \ldots, N_T \text{ e } v = \sum_{j=1}^{N_\sigma} \gamma_j v_j.$$

The orthogonality of V implies that $\|v\|^2 = \sum \gamma_i^2$. Thus, we can write:

$$\frac{q^T B v}{\|v\|} = \frac{e_i^T \Sigma V^T v}{\sqrt{\sum_j \gamma_j^2}} = \frac{\sigma_i \gamma_i}{\sqrt{\sum_j \gamma_j^2}} \geq \sigma_i \geq \beta^2.$$

On the other hand, taking $q = \sum_{j=1}^{N_T} \xi_j u_j$ and indicating with $\gamma = \{\gamma_i\}$ the vector of the coefficients γ_i, we have:

$$\max_{v \in \mathbb{R}^{N_\sigma}} \frac{q^T B v}{\|v\|} = \max_{\gamma \neq 0} \sum_{i=1}^{N_\sigma} \xi_i \frac{\sigma_i \gamma_i}{\sqrt{\sum_j \gamma_j^2}} \geq \beta^2 \sum_{i=1}^{N_T} \frac{\xi_i^2}{\sqrt{\sum_j \xi_j^2}} = \beta^2 \|q\|.$$

□

It is obvious that all these considerations can be extended immediately to the continuous case. Hence, convergence of MFEM requires that matrix $\mathcal{A}$ be invertible for all $\mathcal{T}_h$ uniformly for $h \longrightarrow 0$. Thus, if $\mathcal{A}_h$ is the matrix related to $\mathcal{T}_h$, varying h we have a sequence of problems of the type:

$$\begin{bmatrix} A_h & B_h^T \\ B_h & 0 \end{bmatrix} \begin{bmatrix} u_h \\ p_h \end{bmatrix} = \begin{bmatrix} f_h \\ g_h \end{bmatrix}.$$

For every h, the system is solvable if $\sigma_{min}(B_h) = \beta_h^2 \geq 0$, or:

$$\inf_{q \in \mathcal{W}_h} \sup_{v \in \mathcal{V}_h} \frac{q^T B_h v}{\|q\| \|v\|} > \beta_h^2 > 0 \qquad \forall q \neq 0, \text{ e } \forall v \neq 0,$$

Hence, the spaces $\mathcal{V}_h$ and $\mathcal{W}_h$ need to satisfy the above "inf-sup" condition. If they do, existence and uniqueness and continuous dependence of the solution on the data, or shortly well-posedness, is guaranteed, and the MFEM converge. Note that the error constant in the convergence proofs is proportional to $1/\alpha$ and $1/\beta$. If β decreases for $h \longrightarrow 0$ then convergence will be slower than optimal.
On the other hand, if the spaces $\mathcal{V}_h$ and $\mathcal{W}_h$ do not satisfy the "inf-sup" condition then convergence is not ensured. It may happen that the problem converge for a certain set of data but not for another. In this case different situations may occur:

- the space of vectors v for which $b(p, v) = g$ is empty. This may occur if $N_T > N_\sigma$, i.e., there are more constraints than equations. A typical example is the pair of basis function $\mathcal{P}_1/\mathcal{P}_0$ for the discretization of Stokes equation;

- the kernel of $b(p, v)$ is non empty and the saddle-point system matrix is singular. Typically in this case spurious oscillations $\tilde{p}$ may be generated, as these vectors are such that $b(\tilde{p}, v) = 0$. Often, this situation is known as "mesh-locking";

- the condition that B be of maximum rank is satisfied but the largest singular value tends to zero with h: $\beta_h = \mathcal{O}\left(h^k\right)$. In this case the system is highly ill-conditioned and we may have convergence up to a certain value of h, but after that the rates decrease until convergence is completely lost.

We would like to remark that the easiest way to cure the lack of fulfillment of the "inf-sup" condition is to guarantee that the space $\mathcal{V}_h$ be sufficiently richer than $\mathcal{W}_h$. In other words, the constraints of the problem cannot be too stringent, and enough movement must be allowed in the search space $\mathcal{V}_h$.

3.3.1 More on the solution of the linear system: hybridization

A simple strategy to solve the linear system (3.23) comes from the observation that matrix A is invertible (the form $a(v, w)$ is coercive as it is the discretization of the $\mathcal{L}^2$ scalar product, as already observed above) and we can use (3.26) to express (u, p) in terms of the Schur complement $S = (BA^{-1}B^T)$. We obtain:

$$\begin{aligned} u &= A^{-1}(f - B^T p), \\ Sp &= BA^{-1}f - g. \end{aligned}$$

This is impractical as the inverse of A is a full matrix and both the costs of inversion and of storing are overwhelmingly large. A more efficient way is to proceed with the so-called "hybridization" strategy. The ensuing FE method is called the Mixed-Hybrid FE. The idea is to find some strategy to simplify the inversion of matrix A.
The first observation is that our discrete spaces do not require the continuity of the candidate solutions but only of the normal fluxes across element edges (requirement for belonging to $\mathcal{H}_{\text{div}}$). This last condition is the one that couples all the flux unknowns together. In fact, if we apply Green's lemma on a single element T to eq. (3.6) we obtain a boundary flux term involving the normal fluxes on the element edges (faces). Continuity of the normal fluxes implies that when we sum over all elements these normal fluxes cancel pairwise, since those calculated on the same face from the two neighboring elements are equal and with opposite signs. Thus we are left with the terms on the domain boundary, which are determined by the boundary conditions.
The idea is then to relax the continuity assumption and to re-impose it as a constraint in the linear system. More precisely, we relax the hypothesis that $\mathcal{V}_h \subset \mathcal{H}_{\text{div}}(\Omega)$ and we set $\mathcal{V}_h \subset \mathcal{H}_{\text{div}}(T)$, for all $T \in \mathcal{T}_h$. Thus, working on an element by element, continuity of the normal fluxes is lost and the basis functions $\mathcal{V}_h$ can be defined independently on each element. We re-impose continuity by means of Lagrange multipliers.

To see how this works, we look at the lowest order $\mathcal{RT}_0 - \mathcal{P}_0$ and introduce the corresponding discontinuous space:

$$\widetilde{V}_h = \left\{ v \in \left[\mathcal{L}^2(\Omega)\right]^2 : v|_T \in \mathcal{RT}_0(T) \; \forall T \in \mathcal{T}_h \right\}.$$

Note that $\mathcal{V}_h \subset \widetilde{V}_h$ and that $v \in \mathcal{V}_h$ if and only if $v \in \mathcal{H}_{\text{div}}(\Omega)$. We now put together all the N_σ triangle faces $\sigma \in \mathcal{T}_h$ of the mesh skeleton Γ_h:

$$\Gamma_h = \bigcup_{T \in \mathcal{T}_h} \partial T = \bigcup_{i=1}^{N_\sigma} \sigma_i,$$

and introduce the space of the Lagrange multipliers λ on the set Γ_h:

$$\Lambda_h = \left\{ \mu \in \mathcal{L}^2(\Gamma_h) : \mu|_\sigma \in \mathcal{P}_0(\sigma) \; \forall \sigma \in \Gamma_h \right\}.$$

Testing the mixed system with vector and scalar basis functions, $(v, q) \in \mathcal{RT}_0 \times \mathcal{P}_0$ and applying Green's lemma, we obtain:

$$\sum_{T \in \mathcal{T}_h} \left[\int_T \mu u_h \cdot v \; dx - \int_T p_h \operatorname{div} v \; dx + \int_{\partial T} p_h v \cdot \nu \; ds \right] = 0$$

$$\sum_{T \in \mathcal{T}_h} \int_T \operatorname{div} u_h q \; dx = 0$$

Hence, we can define the bilinear forms:

$$a(v, w) = \sum_{T \in \mathcal{T}_h} \int_T \mu v \cdot w \; dx; \qquad b_h(q, v) = \sum_{T \in \mathcal{T}_h} \int_T q \operatorname{div} v \; dx.$$

for all $q \in \mathcal{W}_h$ and all $v, w \in \widetilde{V}_h$, and the new bilinear form $d(v, \mu)$ defined on the mesh skeleton Γ_h as:

$$d(v, \mu) = -\sum_{T \in \mathcal{T}_h} \int_{\partial T} \mu \; v|_T \cdot \nu \; dx = \sum_{\sigma \in \Gamma_h} \int_\sigma \mu \left[v \cdot \nu \right] \; ds,$$

for all $v \in \widetilde{V}_h$ and all $\mu \in \Lambda_h$, where ν_T is the exterior normal to ∂T and $[v \cdot \nu]$ is the "jump" of the normal component of v across edge σ. Note that $d(v, \mu) = 0$ in every $\sigma \in \Gamma_h$ if and only if $v \in \mathcal{H}_{\text{div}}(\Omega)$ (or $v \in \mathcal{V}_h$). Moreover, since $b(q, v)$ is not defined in $\widetilde{V}_h$, we have to define the "mesh" bilinear form $b_h(q, v)$ built on the triangles (and not on the entire Ω).
Then we can consider the following FEM problem: Find $(u_h, p_h, \lambda_h) \in \widetilde{V}_h \times \mathcal{W}_h \times \Lambda_h$ such that:

$$\begin{aligned} a(u_h, v) + b_h(p_h, v) + d(v, \lambda_h) &= f \qquad \forall v \in \widetilde{V}_h \\ b_h(q, u_h) &= g \qquad \forall q \in \mathcal{W}_h \\ d(u_h, \mu) &= 0 \qquad \forall \mu \in \Lambda_h. \end{aligned}$$

The corresponding algebraic system is given by:

$$\begin{cases} \tilde{A}u + B_h^T p + C^T \lambda &= f \\ B_h u &= g \\ Cu &= 0 \end{cases},$$

with obvious expressions for the matrix elements. Now matrix $\tilde{A}$ is block diagonal with blocks of size $n_\sigma \times n_\sigma$ where n_σ is the number of faces of the element, 3 in the case of triangles. The matrix is thus easily invertible block-by-block with a relatively small computational cost. We can proceed to the block elimination as done before to obtain:

$$u = \tilde{A}^{-1}\left(f - B_h^T p - C\lambda\right),$$

and after substitution we obtain:

$$\begin{cases} B_h \tilde{A}^{-1} B_h^T p + B_h C\lambda = B_h \tilde{A}^{-1} f - g \\ C\tilde{A}^{-1} B_h^T p + C\tilde{A}^{-1} C\lambda = -C\tilde{A}^{-1} f \end{cases}.$$

Matrix $H = B_h \tilde{A}^{-1} B_h^T$ is again block-diagonal and easily invertible. Writing $S = \tilde{A}^{-1} B_h^T$, we have:

$$p = H^{-1}\left[f - S^T g\right],$$

from which, denoting by M the block-diagonal matrix given by: $M = \tilde{A}^{-1} - SH^{-1}S^T$, we have the final system of dimension $N_\sigma \times N_\sigma$ (N_σ being the total number of faces in the mesh) having as unknowns the vector of Lagrange multipliers λ defined on each triangle face:

$$C^T M C\lambda = C^T\left[Mg - SH^{-1}f\right].$$

It is possible to see that this system is symmetric and positive definite and can be solved with PCG. Comparing to the Galerkin FEM (for example $\mathcal{P}_1$ vs. $\hat{\mathcal{RT}}[0] - \mathcal{P}_0$), the size of the system is larger for the MHFEM, being the number of faces in a triangulation approximately 3 times the number of nodes in two dimensions and 7 times in three dimension. On the other hand, the number of nonzero elements per row is on average equal to $2n_\sigma - 1$, independently on the shape of the triangles, and is thus much sparser than Galerkin $\mathcal{P}_1$, increasing the efficiency of any (non-diagonal) PCG preconditioning technique.

Remark 3.6. *We remark that the bilinear form $d(\cdot,\cdot)$ acts on the triangulation skeleton Γ_h, and thus $d(v,\mu)$ tests the trace of p_h on the triangle edges. The Lagrange multipliers λ_h are thus the trace of the pressure on the boundary of the triangles, and thus span a linear distribution of the pressure inside the element (in absence of forcing functions, i.e., $f = 0$), with $p_h|_T$ being the average value. Hence, the lowest order mixed-hybrid finite element can be interpreted*

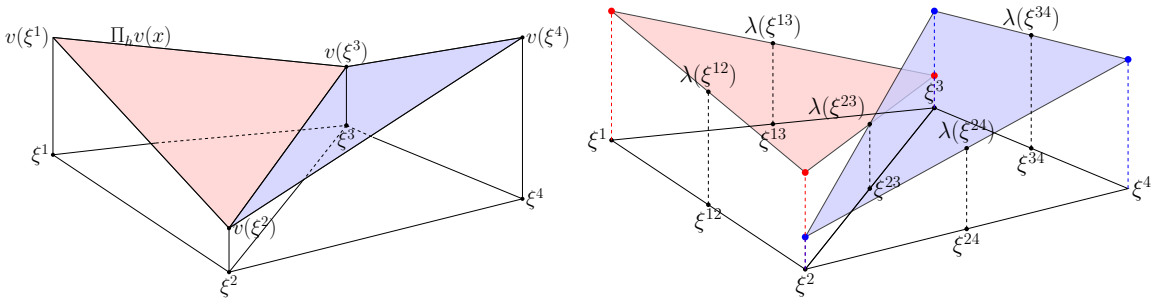


FIGURE 3.4: *Comparison between the conforming $\mathcal{P}_1$ Galerkin (left) and the non-conforming Mixed-Hybrid (right) interpolation of the pressure.*

as a non-conforming $\mathcal{P}_1$ method defined on $\mathcal{T}_h$. Figure 3.4 highlights the intuitive difference between conforming $\mathcal{P}_1$ FEM and the Mixed Hybrid approaches. It is clearly intuitive that, in the case of constant diffusion coefficient, the $\mathcal{P}_1$ Galerkin method guarantees continuity of the tangential component of the pressure gradient not of the normal component. The $\mathcal{RT}_0 - \mathcal{P}_0$ Mixed-Hybrid method, on the contrary, guarantees continuity of the normal but not of the tangential component of th pressure gradient.

3.4 Experimental comparison between Galerkin $\mathcal{P}_1$ and MFEM $\mathcal{RT}_0 - \mathcal{P}_0$ in the solution of elliptic equations

A better understanding of the practical motivations highlighting the usefulness of the mixed finite element approach is obtained by the following considerations on simple test problem. We remark that we are looking here not to the convergence of the scheme for $h \to 0$, a property that is fundamental but cannot be tested in real applications. Rather we are investigating properties of the numerical solution at a fixed h, comparing them with qualitative properties that we expect from the real solution. In this case, we are investigating what is called the "local conservation" properties of the schemes, similarly to what we have done in section 2.11.
Consider the domain and the mesh reported in Figure 3.5. We want to solve equation (3.3) with boundary conditions shown on the Figure. The resulting solution will create a flux vector field $u_h(x) = -a(x)\nabla p_h(x)$ that enters the domain from the inflow boundary (the left edge of the square) and flows towards the outflow boundary, localized in a central portion of the right edge. We note within the domain the presence of two "column-like" internal regions where the diffusion coefficient is $a(x) = 10^{-12}$, much smaller than the background value $a(x) = 1$. The mass flow inside the two pillars is thus impeded, and should be very small. Hence, particle trajectories should leave the inflow boundary, circumnavigate the pillars, and exit from the outflow boundary.
We solve the problem by calculating the flow field $u_h(x)$ with Galerkin $\mathcal{P}_1$ the with mixed hybrid formulation based on $\mathcal{RT}_0 - \mathcal{P}_0$ spaces. Then trajectories are calculated by distribut-

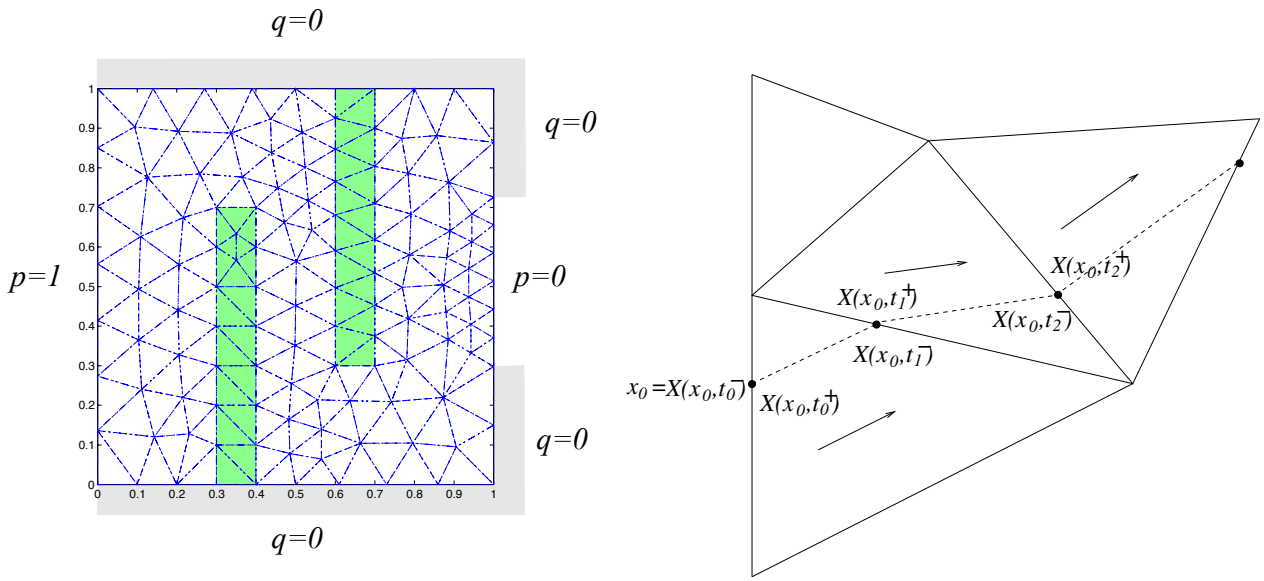


FIGURE 3.5: *Right panel: domain and boundary conditions for the solution of eq.* (3.3)*. The normal flux at the boundary is denoted by* $q = u \cdot n$*. Left panel: graphic representation and notations used in the "particle tracking" procedure.*

ing uniformly 100 particle on the inflow boundary and numerically evaluating the following integral:

$$X(x_0, t) = \int_{t_0}^{t} u_h(X(x_0, \tau))\, d\tau$$

where $X(x_0, t)$ is the position at time t of the particle released at time $t = 0$ at position $x = x_0$, and $u_h(X(x_0, t))$ is the Eulerian velocity at that position. The integral is evaluated by assuming that the velocity vector u_h is constant on each triangle and we discretize the interval $[0, t]$ in subintervals of length h_k that satisfy the CFL (Courant-Friedrichs-Lewy) condition that the particle exactly reaches the boundary of the considered triangle within the time-step h_k ($t_0 = 0$, $t_k = t_{k-1} + h_{k-1}$). The resulting algorithm is exemplified in Figure 3.5. At $k = 0$ we start from the point $X(x_0, 0) = x_0$, which is assumed to be in $\sigma_k \in T_r$, an edge of the Dirichlet boundary where $p = 1$ is imposed. The next point $X(x_0, t_1^-)$ belongs to the boundary of T_r and is found by joining $X(x_0, 0)$ with the boundary of T_r moving along the direction $u_{h,r}$, the (constant) velocity vector in T_r. Next, once the neighboring triangle T_s is identified, we let $X(x_0, t_1^+)$ be the same point $X(x_0, t_1^-)$ but now belonging to T_s. This is the starting point for the iterated procedure. Denoting with $u_h(x_{b_k})$ the velocity vector within the triangle containing $X(x_0, t_k^+)$, the scheme can be written as:

$$X(X(x_0, t_k), t_{k+1}^-) = X(x_0, t_k^+) + \lambda u_h(x_{b_k}),$$

where λ is given by $\left\|X(x_0, t_{k+1}^+) - X(x_0, t_k^+)\right\|$, i.e., the length of the path traveled by the particle within the time step.

Figure 3.6 (top panels) shows the trajectories calculated with the above algorithm using the Galerkin $\mathcal{P}_1$ (left) and the MHFEM $\mathcal{RT}_0 - \mathcal{P}_0$ (right) velocity fields starting from 100 particles uniformly distributed on the inflow boundary. The differences between the two vector fields are evident, but a few comments on these results are notable. Contrary to the MHFEM method, the Galerkin trajectories are not uniformly spaced one-another inside the domain, and converge in several clusters. We recall that convergence of the trajectories indicates the presence of a sink term, while divergence indicates a source. The second observation is that some of the $\mathcal{P}_1$ trajectories exit the domain from no-flow boundaries, obviously violating the properties of the original problem. The numerical flow field is thus obviously non-conservative. The $\mathcal{RT}_0 - \mathcal{P}_0$ flow field, on the other hand, does not show any of these problems, and is everywhere conservative. The lower panel in Figure 3.6 shows a detail of the mesh and the ensuing $\mathcal{P}_1$ velocity field and relevant trajectory. The particle starting from the left inlet point D proceeds along a direction parallel to the elemental velocity until it reaches the opposite triangle edge (point E). Here the trajectory should now follow the direction pointed at by vector w, and thus re-entering the triangle that was just left. This is obviously a contradiction and the trajectory has to stop. In fact, the components of vectors v and w along edge AB have opposite sign, indicating that along edge AB a sink is acting with magnitude $\mid v \cdot n \mid + \mid w \cdot n \mid$, violating the local conservation property. This does not occurs for the $\mathcal{RT}_0 - \mathcal{P}_0$ MHFEM, and is the explanation for the large differences in the trajectories shown in the upper panels of Figure 3.6.

Two final remarks are in order. In the first one we would like to stress again the fact that we are working at a fixed mesh size. At convergence, i.e., $h \to 0$, the Galerkin $\mathcal{P}_1$ converges towards the real solution, which satisfies both local conservation and maximum principles, and evidently all these effects disappear. However, practically always, in practical applications, the mesh size h is determined by the need to follow the geometrical constraints of the domain and of the heterogeneities of the diffusion coefficient, and grid refinement can seldom be performed. For these reasons we look at local properties of the numerical solutions at a fixed mesh size.

The second observation is related to the mass balance error relative to the scalar numerical solution $p_h(x)$. The conservation error resulting from this numerically evaluated field is in most cases negligible when calculated on the correct control volume and not necessarily on the finite element (see discussion in section 2.11). Thus it cannot be mistaken for the error arising from the numerical flow field, a different unknown. This is the reason why, starting from the 1970's, the field of the mixed approach emerged as an active research field, by trying to approximate simultaneously both pressure and velocities.

As a final remark, we note that the computational complexity of the mixed and the mixed-hybrid methods is much higher than Galerkin. In fact, for general triangulations, the number of elements is approximately 1.7 and 3 times the number of nodes for two dimensional and three dimensional triangulations, respectively, while the number of edges or faces is on the average

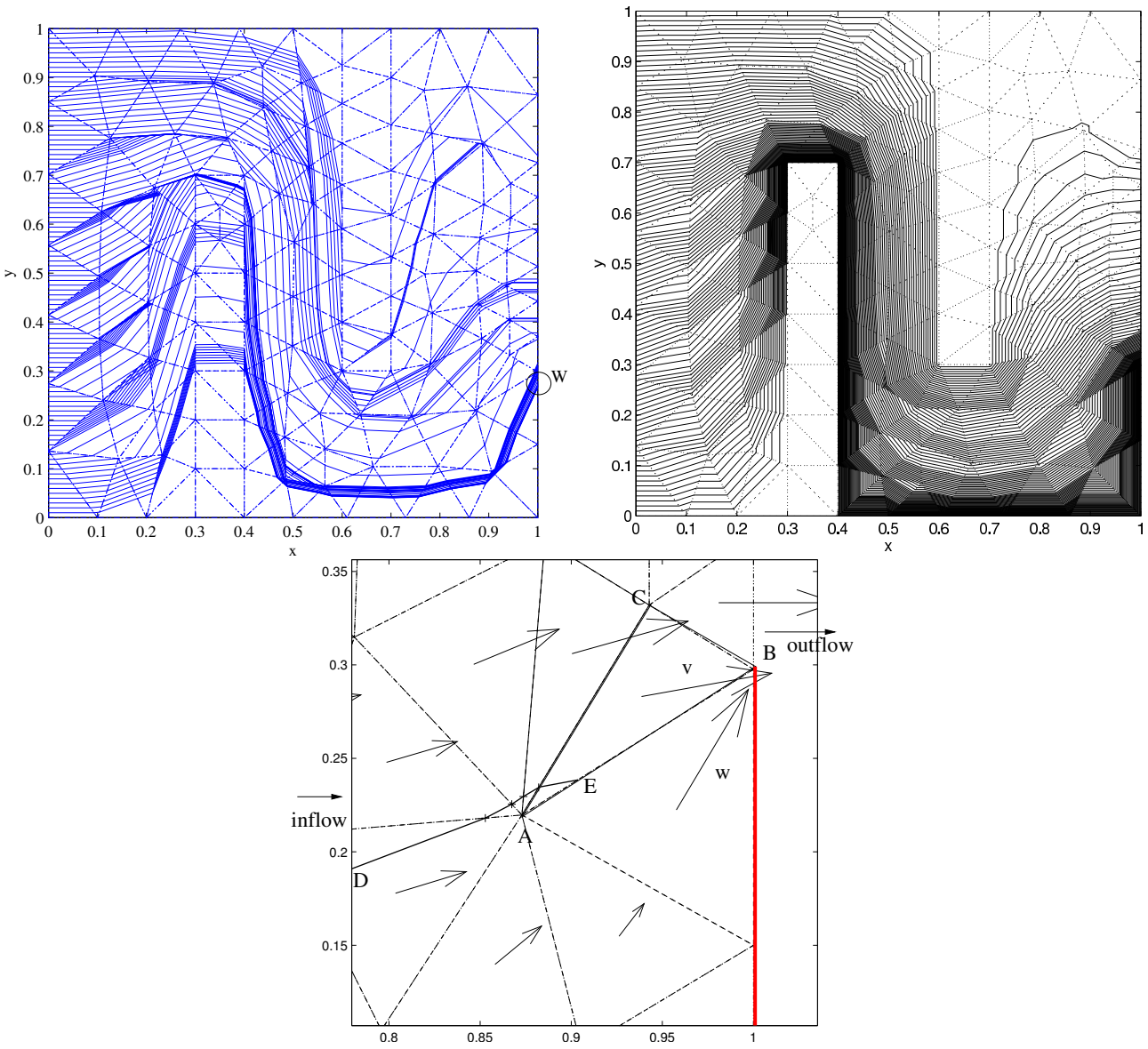


FIGURE 3.6: *Trajectories calculated using the velocity field calculated by Galerkin $\mathcal{P}_1$ (top left panel) and with the MHFEM $\mathcal{RT}_0 - \mathcal{P}_0$ (top right panel). The bottom panel shows a zoom of the Galerkin $\mathcal{P}_1$ velocity field. We note several neighboring triangles where the normal component of the velocity field have opposite sign. This obviously implies the presence of a source or sink term, although in the original problem $f(x) = 0$. These unphysical source/sink terms contribute substantially to the non satisfaction of a local conservation property, and cause small local oscillations that violate the maximum principle of elliptic equations.*

between 3 and 7 times the number of nodes. For this reason, current research concentrates on the development of post-processing techniques that starting from the $\mathcal{P}_1$ velocity, reconstruct a locally conservative (divergence free) field.

3.5 The Stokes equation

We come back to discuss the numerical solution of the Stokes equation (3.1), which we rewrite here:

$$\begin{aligned} -\mu\Delta u + \nabla p &= f && \text{in } \Omega, \\ \operatorname{div} u &= g && \text{in } \Omega, \\ u &= 0 && \text{in } \Gamma, \end{aligned} \tag{3.27}$$

We note first that the pressure appears in this equation under the gradient sign, and is thus defined only up to a constant. Thus we add the constraint:

$$\int_\Omega p \, dx = 0.$$

Moreover, we have that the source term g must have zero mean, as we assume zero velocity on the boundary. In fact, application of the divergence theorem to the the incompressibility equation yields:

$$\int_\Omega \operatorname{div} u \, dx = \int_\Omega g \, dx = \int_\Gamma u \cdot \nu \, ds = 0$$

because $u = 0$ on Γ.
We can then write the following variational formulation:

Problem 3.7 (Stokes variational formulation)**.** Find $(u, p) \in \mathcal{V} \times \mathcal{Q}$ such that:

$$\begin{aligned} a(u, v) + b(p, v) &= F(v) && \forall v \in \mathcal{V}, \\ b(q, u) &= (g, q) && \forall q \in \mathcal{Q}, \end{aligned} \tag{3.28}$$

where

$$\begin{aligned} a(v, w) &= \mu \int_\Omega \nabla v : \nabla w \, dx & b(v, q) &= -\int_\Omega q \operatorname{div} v \\ F(v) &= -\int_\Omega f \cdot v \, dx & (g, q) &= -\int_\Omega gq \, dx \end{aligned} \tag{3.29}$$

and the spaces are given by:

$$\mathcal{V} = [\mathcal{H}_0^1(\Omega)]^d$$
$$\mathcal{Q} = \mathcal{L}_0^2(\Omega) = \left\{ q \in \mathcal{L}^2(\Omega) : \int_\Omega q \, dx = 0 \right\}$$

In this case

$$\mathcal{L}(v,q) = \frac{1}{2}a\left(v,v\right) - F\left(v\right) + b\left(v,q\right).$$

Then, the pair (u,p) must satisfy:

$$\mathcal{L}(u,q) \leq \mathcal{L}(u,p) \leq \mathcal{L}(v,p), \qquad \forall v \in \mathcal{V} \qquad \forall q \in \mathcal{Q},$$

i.e., (u,p) is the saddle point of the Lagrangian $\mathcal{L}(v,q)$. Again note, that in this formulation p can be viewed as the Lagrange multiplier that enforces the constraint $b\left(u,p\right) = 0$.

Remark 3.8. *This is equivalent to look for candidate solutions within the space where the constraint is satisfied. In other words, we can define the space* $\mathcal{W} \subset \mathcal{V}$*:*

$$\mathcal{W} = \left\{v \in \mathcal{V} : b\left(v,q\right) = \left(g,q\right) \quad \forall q \in \mathcal{Q}\right\}.$$

Within this space, assuming the bilinear form $a\left(\cdot,\cdot\right)$ *to be coercive, the linear system* (3.28) *reduces to:*

$$a\left(u,v\right) = F\left(v\right) \qquad \forall v \in \mathcal{V}.$$

Taking $v = u$ *in the previous equation and using Poincaré inequality yields the following:*

$$\left\|u\right\|_{\mathcal{L}^2(\Omega)} \leq C_\Omega \left\|\nabla u\right\|_{\mathcal{L}^2(\Omega)} \leq C \left\|f\right\|_{\mathcal{L}^2(\Omega)}$$

analogous to the stability statement (2.60) *in the Lax-Milgram theorem. The practical strategy is to include the constraint for every function* g *is to solve for* $\tilde{u} = u - u_g$*, where the function* u_g *is such that* $\operatorname{div} u_g = g$*, and thus solve the problem:*

$$a\left(u,v\right) = F\left(v\right) + a\left(u_g,v\right) \qquad \forall v \in \mathcal{V},$$

and proceed as discussed in section 3.1. However, as we have already mentioned in this section, constructing finite element spaces that satisfy the above properties is not easy and seldom used in practical applications.

The well-posedness of the saddle point problem requires that the following two properties are satisfied:

1. the bilinear form $a\left(\cdot,\cdot\right)$ is continuous and coercive, i.e., there exist a constant $\alpha > 0$ such that:

$$\left|a\left(v,v\right)\right| \geq \alpha \left\|v\right\|_{\mathcal{V}}, \qquad \forall v \in \mathcal{W};$$

2. the bilinear form $b\left(\cdot,\cdot\right)$ is continuous and satisfies the inf-sup condition, i.e., there exist a constant $\beta > 0$ such that:

$$\inf_{q\in\mathcal{Q}} \sup_{v\in\mathcal{V}} \frac{b\left(v,q\right)}{\left\|v\right\|_{\mathcal{V}} \left\|q\right\|_{\mathcal{Q}}} \geq \beta.$$

3.5.1 Stable FEM discretizations (Mixed FEM)

Given a regular triangulation $\mathcal{T}_h(\Omega)$ of the domain Ω, the FEM discretization of the Stokes equation is the finite dimensional counterpart of problem 3.7 and reads as follows:

Problem 3.9 (Stokes Galerkin formulation)**.** Find $(u_h, p_h) \in \mathcal{V}_h(\mathcal{T}_h) \times \mathcal{Q}_h(\mathcal{T}_h)$ such that:

$$\begin{aligned} a(u_h, v) + b(p_h, v) &= F(v) \qquad \forall v \in \mathcal{V}_h(\mathcal{T}_h) \subset \mathcal{H}_0^1(\Omega), \\ b(u_h, q) &= (g, q) \qquad \forall q \in \mathcal{Q}_h(\mathcal{T}_h \subset \mathcal{L}^2(\Omega) \end{aligned}$$

leading to the linear system:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

with the matrix elements defined as:

$$\begin{aligned} A = \{a_{ij}\} \qquad & a_{ij} = a(\phi_i, \phi_j) = \mu \int_\Omega \nabla\phi_i : \nabla\phi_j \, dx \qquad && i, j = 1, \ldots, N \\ B = \{b_{ki}\} \qquad & b_{ki} = b(\phi_i, \psi_k) = -\int_\Omega \psi_k \operatorname{div} \phi_i \, dx \qquad && k = 1, \ldots, M; \quad i = 1, \ldots, N \\ f = \{f_i\} \qquad & f_i = -\int_\Omega f \cdot \phi_i \, dx && i = 1, \ldots, N \\ g = \{g_k\} \qquad & g_k = -\int_\Omega g\psi_k \, dx && k = 1, \ldots, M \end{aligned}$$

where $\phi_i(x)$ are vector functions and $\psi_k(x)$ scalar functions forming the bases for $\mathcal{V}_h$ and $\mathcal{Q}_h$, respectively, i.e.:

$$\mathcal{V}_h(\mathcal{T}_h) = \operatorname{Span}(\phi_1, \ldots, \phi_N) \subset \left[\mathcal{H}_0^1(\Omega)\right]^d \qquad \mathcal{Q}_h(\mathcal{T}_h) = \operatorname{Span}(\psi_1, \ldots, \psi_M) \subset \mathcal{L}^2(\Omega).$$

From what we have seen above, this problem is well posed, i.e., the solution exists and it is unique so that it can be solved, if both the following properties are satisfied:

1. the bilinear form $a(\cdot, \cdot)$ is continuous and coercive:

$$| a(v, w) | \leq \gamma \|v\|_{\mathcal{V}_h} \|w\|_{\mathcal{V}_h} \qquad a(v, v) \geq \alpha \|v\|_{\mathcal{V}_h}^2 \text{ for all } v, w \in \mathcal{V}_h;$$

2. the bilinear form $b(\cdot, \cdot)$ is continuous and satisfies the inf-sup condition, i.e.:

$$| b(v, q) | \leq \delta \|v\|_{\mathcal{V}_h} \|q\|_{\mathcal{Q}_h} \qquad b(v, q) \geq \beta \|v\|_{\mathcal{V}_h} \|q\|_{\mathcal{Q}_h} \text{ for all } (v, q) \in \mathcal{V}_h \times \mathcal{Q}_h.$$

When these properties are satisfied, the following analogues of Céa Lemma are valid:

$$\|u-u_h\|_{\mathcal{H}^1(\Omega)} \leq C_1 \|u-v\|_{\mathcal{H}^1(\Omega)} + C_2 \|p-p\|_{\mathcal{L}^2(\Omega)} \qquad \forall (v,q) \in \mathcal{V}_h \times \mathcal{Q}_h$$
$$\|p-p_h\|_{\mathcal{H}^1(\Omega)} \leq C_3 \|u-v\|_{\mathcal{H}^1(\Omega)} + C_4 \|p-p\|_{\mathcal{L}^2(\Omega)} \qquad \forall (v,q) \in \mathcal{V}_h \times \mathcal{Q}_h.$$

The FEM linear system is of the form given in (3.23), and thus, from Theorem 3.13, it is well-posed if $\ker(A) \cap \ker(B) = \{0\}$, or, equivalently as seen in Lemma 3.14, if it matrix B satisfies the discrete inf-sup condition:

$$\max_{v \in \mathbb{R}^{N_\sigma}} \frac{q^T B v}{\|v\|} > \beta^2 \|q\| \qquad \forall q \in \mathbb{R}^{N_T}, q \neq 0.$$

As already discussed in the case of the mixed FEM methods, the inf-sup condition must be checked on the particular couple of spaces $(\mathcal{V}_h, \mathcal{Q}_h)$. In other words, the two FEM spaces cannot be chosen independently of each other. Loosely speaking, we need to allow enough degrees of freedom in $\mathcal{W}_h$ with respect to $\mathcal{Q}_h$ so that we do not impose too many constraint on the saddle point problem. Again, the Fortin criterion comes to help. We have the following result:

Lemma 3.15 (Fortin criterion (for Stokes equation))**.** *If the bilinear form $b(v,q)$ defined in* (3.29) *satisfies the following inf-sup condition:*

$$\inf_{v \in \mathcal{H}^1(\Omega)} \sup_{q \in \mathcal{L}^2(\Omega)} \frac{b(v,q)}{\|v\|_{\mathcal{H}^1(\Omega)} \|q\|_{\mathcal{L}^2(\Omega)}} \geq \beta,$$

then the discrete inf-sup condition:

$$\inf_{q \in \mathcal{V}_h} \sup_{v \in \mathcal{Q}_h} \frac{b(v,q)}{\|v\|_{\mathcal{V}_h} \|q\|_{\mathcal{Q}_h}} \geq \beta,$$

is satisfied if and only if there exist a projection operator $\Pi_h : \mathcal{H}^2 \mapsto \mathcal{V}_h$ such that:

$$b(\Pi_h v, q) = b(v,q) \qquad \forall q \in \mathcal{Q}_h,$$

and:

$$\|\Pi_h v\|_{\mathcal{V}} \leq \gamma_h \|v\|_{\mathcal{V}} \qquad \forall v \in \mathcal{V}.$$

The proof is a simple extension of the proof of Lemma 3.3. Using this lemma we can find the compatibility conditions between the discrete spaces $\mathcal{V}$ and $\mathcal{Q}$, i.e., the relationships that guarantee the well posedness of the discrete problem. In essence, we need to use the solution v of the linear system

$$b(\Pi_h v, q) = b(v,q) \qquad \forall q \in \mathcal{Q}_h$$

to define the operator Π_h that satisfies Fortin's Lemma. This is a difficult task and it is often easier to verify it a posteriori (e.g., by counter examples).

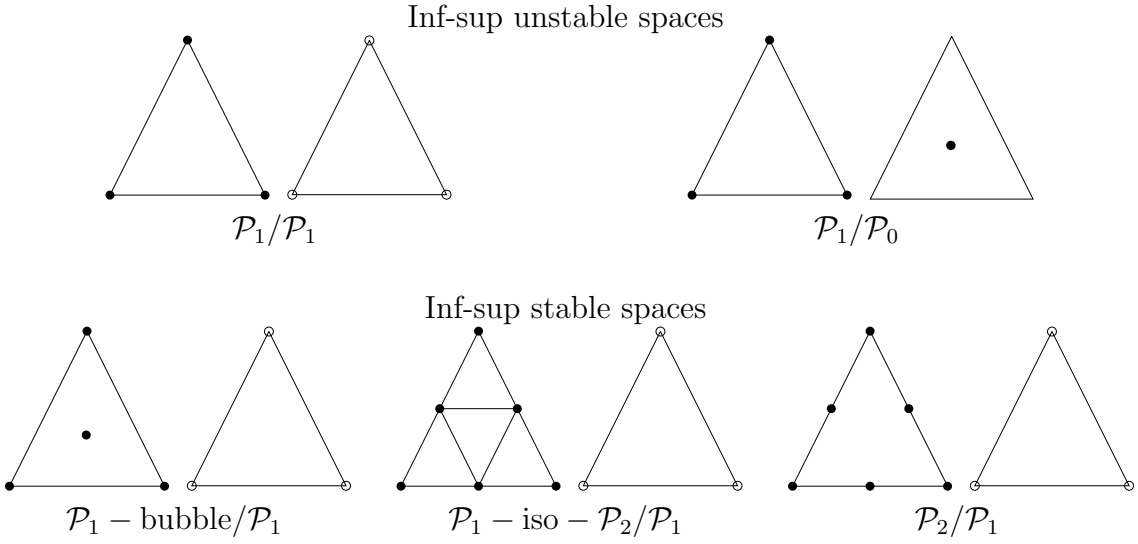


FIGURE 3.7: *Pairs of FEM Stokes spaces. Filled and empty circles indicate positions of velocity and pressure degrees of freedom, respectively. Each pair of triangles show velocity/pressure nodes. The top row displys the inf-sup unstable spaces with the pairs of triangles referring to $\mathcal{P}_1 - /\mathcal{P}_1$, $\mathcal{P}_1 - /\mathcal{P}_0$ spaces. The bottom row shows inf-sup stable spaces with the pairs of triangles referring to $\mathcal{P}_1 -$ bubble$/\mathcal{P}_1$, $\mathcal{P}_1 -$ iso $- \mathcal{P}_2/\mathcal{P}_1$, and $\mathcal{P}_2/\mathcal{P}_1$ (Taylor-Hood) spaces, respectively.*

Remark 3.10. *We note that the constant vector is in the null space of matrix B^T, stating that the pressure is defined up to a constant. In fact, if $p \in \mathbb{R}^N$ is a constant vector, i.e., $p_k = C$, $k = 1, \ldots, N$, we have:*

$$VB^T P = PBV = \left\{ \sum_{i=1}^{M} b\left(\phi_i, p\right) v_i \right\} = \left\{ C \sum_{i=1}^{M} b\left(\phi_i, p\right) v_i \right\} = \int_\Omega C \operatorname{div} v \, dx = C \int_\Gamma v \cdot \nu \, ds = 0$$

because v is zero on Γ.

Inf-sup unstable spaces

The $\mathcal{P}_1/\mathcal{P}_1$ spaces. The first pair of spaces we study is the so called $\mathcal{P}_1/\mathcal{P}_1$ spaces. These spaces are formed by piecewise linear basis functions discussed in Section 2.8.1 for both $\mathcal{V}_h$ and $\mathcal{Q}_h$. Thus, given a regular triangulation $\mathcal{T}_h(\Omega)$, we have:

$$\mathcal{V}_h(\mathcal{T}_h) = \{v \in [\mathcal{C}^0(\mathcal{T}_h)]^d : v|_T \in [\mathcal{P}_1(T)]^d \text{ for every } T \in \mathcal{T}_h, v = 0 \text{ in } \Gamma\},$$
$$\mathcal{Q}_h(\mathcal{T}_h) = \{q \in \mathcal{C}^0(\mathcal{T}_h) : q|_T \in \mathcal{P}_1(T) \text{ for every } T \in \mathcal{T}_h\}.$$

We discuss here a particular example. We take a square $\Omega =]0 \times 1[^2$ with a regular triangulation $\mathcal{T}_h$ formed by rectangular triangles. Take a discrete pressure field defined by nodal values -1, 0, 1 at the three vertices of the triangles. This pressure field has zero mean on each triangle, i.e., $\sum_{i=1}^{3} p_{j,T} = 0$. Since u_h is piecewise linear, its divergence is piecewise constant, and then obviously we have:

$$\int_\Omega p_h \operatorname{div} v \, dx = \sum T \in \mathcal{T}_h \int_T p_h \operatorname{div} v \, dx = \sum T \in \mathcal{T}_h \operatorname{div} v|_T \frac{|T|}{3} \sum_{j=1}^{3} p_j = 0 \qquad \forall v \in \mathcal{V}_h,$$

showing that the bilinear form $b(\cdot, \cdot)$ does not satisfy the inf-sup condition. This result is expected, as the space of the pressure is too rich with respect to the space of the velocity. Note that in this case the null space of B^T is larger than one (when $\operatorname{Ker} B^T = 1$ it is possible to remove the singularity by imposing a null average pressur), and the end effect on the solution is that, when an iterative method is used to solve the linear system, oscillations in the pressure occur.

The $\mathcal{P}_1/\mathcal{P}_0$ spaces. In this case the use of a piecewise constant pressure field implies the immediate satisfaction of the divergence constraint (in weak form). However, also this pair does not satisfy the inf-sup condition. In fact, indicating with N the number of vertices of $\mathcal{T}_h$, with $N = N_I + N_B$, where N_I and N_B are the number of internal and boundary nodes, respectively, and with M the number of triangles. Then, $\operatorname{Dim}(\mathcal{V}_h) = N_I$ and $\operatorname{Dim}(\mathcal{Q}_h) = M - 1$ (remember we need to add the constraint of zero mean to fix the pressure). Using the Euler characteristics of a triangulation of a polygonal domain it can be shown that $M = 2N_I + N_B - 2$, fromm which we deduce that $M - 1 \geq 2(N_I - 1)$. Hence:

$$\operatorname{Dim}\left(\operatorname{Ker} B^T\right) = \operatorname{Dim}(\mathcal{Q}_h) - \operatorname{Dim}\left(\operatorname{Im} B^T\right) \geq M - 1 - \operatorname{Dim}(\mathcal{V}_h) = M - 1 - 2N_I = N_B - 3.$$

Thus there are at least $N_B - 3$ spurious modes for the pressur. In other words, we are imposing at least $N_B - 3$ too many constraints in the divergence equation and B^T is not surjective.

Inf-sup stable spaces

The ($\mathcal{P}_1$−bubble/$\mathcal{P}_1$) (mini element) spaces. The idea is to enrich the $[\mathcal{P}1]^d$ for velocity so that it is sufficiently richer than the pressure space $\mathcal{P}1$. Following this idea, the mini element adds a degree of freedom in the center of gravity of the triangle. In two dimensions we have for the velocities the following:

$$\mathcal{P}_{h,1}^{\beta} = [\mathcal{P}_1(T) \oplus \operatorname{Span}(\beta_T)]^2,$$

where $\beta_T(x)$ is the so called bubble function taking on the value 1 at the gravity center of T and zero at the boundary and it is always $0 \leq \beta_T \leq 1$. Indicating with $\phi_i(x)$ the basis function

for the sdundard $\mathcal{P}_1$ linear FEM space for the reference triangle with vertices in $P_1 = (0,0)$, $P_2 = (1,0)$, and $P_3 = (0,1)$. The $\mathcal{P}_1$ basis functions and the bubble function take then the form:

$$\phi_1(x,y) = 1 - x - y; \quad \phi_2(x,y) = x; \quad \phi_3(x,y) = y; \quad \phi_\beta(x,y) = 27\phi_1(x,y)\phi_2(x,y)\phi_3(x,y).$$

The velocity vector and the pressure are given by:

$$u_h(x,y) = \sum_{i=1}^{3} u_i\phi_i(x,y) + u_\beta\phi_\beta(x,y) \qquad p_h(x,y) = \sum_{i=1}^{3} p_i\phi_i(x,y)$$

where u_i and u_β are two-dimensional vectors containing the x and y components of the nodal (baricentral) velocity vector.

The $\mathcal{P}_i - \mathbf{iso} - \mathcal{P}_2/\mathcal{P}_1$ space. This choice of spaces satisfying the inf-sup condition amounts essentially in choosing $\mathcal{P}_1$ basis functions for both pressure and velocity spaces. The enrichment of the latter space is obtained by using a uniformly refined triangulation obtained by connecting the midpoints of each triangle (see Figure 3.7).

The Taylor-Hood ($\mathcal{P}_k/\mathcal{P}_{k-1}$) spaces. The Taylor-Hood spaces consider both continuous velocity and continuous pressure fields and are of the type $\mathcal{P}_k/\mathcal{P}_{k-1}$, with $k \geq 2$ to satisfy the inf-sup condition. The simplest and most used approach is with $k = 2$, i.e., $\mathcal{P}_2/\mathcal{P}_1$, using quadratic velocity and linear pressure.

3.5.2 Stabilized FEM discretizations

The mixed FEM described in the previous section require the use of different orders of approximation for velocity and pressure. In practical applications, especially when chemically reactive flows are considered, different approximating polynomials may lead to both theoretical and implementation difficulties, as often chemical reactions require the use of derived (interpolated) quantities. In these cases, equal order polynomials are much more beneficial. However, we have seen that in order to being able to solve the saddle point problem arising from the discretization of the Stokes equation we need the satisfaction of the inf-sup condition. The essence of this condition is to make sure that the linear system is not under- or over-constrained, and thus can be solved. Looking at the expression of the saddle-point linear system (3.23), it is intuitive to think that the empty (2,2) block is critical. If we can replace this (2,2) block with an invertible matrix, then the linear system is always solvable as long as matrix A (the (1,1) block) is invertible. The idea of stabilization is to introduce a "consistent" variational crime into the formulation to generate a invertible matrix in the (2,2) blocks.
For example, we can define the following problem:

Problem 3.11 (Stokes Galerkin GLS (satbilized) formulation)**.** Find $(u_h, p_h) \in \mathcal{V}_h(\mathcal{T}_h) \times \mathcal{Q}_h(\mathcal{T}_h)$ such that:

$$\begin{aligned} &a\left(u_h, v\right) + b\left(p_h, v\right) = F\left(v\right) \\ &b\left(u_h, q\right) + c\left(\left(u_h, p_h\right), \left(v, q\right)\right) = 0 \end{aligned} \qquad \forall v \in \mathcal{V}_h(\mathcal{T}_h), \quad q \in \mathcal{Q}_h(\mathcal{T}_h)$$

with the new bilinear form given by:

$$c\left(\left(u_h, p_h\right), \left(v, q\right)\right) = \delta \sum_{T \in \mathcal{T}_h} h_T^2 \int_T (-\mu \Delta u_h + \nabla p_h - f)(-\mu \Delta v + \nabla q)\, dx$$

where δ is an empirical parameter that tunes the amount of stabilization. The above equation is consistent with Stokes variational formulation (3.28) (with $g = 0$) as the term in $c\left(\cdot, \cdot\right)$ is a residual that is equal to zero when we choose $u_h = u$. Thus the consistency of the scheme is preserved.
Using linear basis functions for both velocity and pressure, $\phi_i(x)$ and $\psi_k(x)$ are vector and scalar functions, respectively, that form the bases for $\mathcal{V}_h$ and $\mathcal{Q}_h$, i.e.:

$$\begin{aligned} \mathcal{V}_h(\mathcal{T}_h) &= \left\{v \in \mathcal{C}^0(\Omega) : v|_T \in \left[\mathcal{P}_1(T)\right]^2\right\} = \mathrm{Span}\left(\phi_1, \ldots, \phi_N\right) \subset \left[\mathcal{H}_0^1(\Omega)\right]^2, \\ \mathcal{Q}_h(\mathcal{T}_h) &= \left\{w \in \mathcal{C}^0(\Omega) : w|_T \in \mathcal{P}_1(T)\right\} = \mathrm{Span}\left(\psi_1, \ldots, \psi_N\right) \subset \mathcal{H}_0^1(\Omega). \end{aligned}$$

This stabilized scheme leads to the linear system:

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

with the matrix elements defined as:

$$\begin{aligned} A = \{a_{ij}\} \quad & a_{ij} = a\left(\phi_i, \phi_j\right) = \mu \int_\Omega \nabla \phi_i : \nabla \phi_j \, dx && i, j = 1, \ldots, 2N \\ B = \{b_{ki}\} \quad & b_{ki} = b\left(\phi_i, \psi_k\right) = -\int_\Omega \psi_k \,\mathrm{div}\, \phi_i \, dx && k = 1, \ldots, N; \quad i = 1, \ldots, 2N \\ C = \{c_{km}\} \quad & c_{ki} = c\left(\psi_k, \psi_m\right) = \delta \sum_{T \in \mathcal{T}_h} \int_T \nabla \psi_k \cdot \nabla \psi_m \, dx && k, m = 1, \ldots, N \\ f = \{f_i\} \quad & f_i = -\int_\Omega f \cdot \phi_i \, dx && i = 1, \ldots, 2N \\ g = \{g_k\} \quad & g_k = -\delta \sum_{T \in \mathcal{T}_h} \int_T f \cdot \nabla \psi_k \, dx && k = 1, \ldots, N \end{aligned}$$

4 Finite Volume Methods

In this section we work the details of the Finite Volume family of schemes for the solution of elliptic/parabolic conservation equations.
We will be using most of the theoretical developments described in the previous sections as fundamental building blocks for our numerical approach. We will focus on standard finite volume schemes maintaining and discussing as much as possible multidimensional problems. However, as truly multidimensional methods are still open questions, we will spend much time looking at one-dimensional approximations, without forgetting multidimensional approximations. For the more recent developments we refer the reader to more specialized books.

4.1 The Differential Equation

Problem 4.1. Given a domain $\Omega \in \mathbb{R}^d$, which we assume possesses sufficient regularity, we want to find a vector function $u(x,t) : \Omega \times \mathbb{R}_+ \longrightarrow \mathbb{R}$, that satisfies the following equation:

$$u_t + \operatorname{div} f(u) = 0 \qquad \text{for all } x \in \Omega \text{ and } t \in [0,T], \tag{4.1}$$

$$u(x,0) = u_0 \qquad \text{for all } x \in \Omega \text{ and } t = 0, \tag{4.2}$$

$$u(x,t) = u_d \qquad \text{for all } x \in \partial\Omega_{in} \text{ and } t \in [0,T], \tag{4.3}$$

where $f : \Omega \longrightarrow \mathbb{R}$.

We start by giving some basic definitions of the domain discretizations and its property and define the geometric quantities that will be needed in the development.

4.2 Preliminaries

There are different equivalent ways to develop the Finite Volume Method. Our approach is to rely mainly on Gauss (or Divergence) theorem and the derived integration by parts (Green's Lemma) in order to always focus on the conservation principle that is the foundation of the models of interest. We recall here the most important mathematical results that will be of use in our developments, reminding to standard analysis books for the details of the proofs.

Theorem 4.1 (Gauss or Divergence Theorem)**.** *Given a subset $\Omega \in \mathbb{R}^d$ having piecewise smooth boundary $\Gamma = \partial\Omega$, and a continuously differentiable ($\mathcal{C}^1(\Omega)$) vector field $F(x) \in \Omega$, we have:*

$$\int_\Omega \operatorname{div} F \, dx = \int_\Gamma F \cdot \nu \, ds, \tag{4.4}$$

where ν is the outward unit normal to Γ, dx denotes the volume (surface) measure on Ω and ds the surface measure on Γ and $v \cdot w = \langle v, w \rangle$ denotes the scalar product between vectors v, w of $\mathbb{R}^d$.

Theorem 4.2 (First Green Identity – Green's Lemma)**.** *Let $v, w \in \mathbb{R}^d$ be piecewise continuous vector fields and let Ω and Γ as in the previous theorem, then:*

$$\int_\Omega \nabla v \cdot \nabla w \, dx = \int_\Gamma v \, \nabla w \cdot \nu \, ds - \int_\Omega v \, \operatorname{div} \nabla w \, dx, \tag{4.5}$$

where $\operatorname{div} \nabla = \Delta$ is Laplace differential operator of second derivatives.

Remark 4.2. *In (4.4) the left-hand-side contains the sum of the partial derivatives of the vector field F, and for this reason the hypothesis of the theorem contains the requirement that F be continuously differentiable. The right-hand-side, on the other hand, does not contain any derivative, and thus in principle the integral of the fluxes over the subset Ω could be defined without the requirement that F be $\mathcal{C}^1$. However, the normal to the surface Γ must be well-defined and this is the reason for the requirement that Γ be piecewise smooth. In fact, if the boundary Γ is formed by the union of m smooth surfaces that intersect at boundaries that form $\mathcal{C}^1$ curves, i.e.:*

$$\Gamma = \bigcup_{i=1}^{m} \Gamma_i \qquad \gamma_{ij}(t) = \Gamma_i \cap \Gamma_j \in \mathcal{C}^1 \qquad \text{for all } i \text{ and } j, \qquad \text{then } \int_\Gamma = \sum_{i=1}^{m} \int_{\Gamma_i}.$$

The same argumentation can be made, with the appropriate changes, for Green's Lemma 4.5.

The Finite Volume (FV) scheme can be derived by the following operations:

1. partition the domain Ω into M polygonal "finite volumes" or cells T_i, $i = 1, \ldots, M$ (see next paragraph for a complete definition of cells);

2. integrate equation (4.1) in time and space over the domain Ω and the time interval $[t^k, t^{k+1}]$:

$$\int_{t^k}^{t^{k+1}} \left(\int_\Omega u_t + \operatorname{div} f(u) \, dx \right) \, dt = 0;$$

3. use the linearity property of the integration to write:

$$\int_{t^k}^{t^{k+1}} \left(\int_\Omega u_t + \operatorname{div} f(u) \, dx \right) \, dt = \sum_{i=1}^{M} \int_{t^k}^{t^{k+1}} \left(\int_{T_i} u_t + \operatorname{div} f(u) \, dx \right) \, dt = 0;$$

impose that each term of the sum is zero:

$$\int_{t^k}^{t^{k+1}} \int_{T_i} (u_t + \operatorname{div} f(u) \, dx) \, dt = 0 \qquad i = 1, \ldots, M;$$

4. apply the divergence theorem and use the fact that the cells are of polygonal shape:

$$\int_{t^k}^{t^{k+1}} \left(\int_{T_i} u_t \, dx + \int_{\partial T_i} f(u) \cdot \nu \, dx \right) \, dt = \int_{t^k}^{t^{k+1}} \left(\int_{T_i} u_t \, dx + \sum_{j=1}^{N_\sigma} \int_{\sigma_{ij}} f(u) \cdot \nu_j \, ds \right) \, dt = 0$$

5. exchange the first space integral with the time integral (the functions are assumed to be continuous and the domain of integration is assumed to be constant):

$$\int_{T_i} \int_{t^k}^{t^{k+1}} u_t \, dt \, dx + \int_{t^k}^{t^{k+1}} \left(\sum_{j=1}^{N_\sigma} \int_{\sigma_{ij}} f(u) \cdot \nu_j \, ds \right) \, dt = 0 \qquad i = 1, \ldots, M;$$

6. integrate the first addendum in time:

$$\int_{T_i} u(x, t^{k+1}) \, dx - \int_{T_i} u(x, t^k) \, dx + \int_{t^k}^{t^{k+1}} \left(\sum_{j=1}^{N_\sigma} \int_{\sigma_{ij}} f(u) \cdot \nu_j \, ds \right) \, dt = 0 \qquad i = 1, \ldots, M; \tag{4.6}$$

7. define cell average and the edge flux operators as:

$$\mathcal{A}_T(u(t)) = \frac{1}{\mid T \mid} \int_T u(x,t) \, dx, \tag{4.7}$$

$$\mathcal{G}_\sigma(u(t)) = \frac{1}{\mid \sigma \mid} \int_\sigma f(u(x,t)) \cdot \nu \, ds; \tag{4.8}$$

8. to obtain:

$$\mathcal{A}_{T_i}(u(t^{k+1})) = \mathcal{A}_{T_i}(u(t^k)) - \frac{1}{\mid T_i \mid} \int_{t^k}^{t^{k+1}} \sum_{j=1}^{N_\sigma} \mid \sigma_{ij} \mid \mathcal{G}_{\sigma_j}(u(t)) \, dt = 0 \qquad i = 1, \ldots, M;$$

   we remark that until now we have made no numerical approximations;

9. start the numerical approximation very naturally by approximating the cell average $u_{h,i} \approx \mathcal{A}_{T_i}(u) = \mathcal{A}_h(u)$ and use u_h as unknown of our numerial scheme together with a simple quadrature rule (e.g. left rectangles) to evaluate the remaining time integral to obtain:

$$u_{h,i}^{k+1} = u_{h,i}^k - \frac{\Delta t}{\mid T_i \mid} \sum_{j=1}^{N_\sigma} \mid \sigma_{ij} \mid \mathsf{G}_{h,j}^k = 0 \qquad i = 1, \ldots, M, \tag{4.9}$$

   where $\mathsf{G}_{h,j}^k$ is the numerical approximation of the flux $\mathcal{G}_\sigma(u(t^k))$ at the cell face σ_{ij}.

Remark 4.3. *The first few points of the derivation could have been replaced by a derivation more aware of the definition of a weak formulation given in the previous sections. To this aim we proceed as follows:*

1. *partition the domain Ω into M polygonal "finite volumes" or cells T_i, $i = 1, \ldots, M$ (see next paragraph for a complete definition of cells); define a piecewise smooth test function $\phi(x)$ given by the characteristic function of the i-th cell:*

$$\phi_i(x,t) = \chi_x(x)\chi_t(t), \text{ where } \chi_x(x) = \begin{cases} 1 & \text{if } x \in T_i \\ 0 & \text{otherwise;} \end{cases} \qquad \chi_t(t) = \begin{cases} 1 & \text{if } t \in [t^k, t^{k+1}] \\ 0 & \text{otherwise;} \end{cases}$$

2. *multiply equation* (4.1) *and integrate over the domain Ω and the interval $[t^k, t^{k+1}]$:*

$$\int_{t^k}^{t^{k+1}} \left(\int_\Omega (u_t + \operatorname{div} f(u))\, \phi_i(x,t)\, dx \right) dt = 0;$$

this equation must be satisfied for all functions $\phi_i(x,t)$:

$$\int_{t^k}^{t^{k+1}} \left(\int_{T_i} u_t + \operatorname{div} f(u)\, dx \right) dt = 0, \qquad i = 1, \ldots, M,$$

which is exactly equation (4.6).

At ths point we are left with the task of defining the numerical flux $\mathsf{G}_{h,j}^k$ for each mesh edge. Before going into the development of how to evaluate the numerical flux we need to setup some notation. We would like to stress here that this general Finite Volume setting is what is known as "cell-based". We could derive a "node-based" version more similar to the approach used in the Finite Element Method without adding any complication. Within the same framework we could tackle more complicated PDEs, as for example parabolic or elliptic equations, obtained for example by adding to (4.1) a diffusion term proportional to the second derivatives of $u(x,t)$. For a general discussion on these topics we refer the reader to the specialized Finite Volume literature [8, e.g.] for a complete mathematical theory and to [16] for application to Computational Fluid Dynamics.

4.2.1 Notations

Geometrically we define the mesh $\mathcal{T}_h(\Omega)$ as a finite collection of non-overlapping and non-empty two-dimensional "control volumes" or "cells" generally formed by simplices (e.g, subintervals, triangles, tetrahedra in one-, two-, and three-dimensions, respectively) and denoted with the letter "T" indexed by a Latin subscript, e.g. i, j, k. For example, T_i is the i-th control volume (cell) of the mesh $\mathcal{T} = \{T_i\}$, $i = 1, \ldots, M$, with M being the total number of cells. We assume

that, for every possible choice of h, $\mathcal{T}_h$ covers the domain $\Omega \subset \mathbb{R}^d$ in the sense that for all i and j:

$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}} T, \qquad \text{and} \qquad T_i \cap T_j = \begin{cases} \sigma_{ij} \subset \mathbb{R}^{d-1} & \text{if } T_i \text{ and } T_j, \text{ are neighbors,} \\ \emptyset, & \text{otherwise .} \end{cases}$$

We identify with the symbol σ a mesh face, i.e., the intersection between two neighboring cells, and index it with the indices of the two cells:

$$\sigma_{ij} = T_i \cap T_j,$$

and with the symbol ξ the vertices forming the cells. In a two-dimensional example, two distinct cells are either neighbors, in which case their intersections is the common boundary "edge", or they are far apart, in which case their intersection is empty. Cells, faces, and vertices are identified by global numbers and all are counted only once. Figure 3.2 reports typical efficient data structures that can be used to completely describe the mesh and corresponding quantities in a computer program. For more details consult [18].
Different meshes are parameterized by h, called the "mesh parameter" and defined as the maximum face measure, and by the "mesh diameter" ρ, i.e., the maximum diameter of the circles inscribed in each mesh cell. Letting h_T and ρ_T being the maximum face measure and inscribed circle diameter for the generic cell T we have then:

$$h = \max_{T \in \mathcal{T}} h_T$$
$$\rho = \max_{T \in \mathcal{T}} \rho_T.$$

We require the triangulation to be "regular", i.e., there exists a constant $\beta > 0$ independent from h and ρ and from the triangulation $\mathcal{T}$ such that:

$$\frac{\rho_T}{h_T} \geq \beta \qquad \text{for all } T \in \mathcal{T}.$$

The value of β is a measure of the minimum angle between two consecutive cell edges. The assumption that the triangulation is regular implies that cell angles do not become too small in the limiting process as $h \longrightarrow 0$, so that the problem of numerical interpolation of nodal or edge values is well-posed.

5 Parabolic equations

Consider the model problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \operatorname{div}\left[K(x)\nabla u\right] + f(x,t) && t \in [0,T];\ \ x \in \Omega \subset \mathbb{R}^d, \\ u(x,0) &= u_0(x) && t = 0,\ x \in \Omega, \\ u(x,t) &= u_D(x,t) && t \in [0,T],\ x \in \Gamma_D \subset \partial\Omega, \\ -K(x)\nabla u(x,t)\cdot\nu &= q_N(x,t) && t \in [0,T],\ x \in \Gamma_N \subset \partial\Omega. \end{aligned} \tag{5.1}$$

From a physical point of view, we can interpret this problem as governing the heat flow in a solid characterized by a thermal conductivity given by the function $K(x)$, assumed strictly greater than zero a.e., to guarantee coercivity. We have thus an extra independent variable, time t, which is assumed to vary within an interval $I \subset \mathbb{R}^+$.

5.1 One-dimensional model problem

An intuitive look at the behavior of the solution of eq. (5.1) in a simplified problem is obtained by developing an explicit form of $u(x,t)$. Consider the one-dimensional (in space) problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} && x \in (0,\pi),\ \ t > 0; \\ u(x,0) &= u_0(x) && x \in (0,\pi), \\ u(0,t) &= u(\pi,t) = 0 && t > 0. \end{aligned}$$

The problem is periodic and thus we can use effectively the Fourier transform together with the technique of variable separation to obtain the solution. By simple substitution, it is easy to verify that the following function satisfies the above equation:

$$u(x,t) = \sum_{j=1}^{\infty} u_{0,j}\, \mathrm{e}^{-j^2 t} \sin(jx), \qquad u_{0,j} = \sqrt{2/\pi} \int_0^{\pi} u_0(x) \sin(jx)\, dx, \qquad j = 1, 2, \ldots.$$

where $u_{0,j}$ are the Fourier coefficients of the initial datum $u_0(x)$ expressed in the basis (orthonormal in $\mathcal{L}^2((0,\pi)))$ $\{\sqrt{2/\pi}\ \sin(jx)\}_{j=1}^{\infty}$. The frequency component j (related to the spatial basis $sin(jx)$) is characterized by a temporal scale varying on the order of $\mathcal{O}\left(j^{-2}\right)$. Varying j we have thus a continuous spectrum of components that decay faster and faster in time. As a consequence, the solution becomes more regular as time progresses. Intuitively this is exactly what we would expect from a diffusion-like process. However, for small times, the solution is not necessarily smooth and it may happen that $\|\dot{u}(t)\| = \|\dot{u}(\cdot,t)\|_{\mathcal{L}^2((0,\pi))} \longrightarrow \infty$ for $t \longrightarrow 0$ depending on the initial condition $u(x,0)$. For example, take $u(x,0) = u_0(x) = \pi - x$, then, for an appropriate constant C, we obtain:

$$u_{0,j} = \sqrt{2/\pi} \int_0^{\pi} (\pi - x) \sin(jx)\, dx = \sqrt{2/\pi} \frac{j\pi - \sin(j\pi)}{j^2} = \sqrt{2/\pi} \frac{\pi}{j} = C/j.$$

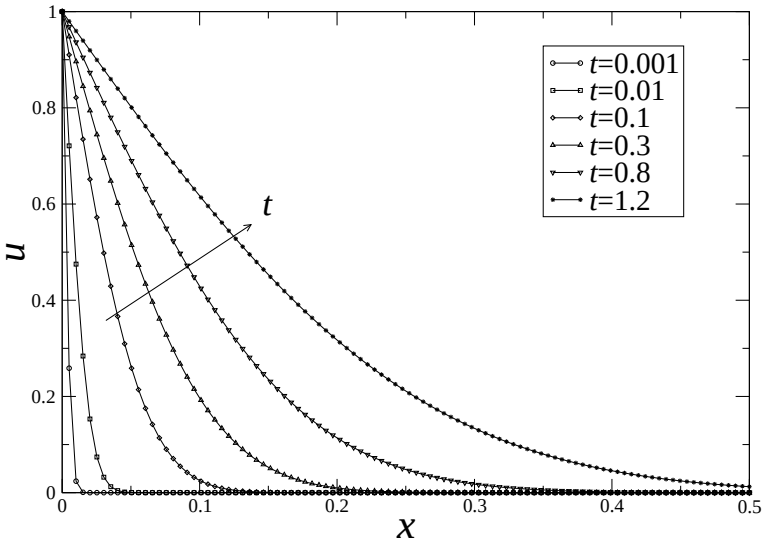


FIGURE 5.1: *Solution of 1D diffusion problem of Example 5.1 as a function of space for different increasing times. The initial steep transient is clearly visible, as well as the smoothing property as time increases.*

For $t \longrightarrow 0$ we have that $\|\dot{u}(t)\| \approx Ct^{-\alpha}$ with $\alpha = 3/4$. Now take $u(x,0) = u_0(x) = \min(x, \pi - x)$, we find:

$$u_{0,j} = \sqrt{2/\pi} \int_0^\pi \min(x, \pi - x) \sin(jx)\, dx = \sqrt{2/\pi} \frac{2\sin(j\pi/2) - \sin(j\pi)}{j^2} = C/j^2,$$

and $\|\dot{u}(t)\| \approx Ct^{-\alpha}$ with $\alpha = 1/4$. In general, if $u_{0,j}$ decays faster than $j^{-2.5}$ for $j \longrightarrow \infty$, then $\|\dot{u}(t)\|$ is bounded for $t \longrightarrow 0$.
The solution will always have a more or less important initial transient where some derivatives may be non smooth. At large enough times, however, the solution will regularize. Note that the presence of time-varying forcing functions may generate important transients also far from the initial time. Some a priory stability estimates can be shown using the energy methods of the next chapter or Parseval inequality. We have:

$$\|u(t)\| \leq \|u_0\|, \qquad t \in (0,T) \tag{5.2}$$

$$\|\dot{u}(t)\| \leq \frac{C}{t} \|u_0\|, \qquad t \in (0,T). \tag{5.3}$$

Example 5.1 (1D diffusion on the half line)**.** Consider the following 1D example:

$$\begin{aligned} \frac{\partial u}{\partial t} &= D\frac{\partial^2 u}{\partial x^2} && x \in (0,\infty)\,; \\ u(0,t) &= u_0 && \forall t; \\ u(\infty,t) &= 0 && \forall t; \\ u(x,0) &= \begin{cases} u_0, & \text{if } x = 0, \\ 0, & \text{if } x \in (0,\infty). \end{cases} \end{aligned}$$

This equation admits an explicit solution given by:

$$u(x,t) = u_0 \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right),$$

where $\text{erfc}(x)$ is the complementary error function. Figure 5.1 shows the solution for $u_o = 1$ at different times. The properties described above are clearly noticeable in this case.

5.2 Variational formulation

We can write now the variational formulation using the separation of variables, assuming for simplicity homogeneous Dirichlet conditions, although the discussion presented in section 2.3.3 applies. We are looking for a function u:

$$\begin{aligned} u : \Omega \times (0,t) &\longrightarrow \mathbb{R} \\ (x,t) &\mapsto u(x,t), \end{aligned}$$

that satisfies the above problem (5.1), assuming $\Gamma_D = \Gamma$; $\Gamma_N = \emptyset$, and $u_D = 0$. We actually can define a mapping

$$\begin{aligned} u : (0,T) &\longrightarrow \mathcal{H}_0^1(\Omega), \\ t &\mapsto u(\cdot,t) \end{aligned}$$

intending with this notation that for a.e. $t \in (0,T)$ the functions $u(t) \in \mathcal{H}_0^1(\Omega)$ and $u'(t) \in \mathcal{H}^{-1}(\Omega)$ (the dual space of $\mathcal{H}_0^1(\Omega)$) are defined as::

$$u(t) := u(x,t), \qquad\qquad u'(t) := \partial_t u(x,t).$$

Thus we can say:

$$u \in \mathcal{L}^2(0,T;\mathcal{H}_0^1(\Omega)) \text{ and } u' \in \mathcal{L}^2(0,T;\mathcal{H}^{-1}(\Omega)),$$

such that u' is the (time) derivative of u in the sense of distributions, i.e., as usual by means of integration by parts, we have that for every $v \in \mathcal{H}_0^1(\Omega)$ and $\phi \in \mathcal{C}_c^\infty(0,T)$:

$$\int_0^T (u'(t),v)_{\mathcal{H}_0^1}\,\phi(t)\,dt = -\int_0^T (u(t),v)_{\mathcal{H}_0^1}\,\phi'(t)\,dt.$$

The weak formulation can be written as:

Problem 5.2. Find $u \in \mathcal{L}^2(0,T;\mathcal{H}_0^1(\Omega))$ and $u' \in \mathcal{L}^2(0,T;\mathcal{H}^{-1}(\Omega))$ such that

$$(u',v) + a\,(u,v) = F\,(v) \qquad \forall v \in \mathcal{H}_0^1(\Omega)$$

where

$$(u',v) = \int_\Omega u'\ v\ dx; \qquad a\,(u,v) = \int_\Omega K(x)\nabla u \nabla v\ dx; \qquad F\,(v) = (f,v)\,.$$

We note that:

$$(u',v) = \left(\frac{du}{dt}, v\right) = \frac{d}{dt}\,(u,v)\,,$$

that the above variational formulation may be written as:

$$\frac{d}{dt}\,(u,v) + a\,(u,v) = (f,v)\,.$$

which, for $v = u(t)$, using the chain rule, becomes:

$$\frac{1}{2}\frac{d}{dt}\,\|u\|^2 + a\,(u,u) = (f,u)\,.$$

Again with $v = u(t)$, using the coercivity of the bilinear form $a\,(\cdot,\cdot)$ (eq. (2.57)), Poincaré inequality (Lemma 2.6), and Cauchy-Schwartz inequality, we obtain the following chain of inequalities:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\,\|u(t)\|^2_{\mathcal{L}^2(\Omega)} + \alpha\,\|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{1}{2}\frac{d}{dt}\,\|u(t)\|^2 + \alpha\left(\|u(t)\|^2_{\mathcal{L}^2(\Omega)} + \|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)}\right)\\
&\leq \frac{1}{2}\frac{d}{dt}\,\|u(t)\|^2 + a\,(u,u) = (f,u) \leq \|f(t)\|_{\mathcal{L}^2(\Omega)}\,\|u(t)\|_{\mathcal{L}^2(\Omega)} \leq C_\Omega\,\|f(t)\|_{\mathcal{L}^2(\Omega)}\,\|\nabla u(t)\|_{\mathcal{L}^2(\Omega)}\,.
\end{aligned}$$

We use now Young's inequality, which can be stated as follows: for every real scalars η and ξ, and every $\epsilon > 0$ we have that:

$$\eta\xi \leq \epsilon\eta^2 + \frac{1}{4\epsilon}\xi^2,$$

with $\epsilon = 1/2\alpha$ to obtain:

$$\frac{1}{2}\frac{d}{dt}\,\|u(t)\|^2_{\mathcal{L}^2(\Omega)} + \alpha\,\|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{C_\Omega^2}{2\alpha}\,\|f(t)\|^2_{\mathcal{L}^2(\Omega)} + \frac{\alpha}{2}\,\|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)}\,,$$

and finally, after simplification:

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

Integration between 0 and t yields:

$$\|u(t)\|^2_{\mathcal{L}^2(\Omega)} + \alpha \int_0^t \|\nabla u(\tau)\|^2_{\mathcal{L}^2(\Omega)} \, d\tau \leq \|u(0)\|^2_{\mathcal{L}^2(\Omega)} + \frac{C_\Omega^2}{\alpha} \int_0^t \|f(\tau)\|^2_{\mathcal{L}^2(\Omega)} \, d\tau.$$

The term on the left-hand side represents the total energy of the system at time t. Intuitively, this energy must be smaller than the initial energy, i.e., the sum of the energy due to u_0 and f. Note that for $\alpha = 1$, the left hand side is exactly the square of the norm of $u(t)$ in $\mathcal{H}^1(\Omega)$. For $f = 0$ we have eq. (5.2).
Another a priori estimate can be obtained by observing that:

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

Then, we can write:

$$\begin{aligned}
&\|u(t)\|_{\mathcal{L}^2(\Omega)} \frac{d}{dt}\|u(t)\|_{\mathcal{L}^2(\Omega)} + \frac{\alpha}{C_\Omega}\|u(t)\|_{\mathcal{L}^2(\Omega)} \|\nabla u(t)\|_{\mathcal{L}^2(\Omega)} \\
&\leq \|u(t)\|_{\mathcal{L}^2(\Omega)} \frac{d}{dt}\|u(t)\|_{\mathcal{L}^2(\Omega)} + \alpha \|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)} \\
&\leq \|u(t)\|_{\mathcal{L}^2(\Omega)} \frac{d}{dt}\|u(t)\|_{\mathcal{L}^2(\Omega)} + \alpha \left(\|u(t)\|_{\mathcal{L}^2(\Omega)} + \|\nabla u(t)\|^2_{\mathcal{L}^2(\Omega)}\right) \\
&\leq \|u(t)\|_{\mathcal{L}^2(\Omega)} \frac{d}{dt}\|u(t)\|_{\mathcal{L}^2(\Omega)} + a\left(u(t), u(t)\right) \\
&= (f, u(t)) \leq \|f\|_{\mathcal{L}^2(\Omega)} \|u(t)\|_{\mathcal{L}^2(\Omega)}
\end{aligned}$$

Assuming $\|u(t)\|_{\mathcal{L}^2(\Omega)} \neq 0$, we have easily:

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

and after integration between 0 and t:

$$\|u(t)\|_{\mathcal{L}^2(\Omega)} \leq \|u(o)\|_{\mathcal{L}^2(\Omega)} + \int_0^t \|f(\tau)\|_{\mathcal{L}^2(\Omega)} \, d\tau,$$

from which, in the case $f = 0$, again we obtain the estimate (5.2).

5.3 FEM formulation

We consider a FEM formulation by using Lagrangian basis functions and apply the Method of Lines (MOL) [17]. Thus, for any $t > 0$, we look for $u_h(t) \in \mathcal{V}_h$ such that:

$$\begin{aligned}
&\left(\frac{du_h(t)}{dt}, v\right) + a\left(u_h, v\right) = (f, v) && \forall v \in \mathcal{V}_h, \\
&\left(u_h(0), v\right) = \left(u_0, v\right).
\end{aligned} \tag{5.4}$$

Separating variables x and t, we can write u_h using the set of Lagrangian basis functions $\{\phi_i\}$ of $\mathcal{V}_h$ as:

$$u_h(t,x) = \sum_{i=1}^{N} u_j(t)\phi_j(x),$$

to yield:

$$\sum_{j=1}^{N} \frac{du_j(t)}{dt}\left(\phi_j,\phi_i\right) + \sum_{j=1}^{N} u_j a\left(\phi_j,\phi_i\right) = \left(f(t),\phi_i\right) \qquad i = 1,\ldots,N.$$

This is an $N \times N$ system of ODE that can be written in compact form:

$$P\dot{u} + Au = b, \tag{5.5}$$

where vector $u = \{u_i(t)\}$ collects all coefficients of $u_h(x,t)$, the mass matrix P has elements given by $p_{i,j} = (\phi_i,\phi_j)$, the stiffness matrix A is as before $a_{ij} = a\left(\phi_i,\phi_j\right)$, the right-hand-side vector has components $b_i = (f(t),\phi_i)$ and the initial solution is projected onto $\mathcal{V}_h$, i.e., $u_i = (u_0,\phi_i)$. Both matrices, P and A, are symmetric and positive definite. Thus, formally, we can invert $P = E^T E$ to yield:

$$\dot{\eta}(t) + \tilde{A}\eta(t) = g(t), \qquad \eta(0) = \eta_0. \tag{5.6}$$

where matrix $\tilde{A} = E^{-T}AE^{-1}$ is symmetric and positive definite and with spectral condition number $\kappa(\tilde{A}) = \mathcal{O}\left(h^{-1}\right)$. The formal ($\mathcal{C}^0$-semigroup) mild solution to this equation is given by:

$$\eta(t) = \mathrm{e}^{-\tilde{A}t}\,\eta_0 + \int_0^t \mathrm{e}^{-\tilde{A}(t-\tau)}\,g(\tau)\,d\tau. \tag{5.7}$$

The system of ODEs given by (5.5) is "stiff", i.e., the eigenvalues of A vary in a large interval, as shown by the fact that $\kappa(\tilde{A})$ is large.
The a priori energy estimates shown for above for the continuous case can be extended to the semi-discrete system (5.4). We have:

$$\|u_h(t)\|^2_{\mathcal{L}^2(\Omega)} + \alpha\int_0^2 \|\nabla u_h(\tau)\|^2_{\mathcal{L}^2(\Omega)}\,d\tau \leq \|u_{0,h}\|^2_{\mathcal{L}^2(\Omega)} + \frac{C^2}{\alpha}\int_0^t \|f(\tau)\|^2_{\mathcal{L}^2(\Omega)}\,d\tau, \tag{5.8}$$

or:

$$\|u_h(t)\|_{\mathcal{L}^2(\Omega)} \leq \|u_{0,h}\|_{\mathcal{L}^2(\Omega)} + \int_0^t \|f(\tau)\|_{\mathcal{L}^2(\Omega)}\,d\tau, \tag{5.9}$$

The semi-discrete problem can be studied in more details in the case of a polygonal domain Ω and using $\mathcal{P}_1$ basis functions (we assume for simplicity $K(x) = 1$). Let $\mathcal{T}_h(\Omega)$ a regular triangulation with mesh parameter h. Then we have:

Theorem 5.1. *Under the above hypothesis, there exists a constant C such that:*

$$\max_{t\in(0,T)} \|u(t) - u_h(t)\|_{\mathcal{L}^2(\Omega)} \leq C\left(1 + |\log(\tfrac{T}{h^2})|\right) \max_{t\in(0,T)} h^2 \|u(t)\|_{H^2(\Omega)},$$

where u is solution of (5.1) and u_h is solution of (5.4).

Proof. We define the following auxiliary problem. Given $t \in (0,T)$, let $\varphi_h : (0,t) \mapsto \mathcal{V}_h$ a function that satisfies:

$$\begin{aligned} &-(\dot{\varphi}_h(\tau), v) + a(\varphi_h(\tau), v) = 0 \qquad\qquad \forall v \in \mathcal{V}_h, \tau \in (0,t), \\ &\varphi_h(t) = e_h(t), \end{aligned} \tag{5.10}$$

where $e_h(\tau) = u_h(\tau) - \tilde{u}_h(\tau)$ and $\tilde{u}_h(\tau)$ is such that:

$$a(u(\tau) - \tilde{u}_h(\tau), v) = 0 \qquad \forall v \in \mathcal{V}_h, \tau \in (0,T).$$

Taking $v = e_h(\tau)$ in the first of (5.10), and setting $\epsilon(\tau) = u(\tau) - \tilde{u}_h(\tau)$, we obtain:

$$\begin{aligned} \|e_h(t)\|^2 &= \int_0^t \left[-(\dot{\varphi}_h(\tau), e_h(\tau)) + a(\varphi_h(\tau), e_h(\tau))\right] d\tau + (\varphi_h(t), e_h(t)) \\ &= \int_0^t \left[(\varphi_h(\tau), \dot{e}_h(\tau)) + a(\varphi_h(\tau), e_h(\tau))\right] d\tau + (\varphi_h(0), e_h(0)) \\ &= \int_0^t \left[(\varphi_h(\tau), \dot{\epsilon}_h(\tau)) + a(\varphi_h(\tau), \epsilon_h(\tau))\right] d\tau + (\varphi_h(0), \epsilon_h(0)) \\ &= -\int_0^t (\dot{\varphi}_h(\tau), \epsilon_h(\tau)) \, d\tau + (\varphi_h(t), \epsilon_h(t)), \end{aligned}$$

from which:

$$\|e_h(t)\|_{\mathcal{L}^2(\Omega)} \leq -\int_0^t (\epsilon_h(\tau), \dot{\varphi}(\tau)) \, ds + (\epsilon_h(\tau), \varphi_h(t)).$$

Using the formal solution (5.7) applied to the auxiliary problem (5.10) and using the inverse estimate given in Lemma 2.28, we arrive at:

$$\begin{aligned} &\|\varphi_h(\tau)\|_{\mathcal{L}^2(\Omega)} \leq \|\epsilon_h(t)\|_{\mathcal{L}^2(\Omega)}, \qquad 0 \leq \tau \leq t \\ &\int_0^t \|\dot{\varphi}_h(\tau)\|_{\mathcal{L}^2(\Omega)} \, d\tau \leq C\left(1 + |\log(\tfrac{t}{h^2})|\right) \|\epsilon_h(t)\|_{\mathcal{L}^2(\Omega)}, \end{aligned}$$

from which we have immediately:

$$\|e_h(t)\|_{\mathcal{L}^2(\Omega)} \leq C\left(1 + |\log(\tfrac{t}{h^2})|\right) \max_{0\leq\tau\leq t} \|\epsilon_h(t)\|_{\mathcal{L}^2(\Omega)}.$$

The proof is completed by noting that $u - u_h = \epsilon_h - e_h$ and using the $\mathcal{L}^2$ estimate of Theorem 2.26 applied to $\epsilon_h(\tau) = u(\tau) - \tilde{u}_h(\tau)$. □

In practice, this tells us that the solution accuracy at time t is independent from the accuracy with which the initial transient is resolved.

5.4 Full discretization

We study in this paragraph the simpler algorithms for time-discretization based on Euler or Crank-Nicolson methods. We refer to the later chapter on the numerical solution of ODEs for more advanced tools. Preliminary, however, we would like to discuss the interaction between temporal and spatial discretization. To do so, we first analyze how the results of section (5.1) can be extended to the fully discrete case. Thus we write the solution to eq. 5.6 in terms of eigenpairs (μ_i, z_i) of matrix $\tilde{A}$. In the case $g(t) = 0$ we have the classical spectral representation of the solution:

$$\eta(t) = \sum_{j=1}^{N} (\eta_0, z_j)\, \mathrm{e}^{-\mu_j t}\, z_j.$$

It is easy to see the the mass matrix has uniform spectrum with eigenvalues all of the order $\mathcal{O}(1)$. Thus the spectral interval of $\tilde{A}$ is that of A, and hence, from Theorem 2.27, we can conclude that $\mu_1 = \mathcal{O}(1)$ and $\mu_n = \mathcal{O}(h^{-2})$. Larger eigenvalues correspond to oscillatory modes (eigenvectors) while smaller eigenvalues correspond to the more regular modes. Hence, the components of the solution $\eta(t)$ are characterized by temporal scales that vary in a large interval confined between $\mathcal{O}(h^{-2})$ and $\mathcal{O}(1)$, a signal that the problem is "stiff". It is then necessary to use implicit schemes to avoid the large restrictions on the temporal integration step forces by stability constraints in case of explicit approaches.

5.4.1 Backward (implicit) Euler scheme

Let $I = (0, T]$ the time interval ($T > 0$) and let $0 = t_0 \leq t_1 \leq \ldots \leq t_M$ a partition of I, where $t_{n+1} = t_n + k_n$ and $I_n = (t_n, t_{n+1})$. We start from the semi-discrete problem (5.4) and substitute the time derivative $du_h(t)dt$ with its incremental ratio. We can write the following problem. find $u_h^n \in \mathcal{V}_h$ such that:

$$\begin{aligned} \left(\frac{u_h^{n+1} - u_h^n}{k_n}, v\right) + a\left(u_h^{n+1}, v\right) &= (f(t_{n+1}, v) && \forall v \in \mathcal{V}_h \;\; n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) && \forall v \in \mathcal{V}_h. \end{aligned} \tag{5.11}$$

This corresponds to applying Backward (implicit) Euler scheme to the ODE system (5.5). We obtain:

$$\left(\frac{1}{k_n} P + A\right) u^{n+1} = \frac{1}{k_n} P u^n + b^{n+1}. \tag{5.12}$$

At every temporal step we need to solve a linear system of dimension $n \times n$ where the system matrix is $M = P/k_n + A$. Intuitively, we could factor matrix M once and for all, and use it effectively to solve the linear system. On the other hand, from what we have seen above, it is

also convenient to increase k_n at every time step once the fast temporal scales are resolved. Then we could think of factorizing separately P and A to form M. This route is feasible as long as the matrix dimensions are not too large, in which case the preconditioned conjugate gradient is often employed, sing as initial condition the solution calculated at the previous time step.
The stability of backward Euler is immediately verified. In fact, taking $v = u_h$ in (5.11) we have for $f(t) = 0$ [7]:

$$\left\|u_h^{n+1}\right\|^2 - \left(u_h^{n+1}, u_h^n\right) + k_n a\left(u_h^{n+1}, u_h^{n+1}\right) = 0.$$

Using Young inequality yields:

$$\frac{1}{2}\left\|u_h^{n+1}\right\|^2 - \frac{1}{2}\left\|u_h^n\right\|^2 + k_n a\left(u_h^{n+1}, u_h^{n+1}\right) \leq 0, \qquad n = 1, \ldots, N.$$

Summing over n we have:

$$\left\|u_h^{n+1}\right\|^2 + 2\sum_{j=1}^{N} k_n a\left(u_h^j, u_h^j\right) \leq \left\|u_h^0\right\|^2,$$

and using the coercivity of $a(\cdot,\cdot)$ we find:

$$\left\|u_h^n\right\| \leq \left\|u_h^0\right\| \leq \|u_0\|, \qquad n = 1, \ldots, N, \tag{5.13}$$

an estimate corresponding to the stability estimate of the semi-discrete system given by (5.8) and (5.9).
Another way to analyze the stability of backward Euler start from the algebraic system (5.12). Assuming $b = 0$, the system becomes:

$$\left(\frac{1}{k_n}P + A\right)u^{n+1} = \frac{1}{k_n}Pu^n.$$

The mass matrix is symmetric and positive definite, and thus invertible, with a condition number of the order of $\mathcal{O}(1)$. Multiplying by k_n and by P^{-1} we obtain formally:

$$u^{n+1} = \left(I + k_n P^{-1}A\right)^{-1} u^n.$$

Matrix $I + k_n P^{-1}A$ is similar to $I + k_n L^{-1}AL^{-T}$, where $P = LL^T$, L being its Choleski factor. We have:

$$\left\|(I + k_n L^{-1}AL^{-T})^{-1}\right\| = \max_{i=1,n}\left[\frac{1}{\lambda_i(I + k_n L^{-1}AL^{-T})}\right] < 1$$

[7] We will always use $\mathcal{L}^2(\Omega)$ norms and for this reason we omit the corresponding subscript.

and since the eigenvalues of $P^{-1}A$ are positive and $k_n > 0$, we readily obtain again (5.13). The backward Euler has a truncation error of the order of $\mathcal{O}(k_n)$. Hence, optimal[8] quadratic convergence is obtained if $k_n = \mathcal{O}(h^2)$, a restriction that is often too strong in practical applications. For this reason the Crank-Nicolson method, based on the trapezoidal quadrature rule, is often used.

5.4.2 Crank-Nicolson method

The Crank-Nicolson scheme (or trapezoidal rule) can be derived by using again a simple incremental ratio to approximate the time derivative and using the weighted arithmetic average of the fluxes (the other terms) at time t_{n+1} and t_n:

$$\begin{aligned}\left(\frac{u_h^{n+1}-u_h^n}{k_n}, v\right) + \frac{1}{2}\left[a\left(u_h^{n+1}, v\right) + a\left(u_h^n, v\right)\right] &= \frac{1}{2}\left[(f(t_{n+1}), v) + (f(t_n), v)\right] \\ &\forall v \in \mathcal{V}_h \quad n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) \qquad \forall v \in \mathcal{V}_h.\end{aligned} \tag{5.14}$$

or in matrix form:

$$\left(\frac{1}{k_n}P + \frac{1}{2}A\right)u^{n+1} = \left(\frac{1}{k_n}P - \frac{1}{2}A\right)u^n + \frac{1}{2}\left(b^{n+1} + b^n\right). \tag{5.15}$$

This scheme has a truncation error of the order of $\mathcal{O}(k_n^2)$, and is unconditionally stable. In fact:

$$\left\|(I + \frac{1}{2}k_n L^{-1}AL^{-T})^{-1}(I - \frac{1}{2}k_n L^{-1}AL^{-T})\right\| = \max_{i,j=1,n}\left|\frac{2I - k_n\lambda_j(L^{-1}AL^{-T})}{2I + k_n\lambda_i(L^{-1}AL^{-T})}\right| < 1.$$

5.4.3 Forward (explicit) Euler scheme

The forward or explicit Euler method is written as follows:

$$\begin{aligned}\left(\frac{u_h^{n+1}-u_h^n}{k_n}, v\right) + a\left(u_h^n, v\right) &= \left(f(t_n), v\right) \\ &\forall v \in \mathcal{V}_h \quad n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) \qquad \forall v \in \mathcal{V}_h.\end{aligned}$$

or in matrix form:

$$\left(\frac{1}{k_n}P\right)u^{n+1} = \left(\frac{1}{k_n}P - A\right)u^n + b^n. \tag{5.16}$$

[8]Optimal convergence is the maximal convergence rate allowable by the interpolation error

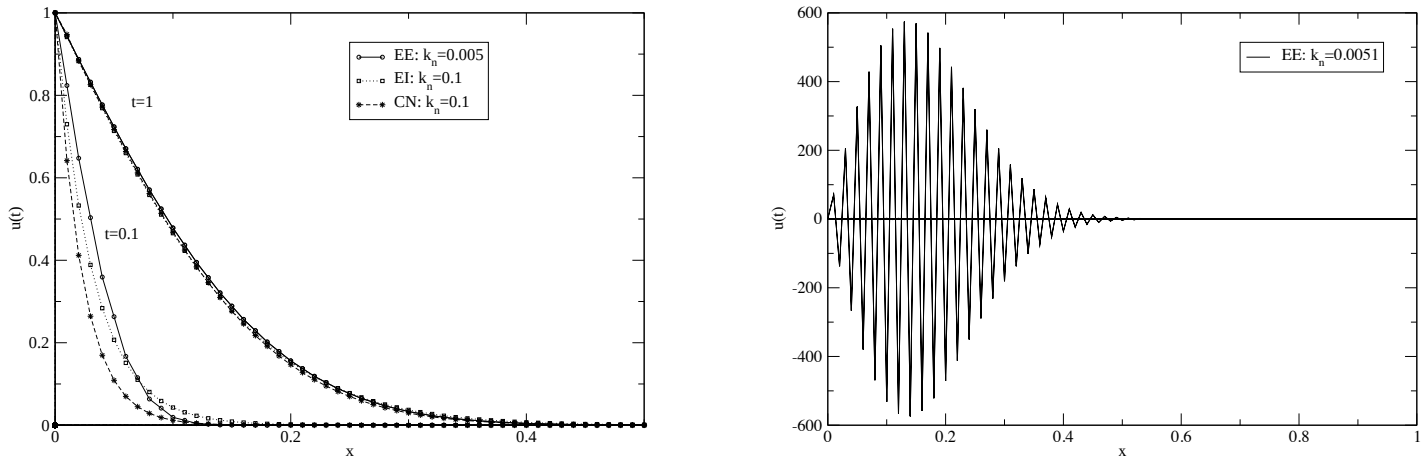


FIGURE 5.2: *Numerical results for explicit Euler (EE), implicit Euler (IE), and Crank-Nicolson (CN) for the solution of the one-dimensional problem using $\mathcal{P}_1$ Galerkin FEM. The left panel shows the solution for the stable cases, while the right panel shows an example of an instable solution.*

The forward Euler scheme has the same accuracy of backward Euler ($\mathcal{O}(k_n)$), but it is only conditionally stable. In fact we have:

$$\left\|(I - k_n L^{-1} A L^{-T})\right\| = \max_{i=1,n} \left|I - k_n \lambda_i (L^{-1} A L^{-T})\right| \leq 1.$$

This inequality is satisfied only if $k_n \leq \frac{2}{\lambda_1(P^{-1}A)} = \mathcal{O}(h^2)$, since from Theorem 2.27 $\lambda_{\min}(A) = \mathcal{O}(h^2)$. The stability condition can be rewritten as:

$$k_n \leq Ch^2 \qquad \text{or } \frac{\sqrt{Ck_n}}{h} \leq 1. \tag{5.17}$$

The constant C depends on the diffusion coefficient $K(x)$, and asserts that the temporal and spatial resolution must be such that the solution variation in passing from t_n to t_{n+1} (approximately of the order of $\sqrt{Dk_n}$) must remain within one cell (of size h). In other words, the scheme must be able to resolve "correctly" the faster components of the solution during the initial transient. In practice, explicit schemes are seldom used because of this stability restriction, and implicit schemes are preferred as the time step restriction due to stability typically offsets the higher cost for the linear system solution.

Experimental results for a one-dimensional model problem. In this section we discuss the numerical results for the three schemes presented above. We use a $\mathcal{P}_1$ Galerkin FEM

formulation for the spatial discretization of the one-dimensional model problem of Example 5.1. Figure 5.2 (left panel) shows the solution obtained for $D = 10^{-2}$ at times $t = 0.1$ e $t = 1.0$, using Galerkin- $\mathcal{P}_1$ with Explicit Euler (EE), implicit Euler (IE), and Crank-Nicolson (CN). We have used $h = 1/100$ and $k_n = 0.1$ for IE and CN, while $k_n = 0.005$ for EE (this value guarantees stability). The right panel shows the results of EE obtained with $k_n = 0.0051$. The classical oscillations due to instability are clearly visible and their amplitude increases for increasing k_n.

6 Pseudospectral Methods: an Overview

6.1 Introduction

Spectral and pseudospectral (PS for short) methods are a very classical approach for solving PDEs. In such a framework, one considers approximations of the form

$$u(x) \approx u_N(x) := \sum_{j=1}^{N} c_j \phi_j(x)$$

to the true solution u built by means of a given (usually very smooth) global *orthogonal basis* $\{\phi_k\}$ of a function space, that we term space of *trial functions*, and takes in account exact differentiation of u_N.
PS methods are characterized by the following very desirable properties

- for an analytic function u the rate of convergence of the truncated expansion u_N to u is *exponential* (indeed this is called *spectral convergence*) instead of linear of polynomial as for finite differences or finite elements methods.
- Even for non-smooth function this approach reveals to be profitable, provided that the singularities are not too strong.
- Due to the fast convergence, in the most of applications a relatively coarse computational grid suffices to achieve a rather good accuracy. This becomes a very beneficial property when the spatial dimension of the problem grows large.

One of the main features of pseudospectral methods is that the trial (and eventually the test) functions are *global*. Consequently, the considered finite dimensional spaces are rather *rigid*; however this turns in a disadvantage only on some classes of problems. The typical issues we need to consider when choosing or implementing a spectral or PS method are

- irregular domains: choice of the basis,
- presence of strong shocks,
- variable resolution requirements in different parts of the domain.

Once the functional space $\mathcal{F}$ the solution u must belong to is known, in order to build the method one needs to chose an orthogonal basis $\{\phi_j\}$ of $\mathcal{F}$. In performing such a choice the following task should be considered [10, 11]:

1. *Rapid convergence.* For smooth functions the truncated expansion needs to converge very rapidly.

2. *Easy differentiation.* For any N and any $\{a_j\}_{j=1,\dots,N} \in \ell^2$ the operation

$$\frac{d}{dx}\left(\sum_{j=1}^{N} a_j \phi_j(x)\right) \;=\; \sum_{j=1}^{N} b_j \phi_j(x)$$

should be computed easily and efficiently.

3. *Easy evaluation and synthesis.* The evaluation of the expansion at a given point and the computation of the coefficients (synthesis) starting with $u(x_i)$, $i = 1, \dots, M$ needs to be easy and fast.

4. *Simple integration.* When considering the weak formulation of differential equations and complementing boundary conditions one needs to compute integrals over the domain or its boundary of the basis functions, a fast and accurate computation algorithm is then required.

5. *Boundary conditions.* In some methods we will need to easy characterize the functional subspace (spanned by combinations of ϕ_j's) satisfying the boundary conditions.

The result of the above requirements is that the typical choices for $\{\phi_k\}$ are

- Fourier basis $\{e^{ik\pi}\}$ for periodic problems
- Orthogonal polynomials (typically arising from a Sturm-Liouville singular operator)for non-periodic boundary conditions.

6.2 Classification of Pseudospectral Methods

In this section we review the generalities on pseudospectral methods in a quite general framework postponing the discussion on specific methods and function spaces to Section **??** and a more detailed discussion on their well posedness and behaviour to Section **??**.
In the sequel we will deal with linear problems of the following type

$$\begin{cases} \mathcal{L}u = f, & \text{in } \Omega \\ \mathcal{B}u = 0, & \text{on } \partial\Omega \end{cases}, \tag{6.1}$$

where $\Omega \subset \mathbb{R}$ is a domain, $\mathcal{L}$ is a linear partial differential operator and $\mathcal{B}$ is a boundary operator (as for instance the trace operator on $\partial\Omega$ or the trace of normal derivatives on $\partial\Omega$). We work on a Hilbert space $\mathcal{H}$ of function defined on Ω but in general $\mathcal{L}$ may be defined only on a *dense* subset $D(\mathcal{L}) \subset \mathcal{H}$, also we will consider the (possibly affine) subspace $D_{\mathcal{B}}(\mathcal{L})$ of $\mathcal{H}$ where the boundary condition $\mathcal{B}u = 0$ is satisfied in the appropriate sense; hence

$$\mathcal{L} : D_{\mathcal{B}}(\mathcal{L}) \subset \mathcal{H} \to \mathcal{H}.$$

The *mild formulation* of the problem (6.1) is then

$$\text{find } u \in D_{\mathcal{B}}(\mathcal{L}): \quad (\mathcal{L}u, v)_{\mathcal{H}} = F(v) := (f, v)_{\mathcal{H}} \ \forall v \in \mathcal{H}.$$

More in general, one considers the associated *weak formulation*

$$a(u, v) = (u, f), \quad \forall v \in \tilde{D}(a),$$

where a bilinear form a on a (possibly larger than $D_{\mathcal{B}}(\mathcal{L})$) (linear or affine) subspace $\tilde{D}(a)$ of $\mathcal{H}$ is introduced starting by $\mathcal{L}$ and performing integration by parts.
To illustrate this standard procedure we consider two model problems: the Poisson Equation with Dirichlet or Neumann boundary conditions in the unit ball $B := \{x \in \mathbb{R}^N : |x| < 1\}$, note that the unit normal on ∂B is x.

$$\begin{cases} \Delta u = f & , \text{ in } B \\ u(x) = 0 & , \forall x \in \partial B \end{cases}. \tag{6.2}$$

$$\begin{cases} \Delta u = f & , \text{ in } B \\ x \cdot \nabla u(x) = 0 & , \forall x \in \partial B \end{cases}. \tag{6.3}$$

For the first integrating by parts we obtain

$$-\int_\Omega \nabla u(x) \cdot \nabla \phi(x) dm(x) + \int_{\partial\Omega} x \cdot \nabla u(x) \phi(x) d\sigma(x) = \int_\Omega f(x)\phi(x) dm(x), \ \forall \phi \in \mathcal{C}_c^\infty(\Omega).$$

Now by the assumption on the support of ϕ the boundary integral vanishes and we get

$$-\int_\Omega \nabla u(x) \cdot \nabla \phi(x) dm(x) = \int_\Omega f(x)\phi(x) dm(x), \ \forall \phi \in \mathcal{C}_c^\infty(\Omega).$$

Notice that the domains of both the bilinear form of the left hand side and the linear form on the right side of this equation can be extended to the whole $H_0^1(\Omega)$; indeed we take $\tilde{D}(a) := \mathcal{C}_c^\infty(\Omega)$, note that the closure of such a space in $H^1(\Omega)$ is precisely $H_0^1(\Omega)$. This justifies the following definition of weak solution of the problem (6.2) $u \in H_0^1(\Omega)$ is a weak solution of problem (6.2) if

$$-\int_\Omega \nabla u(x) \cdot \nabla \phi(x) dm(x) = \int_\Omega f(x)\phi(x) dm(x), \ \forall \phi \in H_0^1(\Omega). \tag{6.4}$$

In contrast, for the Neumann problem the boundary condition is encoded in the equation, not in the space; this is a standard way of treating non-homogeneous conditions. After an integration by parts, the boundary integral term vanishes not only for $\phi \in \mathcal{C}_c^\infty(\Omega)$, it vanishes

for any $\phi \in \mathcal{C}^1(\Omega)$. Thus we set $\tilde{D}(a) := \mathcal{C}^1(\Omega)$, which is dense in $H^1(\Omega)$. The weak solution of problem (6.3) is then any $u \in H^1(\Omega)$ such that

$$-\int_\Omega \nabla u(x) \cdot \nabla \phi(x) dm(x) = \int_\Omega f(x)\phi(x) dm(x), \ \forall \phi \in H^1(\Omega). \tag{6.5}$$

The spectral approximation of problem (6.1) can be split in three families, namely

- Galerkin
- Tau
- Collocation (pseudospectral)

where only the first arises (at least historically and for second order equations) from the weak formulation, while the remaining two from the mild one. However, one may think to this three families of methods as a different consequences of the technique of *weighted residual*: starting by an equation (e.g. the mild or the weak formulation of the equation) one approximately solves it in a finite dimensional subspace V_N imposing the (weighted) residual $\mathcal{L}u_N - f$ to be orthogonal to each function $\psi \in V_N$ (or another subspace W_N).
In the **Galerkin Method** we search $u_N \in V_N$ such that

$$a(u_N, v) = F(v), \ \ \forall v \in V_N, \tag{6.6}$$

here $\{V_N\}$ is a dense *nested* sequence of subspace of $\tilde{D}(a) \cap Z(\mathcal{B})$, i.e., such that $\mathcal{B}v = 0$ $\forall v \in V_N$.
Some variants are available to this scheme: in *Petrov-Galerkin* method one requires the condition (6.6) $\forall v \in W_N$ instead of $\forall v \in V_N$, i.e. the *test functions space* W_N and the *trial functions space* V_N do not necessarily coincide. Instead, in the Galerkin method with *numerical integration* (G-NI, for short) the functional F is replaced by an approximate version F_N given by certain quadrature formulas.
In the **Tau Method** we consider an orthogonal basis $\{\phi_k\}$ of $\mathcal{H}$ (thus not a priori satisfying the boundary conditions $\mathcal{B}\phi_k = 0$) consisting of smooth functions (thus elements of $D(\mathcal{L})$) and set $V_N := \text{Span}(\phi_1, \dots, \phi_N)$, then we search for $u_N := \sum_{k=1}^N c_k \phi_k$ such that

$$\begin{cases} (\mathcal{L}u_N, \phi_k) = (f, \phi_k), & \forall k \in I_N \\ \mathcal{B}u_N = 0 \end{cases}, \tag{6.7}$$

where $I_N \subset \{1, \dots, N\}$ is chosen in a way that makes the above system uniquely solvable.
Finally we introduce the **collocation method (PS)**. Here one requires the approximate solution u_N to satisfy

$$\begin{cases} \mathcal{L}u_N(x_i) = f(x_i), & \forall i \in I_N \\ \mathcal{B}u_N(x_i) = 0, & \forall i \in J_N. \end{cases}, \tag{6.8}$$

One can regard PS methods as a particular instance of spectral one since they can be derived starting by the weighted residual and choosing as test "functions" the Dirac delta functions centred at the collocation points. Note that

$$\langle \mathcal{L}u_N, \delta_{x_i} \rangle = \mathcal{L}u_N(x_i) = f(x_i) = \langle f, \delta_{x_i} \rangle.$$

In contrast, most authors use a dedicated terminology for this method to underline a remarkable practical difference they have with respect to the spectral methods. When considering a spectral method (both for steady state or evolution problems) we solve equations in the *spectral variables/space* c_k to determine the solution; instead, when considering a PS method, we solve equations with respect to the *physical variable/space* $\phi_k(x_i)$.

6.3 Some Classical Example

To illustrate the construction of the methods we present some classical easy examples taken from [19] and [5].

6.3.1 Galerkin Method and its Variants

Let us take in account the linear elliptic partial differential operator

$$\mathcal{L}u := -au'' + bu' + cu \tag{6.9}$$

and, given a periodic function $f \in \mathcal{L}^2[0, 2\pi]$, solve the problem $\mathcal{L}u = f$ by a pure Galerkin method.
We define u_N as the truncated Fourier expansion $u_N(x) := \sum_{k=-N}^{N} \hat{u}_k e^{ikx}$ and we compute the residual $R_N(x) := \mathcal{L}u_N - f = \sum_{k=-N}^{N} \hat{u}_k \mathcal{L}e^{ikx} - f$. The Galerkin method is implemented by solving the equations

$$\left(R_N, e^{ijx}\right)_{\mathcal{L}^2[0,2\pi]} = \sum_{k=-N}^{N} \hat{u}_k \left(\mathcal{L}e^{ikx}, e^{ijx}\right)_{\mathcal{L}^2[0,2\pi]} - \left(f, e^{ijx}\right)_{\mathcal{L}^2[0,2\pi]} = 0, \quad \forall j = -N, \ldots, N.$$

Due to orthogonality and the fact that $\mathcal{L}e^{ikx} = (ak^2 + ibk + c)e^{ikx}$, the above equations turns in

$$\hat{u}_k(ak^2 + ibk + c) = 2\pi \hat{f}_k, \quad \forall k = -N, \ldots, N.$$

Note that actually the computation is even more efficient: since we are assuming u to be a real functions, we need to compute $\hat{u}_k$ only for $k = 0, 1, \ldots, N$ then we can use the relation $\hat{u}_{-k} = \overline{\hat{u}_k}$.
It is also worth to notice that, due to assumed periodicity we treated the above problem without tacking into account any boundary condition; more rigoroursly, when we choose as trial and as test function space the space of Fourier characters we are implicitly solving the problem (6.9) complemented by periodic boundary conditions.

When we introduced the Galerkin method we claimed that it generally arises from the weak formulation of the differential problem; this is not in contrast with the above computations. Indeed we could start by the weak formulation (integrate by parts twice the first member of $\mathcal{L}$ and just once the second, then cancel boundary terms thanks to periodicity)

$$-a\left(u,\phi_j''\right)_{\mathcal{L}^2[0,2\pi]}+b\left(u,\phi_j'\right)_{\mathcal{L}^2[0,2\pi]}+(u,\phi_j)_{\mathcal{L}^2[0,2\pi]}=(f,\phi_j)_{\mathcal{L}^2[0,2\pi]}\,,\quad\forall j=-N,\dots,N$$

and this would lead to the same system of equations above due to the differential properties of ϕ_ks.
To illustrate the **Galerkin method with numerical integration** we introduce the following *advection diffusion reaction equation* complemented by Dirichlet boundary condition at -1 and mixed Robin condition at 1.

$$\begin{cases}\frac{d}{dx}\left(-\nu\frac{du}{dx}+\beta u\right)+\gamma u=f\quad,\ \text{in }]-1,1[\\ u(-1)=0,\\ -\nu\frac{du(1)}{dx}+\beta u(1)=g\end{cases}\quad.\tag{6.10}$$

Since we are dealing with non periodic boundary conditions we choose polynomials as dense subspace of the domain of the differential operator, more precisely we set as basis of such a space the Legendre orthogonal polynomials $\{L_k\}$, i.e. orthogonal polynomials in $[-1,1]$ with respect to Lebesgue measure normalized to get $L_k(1)=1$; such polynomials arise as eigenfunctions of the singular Sturm Liouville operator $\frac{d}{dx}((1-x^2)v')+k(k+1)v=0$.
Recall that the $N-1$ zeros $\{x_1,x_2,\dots,x_{N-1}\}$ of L_N', complemented by the extremal points $x_0=-1$ and $x_1=1$, form the so called Gauss-Lobatto-Legendre nodes of degree N, that are the support of a high (polynomial) precision quadrature formula

$$\int_{-1}^{1}p(x)\sim\sum_{j=0}^{N}p(x_j)w_j,\tag{6.11}$$

where equality holds for any p polynomial of degree at most $2N+1$.
To implement a "pure" Galerkin method with numerical integration we should proceed as follows.

- We enforce the *homogeneous* boundary condition $u(-1)=0$ in a strong sense.

  Using the symmetry of the Legendre basis $L_k(-x)=(-1)^kL_k(x)$, we get $L_k(-1)=(-1)^k$ and thus $p(-1)=\sum_{k=0}^{N-1}(-1)^kc_k(p)$. The boundary condition at -1 can be written $\langle c_k,z_N\rangle=0$, where we set $z_N:=((-1)^k)_{k=0,1,\dots,N-1}$.

  We can compute the subspace V_N of P^{N-1} satisfying this condition. To solve this problem and simultaneously compute an orthogonal basis we can use the Gram Schmidt procedure or the QR algorithm, ending up with a new basis $\{\phi_k\}_{k=0,\dots,N-1}$ of P^{N-1} such that the subspace spanned by $(\phi_1,\dots,\phi_{N-1})$ enjoys the homogeneous boundary condition and ϕ_ks are orthogonal. This task and its possible generalization will be treated more rigorously and more in detail in Section **??**.

- Starting by the remarkable differential properties (see Subsection 7.2) of the Legendre polynomials one can obtain

$$\frac{x^2-1}{k}L'_{k+1}(x) = (2k+1)xL_k(x) - kL_{k-1}(x)$$

and expressing the term (x^2-1) by means of $L_0, L_1(x), L_2(x)$ we derive the coefficients of each L'_j with respect to the basis $\{L_k\}$. Then, using the same change of basis that lead from $\{L_k\}$ to $\{\phi_k\}$, we deduce the coefficients of each ϕ'_j with respect to the basis $\{\phi_k\}$. We can arrange such coefficients $D_{i,j} := \langle \phi'_i, \phi_j \rangle$ in the *differential matrix* D, so that $\frac{d}{dx}\sum_{k=1}^{N-1} c_k\phi_k = \sum_{j=0}^{N-1} Dc_k\phi_j$.

- We consider the mild formulation of the problem: find $u_N = \sum_{k=1}^{N-1} c_k\phi_k$ such that

$$\int_{-1}^{1} \frac{d}{dx}\left(-\nu\frac{d}{dx}u_N(x) + \beta u_N(x)\right)v(x) + dx = \int_{-1}^{1}(f(x) - \gamma u_N(x))v(x)dx, \quad \forall v \in V_N.$$

Then we perform an integration by parts and we impose the condition for $v = \phi_1, \ldots, \phi_N$ getting

$$-\int_{-1}^{1}(\nu\frac{d}{dx}u_N(x) - \beta u_N(x))\phi'_j(x) + [(-\nu\frac{d}{dx}u_N(x) + \beta u_N(x))\phi_j]_{-1}^{1}$$
$$= \int_{-1}^{1}(f(x) - \gamma u_N(x))v(x)dx, \forall j = 1, 2, \ldots, N-1$$

Using the non-homogeneous boundary condition the above equation reduces to

$$-\int_{-1}^{1}(\nu\frac{d}{dx}u_N(x) - \beta u_N(x))\phi'_j(x) = g + \int_{-1}^{1}(f(x) - \gamma, u_N(x))v(x)dx$$
$$\forall j = 1, 2, \ldots, N-1$$

Let us rewrite this last equation using Linear Algebra

$$\langle(-\nu D + \beta I_{N-1})\boldsymbol{c}, D_{:,j}\rangle = g - \gamma c_j + b_j\ , \quad b_j := \int_{-1}^{1} f\phi_j dx$$
$$\forall j = 1, 2, \ldots, N-1,$$

where I_{N-1} denotes the identity matrix, $\boldsymbol{c} := (c_1, \ldots, c_{N-1})$, $D_{:,j}$ is the jth column of D and $\langle\cdot,\cdot\rangle$ is the standard $\mathbb{R}^{N-1}$ duality.

- Here appears the **numerical integration**. We choose to approximate all the above integrals b_js by the quadrature rule (6.11). Thus the final formulation of the problem becomes

$$\langle(-\nu D + \beta I_N)c, D_{:,j}\rangle = g - \gamma c_j + b_j\ , \quad b_j := \sum_{h=1}^{N} f(x_h)\phi_j(x_h)w_h dx$$
$$\forall j = 1, 2, \ldots, N.$$

6.3.2 Collocation Method

We want to illustrate that **Collocation (PS)** *method may arise as a variation on Galerkin Method with NI.* We consider the same problem (**??**) above, but we allow the coefficients ν, β and γ to depend on x.
Starting by the Gauss-Lobatto-Legendre nodes we can form the fundamental Lagrange interpolating polynomials, i.e., polynomials having degree at most N and satisfying $\ell_k(x_j) = \delta_{k-j}$, they can be understood as *discrete delta functions.* Using the differential and recursive properties of Legendre polynomials and the nodes set we are considering, it is possible to prove the following formula

$$\ell_k(x) = \frac{1-x^2}{N(N-1)(x_k - x)} \frac{L'_N(x)}{L_N(x)}.$$

We would like to take as trial and test functions the whole set $\{\ell_k\}$, but since $\ell_0(-1) = 1$ we drop it from the basis in order to satisfy the boundary condition $u(-1) = 0$.
To define a weak formulation of the problem is convenient to introduce the *flux operator* $\mathcal{F}[u](x) := -\nu(x)\frac{d}{dx}u(x) + \beta u(x)$, we impose the mild formulation and then we integrate by parts using the boundary conditions.

$$\int_{-1}^{1} \frac{d}{dx}\mathcal{F}[u]\ell_j dx + \gamma \int_{-1}^{1} u\ell_j dx = \int_{-1}^{1} f\ell_j dx, \quad \forall j = 1, \ldots, N$$

$$-\int_{-1}^{1} \mathcal{F}[u]\frac{d}{dx}\ell_j dx + \gamma \int_{-1}^{1} u\ell_j dx = \int_{-1}^{1} f\ell_j dx, \quad \forall j = 1, \ldots, N$$

so we obtain (substituting u by u_N)

$$\begin{aligned}\int_{-1}^{1} \nu u'_N \ell'_j dx - \beta \int_{-1}^{1} u_N \ell'_j dx + \gamma \int_{-1}^{1} u_N \ell_j = \ldots \\ = \int_{-1}^{1} f\ell_j dx + g\delta_{j-N}, \quad \forall j = 1, 2, \ldots, N.\end{aligned}$$

Finally we use the numerical integration (6.11) to approximate each integral obtaining the following system of equations.

$$\begin{aligned}\sum_{i=0}^{N} \left[\left(\nu \frac{du_N}{dx}\frac{d\ell_j}{dx} \right)(x_i) - \left(\beta u_N \frac{d\ell_j}{dx} \right)(x_i) + (\gamma u_N \ell_j)(x_i) \right] w_i = \\ \sum_{i=0}^{N} (f\ell_j)(x_i) w_i + g\delta_{N-j}, \quad \forall j = 1, \ldots, N.\end{aligned} \tag{6.12}$$

Taking in account that $u_N(x) = \sum_{i=1}^{N} \ell_i(x) c_i(u) = \sum_{i=1}^{N} \ell_i(x) u(x_i)$ and that this implies

$u'_N(x) = \sum_{i=1}^N \ell'_i(x)c_i(u)$, the system of equations (6.12) may be written in the matrix form

$$
\begin{aligned}
&K\boldsymbol{c} = \boldsymbol{b},\ \boldsymbol{c} := (c_1, \dots, c_N),\ \boldsymbol{b} := (b_1, \dots, b_N),\\
&K_{j,l} := \sum_{i=0}^{N} \left[\left(\nu \frac{d\ell_l}{dx}\frac{d\ell_j}{dx} \right)(x_i) \right] w_i - \left(\beta \frac{d\ell_j}{dx} \right)(x_l) + \gamma(x_j) w_j \delta_{|j-l|}\\
&b_j := f(x_j)w_j + g\delta_{N-j}.
\end{aligned}
$$

Let us remark that the combination of the choice of the Lagrange basis (approximated delta functions) and numerical integration leads to a system of linear equation in the variables $u(x_i)$, as we announced for collocation methods. To better interpret this let us introduce the *numerical flux* $\mathcal{F}^N(x) := \sum_{i=0}^N \ell_i(x)\mathcal{F}u(x_i)$, that is the interpolation of the flux.
Now we observe that the left side of equation (6.12) can be rewritten as

$$
-\sum_{i=0}^{N} \left(\mathcal{F}^N \frac{d\ell_j}{dx} \right)(x_i) w_i + \sum_{i=0}^{N} (\gamma u_N \ell_j)(x_i) w_i. \tag{6.13}
$$

Note that the first term is a quadrature rule of precision $2N+1$ applied to a polynomial of degree at most $2N-2$, thus it is an exact integral. We perform integration by parts and we use the boundary conditions to obtain an new formula for the expression in (6.13).

$$
\begin{aligned}
\sum_{i=0}^{N} \left(\frac{d}{dx}\mathcal{F}^N \ell_j \right)(x_i) w_i - \left(\mathcal{F}^N \frac{d\ell_j}{dx} \right)(1) + \left(\mathcal{F}^N \frac{d\ell_j}{dx} \right)(-1) + \sum_{i=0}^{N} (\gamma u_N \ell_j)(x_i) w_i\\
= \sum_{i=0}^{N} \left(\frac{d}{dx}\mathcal{F}^N \ell_j \right)(x_i) w_i - \left(\mathcal{F}^N \frac{d\ell_j}{dx} \right)(1) + \sum_{i=0}^{N} (\gamma u_N \ell_j)(x_i) w_i
\end{aligned}
$$

When we substitute this last expression to the left hand side of (6.12), using the approximate delta function property of the Lagrange basis we simply get

$$
\left(\frac{d\mathcal{F}^N}{dx} + \gamma u_N - f \right)(x_j) = 0,\ \forall j = 1, 2, \dots, N-1
$$

this is precisely a collocation method for the interpolated flux.

6.3.3 Tau Method

Lastly we illustrate the Tau Method when applied to the Poisson equation (6.2) with Dirichlet boundary condition on $Q := [-1, 1]^2$,

$$
\begin{cases} \Delta u = f & , \text{ in } Q \\ u(x) = 0 & , \forall x \in \partial Q \end{cases}. \tag{6.14}
$$

We pick as basis the tensor product space of polynomials $V_N := \mathbb{P}^N \otimes \mathbb{P}^N$ with the tensor product basis

$$\phi_{k,l}(x,y) := \cos(k \arccos(x)) \cos(l \arccos(y)), \quad k,l = 0,1,\dots,N.$$

Note that this basis does not preserve the boundary condition, indeed we enforce them in a weak form by choosing an additional set of test functions. Namely we set

$$\psi_k(x) := \cos(k \arccos(x))$$

and impose

$$\begin{cases} u_N := \sum_{k,l=0}^N c_{k,l}\phi_{k,l}(x,y), & \\ \int_Q \nabla u_N \cdot \nabla \phi_{k,l}\, dx\, dy = \int_Q f\phi_{k,l}\, dx\, dy, & \forall k,l = 0,1,\dots,N-2 \\ \int_{-1}^1 u_N(x,1)\psi_j(x)\, dx = \int_{-1}^1 u_N(x,-1)\psi_j(x)\, dx = 0, & \forall j = 0,1,\dots,N \\ \int_{-1}^1 u_N(1,y)\psi_j(y)\, dy = \int_{-1}^1 u_N(-1,y)\psi_j(y)\, dy = 0, & \forall j = 0,1,\dots,N \end{cases} \tag{6.15}$$

We stress that, despite the number of equations appearing in (6.15) is $(N-1)^2 + 4(N+1) = (N+1)^2 + 4 = dim V_N + 4$, four of such conditions are linear dependent on the other and may be drop out. This phenomena is due to the enforcement of boundary conditions on the corners: each of them is counted twice.
The system (6.15) can be simplified to a system of linear equation with respect to the variables $c_{k,l}$ using the Chebyshev differentiation matrix $D_N = [\langle \phi_{k,l}, \phi_{m,n} \rangle]$ that derives from the relation

$$\phi_k'(x) = \sum_{j=0}^{[(k-1)/2]} \frac{2k}{\eta_{k-1-2j}} \phi_{k-1-2j}(x), \quad \text{where} \quad \eta_m := \begin{cases} 2, & m = 0 \\ 1, & \text{otherwise} \end{cases}. \tag{6.16}$$

Then the orthogonality of the basis lead to a linear system. Note that one needs to compute (analytically, if possible) the right hand side terms $\int_Q f\phi_{k,l}\, dx\, dy$.

6.3.4 Evolution Problems: an example

6.4 Stability Consistency and Convergence of Spectral Methods

We go back to the general linear problem

$$\begin{cases} \mathcal{L}u = f, & \text{in } \Omega \\ \mathcal{B}u = 0, & \text{on } \partial\Omega \end{cases}, \tag{6.17}$$

for $f \in \mathcal{H}$ and a bounded (sufficiently smooth) open set $\Omega \subset \mathbb{R}^d$.
Our aim is to analyse the behaviour of the methods we proposed above and furnish sufficient conditions for their convergence.

6.4.1 Analysis of the Galerkin Method

In this subsection we will assume that there exists an Hilbert space $\mathcal{E}$ (usually called energy space) such that $\mathcal{E}$ is dense in $\mathcal{H}$ and a positive finite constant C such that

$$\|u\|_{\mathcal{H}} \leq C\|u\|_{\mathcal{E}}, \ \ \forall u \in \mathcal{E}.$$

Also we will deal with a sequence of finite dimensional subspaces V_N of the set $D_{\mathcal{B}}(\mathcal{L})$.
The celebrated Lax-Richtmyer Equivalence Theorem states that

- *A consistent scheme is convergent if and only if it is stable.*

In our framework **consistency** reads as follows: *there exists a projection operator* $\Pi_N : D_{\mathcal{B}}(\mathcal{L}) \to V_N$ *such that, at least for a dense subspace of* $D_{\mathcal{B}}(\mathcal{L})$, *we have*

$$\|u - \Pi_N u\|_{\mathcal{E}} \to 0, \ \text{ as } N \to \infty. \tag{6.18}$$

While **stability** *is given by continuous dependence on data, more precisely we require a constant* M *(not depending on* N*) to exist such that*

$$\|u_N\|_{\mathcal{E}} \leq M\|f\|_{\mathcal{H}}. \tag{6.19}$$

Theorem 6.1 (Convergence of Galerkin Method)**.** *Assume that* $\{V_N\}$ *is a dense (w.r.t. the* $\|\cdot\|_{\mathcal{E}}$ *norm) sequence of subspace of* $D_{\mathcal{B}}(\mathcal{L}) \subset \mathcal{E} \subset \mathcal{H}$, *moreover the following two hypothesis hold true.*

i) (Coercivity.) There exists a positive finite constant α *(not depending on* N*) such that*

$$\alpha\|u\|_{\mathcal{E}} \leq (\mathcal{L}u, u)_{\mathcal{H}}, \ \ \forall u \in V_N. \tag{6.20}$$

ii) (Continuity.) There exists a positive finite constant M *(not depending on* N*) such that*

$$|\,(\mathcal{L}u, v)_{\mathcal{H}}\,| \leq M\|u\|_{\mathcal{E}}\|v\|_{\mathcal{E}}, \ \ \forall u, v \in V_N. \tag{6.21}$$

Then the Galerkin method based on V_N *is convergent, i.e., denoting by* u_N *the approximate solution, we have*

$$\|u - u_N\|_{\mathcal{H}} \to 0, \quad \text{as } N \to \infty.$$

Proof. The combination of (6.20) and the equation itself leads to *stability* of the Galerkin approximation scheme. Precisely, we have

$$\|u_N\|_{\mathcal{E}}^2 \leq \frac{(\mathcal{L}u_N, u_N)_{\mathcal{H}}}{\alpha} = \frac{1}{\alpha}\|f\|_{\mathcal{H}}\|u_N\|_{\mathcal{H}} \leq \frac{C}{\alpha}\|f\|_{\mathcal{H}}\|u_N\|_{\mathcal{E}}.$$

On the other hand the density assumption on the subspaces V_N gives us the *consistency* of the method. Indeed one can set

$$\begin{aligned}\Pi_N :& D_{\mathcal{B}}(\mathcal{L}) \longrightarrow V_N \\ & u \longmapsto \tilde{u}_N : \|u - \tilde{u}_N\|_{\mathcal{E}} = \inf_{v \in V_N} \|u - v\|_{\mathcal{E}},\end{aligned}$$

that is the (best approximation) orthogonal projection onto V_N with respect to the $\mathcal{E}$ norm. By this choice of Π_N, consistency is precisely the density assumption.
Since by the hypothesis (6.21) the bi-linear operator $\mathcal{L}$ is continuous with respect to the $\|\cdot\|_{\mathcal{E}}$ we could apply the Lax-Richtmeyer Equivalence Theorem to get $\|u - u_N\|_{\mathcal{E}} \to 0$; note that this convergence is stronger than the convergence in $\mathcal{H}$. □

Remark 6.1 (Cea's Lemma)**.** *It is probably more instructive to perform the explicit computations leading to the convergence result instead of relying on the Equivalence Theorem.*
First define $r_N := u_N - \Pi_N u$, note that, using the property of u_N, for any $v \in V_N$ we have

$$(\mathcal{L} r_N, v)_{\mathcal{H}} = (\mathcal{L} u_N, v)_{\mathcal{H}} - (\mathcal{L}\Pi_N u, v)_{\mathcal{H}} = (\mathcal{L} u, v)_{\mathcal{H}} - (\mathcal{L}\Pi_N u, v)_{\mathcal{H}} = (\mathcal{L}(u - \Pi_N u), v)_{\mathcal{H}},$$

thus, taking $v := r_N$ and applying both continuity and coercivity, we get

$$\alpha \|r_N\|_{\mathcal{E}}^2 \leq |\,(\mathcal{L} r_N, r_N)_{\mathcal{H}}\,| = |\,(\mathcal{L}(u - \Pi_N u), r_N)_{\mathcal{H}}\,| \leq M\alpha \|r_N\|_{\mathcal{E}} \alpha \|u - \Pi_N u\|_{\mathcal{E}}.$$

Finally, we use triangle inequality

$$\|u - u_N\|_{\mathcal{E}} \leq \|u - \Pi_N u\|_{\mathcal{E}} + \|u_N - \Pi_N u\|_{\mathcal{E}} \leq \left(1 + \frac{M}{\alpha}\right) \|u - \Pi_N u\|_{\mathcal{E}} \to 0.$$

This last inequality makes Theorem 6.1 even more precise: the error of the Galerkin Method is of the same order of the best approximation error with respect to the norm used in providing the stability. This result is known as Cea's Lemma.

Remark 6.2. *The assumption* (6.20) *on the coercivity of $\mathcal{L}$ is not adapted to all type of problems -indeed this assumption is modelled on uniform elliptic operators- and may be replaced in the statement of Theorem 6.1 by other assumptions (leading to an equivalent conclusion) as for instance the* inf sup condition. *The reader is invited to compare Theorem 6.1 with* **??**.

In order to illustrate the convergence of Galerin method, let us go back to the **Poisson Equation** (6.2) with Dirichlet boundary condition

$$\begin{cases} \Delta u = f & , \text{ in } Q \\ u(x) = 0 & , \forall x \in \partial Q \end{cases}, \tag{6.22}$$

where $Q :=]-1,1[^2$.
In this setting we pick $\mathcal{H} := L^2(Q)$, $\mathcal{E} := H_0^1(Q)$ and consider the weak formulation of this problem

$$\begin{cases} a(u_N, v) = (f, V)_{L^2} & , \text{ for any } v \in H_0^1(Q) \\ u \in H_0^1(Q) & , \end{cases}, \tag{6.23}$$

where $a_{L^2}(u_N, v) := \int_Q \nabla u_N \cdot \nabla v \, dx \, dy$.
We choose, $V_N := \{p \in P^N \times P^N, p|_{\partial Q} = 0\}$ and we choose as a basis of V_N the tensor product basis generated by elements of the type $\phi_k(x)\phi_j(y)$, where

$$\phi_k(x) := \begin{cases} L_0(x) - L_k(x), & k \text{ even} \\ L_1(x) - L_k(x), & \text{otherwise} \end{cases}$$

and L_k denotes the Legendre polynomial of degree k.
Now we start to check the assumptions of Theorem 6.1. Stability follows by the famous Poincar inequality (we can pick 1 as coercivity constant)

$$\|u_N\|_{H_0^1}^2 = (\nabla u_N, \nabla u_N)_{L^2} = a(u_N, u_N).$$

Similarly the continuity of the bilinear form easily follows from the fact that $a(\cdot, \cdot)$ is the inner product of H_0^1.
On the other hand we consider the consistency, so we need to show the denseness of V_N in H_0^1, indeed we introduce a more precise result (without proof) that will be useful to estimate precisely the rate of convergence.
First, let us introduce the following family of seminorms

$$|u|_{H^{m,N}} := \left(\sum_{k=\min(m,N+1)}^{m} \sum_{i}^{d} \|\partial_{x_i}^k u\|_{L^2} \right)^{1/2}.$$

The following inequality, which we will refer to as *approximation inequality*, is of fundamental importance to our aims.

$$\|u - \Pi_N u\|_{H^k} \leq C_{\Omega,k,m} N^{k-m} |u|_{H^{m,N}}, \tag{6.24}$$

where Π_N is the best approximation projector onto $P^N \cap \{f : f|_{\partial Q} = 0\}$ with respect to the $H_0^1(Q)$ norm.
Now we want to estimate the error $e_n := u - u_n$, by the Cea's Lemma (see Remark 6.1) and the approximation inequality (6.24) it follows immediately that

$$\|e_n\|_{H_0^1} \leq 2\|u - \Pi_N u\|_{H_0^1} \leq 2C_{\Omega,1,0} N^k |u|_{H^{0,N}} = 2C_{\Omega,1,0} N \|u\|_{L^2} \leq 2C_{\Omega,k,m} N \|f\|_{L^2}. \tag{6.25}$$

Here the very last inequality follows by the stability.
We want to show that the so called *elliptic regularity* or the *Moser* trick allows to show that the convergence of the function (no weak derivatives are consider) u_N to u is, in an L^2 sense, faster than this rate. To this aim we consider the dual problem

$$\begin{cases} \Delta\phi = e_N & , \text{ in } Q \\ \phi(x) = 0 & , \forall x \in \partial Q \end{cases}. \tag{6.26}$$

Together with its weak formulation

$$\begin{cases} a(\phi, v) = (f, v)_{L^2} & , \text{ for any } v \in H_0^1(Q) \\ \phi \in H_0^1(Q) & , \end{cases}. \tag{6.27}$$

Now the elliptic regularity theory furnish the estimate

$$\|\phi\|_{H^2} \leq C\|e_N\|_{L^2},$$

the solution of this problem is indeed more regular.
On the other hand, integration by parts and the fact that ϕ solves the above problem lead to

$$\|\phi\|_{H^2}^2 \leq C\|e_N\|_{L^2}^2 = Ca(\phi, e_N).$$

Now notice that $a(e_N, \Pi_N\phi) = 0$ since both $a(u_N, \Pi_N\phi)$ and $a(u, \Pi_N\phi)$ are equal to $(f, \Pi_N\phi)_{L^2}$.
Thus we get

$$\|e_N\|_{L^2}^2 \leq Ca(\phi - \Pi_N\phi, e_N) \leq C\|e_N\|_{H^1}\|\phi - \Pi_N\phi\|_{H^2}.$$

Finally we use the approximation inequality (6.24) to get

$$\begin{aligned} &\|e_N\|_{L^2}^2 \\ \leq &Ca(\phi - \Pi_N\phi, e_N) \leq C\|e_N\|_{H^1}C_{\Omega,k,m}\|\phi - \Pi\phi\|_{H^2} \\ \leq &C\|e_N\|_{H^1}C_{\Omega,2,0}N^2\|\phi\|_{H^2} \leq CN^2\|e_N\|_{H^1}^2 \end{aligned}$$

The combination with equation (**??**) leads to

$$\|e_N\|_{L^2} \leq C'N\|e_N\|_{H^1} \leq C''N^2\|f\|_{L^2}.$$

6.4.2 Analysis of the Collocation Method

Throughout this subsubsection we assume Ω to be the interior of a parallepiped $\overline{\Omega} = [a_1, b_1] \times \cdots \times [a_d, b_d]$ and denote by $P^N(\Omega)$ the set of tensor product of polynomials having degree at most N in each variable (separately).
We assume also that for any $N > 0$ a Gauss-Radau quadrature formula

$$\int_\Omega f(x)w(x)\, dx^{(1)} \dots\, dx^{(d)} \approx \sum_{j_1=0}^{N}\sum_{j_2=0}^{N}\cdots\sum_{j_d=0}^{N} f(x_{j_1}, \dots, x_{j_d})\tilde{w}_{j_1}\dots\tilde{w}_{j_d}$$

is given with precision $2N-1$, i.e. the above integral is computed exactly if $f \in P^{2N-1}$.
Let us introduce the discrete scalar product

$$(u,v)_N := \sum_{j_1=0}^{N}\sum_{j_2=0}^{N}\cdots\sum_{j_d=0}^{N} u(x_{j_1},\ldots,x_{j_d})v(x_{j_1},\ldots,x_{j_d})\tilde{w}_{j_1}\ldots\tilde{w}_{j_d},\ \ \forall u,v\in P^N$$

and denote by $\|u\|_N$ the corresponding norm (notice that $\|u\|_N$ implies that u is the zero polynomial since the set of considered points is unisolvent of degree N, thus $\|\cdot\|_N$ is a norm). It is convenient to rearrange the indexes of our nodes of integration, precisely we enumerate them in a way such that $x_k = (x_{j_1},\ldots,x_{j_d}) \in \partial\Omega$ for any $k = 1,\ldots M_b(N)$ while $x_k \in \Omega$ for $k = M_b(N)+1,\ldots,(N+1)^d$. Also we rearrange the weights setting

$$w_k := \tilde{w}_{j_1}\cdot\cdots\cdot\tilde{w}_{j_d},\ \text{where } x_k = (x_{j_1},\ldots,x_{j_d}).$$

We define the trial function space

$$V_N := \{u \in P^N(\Omega) : \mathcal{B}(u)(x_k) = 0, \forall k = 1,\ldots M_b(N)\},$$

and the test function space

$$W_N := \{v \in P^N(\Omega) : v(x_k) = 0, \forall k = 1,\ldots M_b(N)\}.$$

We introduce the operator

$$\begin{aligned} &Q_N : \mathcal{C}^0(\overline{\Omega}) \longrightarrow W_N \\ &\quad f \longmapsto Q_N v : \begin{cases} Q_N v(x_k) = 0, & \forall k \in \{1,\ldots,M_b(N)\} \\ Q_N v(x_k) = v(x_k) & \forall k \in \{M_b(N)+1,\ldots,(N+1)^d\} \end{cases}, \end{aligned}$$

and the operator

$$\begin{aligned} &\mathcal{L}_N : P^N(\Omega) \longrightarrow P^N \\ &\quad u \longmapsto \mathcal{L}_N u \end{aligned}$$

that approximates $\mathcal{L}$ by interpolating varying coefficients and taking the derivation by interpolation.
Also we need the existence of a projection operator acting on a dense subspace $\mathcal{F}$ of $D_{\mathcal{B}}(\mathcal{L})$,

$$\begin{aligned} &\Pi_N : \mathcal{F} \longrightarrow V_N \cap D_{\mathcal{B}}(\mathcal{L}) \\ &\quad u \longmapsto \Pi_N u, \quad (\Pi_n u \equiv u,\ \text{if } u \in V_N \cap D_{\mathcal{B}}(\mathcal{L})). \end{aligned}$$

The collocation method approximate solution is defined by solving

$$\begin{aligned} &\text{Find } u_N \in V_N \text{ such that} \\ &(\mathcal{L}_N u, \ell_k)_N = (f,\ell_k)_N, \forall k = M_b(N)+1,\ldots,(N+1)^2, \end{aligned} \tag{6.28}$$

where $\ell_k(x)$ is the Lagrange basis w.r.t. the nodes x_k. Notice that the set $\{\ell_k, k = M_b(N) + 1, \dots, (N+1)^2\}$ generates W_N, indeed it is an orthogonal basis of W_N with respect to the scalar product $(\cdot,\cdot)_N$.
Due to the definition of the operator Q_N we can rewrite equation (6.28) as

$$\begin{aligned} &\text{Find } u_N \in V_N \text{ such that} \\ &Q_N(\mathcal{L}_N u - f) = 0, \text{ in } W_N, \end{aligned} \tag{6.29}$$

that is $Q_N(\mathcal{L}_N u - f)(x_k) = 0, \forall k = M_b(N) + 1, \dots, (N+1)^2$.

Proposition 6.2 (Stability of Collocation Methods)**.** *Assume that the collocation scheme is* coercive*, i.e., there exists a strictly positive constant α such that for any $N > 0$ we have*

$$\alpha \|u\|_{\mathcal{E}}^2 \leq (Q_N \mathcal{L}_N u, u)_N, \quad \forall u \in V_N \tag{6.30}$$

and that there exists a positive finite constant N such that

$$\|u\|_N := (u,u)_N^{1/2} \leq C\|u\|_{\mathcal{E}}. \tag{6.31}$$

Then the collocation scheme is stable in the sense that for each N we have

$$\|u_N\|_{\mathcal{E}} \leq \frac{C}{\alpha}\|f\|_N \tag{6.32}$$

Proof. Simply notice that

$$\|u_N\|_{\mathcal{E}}^2 \leq \frac{1}{\alpha}(Q_N \mathcal{L}_N u, u)_N = \frac{1}{\alpha}(Q_N f, u)_N \leq \frac{1}{\alpha}\|Q_N f\|_N \|u_N\|_N \leq \frac{C}{\alpha}\|Q_N f\|_N \|u_N\|_{\mathcal{E}},$$

we used that Q_N is the projection operator upon W_N □

To explain where the consistency requirements arise from we show first how to compute an error bound for collocation schemes. In the following computation we assume the bilinear form $(\mathcal{L}\cdot,\cdot)_{\mathcal{E}}$ to be continuous with norm $0 < M < \infty$. Let us set $r_N := \Pi_N u - u_N$, we have, $\forall v \in V_N$

$$\begin{aligned} &(Q_N \mathcal{L}_N r_N, v)_N = (Q_N \mathcal{L}_N u_N, v)_N - (Q_N \mathcal{L}_N \Pi_N u, v)_N \\ &= (Q_N \mathcal{L}_N u_N, v)_N - (\mathcal{L}u, v)_{\mathcal{E}} + (\mathcal{L}u, v)_{\mathcal{E}} - (Q_N \mathcal{L}_N \Pi_N u, v)_N \\ &= (Q_N \mathcal{L}_N u_N, v)_N - (f, v)_{\mathcal{E}} + (\mathcal{L}(u - \Pi_N u), v)_{\mathcal{E}} + (\mathcal{L}\Pi_N u, v)_{\mathcal{E}} - (Q_N \mathcal{L}_N \Pi_N u, v)_N. \end{aligned}$$

Now we pick $v := r_N$ and we get

$$\begin{aligned} &\alpha\|r_N\|_{\mathcal{E}}^2 \leq (Q_N \mathcal{L}_N r_N, r_N)_N \\ &= (Q_N \mathcal{L}_N u_N, r_N)_N - (f, r_N)_{\mathcal{E}} + (\mathcal{L}(u - \Pi_N u), r_N)_{\mathcal{E}} + (\mathcal{L}\Pi_N u, r_N)_{\mathcal{E}} - (Q_N \mathcal{L}_N \Pi_N u, r_N)_N \\ &\leq |\,(Q_N \mathcal{L}_N u_N, r_N)_N - (f, r_N)_{\mathcal{E}}\,| + M\|u - u_N\|_{\mathcal{E}}\|r_N\|_{\mathcal{E}} + |\,(\mathcal{L}\Pi_N u, r_N)_{\mathcal{E}} - (Q_N \mathcal{L}_N \Pi_N u, r_N)_N\,|. \end{aligned}$$

Finally we use triangle inequality to get

$$\begin{aligned} \|u-u_N\|_{\mathcal{E}} \leq \left(1+\frac{M}{\alpha}\right)\|u-\Pi_N\|_{\mathcal{E}} + \\ \frac{|\left(\mathcal{L}\Pi_N u, r_N\right)_{\mathcal{E}} - \left(Q_N\mathcal{L}_N\Pi_N u, r_N\right)_N|}{\alpha\|r_N\|_{\mathcal{E}}} + \frac{|\left(Q_N f, r_N\right)_N - \left(f, r_N\right)_{\mathcal{E}}|}{\alpha\|r_N\|_{\mathcal{E}}} \end{aligned} \tag{6.33}$$

Therefore we formulate the following consistency requirements: we say that the collocation method is consistent if the following holds true.

$$\|u-\Pi_N u\|_{\mathcal{E}} \to 0, \quad \text{as } N\to\infty, \forall u\in\mathcal{F}. \tag{6.34}$$

$$\sup_{v\in V_N} \frac{|\left(Q_N f, v\right)_N - \left(f, v\right)_{\mathcal{E}}|}{\alpha\|v\|_{\mathcal{E}}} \to 0, \quad \text{as } N\to\infty, \forall u\in\mathcal{F}. \tag{6.35}$$

$$\sup_{v\in V_N} \frac{|\left(\mathcal{L}\Pi_N u, v\right)_{\mathcal{E}} - \left(Q_N\mathcal{L}_N\Pi_N u, v\right)_N|}{\alpha\|v\|_{\mathcal{E}}} \to 0, \quad \text{as } N\to\infty, \forall u\in\mathcal{F}. \tag{6.36}$$

It follows by (6.33) that if the above consistency conditions are satisfied the method is convergent.

6.4.3 Analysis of the Tau Method

Let us start for simplicity assuming that $d=1$, i.e. $\Omega =]-1,1[$, possibly applying an affine transformation.
We consider a weight function $w:\Omega\to\mathbb{R}$ and form a basis of orthogonal polynomials ϕ_k such that ϕ_k has degree k and $\int_{-1}^{1}\phi_k(x)\phi_j(x)w(x)dx = c_k\delta_{j,k}$.
Let m be the number of imposed boundary conditions (typically we will deal with $m=2$) we set $W_N := \text{Span}\left(\phi_0,\phi_1,\ldots,\phi_{N-m}\right)$ and $V_N := \{u\in\text{Span}\left(\phi_0,\ldots,\phi_N\right), \mathcal{B}u(\pm1)=0\}$. Then the Tau Method can be written as

$$\begin{aligned} &\text{Find } u_N\in V_N \text{ such that} \\ &\left(\mathcal{L}u, v\right) = \left(f, v\right), \forall v\in W_N. \end{aligned} \tag{6.37}$$

Things becomes a little more complicated when $d>1$. We introduce the lattice of indexes

$$I^{(N)}_{\text{int}} := \{\boldsymbol{k}=(k_1,\ldots,k_d) | 0\leq k_j\leq N-m_j, j=1,2,\ldots,d\} \subset I^{(N)} = \{0,1,\ldots,N\}^d,$$

where m_j is the number of the boundary conditions imposed on the couple of faces $S_i^{\pm} := \{x_i=\pm1\}$; also we set $I^{(N)}_{\text{b}} := I^{(N)}\setminus I^{(N)}_{\text{int}}$.
Accordingly we denote by $\phi_{\boldsymbol{k}}(\boldsymbol{x})$ the functions

$$\phi_{k_1}(x_1)\cdot\cdots\cdot\phi_{k_d}(x_d), \quad \text{where } \boldsymbol{x} := (x_1,x_2,\ldots,x_d).$$

We set $V_N := \text{Span}\left(\phi_{\boldsymbol{k}}(\boldsymbol{x}), \boldsymbol{k} \in I^{(N)}\right)$, while the test function space is $W_N := \text{Span}\left(\phi_{\boldsymbol{k}}(\boldsymbol{x}), \boldsymbol{k} \in I^{(N)}_{\text{int}}\right)$. The differential equation is then expressed as

$$\begin{aligned} &\text{Find } u_N \in V_N \text{ such that} \\ &(\mathcal{L}u, \phi_{\boldsymbol{k}}) = (f, \phi_{\boldsymbol{k}}), \forall \boldsymbol{k} \in I^{(N)}_{\text{int}}. \end{aligned} \tag{6.38}$$

In order to impose the boundary conditions we consider the scalar product on $\partial\Omega$ defined as follows. First we denote by $w_i(\boldsymbol{x}) := \prod_{i \neq j=1}^{d} w(x_j)$ then set

$$(u, v)_{\partial\Omega} := \sum_{i=1}^{d} \left(\int_{S_i^+} u(\boldsymbol{x}) v(\boldsymbol{x}) w_i(\boldsymbol{x}) d\sigma(\boldsymbol{x}) + \int_{S_i^-} u(\boldsymbol{x}) v(\boldsymbol{x}) w_i(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \right).$$

Boundary conditions are imposed as

$$(\mathcal{B}u, \phi_{\boldsymbol{k}})_{\partial\Omega} = 0, \quad \forall \boldsymbol{k} \in I^{(N)}_{\text{b}}. \tag{6.39}$$

Thus the multidimensional Tau Method reads as

$$\begin{aligned} &\text{Find } u_N \in V_N \text{ such that} \\ &(\mathcal{L}u, \phi_{\boldsymbol{k}}) = (f, \phi_{\boldsymbol{k}}), \forall \boldsymbol{k} \in I^{(N)}_{\text{int}} \\ &(\mathcal{B}u, \phi_{\boldsymbol{k}})_{\partial\Omega} = 0, \quad \forall \boldsymbol{k} \in I^{(N)}_{\text{b}}. \end{aligned} \tag{6.40}$$

Since in this setting the test and trial function spaces are different, to prove the convergence we rely on a **inf-sup** type condition instead of a coercivity one.

Proposition 6.3. *Let $\mathcal{E} \subset \mathcal{H}$ be a Hilbert space such that $D_{\mathcal{B}}(\mathcal{L})$ is dense in $\mathcal{E}$. Let $\mathcal{F}$ be a dense Hilbert subspace of $\mathcal{H}$ with continuous embedding (i.e., $\exists 0 < C < \infty$ such that $\|v\|_{\mathcal{H}} \leq C\|v\|_{\mathcal{F}}$).*
Assume that there exists a finite strictly positive constant α such that

$$\alpha \leq \inf_{u \in V_N} \sup_{v \in W_N} \frac{(\mathcal{L}u, v)}{\|u\|_W \|v\|_V}. \tag{6.41}$$

Then the Tau method is stable in the sense that

$$\|u_N\|_W \leq \frac{C}{\alpha} \|f\|. \tag{6.42}$$

The proof is very similar to the Galerkin coercive case.
As in the Galerkin case if we furthermore assume the consistence of the method we get the convergence (and indeed a version of the Cea's Lemma) for free. For instance one can assume that there exists a dense subspace W_1 of $D_{\mathcal{B}}(\mathcal{L})$ and a sequence of projection operator (best

W norm projection in general works) $\Pi_N : \mathcal{H} \to V_N$ such that $\lim_N \|u - \Pi_N u\|_W \to 0$ for any $u \in W_1$.
Let us consider the stability and convergence analysis of a concrete example of Tau Method for solving the following problem

$$\begin{cases} \mathcal{L}u := -\frac{d^2}{dx^2}u + u = f & \text{in }]-1,1[\\ \frac{d}{dx}u(-1) = \frac{d}{dx}u(1) = 0 \end{cases}. \tag{6.43}$$

We set $w(x) := \frac{1}{\pi\sqrt{1-x^2}}$ and consider the related Chebyshev orthogonal polynomials $T_k(x)$. The test and trial spaces are defined as

$$V_N := \{u(x) := \sum_{k=0}^{N} c_k(u)T_k(x), \frac{d}{dx}u(-1) = \frac{d}{dx}u(1) = 0\},$$

$$W_N := \{v(x) := \sum_{k=0}^{N-2} c_k(v)T_k(x)\}.$$

We want to show that the inf sup condition (6.41) holds true. To this aim, introduce the L^2_w projection operator Π_{N-2} and note that

$$\Pi_{N-2}\mathcal{L}u = \mathcal{L}u - (u - \Pi_{N-2}u).$$

Now observe that (Parseval Identity)

$$\begin{aligned} &\sup_{v\in W_N} \frac{(\mathcal{L}u, v)}{\|v\|_V} = \frac{(\mathcal{L}u, \Pi_{N-2}\mathcal{L}u)}{\|\Pi_{N-2}\mathcal{L}u\|_V} = \frac{(\mathcal{L}u, \mathcal{L}u - (u - \Pi_{N-2}u))}{\|\Pi_{N-2}\mathcal{L}u\|_V} \\ =&\frac{1}{\|\Pi_{N-2}\mathcal{L}u\|_V}\left(\|\mathcal{L}u\|^2 - (\mathcal{L}u, u - \Pi_{N-2}u)\right) \\ \geq&\frac{\|\mathcal{L}u\|^2 - \|\mathcal{L}u\|\|u - \Pi_{N-2}u\|}{\|\Pi_{N-2}\mathcal{L}u\|_V} \end{aligned}$$

Now we use the a *priori estimate* $\|u\|_{H^2_w} \leq C\|\mathcal{L}u\|$ and the *approximation inequality* $\|u - \Pi_{N-2}u\| \leq C'N^{-2}\|u\|_{H^2_w}$ to get

$$\sup_{v\in W_N} \frac{(\mathcal{L}u, v)}{\|v\|_V} \geq \frac{1}{2CC'}\|u\|_{H^2_w}, \quad \forall N > \sqrt{2CC'}.$$

7 Tools from Approximation Theory

7.1 Elements of Fourier Analysis

7.1.1 Fourier Series

We denote by $\{\phi_k\}_{k\in\mathbb{Z}}$ the set of *Fourier Characters* $\phi_k := e^{ikx}$. The basis of Fourier Analysis are contained in the following theorem.

Theorem 7.1. *Fundamental Th. of Fourier Analysis. The sequence* $\{\phi_k\}$ *is a complete orthogonal system in* $L^2(\mathbb{T})$. *The* analysis operator *(transform)*

$$\begin{aligned} L^2(\mathbb{T}) &\longrightarrow && \ell^2(\mathbb{C}) \\ u &\mapsto && \hat{u} = \{\hat{u}_k\}_{k\in\mathbb{Z}} := \{\frac{1}{2\pi}\int_0^{2\pi} u(x)\overline{\phi}_k(x)\}_{k\in\mathbb{Z}} \end{aligned}$$

is an isometry *of* $L^2(\mathbb{T})$ *onto* $\ell^2(\mathbb{C})$, *i.e.,*

$$\|u\|_{L^2(\mathbb{T})} = \|\hat{u}\|_{\ell^2(\mathbb{C})} \quad \textit{(Parseval Identity).}$$

The inverse of this application is the synthesis operator

$$\hat{u} \mapsto \sum_{k=-\infty}^{\infty} \hat{u}_k\phi_k,$$

where the convergence of the series has to be intended in the $L^2(\mathbb{T})$ *sense.*

Remark 7.1. *The analysis makes sense even if* u *is merely* L^1 *instead of* L^2, *it turns out to be a* norm decreasing operator *of* L^1 *on* ℓ^∞.

Definition 7.2. Truncated series. We denote by $S_N u(x)$ the truncated Fourier series up to $|k| \le N$, i.e.,

$$S_N u(x) := \sum_{k=-N}^{N} \hat{u}_k\phi_k \in P^N(\mathbb{T}).$$

Here $P^N(\mathbb{T}) := \text{Span}\,(\phi_k, |k| \le N)$ is the space of trig-polynomials of degree at most N.

Remark 7.2. *Note that by the Pyhagorean Theorem* $S_N u(x)$ *is the best* $L^2(\mathbb{T})$ *approximation to* u *belonging to* $P^N(\mathbb{T})$.

Due to Theorem 7.1 $S_N u \to u$ in $L^2(\mathbb{T})$. A antural question one may ask is weather the convergence holds point-wise or even, possibly under additional conditions, uniformly. We postpone this for a while. For the moment we just cite that Kolmogorov invented a technique to build examples of bounded continuous functions on $\mathbb{T}$ such that their Fourier series is almost everywhere divergent!
The other natural question to ask (especially to our aims, cfr. Ch.…) is how farst the Fourier coefficients of a differentiable function u decrease, in other words, how fast is the Fourier truncated series convergent.

Proposition 7.3. *Coefficients decay.*

1. *Let $u \in \mathcal{C}^m([0, 2\pi])$ with $u^{(j)}$ periodic for any $0 \leq j \leq m-1$, then*

$$\hat{u}_k = \mathcal{O}(k^{-m}).$$

2. *Let $u \in \mathcal{C}^\infty([0, 2\pi])$ be periodic together with all its derivatives, then*

$$\hat{u}_k = \mathcal{O}(k^{-M}),\ \forall M \in \mathbb{N}, \quad \textit{(spectral convergence).}$$

To prove the above result one needs simply to notice that

$$2\pi\hat{u}_k = \int_0^{2\pi} u(x)\phi_k(x)\,dx = -(ik)^{-1}(u(0^+)-u(2\pi^-))+(ik)^{-1}\int_0^{2\pi} u'(x)\phi_k(x)\,dx = (ik)^{-1}\hat{u'}\ldots$$

Remark 7.3. *It is quite relevant to note that the same argument above can be used to prove that, for a differentiable periodic function u with periodic derivative*

$$(S_N u)'(x) = \sum_{k=-N}^{N} \hat{u}_k \phi_k'(x) = \sum_{k=-N}^{N} ik\hat{u}_k\phi_k(x) = \sum_{k=-N}^{N} \hat{u'}\phi_k(x) = S_N(u')(x).$$

That is, projection commutes with differentiation.
We can note also that the Fourier basis offers the easiest example of differentiation matrix. The function: the function $S_N u$ is represented by its coefficients $\hat{u}|_N$, the coefficients $\hat{u'}|_N$ of $S_N(u')$ can be determined by left multiplying the vector $\hat{u}$ by the differentiation matrix *$D_N = diag(-iN, -i(N-1), \ldots, 0, i, \ldots, iN)$, that is*

$$\hat{u'}|_N = D_N \hat{u}|_N.$$

We now take into account the issue of point-wise convergence of Fourier series. As a matter of fact, the fundamental results and techniques in this very classical area of Analysis are very difficult and the (possibly multi-dimensional) modern Fourier transform and analysis heavily rely on these results.

We recall that a function $f :]0, 2\pi[\to \mathbb{R}$ is said to be of *bounded variation*, $f \in BV(]0, 2\pi[)$ in symbols, if

$$\sup_{0<x_1<\ldots\cdots<x_M<2\pi} \sum_{i=1}^{M-1} |f(x_{i+1}) - f(x_i)| < \infty.$$

By definition this is the variation (i.e., the length) of the graph of f. This definition is usually founded in the context of Riemann integration, the equivalent definition in the context of Lebesgue integration is $f \in L^1(]0, 2\pi[)$ such that

$$\sup_{\psi \in \mathcal{C}^1_c(]0,2\pi[), |\psi| \leq 1} \int \psi' f \, dx < \infty.$$

Probably the most useful results regarding the point-wise convergence of Fourier series are the ones we collect in the following theorem.

Theorem 7.4. *Point-wise and uniform convergence.*

- *Let $f \in BV(]0, 2\pi[)$ be periodic, then $S_N f(x) \to \frac{1}{2}(f(x^+) + f(x^-))$.*
- *If f is furthermore continuous, the the convergence is uniform in x.*

In the case of jump discontinuities, note that this is compatible with the assumption $f \in BV(]0, 2\pi[)$ the sequence of Fourier truncates exhibits a oscillatory behaviour near the discontinuity, whose frequency increases as $N \to \infty$. This issue is termed the *Gibbs Phenomenon.*
Let us consider the subspace $H^m_{per}(0, 2\pi)$ of $H^m(0, 2\pi)$ consisting of square integrable functions that are periodic together with their first $m - 1$ weak derivatives and have square integrable m derivatives. We use the classical Sobolev norm

$$\|u\|_{H^m(0,2\pi)} := \left(\sum_{l=0}^{m} \int_0^{2\pi} |u^{(l)}(x)|^2 \, dx \right)^{1/2}.$$

We can show, just working with smooth approximations to u, that for any $u \in H^m_{per}(0, 2\pi)$ we have $u^{(1)} = \sum_{k=-\infty}^{\infty} ik\hat{u}_k \phi_k$, therefore, using Parseval Identity, the Sobolev norm $\|u\|_{H^m(0,2\pi)}$ is indeed equivalent to the norm

$$\|u\|_m := \left(\sum_{k=-\infty}^{\infty} (1 + |k|^{2m}) |\hat{u}_k|^2 \right)^{1/2}.$$

Even much more can be said: *we can characterize the space $H^m_{per}(0, 2\pi)$ as the set of periodic square integrable functions for which we can differentiate the Fourier series term-wise.*

Proposition 7.5. *Truncation Error. Let $u \in H^m_{per}(0, 2\pi)$, then the following error bounds holds.*

$$\|u - S_N u\|_{L^2} \leq CN^{-m}\|u^{(m)}\|_{L^2} \leq CN^{-m}\|u\|_{H^m}. \tag{7.1}$$

$$\|u - S_N u\|_{H^l} \leq CN^{l-m}\|u^{(m)}\|_{L^2} \leq CN^{l-m}\|u\|_{H^m}, \quad \forall l = 0, 1, \ldots, m-1. \tag{7.2}$$

Proof.

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\|u - S_N u\|_{L^2} = \left(\sum_{|k|>N} |\hat{u}_k|^2\right)^{1/2} = \left(\sum_{|k|>N} \frac{1}{|k|^{2m}}|k|^{2m}|\hat{u}_k|^2\right)^{1/2} \\
\leq & N^{-m}\left(\sum_{|k|>N} |k|^{2m}|\hat{u}_k|^2\right)^{1/2} \leq CN^{-m}\|u^{(m)}\|_{L^2} \leq CN^{-m}\|u\|_{H^m}.
\end{aligned}$$

□

7.1.2 Fourier Interpolation

7.2 Polynomial Approximation, Interpolation and Quadrature

8 A pseudo-spectral solution to the Stokes Problem

8.1 The Method

8.1.1 Generalities

We are interested in setting up a pseudo-spectral method for the following Stokes Problem

$$\begin{cases} \Delta \boldsymbol{u} - \sigma \boldsymbol{u} - \nabla p = \boldsymbol{f} & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \\ \boldsymbol{u}|_\Gamma = \boldsymbol{u}_\Gamma & \text{on } \Gamma := \partial\Omega \end{cases}, \tag{S-I}$$

where $\sigma \geq 0$, $\Omega \subset \mathbb{R}^2$ is a bounded domain, $\boldsymbol{u} : \bar{\Omega} \to \mathbb{R}^2$ is the velocity field and $p : \bar{\Omega} \to \mathbb{R}$ is the pressure.
It is worth to stress that this particular instance of the Stokes Problem may arise when one deals with a Stokes evolution equation and uses numerical integration: a problem like (S-I) must be solved aty each time step.
First we observe that $\nabla \cdot \Delta \boldsymbol{u} = \Delta \nabla \cdot \boldsymbol{u} = \Delta(0) = 0$ thus the first two equations in (S-I) implies $\Delta p = \nabla \cdot \boldsymbol{f}$. Also we introduce the *Helmotz operator* $H_\sigma := \Delta - \sigma Id$ of parameter σ so we can re-write (S-I) in the form

$$\begin{cases} H_\sigma \boldsymbol{u} - \nabla p = \boldsymbol{f} & \text{in } \Omega \\ \Delta p = \nabla \cdot \boldsymbol{f} & \text{in } \Omega \\ \boldsymbol{u}|_\Gamma = \boldsymbol{u}_\Gamma & \text{on } \Gamma := \partial\Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{on } \Gamma \end{cases}, \tag{S-II}$$

It is rather important to note that (at least at this stage, i.e., no discretization procedure performed so far) the latter formulation implies the first one. For (let us reason in term of classical solutions for simplicity), set $Q := \nabla \cdot \hat{\boldsymbol{u}}$, where $\hat{\boldsymbol{u}}$ solves (S-II). It follows that Q solves

$$\begin{cases} \Delta Q = \sigma Q & \text{in } \Omega \\ Q = 0 & \text{on } \Gamma \end{cases},$$

thus $\nabla \cdot \hat{\boldsymbol{u}} = Q \equiv 0$ in Ω, that is $\hat{\boldsymbol{u}}$ is *divergence-free* and thus solves (S-I) as well.

8.1.2 Rough Chebyshev-Chebyshev discretization

We focus on the very simple case of $\Omega =]-1, 1[^2$. This leads to use Chebyshev polynomials as basis functions. Let us denote by $\mathbb{P}_2^N$ the space of tensor product polynomials of degree at most N in each of the two variables separately.

We will use the Chebyshev-Lobatto quadrature nodes

$$x_i = \cos\left(\frac{\pi i}{N}\right), \quad i = 0, 1, \ldots, N$$
$$y_j = \cos\left(\frac{\pi j}{N}\right), \quad j = 0, 1, \ldots, N$$

and define the following computational domains

$$\begin{aligned}
\Omega_N &:= \{(x_i, y_j) : (i,j) \in \{1, 2, \ldots, N-1\} \times \{1, 2, \ldots, N-1\}\},\\
\overline{\Omega}_N &:= \{(x_i, y_j) : (i,j) \in \{0, 1, \ldots, N\} \times \{0, 1, \ldots, N\}\},\\
\overline{\Omega}_N^{\text{int}} &:= \{(x_i, y_j) \in \overline{\Omega}_N : \, (i,j) \notin \{0, N\} \times \{0, N\}\},\\
\Gamma_N &:= \{(x_i, y_j) : (i,j) \in \{0, N\} \times \{0, 1, \ldots, N\} \cup \{0, 1, \ldots, N\} \times \{0, N\}\},\\
\Gamma_N^{\text{int}} &:= \{(x_i, y_j) \in \Gamma_N : \, (i,j) \notin \{0, N\} \times \{0, N\}\},
\end{aligned}$$

corresponding respectively to

- interior nodes
- all nodes
- all nodes but the corners
- boundary nodes
- all boundary nodes but the corners.

The resulting discretized problem becomes: *find* $\boldsymbol{u}^N \in \mathbb{P}_2^N \times \mathbb{P}_2^N$ *and* $p^N \in \mathbb{P}_2^N$ *such that, denoting by* $\boldsymbol{f}^N in \mathbb{P}_2^N \times \mathbb{P}_2^N$ *the interpolating polynomial of* $\boldsymbol{f}$ *on* $\overline{\Omega}_N$*, we have*

$$\begin{cases}
H_\sigma \boldsymbol{u}^N - \nabla p^N = \boldsymbol{f}^N & \text{in } \Omega_N \\
\Delta p^N = \nabla \cdot \boldsymbol{f}^N & \text{in } \Omega_N \\
\boldsymbol{u}_N|_\Gamma = \boldsymbol{u}_\Gamma & \text{on } \Gamma_N^{\text{int}} \\
\nabla \cdot \boldsymbol{u}_N = 0 & \text{on } \Gamma_N^{\text{int}}
\end{cases}, \tag{S-CC}$$

We stress that, as usual in PS methods, the derivatives are taken in the interpolation sense, that is one first interpolates than takes derivatives.

8.1.3 A variational crime

In this discretization procedure a remarkable phenomena comes into play: even if any classical solution $\boldsymbol{u}$ of (S-II) is a divergence free vector field, its approximation $\boldsymbol{u}^N$ does not need to fulfil the same property. This is a consequence of imposing the differential equations in the strong sense point-wise. Remember that, when showing that a solution to (S-II) needs to be divergence free, we used that the laplacian of an *identically vanishing* function is zero. Here equations are satisfied only on a finite sets of points so we can't perform the same reasoning: in general $\nabla \cdot \boldsymbol{u}_N \not\equiv 0$ on Ω.

Recall that the divergence free condition represent the conservation of the mass, hence trying to preserve such a condition seems very reasonable. More concretely, as often occurs, the lack of physical meaning of the discretized equation (S-CC) boils down to numerical instability in solving the steady state problem (S-I) and even worse behaviour of the final numerical solution if (S-I) arises as a intermediate problem in the time integration of an evolution equation. We need to overcome this problem.

We introduce a *numerical flux* $\boldsymbol{B}^N \in \mathbb{P}_2^N \times \mathbb{P}_2^N$ such that its normal component at the boundary is precisely the normal component of the residual, more precisely we replace the problem (S-CC) with

$$\begin{cases} H_\sigma \boldsymbol{u}^N - \nabla p^N = \boldsymbol{f}^N & \text{in } \Omega_N \\ \Delta p^N = \nabla \cdot \boldsymbol{f}^N - \nabla \cdot \boldsymbol{B}^N & \text{in } \Omega_N \\ \boldsymbol{B}^N = 0 & \text{in } \Omega_N \\ \boldsymbol{u}_N|_\Gamma = \boldsymbol{u}_\Gamma & \text{on } \Gamma_N^{\text{int}} \\ \boldsymbol{B}^N \cdot \nu = (H_\sigma \boldsymbol{u}^N - \nabla p^N - \boldsymbol{f}^N) \cdot \nu & \text{on } \Gamma_N^{\text{int}} \end{cases} . \tag{S-CC'}$$

Here ν is the outer unit normal.

At this stage it is not clear why this *variational crime* should enforce the divergence free condition on $\boldsymbol{u}_N$, indeed this will surface out in a while when presenting the *Influence Matrix Method* that we use to solve (S-CC').

8.1.4 Influence Matrix

We split (S-CC') in two sub problems. More precisely, we assume that

$$\boldsymbol{u}_N = \tilde{\boldsymbol{u}}^N + \bar{\boldsymbol{u}}^N, \quad p_N = \tilde{p}^N + \bar{p}^N, \quad \boldsymbol{B}^N = \tilde{\boldsymbol{B}}^N + \bar{\boldsymbol{B}}^N,$$

where each $\tilde{}$ and $\bar{}$ function solves respectively the $\tilde{}$ or the $\bar{}$ problem below.

$$\begin{cases} \Delta\tilde{p}^N = -\nabla \boldsymbol{f}^N & \text{in } \Omega_N \\ \tilde{p}^N = 0 & \text{on } \Gamma_N^{\text{int}} \end{cases}, \qquad (\tilde{}\text{problem I})$$

$$\begin{cases} H_\sigma \tilde{\boldsymbol{u}}^N = \boldsymbol{f}^N + \nabla\tilde{p}^N & \text{in } \Omega_N \\ \tilde{\boldsymbol{u}}^N = \boldsymbol{u}_\Gamma & \text{on } \Gamma_N^{\text{int}} \end{cases}, \qquad (\tilde{}\text{problem II})$$

$$\begin{cases} \tilde{\boldsymbol{B}}^N = 0 & \text{in } \Omega_N \\ \tilde{\boldsymbol{B}}^N \cdot \nu = (H_\sigma \tilde{\boldsymbol{u}}^N - \boldsymbol{f}^N - \nabla p^N) \cdot \nu & \text{on } \Gamma_N^{\text{int}} \\ \tilde{\boldsymbol{B}}^N \cdot \tau = 0 & \text{on } \Gamma_N^{\text{int}} \end{cases}, \qquad (\tilde{}\text{problem III})$$

Here τ is the unit tangent.

$$\begin{cases} \Delta\bar{p}^N = -\nabla(\tilde{\boldsymbol{B}}^N + \bar{\boldsymbol{B}}^N) & \text{in } \Omega_N \\ H_\sigma \bar{\boldsymbol{u}}^N = \nabla\bar{p}^N & \text{in } \Omega_N \\ \bar{\boldsymbol{u}}^N = 0 & \text{on } \Gamma_N^{\text{int}} \\ \nabla \cdot \bar{\boldsymbol{u}}^N = -\nabla \cdot \tilde{\boldsymbol{u}}^N & \text{on } \Gamma_N^{\text{int}} \\ \bar{\boldsymbol{B}}^N \cdot \nu = (H_\sigma \bar{\boldsymbol{u}}^N - \nabla\bar{p}^N)\nu & \text{on } \Gamma_N^{\text{int}} \end{cases} \qquad (\bar{}\text{problem})$$

First, notice that the $\tilde{}$ problems I, II and III can be considered separately in this order. We compute $\tilde{p}^N$ solving ($\tilde{}$problem I), then we compute the gradient of $\tilde{p}^N$, we plug it in ($\tilde{}$problem II) and we compute $\tilde{\boldsymbol{u}}^N$ by solving it; finally $\tilde{\boldsymbol{B}}^N \cdot \nu$ is computed and stored. Recall that the solution to our final problem consists in finding $\boldsymbol{u}^N$ and p^N, while the numerical flux is an additional variable whose tangential or normal values we are going to compute only when these are required in order to compute $\boldsymbol{u}^N$ or p^N.
Now we consider ($\bar{}$problem). We use a quite unusual method,termed *superposition method*. In this technique each function ϕ of the set $\{\bar{p}^N, \bar{\boldsymbol{u}}^N, R^N := \bar{\boldsymbol{B}}^N \cdot \nu, \boldsymbol{B}\}$ is expanded in the form

$$\phi = \sum_{l=1}^{2L} \xi_l \phi_l, \quad L := \text{Card}\Gamma_N^{\text{int}} = 2(2N-2), \tag{8.1}$$

but the coefficients are fixed: what varies are the ϕ_l themselves.
More precisely, we consider two families of problems, each of them of L problems.
For $l = 1, 2, \ldots, L$ solve

$$\begin{cases} \Delta\bar{p}_l^N = 0 & \text{in } \Omega_N \\ \bar{p}_l^N(\eta_m) = \delta_{l,m} & \eta_m \in \Gamma_N^{\text{int}} \end{cases} \tag{8.2}$$

$$\begin{cases} \Delta\bar{\boldsymbol{u}}_l^N = \nabla\bar{p}_l^N & \text{in } \Omega_N \\ \bar{\boldsymbol{u}}_l^N = 0 & \text{on } \Gamma_N^{\text{int}} \end{cases} \tag{8.3}$$

$$\boldsymbol{B}_l^N = 0 \quad \text{in } \overline{\Omega}_N \tag{8.4}$$

For $l = L+1, L+2, \dots, 2L$ solve

$$\begin{cases} \boldsymbol{B}_l^N = 0 & \text{in } \Omega_N \\ \boldsymbol{B}_l^N \cdot \nu(\eta_m) = \delta_{L+m,l} & \eta_m \in \Gamma_N^{\text{int}} \end{cases} \tag{8.5}$$

$$\begin{cases} \Delta \bar{p}_l^N = -\nabla \boldsymbol{B}_l^N \cdot & \text{in } \Omega_N \\ \bar{p}_l^N = 0 & \text{in } \Omega_N \end{cases} \tag{8.6}$$

$$\begin{cases} H_\sigma \bar{\boldsymbol{u}}^N = \nabla \bar{p}^N & \text{in } \Omega_N \\ \bar{\boldsymbol{u}}^N = 0 & \text{on } \Gamma_N^{\text{int}} \end{cases} \tag{8.7}$$

Again, one first solves (8.2)to determine $\bar{p}_l^N$, $l \in \{1, \dots, L\}$, computes the gradient of the pressure and solves (8.3) by plugging it into the equation. This leads to compute $\bar{\boldsymbol{u}}_N^l$, $l \in \{1, \dots, L\}$, then we compute its laplacian and finally the values $\boldsymbol{R}_l^N = (\Delta \bar{\boldsymbol{u}}_N^l - \nabla \bar{p}_l^N) \cdot \nu$ at the nodes in Γ_N^{int}.

Now consider the case $l > L$. It is worth to stress that (8.5) fully characterize the values of $\nabla \cdot \boldsymbol{B}_l^N$ at points of Ω_N, this is a consequence of the tensor product construction of both the domain and the functions space we took into account: notice that the normal derivatives are polynomial of only one variable. In order to get convinced is convenient to draw a picture of $\overline{\Omega}_3$ (the easiest example), pick any l, m and write the values of $\boldsymbol{B}_l^N$ given by (8.5) and try to compute $\nabla \cdot \boldsymbol{B}_l^N(x_2, y_2)$ (the only node in Ω_3).

After determining the value of $\nabla \cdot \boldsymbol{B}_l^N$ at Ω_N by solving (8.5), we plug it into (8.6), we solve it and we compute $\bar{p}_l^N$ and $\boldsymbol{R}_l^N(\eta_m)$ for $l = L+1, L+2, \dots, 2L$, $m = 1, \dots, L$.

Finally we compute the gradient of $\bar{p}_l^N$, we plug it in (8.7) and we solve it to determine $\bar{\boldsymbol{u}}_l^N$ and $\boldsymbol{R}_l^N$ for $l > L$.

Now note that any set of functions $\{\bar{p}^N, \bar{\boldsymbol{u}}^N, \boldsymbol{R}^N, \boldsymbol{B}\}$ defined by a *superposition* of the form (8.1), that is for any $\boldsymbol{\xi} \in \mathbb{R}^{2L}$, satisfy the first three equations in ($^-$problem); we want to determine $\boldsymbol{\xi}$ by imposing the remaining two equations.

$$\begin{cases} \sum_{l=1}^{2L} \xi_l \nabla \cdot \bar{\boldsymbol{u}}_l^N(\eta_m) = -\nabla \cdot \tilde{\boldsymbol{u}}^N(\eta_m) & \eta_m \in \Gamma_N^{\text{int}} \\ \sum_{l=1}^{2L} \xi_l (\boldsymbol{B}_l^N \cdot \nu - \boldsymbol{R}_l^N)(\eta_m) = (\tilde{\boldsymbol{B}}^N \cdot \nu)(\eta_m) & \eta_m \in \Gamma_N^{\text{int}} \end{cases} \tag{8.8}$$

This system may be written in the form $\mathcal{I}^N \boldsymbol{\xi} = c$, where the matrix

$$\mathcal{I}^N := \begin{bmatrix} \nabla \cdot \bar{\boldsymbol{u}}_1^N(\eta_1) & \nabla \cdot \bar{\boldsymbol{u}}_2^N(\eta_1) & \dots & \dots & \dots & \nabla \cdot \bar{\boldsymbol{u}}_{2L}^N(\eta_1) \\ \nabla \cdot \bar{\boldsymbol{u}}_1^N(\eta_2) & \nabla \cdot \bar{\boldsymbol{u}}_2^N(\eta_2) & \dots & \dots & \dots & \nabla \cdot \bar{\boldsymbol{u}}_{2L}^N(\eta_2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \nabla \cdot \bar{\boldsymbol{u}}_1^N(\eta_L) & \nabla \cdot \bar{\boldsymbol{u}}_2^N(\eta_L) & \dots & \dots & \dots & \nabla \cdot \bar{\boldsymbol{u}}_{2L}^N(\eta_L) \\ (\boldsymbol{B}_1^N \cdot \nu - \boldsymbol{R}_1^N)(\eta_1) & (\boldsymbol{B}_2^N \cdot \nu - \boldsymbol{R}_2^N)(\eta_1) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N \cdot \nu - \boldsymbol{R}_{2L}^N)(\eta_1) \\ (\boldsymbol{B}_1^N \cdot \nu - \boldsymbol{R}_1^N)(\eta_2) & (\boldsymbol{B}_2^N \cdot \nu - \boldsymbol{R}_2^N)(\eta_2) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N \cdot \nu - \boldsymbol{R}_{2L}^N)(\eta_2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ (\boldsymbol{B}_1^N \cdot \nu - \boldsymbol{R}_1^N)(\eta_L) & (\boldsymbol{B}_2^N \cdot \nu - \boldsymbol{R}_2^N)(\eta_L) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N \cdot \nu - \boldsymbol{R}_{2L}^N)(\eta_L) \end{bmatrix}$$

is the *Influence Matrix* and

$$c = \begin{pmatrix} -\nabla \cdot \tilde{\boldsymbol{u}}^N(\eta_1) \\ -\nabla \cdot \tilde{\boldsymbol{u}}^N(\eta_2) \\ \vdots \\ -\nabla \cdot \tilde{\boldsymbol{u}}^N(\eta_L) \\ (\tilde{\boldsymbol{B}}^N \cdot \nu)(\eta_1) \\ (\tilde{\boldsymbol{B}}^N \cdot \nu)(\eta_L) \\ \vdots \\ (\tilde{\boldsymbol{B}}^N \cdot \nu)(\eta_L) \end{pmatrix}.$$

Unfortunately a problem may arise when solving the influence matrix system: in general the matrix $\mathcal{I}^N$ may be not invertible; experimentally four zero singular values appear. Two situations may occur: if c lies in the image of $\mathcal{I}^N$ the system has solution(s), otherwise there exists no solution.

We present two methods to cope with this issue, the firs one, due to ..., is a spectral regularizing technique that leads to recover a true solution if there exists one. The second is a mixed approach between exact solution and least squares approximation.

Regularization method. We compute the factorization

$$\mathcal{I}^N = U^T \Sigma V, \;\; \Sigma = \mathbb{I}\sigma, \;\; \sigma = (\sigma_1, \dots, \sigma_{2L-4}, 0, 0, 0, 0)^T$$

and we replace $\mathcal{I}^N$ by the regularized matrix

$$\mathcal{I}_\lambda^N := U^T \Sigma_\lambda V, \;\; \Sigma_\lambda = \mathbb{I}\sigma_\lambda, \;\; \sigma = (\sigma_1, \dots, \sigma_{2L-4}, \lambda, \lambda, \lambda, \lambda)^T.$$

Then we solve the system $\mathcal{I}_\lambda^N \boldsymbol{\xi} = c$. Note that, if a "true" solution $\hat{\boldsymbol{\xi}}$ of the original system exists, then $\hat{\xi}$ is a solution of the regularized system as well.

Mixed approach. We divide the influence matrix $\mathcal{I}^N \in M_{2L\times 2L}(\mathbb{R})$ in two sub-matrices $A, B \in M_{L\times L}(\mathbb{R})$

$$\mathcal{I}^N = \left[\begin{array}{c} A \\ B \end{array}\right],$$

$$A := \left[\begin{array}{cccccc} \nabla\cdot\bar{\boldsymbol{u}}_1^N(\eta_1) & \nabla\cdot\bar{\boldsymbol{u}}_2^N(\eta_1) & \dots & \dots & \dots & \nabla\cdot\bar{\boldsymbol{u}}_{2L}^N(\eta_1) \\ \nabla\cdot\bar{\boldsymbol{u}}_1^N(\eta_2) & \nabla\cdot\bar{\boldsymbol{u}}_2^N(\eta_2) & \dots & \dots & \dots & \nabla\cdot\bar{\boldsymbol{u}}_{2L}^N(\eta_2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \nabla\cdot\bar{\boldsymbol{u}}_1^N(\eta_L) & \nabla\cdot\bar{\boldsymbol{u}}_2^N(\eta_L) & \dots & \dots & \dots & \nabla\cdot\bar{\boldsymbol{u}}_{2L}^N(\eta_L) \end{array}\right]$$

$$B := \left[\begin{array}{cccccc} (\boldsymbol{B}_1^N\cdot\nu - \boldsymbol{R}_1^N)(\eta_1) & (\boldsymbol{B}_2^N\cdot\nu - \boldsymbol{R}_2^N)(\eta_1) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N\cdot\nu - \boldsymbol{R}_{2L}^N)(\eta_1) \\ (\boldsymbol{B}_1^N\cdot\nu - \boldsymbol{R}_1^N)(\eta_2) & (\boldsymbol{B}_2^N\cdot\nu - \boldsymbol{R}_2^N)(\eta_2) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N\cdot\nu - \boldsymbol{R}_{2L}^N)(\eta_2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ (\boldsymbol{B}_1^N\cdot\nu - \boldsymbol{R}_1^N)(\eta_L) & (\boldsymbol{B}_2^N\cdot\nu - \boldsymbol{R}_2^N)(\eta_L) & \dots & \dots & \dots & (\boldsymbol{B}_{2L}^N\cdot\nu - \boldsymbol{R}_{2L}^N)(\eta_L) \end{array}\right]$$

We solve exactly the problem $A\boldsymbol{\xi} = (c_1, \dots, c_L)^T$. To do that we use the QR factorization

$$A = R^T Q^T = \left[\begin{array}{c} R_1^T \\ \mathbb{O} \end{array}\right] Q^T.$$

We get

$$\boldsymbol{\xi}^1 := (\xi_1, \dots, \xi_M)^T = Q R_1^{-T} c_1, \quad M := \text{Rank } A.$$

Then we split the matrix B as $B = [B_1, B_2]$ with $B_1 \in M_{L,M}(\mathbb{R})$ thus the condition $B\boldsymbol{\xi} = c_2$ is equivalent to

$$B_1\boldsymbol{\xi}^1 + B_2\boldsymbol{\xi}^2 = c_2 \iff B_2\boldsymbol{\xi}^2 = c_2 - B_1\boldsymbol{\xi}^1.$$

This last equation is finally solved with respect to the variable $\boldsymbol{\xi}^2$ in the least squares sense.

8.2 Implementation

A Finite difference discretization of the convection diffusion equation.

The finite difference method for the discretization of the constant coefficient convection-diffusion equation in one dimension proceeds as follows. Consider a uniform partition of the interval $I = [0, 1]$ into n subintervals of length h and whose endpoints are given by x_i $x_{i+1} = x_i + h$, $i = 0, 1, \ldots, n + 1$. Let u_i be the numerical approximation of the solution u at point x_i: $u_i \approx u(x_i)$. Using Taylor series expansions we have:

$$u_{i+1} = u_i + hu'_i + \frac{h^2}{2}u''_i + \ldots \tag{A.1}$$

$$u_{i-1} = u_i - hu'_i + \frac{h^2}{2}u''_i + \ldots. \tag{A.2}$$

Summing the two equations and neglecting the higher order terms we recover the following approximation of the second derivative at the i-th node:

$$u''_i \approx \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2}.$$

Subtracting the equations (A.1) and (A.2) we obtain a centered approximation for the the first derivative in x_i:

$$u'_i \approx \frac{u_{i+1} - u_{i-1}}{2h},$$

Using these to approximation in the differential equation (2.50) we obtain the difference equation:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0,$$

that can be written for all $i = 1, \ldots, n$. This equation coincides with (2.51) and is an approximation of order of accuracy $\mathcal{O}\left(()\, h^2\right)$.
The first derivative can be expressed using a first order approximation that uses the forward (eq. (A.1)) or the backward (eq. (A.2)) Taylor expansions. We then obtain:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0,$$
$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0.$$

B Finite difference operators

We introduce in this section the classical difference operators that are often used in developing finite difference approximations to differential equations. The reader is referred to general books on numerical analysis, such as Stoer and Bulirsch [25] for more details.
Let $\{u_n\}$ be a sequence of real numbers. The shift, forward difference, and backward difference operators E, ∇, Δ are defined as:

$$\mathrm{E} : u_n \mapsto u_{n+1}; \qquad \mathrm{E}^{-1} : u_{n+1} \mapsto u_n; \qquad \Delta : u_n \mapsto (u_{n+1} - u_{jndx}); \qquad \nabla : u_n \mapsto (u_n - u_{n-1}).$$

Noting that the inverse of the shift operator exists, we can write immediately the relationships:

$$\Delta = \mathrm{E} - I = E\nabla; \qquad \nabla = I - \mathrm{E}^{-1}; \qquad \mathrm{E} = (I - \nabla)^{-1};$$

Hence, we can write, for any positive integer m:

$$\Delta^m u_n = (\mathrm{E} - I)^m u_n = \sum_{j=0}^{m} (-1)^j \binom{m}{j} u_{n+m-j}$$

$$\nabla^m u_n = (I - \mathrm{E}^{-1})^m u_n = \sum_{j=0}^{m} (-1)^j \binom{m}{j} u_{n-j}$$

Assuming that all derivatives of u exists in the appropriate sense, we can write formally

$$\begin{aligned} \mathrm{E}^s u_n &= u(t_n + sh) = u_n + sh\frac{du_n}{dt} + \frac{1}{2}(sh)^2 \frac{d^2 u_n}{dt^2} + \dots \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (sh)^k \frac{d^k}{dt^k} u_n = \sum_{k=0}^{\infty} \frac{1}{k!} \left(sh\frac{d}{dt} \right)^k u_n \\ &= e^{(sh\frac{d}{dt})^k} u_n. \end{aligned}$$

C Numerical solution of Ordinary Differential Equations

In this chapter we address the problem of solving systems of ODEs (Ordinary Differential Equations). It is an elementary introduction to a vast subject matter that has reached nowadays a mature development. These notes are a personal re-elaboration and synthesis of the material presented in [23]. For more details we refer to classical textbooks such as [12, 13, 14, 1].

C.1 The Cauchy problem

The Cauchy problem looks for the solution $y(t)$ of a first-order differential equation in which the derivative of $y(t)$ is equal to a given function $f(t, y(t))$. Moreover, the function $y(t)$ satisfies an initial condition$y(t_0) = y_0$. Consider the following:

Problem C.1 (Cauchy Problem)**.** Find $y(t) \in \mathcal{C}^1(I)$, $I = [t_0, T]$ $(0 < T < \infty)$, such that:

$$\begin{aligned} &y'(t) = f(t, y(t)) \qquad t \in I, \\ &y(t_0) = y_0, \qquad \text{Initial Condition (IC)} \end{aligned} \tag{C.1}$$

where $f(t, y) : S \to \mathbb{R}$, $S = I \times (-\infty, +\infty)$.

The solution this problem is a function $y(t) : I \to \mathbb{R}$ which can be written in the following integral form:

$$y(t) = y_0 + \int_{t_0}^{t} f\left(\tau, y(\tau)\right) d\tau. \tag{C.2}$$

We can give the following "local" interpretation of the Cauchy problem. Let δ and η two positive real numbers, and let $t_0 \in I$ and $y_0 \in \mathbb{R}$. We form the intervals $K = [t_0 - \delta, t_0 + \delta]$ and $J = [y_0 - \eta, y_0 + \eta]$, and we let $I \times J \subseteq \Sigma \subset \mathbb{R}$. We assume that the function $f : \Sigma \to \mathbb{R}$ is Lipschitz[9] in Σ, i.e., there exist a constant $L > 0$ such that

$$| f(t, y_1) - f(t, y_2) | \leq L | y_1 - y_2 |$$

[9] A Lipschitz function f is continuous in J, but not all continuous functions are Lipschitz. For example, $g(y) = | y |^{1/3}$ is continuous but not Lipschitz in $[-1, 1]$, as

$$\frac{g(y) - g(x)}{y - x} \sim x^{-2/3} \qquad \text{as } x \text{ approaches } 0.$$

In particular, a function $g \in \mathcal{C}^1(J)$ and such that a constant $K > 0$ exists so that $| g'(y) | \leq K$ for all $y \in J$ is Lipschitz in J. In fact:

$$| g(y_1) - g(y_2) | = | g'(\xi)(y_1 - y_2) | \leq K | y_1 - y_2 |.$$

The function $g(y) = | y | \notin \mathcal{C}^1(I)$ but is evidently Lipschitz in I.

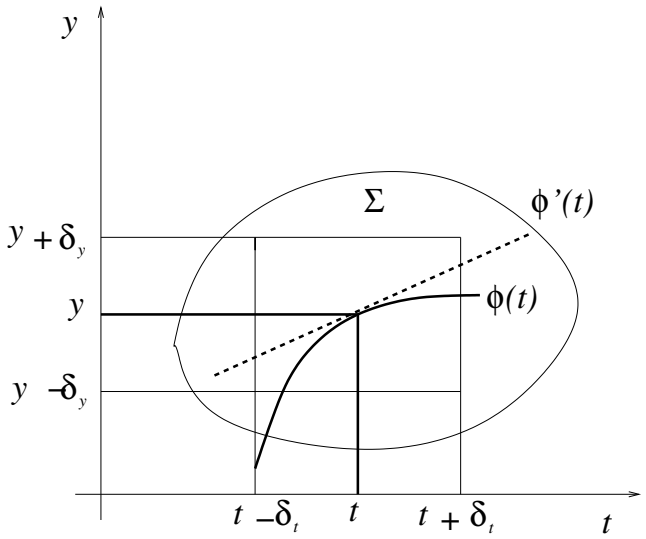


FIGURE C.1: *Geometric interpretation of* $y := \phi(t)$ *solution of the Cauchy problem* (C.1).

for every $t \in I$ and for every y_1 and $y_2 \in J$.
Let $M = \max_{t \in I, y \in J} \mid f(t,y) \mid$ and let $M\delta < \eta$. Then there exists one and only one $\phi : K \to \mathbb{R}$ such that:

1. $\phi(t)$ and $\phi'(t)$ are continuous for all t in I;
2. $\phi(t_0) = y_0$;
3. $\phi(t)$ is in J for all $t \in I$;
4. $\phi'(t) = f(t, \phi(t))$ for all t is in$= I$.

The function $\phi(t)$ is called solution or integral of the Cauchy problem (C.1), and hence $y(t) := \phi(t)$. An intuitive geometric picture is given in Figure C.1. For each point $t \in I$, the derivative of $\phi(t)$ is equal to the value assumed by f at that point.
This discussion can be made formal by recalling (without proof) Picard's theorem, which can be stated in all generality for systems of ODEs given by Problem C.1, when $y : I \mapsto \mathbb{R}^m$, $f : I \times \mathbb{R}^m \mapsto \mathbb{R}^m$, and $\|\cdot\|$ is a vector norm, e.g., the Euclidean norm:

Theorem C.1 (Picard)**.** *Let* $f(\cdot,\cdot)$ *be a continuous function of* (t,y) *in the region* U *containing the parallelepiped:*

$$R = \{(t,y) : t_0 \le t \le T, \|y - y_0\| \le Y\},$$

and Lipschitz in the variable y*, i.e., for any* $(t,y_1) \in R$ *and* $(t,y_2) \in R$*:*

$$\|f(t,y_1) - f(t,y_2)\| \le L \|y_1 - y_2\|. \tag{C.3}$$

Assume the constant $M = \max[\|f(t,y)\| : (t,y) \in R]$ *is such that:*

$$M(T - t_0) \leq Y.$$

Then there exists a unique function $t \mapsto y(t)$*, continuous and differentiable in* I *that satisfies* (C.1) *and solves Problem C.1.*

In the case of systems, a sufficient condition for which f is Lipschitz is to require that the Jacobian matrix of f has bounded norm:

$$\left\| \frac{\partial f}{\partial y} \right\| \leq L \tag{C.4}$$

where the matrix norm must be compatible with the employed vector norm. The converse is not true. For example, the function $y \mapsto f(y) = (\mid y_1 \mid, \ldots, \mid y_m \mid)^T$ with $t_0 = 0$ and $y_0 = 0$ satisfies (C.3) but not (C.4) since f is not differentiable at $y = 0$.
Similarly to what we have done for PDEs in Definition 1.1, we can give a definition of well-posedness for a Cauchy problem:

Definition C.2 (Well posedness)**.** The Cauchy Problem (C.1) is *well posed* if its solution exists, it is unique, and it depends continuously on the initial data and on the flux function f.

We can render the statement of continuous dependence upon the data more precise by looking at the stability of the Cauchy problem:

Definition C.3 (Stable problem (Lyapunov))**.** The Cauchy problem (C.1) is *stable* if for each perturbation $(\delta_0, \delta(t))$, such that $\delta(t)$ is a continuous function in I, and $\mid \delta_0 \mid < \epsilon$ and $\mid \delta(t) \mid < \epsilon$ for each $t \in I$, the solution $z(t)$ of the perturbed problem:

$$\begin{aligned} &z'(t) = f(t, z(t)) + \delta(t) \qquad t \in I, \\ &z(t_0) = y_0 + \delta_0, \end{aligned} \tag{C.5}$$

is such that:

$$\|y(t) - z(t)\| < K\epsilon \qquad \forall t \in I,$$

with the constant K depends only on the problem data (t_0, y_0, and f) and not on ϵ.

Given the Cauchy problem (C.1), if $f(t,y)$ is continuous and uniformly Lipschitz in y for $t \in I$ and $y \in \mathbb{R}$, then the Cauchy problem is well posed and we have the following:

Theorem C.4. *Under the hypotheses of Picards theorem, the (unique) solution* y *to the Cauchy problem C.1 is stable in* I *(assuming* $-\infty < t_0 < T < +\infty$*).*

Proof. We sketch the proof for $y \in \mathbb{R}$. If $w(t) = y(t) - z(t)$ then its derivative is $w'(t) = f(t, y(t)) - f(t, z(t)) + \delta(t)$ and we have:

$$w(t) = w(0) + \int_{t_0}^{t} w'(\tau)\, d\tau = \delta_0 + \int_{t_0}^{t} [f(\tau, y(\tau)) - f(\tau, z(\tau))]\, d\tau + \int_{t_0}^{t} \delta(\tau)\, d\tau$$

Applying Gronwall lemma[10], we obtain:

$$\begin{aligned} \mid w(t) \mid &\leq (1 + \mid t - t_0 \mid)\, \epsilon + L \int_{t_0}^{t} \mid w(\tau) \mid d\tau \\ &\leq \epsilon\, (1 + \mid t - t_0 \mid)\, e^{L \mid t - t_0 \mid}, \qquad \forall t \in I \\ &\leq C\epsilon, \qquad C = (1 + M_t) e^{M_t L}, \qquad M_t = \max_t \mid t - t_0 \mid. \end{aligned}$$

□

Example C.2.

$$y'(t) = -y(t) + t \qquad t \in I = [0, 1],$$

$$y(t_0) = 1,$$

with solution given by: $y(t) = 2e^{-t} + t - 1$.

$$z'(t) = -z(t) + t + \delta \qquad t \in I = [0, 1],$$

$$z(t_0) = 1 + \delta_0,$$

whose solution is: $z(t) = (2 + \delta_0 - \delta)e^{-t} + t + \delta - 1$. Then:

$$\mid y(t) - z(t) \mid = \mid (\delta - \delta_0)e^{-t} - \delta \mid \leq \mid \delta \mid (1 - e^{-t}) + \mid \delta_0 \mid e^{-t} \leq \epsilon,$$

for every $t \in I$.

C.2 One step linear methods

Explicit solutions to the Cauchy (C.1) are available only for specific $f(t, y)$. In general, numerical solutions are sought. Thus, we want to find an algorithm that approximates $y(t)$ for

[10] Gronwall lemma is important and can be stated as follows:

Lemma C.5 (Gronwall)**.** *Let $g(t)$ be an integrable non-negative function in I, and let $\varphi(t)$ and $\psi(t)$ be two continuous functions in I, with ψ non-decreasing. If $\varphi(t)$ satisfies:*

$$\varphi(t) \leq \psi(t) + \int_{t_0}^{t} g(\tau)\varphi(\tau)\, d\tau, \qquad \forall t \in I,$$

then:

$$\varphi(t) \leq \psi(t) e^{\left(\int_{t_0}^{t} g(\tau)\, d\tau\right)} \qquad \forall t \in I.$$

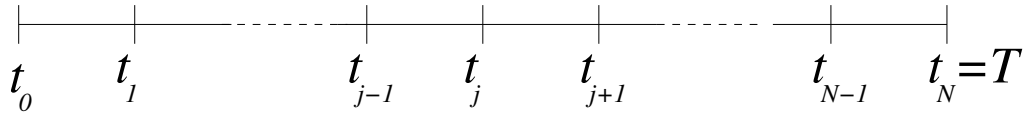


FIGURE C.2: *Discretization of the interval $I = [t_0, T]$ in N subinterval of size h.*

some discrete set $t_n \in I$. To this aim, we *discretize* the interval I, i.e., we subdivide I in subintervals of size h. The discretization is identified by the endpoints of the subintervals that are denoted with $t_n = t_0 + nh$, $j = 0, 1, \ldots, N$, with $h = (T - t_0)/N$ (see Figure C.2). We denote with y_n the numerical solution at step t_n: $y_n \approx y(t_n)$. Thus, our numerical schemes will approximate $y(t)$ pointwise, i.e., they will look for y_n, $j = 1, \ldots, N$.

C.2.1 Forward (explicit) Euler method.

The Taylor expansion of $y(t + h)$ is

$$y(t + h) = y(t) + hy'(t) + \frac{1}{2}h^2 y''(t) + \mathcal{O}\left(h^3\right)$$

Neglecting the second order terms we have the following approximation:

$$y'(t) \approx \frac{y(t + h) - y(t)}{h}.$$

Thus we can write the original ODE at $t = t_n$ as:

$$y'(t_n) = f\left(t_n, y(t_n)\right); \qquad y'(t_n) \approx \frac{y(t_{n+1}) - y(t_n)}{h},$$

to yield:

$$y(t_{n+1}) \approx y(t_n) + hf\left(t_n, y(t_n)\right).$$

Starting from the initial condition $y(t_0) = y_0$ we define the *forward Euler* method as:

$$y_{n+1} = y_n + hf(t_n, y_n), \qquad n = 0, 1, \ldots \tag{C.6}$$

This scheme is also known as *explicit Euler method*, since the computation of y_n requires only the evaluation of the function $f(t, y)$ at the previous point (t_{n-1}, y_{n-1}).

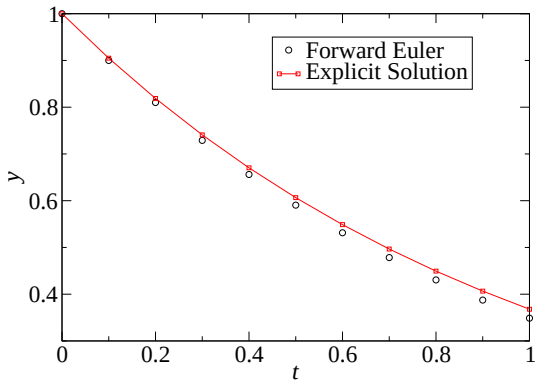
ALGORITHM FORWARD_EULER
Input: t_0, y_0, N, h;
FOR $j = 0, 1, \ldots, N-1$
$1. t_n = t_0 + nh$
$2. f_n = f(t_n, y_n)$
$3. y_{n+1} = y_n + hf_n$
END FOR

Example C.3. Let us apply the Explicit Euler method above to the Cauchy problem defined with $f(t, y) = -y(t)$, $t_0 = 0$, $T = 1$, and $y_0 = 1$. The explicit solution is easily found by separation of variables and is given by $y(t) = e^{-t}$. Substituting $f = -y$ in (C.6), we have:

$$y_{n+1} = (1 - h)y_n$$

Using an integration step $h = 0.1$, from which $N = 10$, we obtain the following table:

j	y_n	$y(t_n)$
0	1	1
0.1	0.9	0.904837418
0.2	0.81	0.818730753
0.3	0.729	0.740818221
0.4	0.6561	0.670320046
0.5	0.59049	0.60653066
0.6	0.531441	0.548811636
0.7	0.4782969	0.496585304
0.8	0.43046721	0.449328964
0.9	0.387420489	0.40656966
1.0	0.34867844	0.367879441



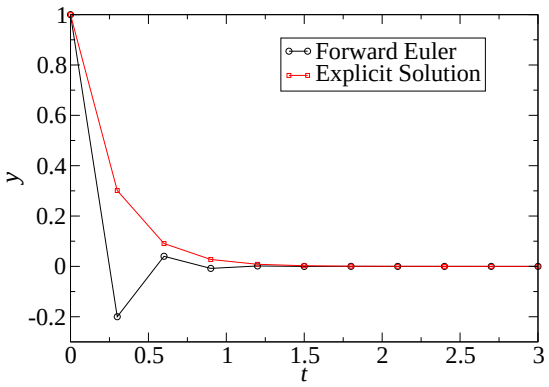
We can see that the forward difference scheme produces a good approximation of the explicit solution.

Example C.4. As second example, consider the Cauchy problem with $f(t, y) = -5y(t)$, $t_0 = 0$, $T = 3$, and $y_0 = 1$. The explicit solution is now $y(t) = e^{-5t}$. Using Algorithm (C.6) with $f = -5y$ we obtain:

$$y_{n+1} = (1 - 5h)y_n.$$

Using $h = 0.3$ ($N = 10$) we have the following table:

j	y_n	$y(t_n)$
0	1	1
0.3	-0.2	0.301194212
0.6	0.04	0.090717953
0.9	-0.008	0.027323722
1.2	0.0016	0.008229747
1.5	-0.00032	0.002478752
1.8	6.4E-05	0.000746586
2.1	-0.0000128	0.000224867
2.4	0.00000256	6.77287E-05
2.7	-5.12E-07	2.03995E-05
3	1.024E-07	6.14421E-06



In this case the forward Euler scheme yields a numerical solution that oscillates, resulting in a less accurate estimate of the explicit solution. In fact, in this case, the scheme is *not stable*, although the original Cauchy problem is Lyapunov-stable. The experimental observation of the results of this and the previous examples suggest that Forward Euler is stable under some restricting conditions on the step h. This restriction can by found by analyzing the convergence of the scheme, which will be done in the next section.

C.2.2 Backward (implicit) Euler method.

Writing the original ODE at $t = t_{n+1}$ as:

$$y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1})). \tag{C.7}$$

we can use the Taylor expansion of $y(t_n) = y(t_{n+1} - h)$ to yield:

$$y(t_n) = y(t_{n+1} - h) = y(t_{n+1}) - hy'(t_{n+1}) + \frac{h^2}{2}y''(t_{n+1}) + \mathcal{O}\left(h^3\right).$$

Again, neglecting $\mathcal{O}\left(h^2\right)$ terms we can write:

$$y'(t_{n+1}) \approx y'_{n+1} = \frac{y_{n+1} - y_n}{h},$$

and, after substitution in (C.7), we obtain the Backward Euler (implicit) scheme:

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) \qquad j = 0, 1, \ldots, N-1. \tag{C.8}$$

The scheme is termed implicit as the unknown, i.e., the value y_{n+1}, appears in both left and right hand sides of the equal sign and is contained as an argument of the function $f(t, y)$, which thus in general cannot be evaluated explicitly. In this case we would need to use a Newton-Raphson like iteration to evaluate y_{n+1} at every step. We have then the following algorithm:

ALGORITHM BACKWARD_EULER
Input: t_0, y_0, N, h;
FOR $j = 0, 1, \ldots, N-1$
$\quad$ 1.$t_{n+1} = t_n + h$
$\quad$ 2.Solve $y_{n+1} - y_n - hf(t_{n+1}, y_{n+1}) = 0$
END FOR

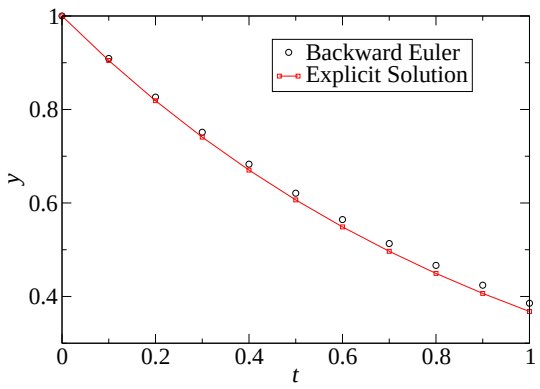
We solve the same problems as before but using Backward Euler. Note that, since $f(t, y)$ is linear in y, the scheme is effectively explicit.

Example C.5. Case $f(t, y) = -y(t)$, with $t_0 = 0$, $T = 1$ and $y_0 = 1$. Using $f = -y$ in (C.8), we have:

$$y_{n+1} = \frac{1}{1+h} y_n.$$

Let $h = 0.1$, from which $N = 10$. The solution is then given in the following table:

j	y_n	$y(t_n)$
0	1	1
0.1	0.909090909	0.904837418
0.2	0.826446281	0.818730753
0.3	0.751314801	0.740818221
0.4	0.683013455	0.670320046
0.5	0.620921323	0.60653066
0.6	0.56447393	0.548811636
0.7	0.513158118	0.496585304
0.8	0.46650738	0.449328964
0.9	0.424097618	0.40656966
1	0.385543289	0.367879441



We can see that the Backward Euler approximation is similar to the Forward Euler numerical solution, but it overestimates the real solution, contrary to the Explicit Euler case.

Example C.6. We try now the case $f(t, y) = -5y(t)$, with $t_0 = 0$, $T = 3$ and $y_0 = 1$:

$$y_{n+1} = \frac{1}{1+5h} y_n,$$

and using $h = 0.3$ ($N = 10$) we have:

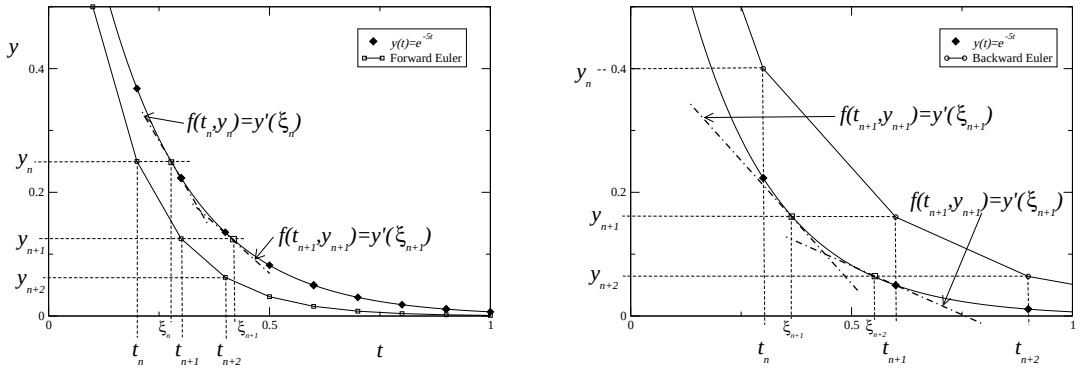
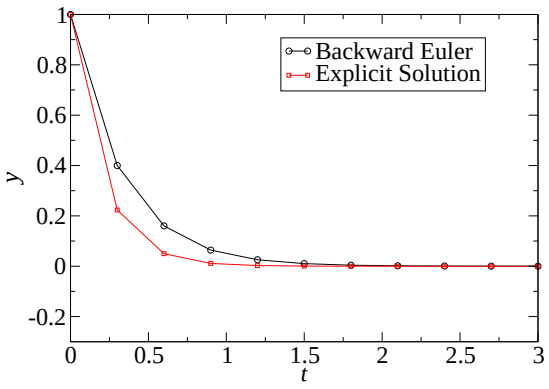


FIGURE C.3: *Geometric interpretation of the Forward (left) and Backward (right) Euler schemes for the differential equation $y' = -5y$. The dotted and dashed-dotted lines represents $f(t, y)$ evaluated at n and $n + 1$ for the Forward and Backward Euler schemes, respectively.*

j	y_n	$y(t_n)$
0	1	1
0.3	0.4	0.22313016
0.6	0.16	0.049787068
0.9	0.064	0.011108997
1.2	0.0256	0.002478752
1.5	0.01024	0.000553084
1.8	0.004096	0.00012341
2.1	0.0016384	2.75364E-05
2.4	0.00065536	6.14421E-06
2.7	0.000262144	1.37096E-06
3	0.000104858	3.05902E-07



In this case, Backward Euler produces estimates that are less accurate with respect to the previous case, but, contrary to the Forward Euler case, there are no oscillations. In fact, Backward Euler has the same order of accuracy of Forward Euler and is always stable, independently of the step size h.

Let us look at the geometrical interpretation of the Forward and Backward Euler schemes reported in Figure C.3. Looking at the implicit method of eq. (C.8) we see that the point (t_{n+1}, y_{n+1}) lies on the line with direction $y'_{n+1} = (y_{n+1} - y_n)/h = f(t_{n+1}, y_{n+1})$ passing by (t_n, y_n), while for the explicit scheme the line goes through the same point but has slope given by $y'_n = (y_{n+1} - y_n)/h = f(t_n, y_n)$. Thus, in the Backward Euler method the next point is obtained using the slope evaluated at t_{n+1} given in this case by $y'(t) = f(t, y) = -5y$, which

is decreasing with t ($y(t) = e^{-5t}$), thus giving reason to the overestimation of the real solution (Figure C.3, right). In the case of Forward Euler, similar considerations give reason to the underestimation (Figure C.3, left).

C.2.3 Crank-Nicolson Method.

For both Euler schemes, we used a Taylor expansion and neglected all the terms of the order $\mathcal{O}(h^2)$. The Crank-Nicolson method tries to neglect terms of greater order. To do so, we see that subtracting the two forward and backward Taylor series we can exploit cancellation of terms. More precisely, we have:

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n) + \frac{h^3}{6}y'''(t_n) + \mathcal{O}\left(h^4\right)$$
$$y(t_n) = y(t_{n+1}) - hy'(t_{n+1}) + \frac{h^2}{2}y''(t_{n+1}) - \frac{h^3}{6}y'''(t_{n+1}) + \mathcal{O}\left(h^4\right),$$

from which we have immediately:

$$y'(t_{n+1}) + y'(t_n) \approx y'_{n+1} + y'_n = \frac{y_{n+1} - y_n}{h}.$$

Writing eq. (C.1) as an average between times t_{n+1} and t_n, we have the Crank-Nicolson scheme:

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\right]. \tag{C.9}$$

The method is clearly implicit, and we have the following algorithm:

ALGORITHM CRANK-NICOLSON
Input: t_0, y_0, N, h;
FOR $j = 0, 1, \ldots, N-1$
$1. t_{n+1} = t_n + h$
$2. f_n = f(t_n, y_n) = 0$
3.Solve $y_{n+1} - y_n - 0.5h\left[f(t_{n+1}, y_{n+1}) + f_n\right] = 0$
END FOR

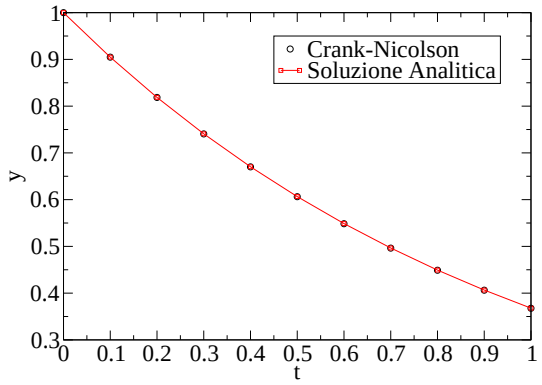
We show now the same results as before obtained with the Crank-Nicolson scheme. We note that, as occurred with Backward Euler, since $f(t, y)$ is linear in y, so is in this case the overall scheme.

Example C.7. Case of $f(t, y) = -y(t)$, with $t_0 = 0$, $T = 1$ and $y_0 = 1$. Using $f = -y$ in (C.9), we have:

$$y_{n+1} = \frac{2 - h}{2 + h}y_n.$$

Let $h = 0.1$, from which $N = 10$. We obtain:

j	y_n	$y(t_n)$
0	1	1
0.1	0.904761905	0.904837418
0.2	0.818594104	0.818730753
0.3	0.740632761	0.740818221
0.4	0.670096308	0.670320046
0.5	0.606277612	0.60653066
0.6	0.548536887	0.548811636
0.7	0.496295278	0.496585304
0.8	0.449029061	0.449328964
0.9	0.406264389	0.40656966
1	0.367572542	0.367879441



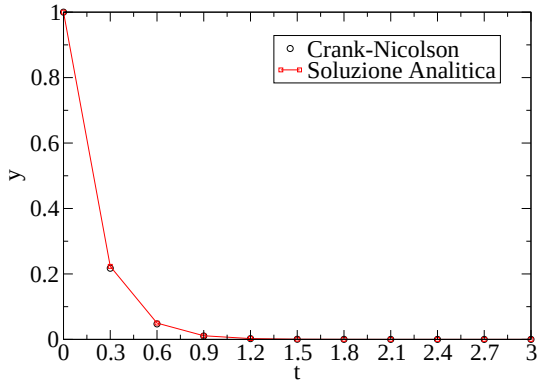
The CN method gives approximations of the solution that are more accurate than both Euler schemes. The reason is due to the fact that we have neglected terms in the Taylor expansions that were of higher order with respect to those neglected for the Euler approaches.

Example C.8. Case $f(t, y) = -5y(t)$, with $t_0 = 0$, $T = 3$ and $y_0 = 1$. Setting $f = -5y$ in (C.6), yields:

$$y_{n+1} = \frac{2 - 5h}{2 + 5h} y_n.$$

Using $h = 0.3$, from which $N = 10$, we have:

j	y_n	$y(t_n)$
0	1	1
0.3	0.217391304	0.22313016
0.6	0.047258979	0.049787068
0.9	0.010273691	0.011108997
1.2	0.002233411	0.002478752
1.5	0.000485524	0.000553084
1.8	0.000105549	0.00012341
2.1	2.29454E-05	2.75364E-05
2.4	4.98813E-06	6.14421E-06
2.7	1.08438E-06	1.37096E-06
3	2.35734E-07	3.05902E-07



Also in this case the accuracy of Crank-Nicolson is better than the Euler Schemes, and there are no oscillations, suggesting that it is stable. One may think, rightly, that it is the implicitness of the formulation that makes the schemes unconditionally stable. We will see this more precisely in the next sections.

Remark C.9 (Derivation of the schemes from quadrature formulas)**.** *The derivation of the studied one-step schemes can be effectively done from the point of view of integration, instead of differentiation. Thus, we integrate the ODE between t_n and t_{n+1}, to obtain:*

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(\tau, y(\tau))\, d\tau$$

Now, the integral on the right-hand-side can be approximated by a quadrature formula. The Forward and Backward Euler schemes are easily obtained by using a quadrature formula based on piecewise constant interpolation using information on the endpoints of the interval. Thus we build rectangles that have a base of $h = t_{n+1} - t_n$ a height of $y(t_n)$ and $y(t_{n+1})$ for Forward and Backward Euler, respectively. Crank-Nicolson method is obtained by means of the trapezoidal rule. We have then:

$$\begin{aligned}
&y(t_{n+1}) \approx y(t_n) + hf(t_n, y(t_n)),\\
&y(t_{n+1}) \approx y(t_n) + hf(t_{n+1}, y(t_{n+1})),\\
&y(t_{n+1}) \approx y(t_n) + \frac{h}{2}\left[f(t_{n+1}, y(t_{n+1})) + f(t_n, y(t_n))\right],
\end{aligned}$$

The scheme are readily obtained by using y_n in place of $y(t_n)$.

C.2.4 Explicit Runge-Kutta methods

The numerical schemes presented in the previous sections are part of the more general Runge-Kutta family of methods for the solution of ODEs of higher order. In the m-stage Runge-Kutta method the numerical solution of the Cauchy problem can be written as:

$$y_{n+1} = y_n + \omega_1 Y_1' + \omega_2 Y_2' + \ldots + \omega_m Y_m', \tag{C.10}$$

where $Y_1', \ldots, Y_m'$ are defined from the general expression:

$$Y_s' = hf\left(t + \alpha_s\, h, y_j + \sum_{i=1}^{m-1} \beta_{s,i} Y_i'\right), \qquad s = 1, \ldots, m.$$

It is simple to show that the forward difference scheme is obtained with $m = 1$ and imposing $\omega_1 = 1$, $\alpha_1 = 0$, and $\beta_{1,1}$=0. In a similar way, the backward difference is obtained with $\omega_1 = 1$, $\alpha_1 = 1$, and $\beta_{1,1}$=1.
In the explicit Runge-Kutta methods the coefficients Y_s' can be derived explicitly and have

the form:

$$\begin{aligned}
Y_1' &= hf(t_n, y_n) \\
Y_2' &= hf(t_n + \alpha_2 h, y_n + \beta_{2,1} Y_1') \\
Y_3' &= hf(t_n + \alpha_3 h, y_n + \beta_{3,1} Y_1' + \beta_{3,2} Y_2') \\
&\vdots \\
Y_m' &= hf(t_n + \alpha_m h, y_n + \beta_{m,1} Y_1' + \beta_{m,2} Y_2' + \ldots + \beta_{m,m-1} Y_{m-1}').
\end{aligned}$$

The coefficients $w_1, \ldots, w_m$, $\alpha_0, \alpha_1, \ldots$ and $\beta_{2,1}, \beta_{3,1}, \beta_{3,2}, \ldots$ are computed using the Taylor series. In the following we derive the computation of these coefficients for $m = 2$:

$$y_{n+1} = y_n + \omega_1 Y_1' + \omega_2 Y_2'. \tag{C.11}$$

Let $y(t)$ be the solution of the Cauchy problem. For the chain rule, the second derivative of $y(t)$ satisfies:

$$y''(t) = \frac{d}{dt}[y'(t)] = \frac{d}{dt}[f(t, y(t)] = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) y'(t).$$

Thus, the Taylor series of $y(t_n + h)$ can be written as:

$$y(t_n+h) = y(t_n) + h\ f(t_n, y(t_n)) + \frac{1}{2}h^2 \left[\frac{\partial f}{\partial t}(t_n, y(t_n)) + f(t_n, y(t_n))\frac{\partial f}{\partial y}(t_n, y(t_n))\right] + R_1, \tag{C.12}$$

where R_1 is the residual term. Now we write Y_2' using the Taylor series of f in $(t_n, y(t_n))$:

$$Y_2' = h\left[f(t_n, y(t_n)) + \alpha_2 h \frac{\partial f}{\partial t}(t_n, y(t_n)) + \beta_{2,1} Y_1' \frac{\partial f}{\partial y}(t_n, y(t_n)) + R_2\right]$$

where R_2 is the residual term, $R_2 = \mathcal{O}(h^2)$.
Eq.(C.11) becomes:

$$y_{n+1} = y_n + (\omega_1 + \omega_2) h f(t_n, y_n) + \omega_2 \alpha_2 h^2 \frac{\partial f}{\partial t}(t_n, y_n) + \omega_2 \beta_{2,1} h^2 \frac{\partial f}{\partial y}(t_n, y_n) + \omega_2 h R_2. \tag{C.13}$$

Imposing that (C.12) equals (C.13), we obtain the conditions for the parameters ω_1, ω_2, α_2 and $\beta_{2,1}$:

$$\begin{cases} \omega_1 + \omega_2 = 1 \\ \alpha_2 \omega_2 = 1/2 \\ \beta_{2,1} \omega_2 = 1/2 \end{cases} \tag{C.14}$$

Since this system has four unknowns and three equations, it admits infinite solutions. For example:

- *Heun* formula: $\omega_1 = \omega_2 = 1/2$, $\alpha_2 = \beta_{2,1} = 1$.
- *Runge* formula (RK2): $\omega_1 = 0$, $\omega_2 = 1$, $\alpha_2 = \beta_{2,1} = 1/2$.

The order of accuracy of the schemes with $m = 2$ can be derived as follows. Note that, dividing (C.13) by h, the residual term is $\omega_2 R_2$, which is proportional to h^2. Thus the schemes with $m = 2$ are of the second order accuracy. We will see in the next section a more detailed definition of order of accuracy.

Remark C.10. *Heun and Runge formula are written explicitly as:*

$$y_{n+1} = y_n + \frac{1}{2}\left[h\ f(t_n, y_n) + f(t_n + h, y_n + h\ f(t_n, y_n))\right]$$

and

$$y_{n+1} = y_n + h\ f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}\ f(t_n, y_n)\right),$$

respectively. The former can be interpreted as an explicit version of Crank-Nicolson, where we have resolved the implicitness deriving by the use of the trapezoidal rule by estimating y_{n+1} with one step of Forward Euler. Analogously, Runge's scheme can be thought of as the use of the mid-point quadrature rule (one-point Gaussian quadrature where the Gauss point is the center of the integration interval) where $y_{n+1/2}$ is again evaluated using Forward Euler with a step-size $h/2$. Thus we have:

$$\begin{aligned} y^*_{n+1} &= y_n + h\ f(t_n, y_n) \\ y_{n+1} &= y_n + \frac{1}{2}h\left[\ f(t_n, y_n) + f(t_n + h, y^*_{n+1})\right] \end{aligned} \qquad \textit{Heun} \qquad \text{(C.15)}$$

$$\begin{aligned} y^*_{n+1/2} &= y_n + \frac{h}{2}\ f(t_n, y_n) \\ y_{n+1} &= y_n + h\ f\left(t_n + \frac{h}{2}, y^*_{n+1/2}\right) \end{aligned} \qquad \textit{Runge} \qquad \text{(C.16)}$$

The *Runge-Kutta* scheme (RK4) is largely used. It is obtained with an analogous procedure, but with $m = 4$ and $\omega_1 = \omega_4 = 1/6$, $\omega_2 = \omega_3 = 2/6$:

$$y_{n+1} = y_n + \frac{1}{6}Y'_1 + \frac{2}{6}Y'_2 + \frac{2}{6}Y'_3 + \frac{1}{6}Y'_4, \qquad \text{(C.17)}$$

and the coefficients Y'_1, Y'_2, Y'_3 and Y'_4 are:

$$\begin{cases} Y'_1 = h\ f(t_n, y_n) \\ Y'_2 = h\ f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}Y'_1) \\ Y'_3 = h\ f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}Y'_2) \\ Y'_4 = h\ f(t_n + h, y_n + Y'_3) \end{cases}.$$

The order of accuracy of RK4 is 4, equal to the number of stages.
It is intuitive to assume that the order of accuracy for a m-stage RK method is $p = m$. This is true for $m = 1, 2, 3, 4$, not not so anymore for $m \geq 5$. Runge-Kutta methods were extensively analyzed by John Butcher, who demonstrated that the order of accuracy for RK methods with $m = 5, 6, 7, 8, 9$ stages is $p = 4, 5, 6, 6, 7$, respectively, and for $m \geq 10$ the highest order is $p \leq m - 2$. The suboptimality of high order RK methods leads to the question whether it is possible to find efficient high-order methods.

C.2.5 Adams methods

An intuitive idea to extend RK methods comes from the reinterpretation of RK schemes as quadrature rules. In fact, we can write:

$$y(t_{n+1}) = y(t_n) + h \int_0^1 y'(t_n + \alpha h)\, d\alpha. \tag{C.18}$$

Then we approximate $y'(t) = f(t, y(t))$ by quadrature formula with weights ω_s and quadrature nodes β_s. We have:

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{s=1}^{m} \omega_s Y'_s \\ Y'_s &= f(t_n + \alpha_s h, Y_s) \end{aligned}$$

where $Y_s \simeq y(t_n + \alpha_s h)$. These intermediate values can be approximated with another quadrature formula approximating the integral from t_n to $\alpha_s h$;

$$Y_i = y_n + h \sum_{j=1}^{s} \beta_{ij} Y'_j, \qquad i = 1, \ldots, s.$$

Starting from (C.18), Adams methods replace the integrand with the interpolating polynomial $\pi_k(t)$ passing through points (t_{n-j}, f_{n-j}), $j = 0, \ldots, k$, where we have simplified the writing by setting $f_n = f(t_n, y_n)$. Using Newton interpolation formula with a constant stepsize h we obtain readily:

$$y_{n+1} = y_n + h f_n + \sum_{i=1}^{k-1} \gamma_i \nabla^i f_n,$$

where we have denoted with $\nabla f_n = (1 - \mathrm{E}) f_n = f_n - f_{n-1}$ the backward difference operator, and $\nabla^i f_n = \nabla^{i-1} f_n - \nabla^{i-1} f_{n-1}$, defined in Section B. Note that the interpolating polynomial is used in "extrapolation" mode, i.e., not with the interpolation interval. This is remedied by adding the additional point $(t_{n+1}, y(t_{n+1}))$ and use a few iterations of a Picard iteration to solve the nonlinearity.

C.3 Errors and Convergence

In this section we want to study the convergence of the sequence of numerical approximates y_n towards the real solution $y(t_n)$. Thus, we define the error as the difference between the numerical and real solutions, $e(t_n) = y_n - y(t_n)$, and say that convergence requires that the error tends to zero as $h \longrightarrow 0$. From the experiences collected in the previous practical examples, we can pose the following questions given a numerical scheme:

1. does the numerical solution tend to the real solution (i.e., the error tends to zero) as $h \to 0$?
2. How fast does the error go to zero?
3. what is the maximum value of h that returns a numerical solution close in some sense to the real solution up to a prescribed tolerance?

The error is defined for all $t_n \in I$, we can ask for example that the infinity norm of the error (the maximum absolute value) tend to zero. In this case we would have:

Definition C.6 (Convergence)**.** A numerical scheme for the solution of the Cauchy problem (C.1) is *convergent* if:

$$\lim_{h\to 0} \max_{0\leq j\leq N} \mid y_n - y(t_n) \mid = 0$$

But we can use any equivalent functional norm:

$$\lim_{h\to 0} \|y_h - y(t)\| = 0$$

where $y_h = \{y_0, y_1, \ldots, y_N\}$. observing that $y \in \mathcal{C}^1(I)$, at least, we can use, for example, the $\mathcal{L}^2(I)$ norm given by:

$$\|g(t)\| = \left[\int_I \mid g(t) \mid^2 dt\right]^{\frac{1}{2}}$$

C.3.1 Experimental convergence

Before discussing theoretical convergence, we look at it experimentally. We consider the following (relative) error norm, which assumes the use of a constant integration step $h = t_{n+1} - t_n$:

$$e_{h,xx} = \|y_h - y(t)\| \approx \frac{\left[h \sum_{j=0}^{N=1} \mid y_n - y(t_n) \mid^2\right]^{\frac{1}{2}}}{\left[h \sum_{j=0}^{N=1} \mid y(t_n) \mid^2\right]^{\frac{1}{2}}},$$

h	$e_{h,FE}$	$e_{h,BE}$	$e_{h,CN}$	$\frac{e_{h,FE}}{e_{h-1,FE}}$	$\frac{e_{h,BE}}{e_{h-1,BE}}$	$\frac{e_{h,CN}}{e_{h-1,CN}}$
0.2	5.75E-02	7.42E-02	2.62E-03	–	–	–
0.1	2.49E-02	3.62E-02	6.21E-04	2.31	2.05	4.23
0.05	1.16E-02	1.78E-02	1.51E-04	2.15	2.03	4.11
0.025	5.60E-03	8.28E-03	3.48E-05	2.07	2.15	4.33

TABLE C.1: *Convergence of the methods Forward (FE) and Backward (BE) Euler and Crank-Nicolson (CN) methods. The error norm and its ratio is shown for successively refined mesh subdivisions h. The ratio between the error norms at consecutive refinements approaches 2 for both Euler scheme and 4 for Crank-Nicolson, indicating linear ($\mathcal{O}(h)$) and quadratic ($\mathcal{O}(h^2)$) convergence, respectively.*

where $xx = FE,\ BE,\ CN$ per Forward (Explicit) Euler, Backward (Implicit) Euler and Crank-Nicolson. We compute this norm for different values of h, for example in a geometric sequence h =0.2, 0.1, 0.05, 0.025. We obtain the following table:
Table C.1 shows the experimental results obtained on the test case of example C.3. From the ratio between the error norms of two successively refined discretizations, we see that both Euler schemes show first order convergence ($\mathcal{O}(h)$) while Crank-Nicolson shows second order convergence ($\mathcal{O}(h^2)$). We have learned by now that convergence rates typically depend on the scheme and on the regularity of the real solution and we will need to justify theoretically the experimental convergence rates.

C.3.2 Consistency and truncation error.

We observe first that, as expected, the experimental convergence rates follow the pattern of the order of the terms that were neglected in the Taylor series expansion defining the different methods. We say that the order convergence depends on the *truncation error*. However, we need to distinguish between the truncation error, which occurs at each step, and its propagation through the steps needed to evaluate the solution at a fixed time. More precisely, the former is the error arising from a one-step calculation (from t_n to t_{n+1}), assuming that the initial solution is exact ($y_n = y(t_n)$). The latter, is the propagation (or global) error arising from the accumulation of the truncation errors at previous steps. Hence, we can write:

$$e_{n+1} = y(t_{n+1}) - y_{n+1} = y(t_{n+1}) - y^*_{n+1} + y^*_{n+1} - y_{n+1} = \tau_{n+1} + \varepsilon_{n+1}$$

where y^*_{n+1} is the approximation calculated by the numerical scheme starting from the exact solution $y(t_n)$. The accumulation (or propagation error) and the truncation error are embodied by the concepts of stability and consistency, respectively. We analyze in this section the latter, leaving the former for the next section.
Let us write the previous schemes in abstract form as (all the one-step methods can be written

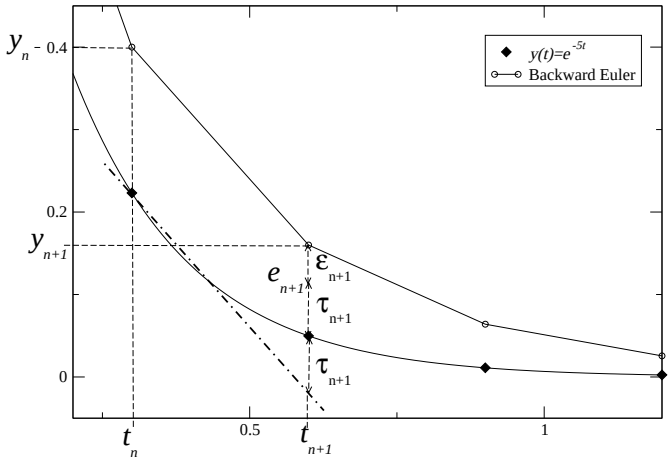


FIGURE C.4: *Geometric exemplification of the truncation error $\tau_{j+1}(h)$ for Forward Euler applied to example C.4.*

as in the following):

$$y_{n+1} = y_n + h\Phi(t_n, y_n; h), \qquad 0 \leq j \leq N-1, \tag{C.19}$$

where the function $\Phi(\cdot, \cdot; h)$ identifies the particular scheme[11] Substituting in the previous equation the real solution $y(t)$ we obtain:

$$y(t_{n+1}) = y(t_n) + h\Phi(t_n, y(t_n); h) + \epsilon_{n+1}, \tag{C.20}$$

where ϵ_{n+1} is the residual at step $n+1$. We can rewrite this residual as:

$$\epsilon_{n+1} = h\tau_n(h)$$

where $\tau_{n+1}(h)$ is the Truncation Error (TE), given formally by:

$$\tau_n(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - \Phi(t_n, y(t_n); h). \tag{C.21}$$

The maximum value of the TE is denoted by $\tau(h)$:

$$\tau(h) = \max_{0 \leq n \leq N-1} \tau_{n+1}(h). \tag{C.22}$$

[11]For example, Forward Euler is given by: $\Phi(t_n, y_n; h) = f(t_n, y_n)$.

Assume that the function Φ is such that:

$$\lim_{h\to 0} \Phi(t_n, y(t_n); h) = f(t_n, y(t_n)) \qquad j = 1, 2, \ldots, N.$$

Recalling the expression of the Taylor expansion by which $y(t_{n+1}) - y(t_n) = hy'(t_n) + O(h^2)$, we see immediately that the previous equation implies that $\tau_{n+1}(h) \to 0$ for h that tends to zero, from which we have immediately:

$$\lim_{h\to 0} \tau(h) = 0,$$

which implies that the scheme (C.19) is *consistent*, in the general sense already specified in Section 2.1.4. We say that the scheme has *order of convergence* or of *accuracy* equal to p if:

$$\tau(h) = O(h^p).$$

From eq. (C.21) we see then that the global (propagation) error is of order p.

Example C.11 (Error analysis of Forward Euler.)**.** Assuming f sufficiently continuous in both arguments so that we can assume y'' to be bounded in the interval $[t_0, T_N]$, using Tailor series expansion of f and noting that $y' = f(x, y)$, subtracting equations (C.19) from (C.20) we obtain:

$$e_{n+1}(h) = y(t_{n+1}) - y^*_{n+1} = y(t_n) + hy'(t_n)) + \frac{h^2}{2}| \, y''(\xi) \, | - y_n - hf(x_n, y(t_n)) = \frac{h^2}{2}| \, y''(\xi) \, |$$

with $\xi \in I$ an appropriate point (arising from application of Lagrange, or mean value, theorem). We can define the truncation error for the Forward Euler scheme as:

$$\tau_n(h) = \frac{y(t_{n+1}) - y(t_n)}{h} - f(x_n, y(x_n))$$

so that we find:

$$| \, \tau_n(h) \, | = \frac{h}{2}| \, y''(\xi) \, | \leq \frac{hM}{2}, \qquad M = \max_{\eta\in[t_0,T_M]} | \, y''(\eta) \, |.$$

Thus, the Forward Euler scheme is first order accurate, as experimented in the test cases shown in table C.1. Figure C.4 shows a geometrical exemplification of the differences between the TE and the propagation error. In a similar fashion, it can be verified that both Euler schemes are formally consistent and first order accurate, while Crank-Nicolson is second order accurate, thus confirming the experimental observations reported in table C.1.
Convergence can be analyzed as follows. Define the total error $e_n \stackrel{\text{def}}{=} y(t_n) - y_n$. Then, the error equation becomes:

$$e_{n+1} = e_n + h \, [f(x_n, y(x_n)) - f(x_n, y_n)] + h\tau_n(h)$$

Using the assumption of Lipschitz continuity of f with respect to the second argument we have:

$$\mid e_{n+1} \mid \leq (1 + hL) \mid e_n \mid + h\tau_n. \tag{C.23}$$

Letting $\tau = \max_{0 \leq n \leq (M-1)} \mid \tau_n \mid$, the solution of the above difference inequality is:

$$\mid e_n \mid \leq \frac{\tau}{L} \left[(1 + hL)^n - 1 \right] + (1 + hL)^n \mid e_0 \mid \leq \frac{\tau}{L} (e^{L(t_n - t_0)} - 1) + e^{L(t_n - t_0)} \mid e_0 \mid. \tag{C.24}$$

Remark C.12. *To find the solution of the above inequality* (C.23) *we can proceed either by induction or using an approach similar to the technique for finding solutions to linear constant coefficients ODEs. Using the latter approach, using the equal sign in place of the* $\leq$ *sign, we add to the general solution of the homogeneous equation* $\tilde{e}_n$ *one particular solution of the complete equation* $e^* n$*. Thus, looking at the homogeneous equation, we let* $\tilde{e})_n = \lambda^n$*, to yield immediately:*

$$\tilde{e}_n = C_1 (1 + hL)^n.$$

Now, we look for a particular solution of the form $e_n^* = An + B$*. Substituting into the complete equations we have:*

$$An + A + B = (An + B)(1 + hL) + h\tau.$$

Equating the polynomials of same order in n *we get* $A = 0$ *and* $B = -\tau/L$*. Imposing* $e_n = e_0$ *for* $n = 0$*, we obtain immediately* (C.24)*.*

Hence, the error of FE can be estimated by:

$$\mid e_n \mid \leq e^{L(t_n - t_0)} \mid e_0 \mid + \frac{\tau}{L} \left[(e^{L(t_n - t_0)} - 1) \right] \leq e^{L(t_n - t_0)} \mid e_0 \mid + \frac{Mh}{2L} \left[(e^{L(t_n - t_0)} - 1) \right],$$

which tells us that, if there is no round-off error so that the initial solution is represented exactly ($e_0 = 0$), then the error tends to zero as $h \longrightarrow 0$ and $n \longrightarrow \infty$ so that $t_n = t_0 + nh \longrightarrow t$, and the scheme converges to the exact solution. If we consider round-off errors, we can argue that the accuracy with which the scheme approximate the solution is bound by the time-propagation of the initial error, in this case bounded by the exponential of the Lipschitz constant of f.

Example C.13 (the θ-method)**.** The Forward and Backward Euler and the Crank-Nicolson methods can be grouped into one single method generally called the θ method. In fact we have:

$$y_{n+1} = y_n + h \left[(1 - \theta) f(t_n, y_n) + (\theta) f(t_{n+1}, y_{n+1}) \right].$$

For $\theta = 0$ we recover FE, for $\theta = 1$ BE, and $\theta = 1/2$ returns CN. Using the same technique as before we see that the error behaves as:

$$\mid e_n \mid \leq e^{L\frac{t_n - t_0}{1-\theta L h}} \mid e_0 \mid + \frac{h}{L}\left(\mid \tfrac{1}{2} - \theta \mid M + \frac{1}{3}\tilde{M}h \right) \left[(e^{L(t_n - t_0)} - 1) \right],$$

where $\tilde{M} = \max_{t\in[t_0,t_n]} \mid y'''(t) \mid$. From the previous equation we see that, in the case $e_0 = 0$, the global error is $\mid e_n \mid \leq \mathcal{O}(h^2)$ for $\theta = 1/2$ and $\mid e_n \mid \leq \mathcal{O}(h)$ for $\theta \neq 1/2$. Finally, we note that if we use $\theta = 1/2$ and employ Forward Euler to render explicit the Crank-Nicolson scheme (see Heun scheme (C.15)), we again obtain second order convergence $\mid e_n \mid \leq \mathcal{O}(h^2)$.

Explicit one-step methods. We can specialize precisely the above discussion for a general one-step method of the form:

$$y_{n+1} = y_n + h\Phi(t_n, y_n; h), \quad n = 0, \ldots, N-1, \tag{C.25}$$

where Φ is assumed to be a continuous function of its arguments. We define the global error and the truncation error as:

Definition C.7 (Global error and truncation error)**.** Given the explicit one-step method (C.25), the global error (GE) at time t_n is given by:

$$e_n \stackrel{\text{def}}{=} y(t_n) - y_n \tag{C.26}$$

and the truncation error (TE) is:

$$\tau_n \stackrel{\text{def}}{=} \frac{y(t_{n+1}) - y(t_n)}{h} - \Phi(t_n, y(t_n); h).$$

We can bound GE in terms of TE as the following theorem states.

Theorem C.8. *Given the explicit one-step method* (C.25)*, assume the scheme function* Φ *is continuous and Lipschitz in the second argument, i.e. it satisfies the hypothesis of Picard's theorem:*

$$\mid \Phi(t, y_1; h) - \Phi(t, y_2; h) \mid \leq L \mid y_1 - y_2 \mid \qquad \textit{for } (t, y) \in R. \tag{C.27}$$

Then, if $\mid y_n - y_0 \mid \leq Y$ *we have for* $n = 0, \ldots, N$*:*

$$\mid e_n \mid \leq e^{L(t_n - t_0)} \mid e_0 \mid + \left[\frac{e^{L(t_n - t_0)}}{L} \right] \tau$$

where $\tau = \max\limits_{0\leq n\leq N-1} \mid \tau_n(h) \mid$.

Proof. The proof is a immediate extension of the proof given in Example C.11. □

We can now proceed and define what we mean by consistency. We have then:

Definition C.9 (Consistency)**.** A one-step explicit method is consistent with the Cauchy problem C.1 if for any $\epsilon > 0$ there exists a step-size $h(\epsilon) > 0$ for which $\mid \tau_n \mid < \epsilon$ and $0 < h < h(\epsilon)$ and for every $(t_n, y(t_n))$ and $(t_{n+1}, y(t_{n+1}))$.

Assuming enough regularity of both y and Φ, we can restate the consistency condition as:

$$\lim_{h \longrightarrow 0} \tau_n = y'(t_n) - \Phi(t_n, y(t_n); h) = 0$$

from which we obtain immediately the necessary and sufficient condition for consistency:

$$\Phi(t, y; 0) = f(t, y) \tag{C.28}$$

We can now precisely specify the conditions for the convergence of the explicit one-step methods:

Theorem C.10. *Given an explicit one-step method of the form* (C.25)*, and assume that the solutions of the Cauchy problem C.1 and of the one-step method both lies in the region R of the Picard's theorem for some $0 < h < h_0$. Moreover, assume the function $\Phi(\cdot, \cdot; \cdot)$ satisfies the consistency condition* (C.28)*, is uniformly continuous in R and satisfies the Lipschitz condition* (C.27) *on $R \times [0, h_0]$.*
Then the sequence (y_n) of approximates generated by the explicit one-step method (C.25) *with successively smaller values of $h < h_0$ converges to the solution of the Cauchy problem C.1 in the sense that:*

$$\mid y(t_n) - y_n \mid \longrightarrow 0 \qquad \textit{for } h \longrightarrow 0, t_n \longrightarrow t \in [t_0, T].$$

Proof. Let $h = \frac{T - t_0}{N}$, with $N > 0$ and $h < h_0$. We assume that there is no round-off error, so that $y(t_0) = y_0$ and $e_0 = 0$. From Theorem C.8 we have:

$$\mid e_n \mid = \mid y(t_n) - y_n \mid \leq \left[\frac{e^{L(t_n - t_0)}}{L} \right] \max_{0 < n < N-1} \mid \tau_n \mid.$$

We need to work now on $\tau_n(h)$. Using the consistency condition (C.28), and the continuity of y and y', we can write:

$$\begin{aligned} \tau_n &= \frac{y(t_n) - y_n}{h} - \Phi(t_n, y(t_n); h) + \Phi(t_n, y(t_n); 0) - f(t_n, y(t_n)) \\ &= y'(\xi) - y'(t_n) + \Phi(t_n, y(t_n); 0) - f(t_n, y(t_n)) \end{aligned}$$

where $\xi \in [t_n, t_{n+1}]$ comes from the application of the mean value theorem. From the uniform continuity of y' and Φ we have that

$$\mid y'(\xi) - y'(t_n) \mid \leq \frac{1}{2}\epsilon \qquad h < h_1(\epsilon), \forall n,$$

and

$$\mid \Phi(t_n, y(t_n; 0) - \Phi(t_n, y(t_n); h) \mid \leq \frac{1}{2}\epsilon \qquad h < h_2(\epsilon), \forall n.$$

Hence we find:

$$\mid \tau_n \mid \leq \epsilon,$$

from which we can write:

$$\mid e_n \mid \leq \left[\frac{e^{L(t_n - t_0)}}{L}\right] \epsilon \longrightarrow 0 \text{ as } h \longrightarrow 0.$$

Moreover, from the uniform continuity of y we have immediately:

$$\mid y(t) - y_n \mid \leq \mid y(t) - y(t_n) \mid + \mid y(t_n) - y_n \mid \longrightarrow 0 \text{ as } h \longrightarrow 0.$$

□

We can now give a formal definition of order of convergence, coming directly from the previous theorem.

Definition C.11 (Order of accuracy)**.** The explicit one-step method (C.25) is said to have order of accuracy p if there exist constants K and h_0 such that:

$$\mid \tau_n \mid \leq Kh^p \qquad \text{for } 0 < h < h_0.$$

From this definition it is immediate to see that both Euler schemes are of order $p = 1$, while Crank-Nicolson, Runge, and Heun are of order $p = 2$. With somewhat lengthy calculations, it can be shown that the 4-stage Runge-Kutta scheme is of order $p = 4$.

C.3.3 Stability

The stability of a scheme is related to the well-posedness of the Cauchy problem as discussed in Definition C.2 and its stability (Definition C.3). In practice, we need to understand how the local errors, which we necessarily observe in any numerical scheme because of truncation, behave under the different conditions. In other words, we require that local errors do not grow with increasing applications of the numerical scheme. The number of times we apply

the numerical scheme may increase because we increase time or because we decrease h. To accommodate these two different views, we need to define different kinds of stability .
Assume we are looking for the solution in a bounded time interval $I = [t_0, T]$, we concentrate on a given time $t_n \in I$ and look for small perturbations of the numerical solution as h tends to zero. Since $t_n = t_0 + nh$, we let n tend to infinity so that

$$\lim_{\substack{h\to 0 \\ n\to\infty}} t_n = t$$

We can introduce the following definition:

Definition C.12 (Zero stability)**.** A one-step scheme (eq. (C.19)) is *zero stable* if perturbations of the solution remain bounded as h tends to zero.
Equivalently, the one step scheme (C.19) for the numerical solution of (C.1) is *zero stable* if there exists $h_0 > 0$ and a constant $C > 0$ such that for each $h \in (0, h_0]$ and $\epsilon > 0$, for $\mid \delta_n \mid \leq \epsilon$ we have that:

$$\mid y_n - z_n \mid \leq \epsilon \qquad \forall 0 \leq n \leq N,$$

where y_n is the numerical solution of Problem (C.1), and z_n is the numerical solution of Problem (C.5).

The explicit scheme (C.19) whenever $\Phi(t, y; h)$ is Lipschitz continuous, is zero stable. In fact, a Lipschitz function $\Phi(t, y; h)$ is characterized by:

$$\mid \Phi(t_k, y_k; h) - \Phi(t_k, z_k; h) \mid \leq \Lambda \mid y_k - z_k \mid.$$

Let $w_k = y_k - z_k$, then we obtain:

$$w_{k+1} = w_k + h\left[\Phi(t_k, y_n; h) - \Phi(t_k, z_n; h)\right] + h\delta_k.$$

Summing for $k = 0, 1, \ldots, n$:

$$w_{n+1} = w_0 + h\sum_{k=0}^{n} \delta_k h \sum_{k=0}^{n} \left[\Phi(t_k, y_n; h) - \Phi(t_k, z_n; h)\right].$$

From the fact that Φ is Lipschitz we have:

$$\mid w_{n+1} \mid \leq \mid w_0 \mid + h\sum_{k=0}^{n} \delta_{k+1} + h\Lambda \sum_{k=0}^{n} \mid w_k \mid$$

Using the discrete version of Gronwall Lemma [12] we readily find:

$$\mid w_n \mid \leq \epsilon\,(1 + hn)\, e^{\Lambda h n}.$$

The proof concludes noting that $hn \leq T$.
A similar argument leads to the following:

12

Theorem C.14 (Lax-Richtmeyer-Dahlquist Theorem)**.** *A scheme that is zero stable and consistent is convergent.*

Note that the Euler methods are zero stable because Problem (C.1) is well posed. Since they are consistent, then they are also convergent.

Example C.14 (Convergence analysis of Forward Euler method)**.** As done before, we define the error at time t_{n+1} as $e_{n+1} = y(t_{n+1}) - y_{n+1}$ and denote with y^*_{n+1} the solution obtained after one step of FE starting from the exact solution:

$$y^*_{n+1} = y(t_n) + hf(t_n, y(t_n)).$$

Then we can write:

$$e_{n+1} = \left(y(t_{n+1}) - y^*_{n+1}\right) + \left(y^*_{n+1} - y_{n+1}\right).$$

The first term on the right-hand-side is the residual (h times the truncation error), while the second term takes care of the accumulation of this error in time. Hence:

$$| e_{n+1} | \leq h| \tau_{n+1} | + (1 + hL)| e_n |.$$

Noting that $e_0 \leq \tau(h)$ and bounding the truncation error with the global one, we obtain by recurrence:

$$| e_{n+1} | \leq [1 + (1 + hL) + \ldots + (1 + hL)^n] h| \tau(h) | \leq \frac{e^{L(t_{n+1}-t_0)} - 1}{L}\tau(h),$$

since $(1 + hL) \leq e^{hL}$ and $t_{n+1} = t_0 + (n + 1)h$. From the consistency of Forward Euler we have:

$$\tau_{n+1}(h) = \frac{h}{2}y''(\xi), \qquad \xi \in [t_n, t_{n+1}],$$

and using $M = max_{\xi \in I}| y''(\xi) |$ we have convergence of the Forward Euler scheme with first-order ($p = 1$) accuracy:

$$| e_n | \leq \frac{e^{L(t_{n+1}-t_0)} - 1}{L}\frac{M}{2}h.$$

Lemma C.13 (Discrete Gronwall Lemma)**.** *Let K_n a non-negative sequence and φ_n a sequence such that, given $\varphi_0 \leq g_0$:*

$$\varphi_n \leq g_0 + \sum_{k=0}^{n-1} p_k + \sum_{k=0}^{n-1} K_k\varphi_k.$$

If $g_0 \geq 0$ and $p_n \geq 0$ for every $n \geq 0$, then:

$$\varphi_n \leq \left(g_0 + \sum_{k=0}^{n-1} p_k\right) e^{\left(\sum_{k=0}^{n-1} K_k\right)}.$$

Remark C.15. *Up to now our analysis assumed that no round-off error (finite precision) occurs. Including this sort of errors in the analysis is straight forward, but if we do it we cannot conclude that the error tends to zero as h tends to zero. In fact, an extra term proportional to* $1/h$ *appears in the expression for the global error (generally with a small constant). Hence, there will be an optimal value of* h^* *that will minimize the global error, and below which the error will start increasing, signaling the dominance of the round-off error over the truncation error. In practice, most of the times* h^* *is very small and the effects of round-off are negligible for practical values of h.*

C.3.4 Absolute stability

Zero stability does not explain the behavior of Forward Euler for the Example C.4, where the scheme was stable for $h < 0.2$. To see this, we need to look at what is called *absolute stability*, that emphasizes stability for fixed h and increasing t. However, this stability concept depends strongly on $f(t, y(t))$, and is customary checking stability against the so called *test equation*:

$$\begin{aligned} y' &= \lambda y \\ y(0) &= 1 \end{aligned} \tag{C.29}$$

whose explicit solution is given by $y(t) = e^{\lambda t}$. Although in general $\lambda \in \mathbb{C}$, we will assume for simplicity $\lambda \in \mathbb{R}$. It is clear that the solution will tend to zero for t that tends to infinity for $\lambda < 0$, otherwise it will tend to infinity. Thus, for $\lambda < 0$, stability implies that perturbation must become smaller and smaller and tend to zero asymptotically. In the case $\lambda > 0$ the definition of stability will have to take into consideration the fact that the solution is unbounded for $t \longrightarrow \infty$. Hence, it is sufficient to require that perturbations remain increase in time but not too much, so that the accuracy of the numerical solution is not overshadowed. We will consider in what follows only the case $\lambda < 0$, and refer to more specialized textbooks for in depth analysis.
We can then define:

Definition C.15 (Absolute stability)**.** The one-step scheme (C.19) is *absolutely stable* if, when applied to the test equation (C.29) with $\lambda < 0$, the numerical solution y_n satisfies:

$$\lim_{tn \to \infty} \mid y_n \mid \to 0$$

Remark C.16. *Note that the test equation resemble a linearized version of the general Cauchy problem* (C.1)*:*

$$\begin{aligned} y'(t) &= \frac{\partial f}{\partial y}(t, y(t))y(t) \\ y(0) &= 1 \end{aligned}$$

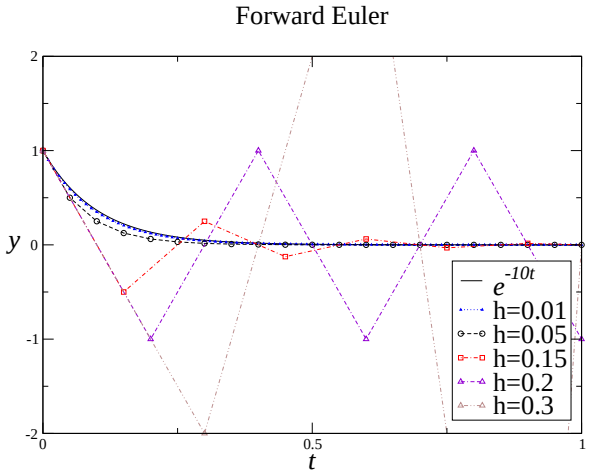


FIGURE C.5: *Numerical solution of the test equation with* $\lambda = -10$ *obtained with Forward Euler for different values of* h *compared with the real solution.*

from which the test equation is obtained by setting $\lambda = \frac{\partial f}{\partial y}(y, y(t))$.

Absolute stability of Euler and Crank-Nicolson schemes We study now the absolute stability for the one-step methods studied so far.

Forward Euler. Applying Forward Euler to the test equation (C.29) we obtain:

$$y_{n+1} = y_n + \lambda h y_n = (1 + h\lambda) y_n.$$

By recursion, starting from $n = 0$, we have immediately:

$$y_n = (1 + h\lambda)^n y_0,$$

that tends to zero only if:

$$| \, 1 + h\lambda \, | < 1.$$

Thus the condition for which the Forward Euler scheme is absolutely stable is:

$$0 < h < \frac{2}{-\lambda} \qquad \lambda < 0.$$

In the case $\lambda = -1$ (Example C.3) the scheme is stable if $h < 2$. In the case $\lambda = -5$ (Example C.4) the scheme is stable if $h < 0.4$. In this case the solution tends to zero, following

the real solution. More precisely, the numerical solution will tend to zero monotonically for $0 < 1 + h\lambda < 1$, which implies $0 < h < 0.2$, while it will oscillate around the true solution (overshooting and undershooting it at successive iterations) for $-1 < (1 + h\lambda) < 0$), i.e., $0.2 < h < 0.4$ In the case $\lambda = -10$, the scheme is stable only if $h < 0.2$, and the numerical solution oscillates if $h < 0.1$, as can be seen from Figure C.5, where the results obtained in this last case for different values of h are shown.

Backward Euler. Application of this scheme to the test equation (C.29) yields:

$$y_{n+1} = y_n + \lambda h y_{n+1} = \frac{1}{1 - h\lambda} y_n.$$

As done before, starting from $n = 0$, we obtain the stability condition:

$$| \tfrac{1}{1-h\lambda} | < 1,$$

which is always verified for all values of h assuming, as before, $\lambda < 0$. The scheme is thus *unconditionally* absolutely stable. It is easy to see that the scheme will always be monotone.

Crank-Nicolson. As mentioned before, Crank-Nicolson is an implicit scheme and it should be unconditionally absolutely stable. In fact:

$$y_{n+1} = y_n + \frac{\lambda h}{2} (y_{n+1} + y_n) = \frac{2 + h\lambda}{2 - h\lambda} y_n,$$

from which the absolute stability condition becomes:

$$\left| \frac{2 + h\lambda}{2 - h\lambda} \right| < 1,$$

always satisfied for all values of $h > 0$ (recall that a negative h would make no sense). On the other hand, the monotonicity of the Crank-Nicolson scheme will be ensured only for $h < 2/| \lambda |$.

C.3.5 Practical implementation of implicit schemes

We now look at the implementation of Backward Euler scheme. The algorithm is:

ALGORITHM IMPLICIT_EULER
Input: t_0, y_0, N, h;
FOR $j = 0, 1, \ldots, N - 1$
    1.$t_{n+1} = t_n + h$
    2.Solve $y_{n+1} - y_n - hf(t_{n+1}, y_{n+1}) = 0$
END FOR

Step # 2 requires the solution of a nonlinear equation for the unknown y_{n+1}. This step is typically implemented using a Newton-like scheme. For simplicity let x be our unknown ($x := y_{n+1}$). The problem becomes:
Find the zero of the function $g(x)$, i.e., find the solution of the nonlinear equation $g(x) = 0$, where:

$$g(x) = x - y_n - hf(t_{n+1}, x) = 0.$$

We can use for example the Newton or Picard methods. Let us study in details what happens with the Picard method. To this aim, we indicate with r the nonlinear (Picard) iteration index. Thus we can write:

$$x^{(r+1)} = g(x^{(r)}) = x^{(r)} - y_n - hf(t_{n+1}, x^{(r)}).$$

It is well known that this iterative scheme converges if $\mid g'(x) \mid < 1$ for $x \in I_\xi$, where I_ξ is a neighborhood of the fixed point ξ. The derivative of g is:

$$g'(x) = 1 - h\frac{\partial f}{\partial x}(t_{n+1}, x),$$

which entails the following convergence condition:

$$\left|1 - h\frac{\partial f}{\partial x}(t_{n+1}, x)\right| < 1.$$

We see that this is equivalent to the restriction on h imposed by the absolute stability for the Forward Euler scheme:

$$h\left|\frac{\partial f}{\partial x}(t_{n+1}, x)\right| < 2.$$

If instead of the Picard method we use Newton's scheme we have no restriction on h for the convergence of the nonlinear iteration, assuming the initial guess is sufficiently close to the final solution. This study can be repeated, with the appropriate adjustments, also for the Crank-Nicolson scheme.

Remark C.17. *For other schemes, such as Runge-Kutta, in general the rule is that explicit schemes are conditionally stable, while implicit schemes are unconditionally stable.*

Remark C.18. *In practical applications, the restriction due to stability is often very strong, and it is often preferable (i.e., computationally less expensive) to use an implicit scheme together with Newton's methods. However, this conclusion must be verified case by case, and depends strongly by the behavior of $f(t, (y(t))$.*

C.4 Linear multistep methods

The methods presented in the previous sections compute y_{n+1} using only one previously-computed point, (t_n, y_n). These methods are called single-step methods. It is intuitive that multi-step methods can be obtained in an attempt to achieve higher accuracy of the approximate solution y_{n+1} using m previously computed points. Indeed, for example, we can extend eq.(C.18) on the interval $[t_{n-1}, t_{n+1}]$ and use Simpson's Rule to approximate the integral:

$$\begin{aligned} y(t_{n+1}) &= y(t_{n-1}) + h \int_{t_{n-1}}^{t_{n+1}} y'(t)\, dt, \\ &\simeq y(t_{n-1}) + \frac{h}{3} \left[f(t_{n-1}, y(t_{n-1})) + 4f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1})) \right) \end{aligned}$$

This procedure can be generalized to a general m-step method. This class of schemes are called linear multistep method and have as particular case the one-step methods described above, and can be implicit or explicit. A linear multistep method (LMM) can be written as:

$$\begin{aligned} a_m y_{n+m} + a_{m-1} y_{n-1} + \ldots + a_0 y_n = \\ h \left(b_m f(t_{n+m}, y_{n+m}) + b_{m-1} f(t_{n+m-1}, y_{n+m-1}) + \ldots + b_0 f(t_n, y_n) \right) \end{aligned}$$

where $a_0, \ldots a_m$ and $b_0, \ldots, b_m$ are fixed coefficients. In a more compact form, an LMM scheme can be written as:

$$\sum_{k=0}^{m} a_k y_{n+k} = h \sum_{k=0}^{m} b_k f(t_{n+k}, y_{n+k}). \tag{C.30}$$

Note that to apply a multistep method we need m initial points,

$$y(t_0) = y_0, \ldots, y(t_{m-1}) = y_{m-1}.$$

Example C.19.

- Backward Euler (implicit one-step method, $m = 1$): $a_1 = 1$, $a_0 = -1$; $b_0 = 0$, $b_1 = 1$:

$$y_{n+1} = y_n + h f(t_{n+1}, y_{n+1});$$

- Forward Euler (explicit one-step method, $m = 1$) $a_m = 1$, $a_{m-1} = -1$, $a_{m-2} \ldots = a_0 = 0$; $b_0 = 1$, $b_1 = 0$:

$$y_{n+1} = y_n + h f(t_n, y_n);$$

- Crank-Nicolson (implicit one-step method, $m = 1$): $a_1 = 1$, $a_{m0} = -1$; $b_0 = 1/2$, $b_1 = 1/2$:

$$y_{n+1} = y_n + \frac{h}{2} \left[f(t_n, y_n) + f(t_{n+1}, y_{n+1}) \right];$$

- Adams-Bashforth (explicit four-step method, $m = 4$): $a_4 = 1$, $a_3 = -1$, $a_2 = a_1 = a_0 = 0$; $b_4 = 0$, $b_3 = 55/24$, $b_2 = -59/24$, $b_1 = 37/24$, $b_0 = -9/24$:

$$y_{n+4} = y_{n+3} + \frac{h}{24}\left[55f(t_{n+3}, y_{n+3}) - 59f(t_{n+2}, y_{n+2}) + 37f(t_{n+1}, y_{n+1}) - 9f(t_n, y_n)\right]$$

  Adams-Moulton (implicit four-step method): $a_4 = 1$, $a_3 = -1$, $a_2 = a_1 = a_0 = 0$; $b_4 = 9/24$, $b_3 = 19/24$, $b_2 = -5/24$, $b_1 = -9/24$, $b_0 = 0$:

$$y_{n+4} = y_{n+3} + \frac{h}{24}\left[9f(t_{n+4}, y_{n+4}) + 19f(t_{n+3}, y_{n+3}) - 5f(t_{n+2}, y_{n+2}) - 9f(t_{n+1}, y_{n+1})\right]$$

C.4.1 Convergence of LMMs

Truncation error and zero-stability. The definition of the truncation error for an LMM is a straight forward generalization of the definition of TE for one=step methods.

Definition C.16 (Truncation error of LMM)**.** Given the linear multistep method of eq. (C.30), the *truncation error* (TE) is defined as:

$$\tau_n = \frac{\sum_{k=0}^{m} a_k y(t_{n+k}) - h\sum_{k=0}^{m} b_k f(t_{n+k}, y(t_{n+k}))}{h\sum_{k=0}^{m} b_k}.$$

Again, this definition can be interpreted as the residual obtained by substituting the real solution $y(t_n)$ in place of the numerical approximation in the expression of the LM scheme, written in terms of direct approximation of $y' = f(t, y)$. It coincides with the global error. Hence we can state the following.

Definition C.17 (Consistency of LMM)**.** A LMM of the form (C.30) is *consistent* with the Cauchy Problem C.1 if the TE tends to zero in the limit for h that tends to zero:for all $\epsilon > 0$ there exists $h(\epsilon) > 0$ such that

$$\mid \tau_n \mid < \epsilon \qquad \text{for all } 0 < h < h(\epsilon)$$

and for all $m + 1$ points $(t_n, y(t_n)$ $(t_{n+1}, y(t_{n+1})$ $\ldots$, $(t_{n+m-1}, y(t_{n+m-1})$.

We can define the zero-stability using the same concept as before, but taking now into account that we have m initial solutions to start the full method.

Definition C.18 (Zero stability of LMM)**.** An m-step LMM of the form (C.30) is *zero-stable* if there exists a constant K such that, given the sequences $\{y_n\}$ and $\{\hat{y}_n\}$ generated by the

same scheme but for different starting conditions $y_0, y_1, \ldots, y_{m-1}$ and $\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{m-1}$, then we have:

$$\mid y_n - \hat{y}_n \mid \leq K \max \{\mid y_0 - \hat{y}_0 \mid, \mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_{m-1} - \hat{y}_{m-1} \mid\}$$

for all $t_n \leq T$ and $h \longrightarrow 0$.

It is typical to define LMM characteristic polynomials as follows.

Definition C.19 (First and Second Characteristic Polynomials of LMM)**.** Given the LMM of (C.30), the *first and second characteristic* polynomials are given by:

$$\rho(z) = \sum_{k=0}^{m} a_k z^k \qquad a_m \neq 0;$$
$$\sigma(z) = \sum_{k=0}^{m} b_k z^k \qquad a_0^2 + b_0^2 \neq 0$$

Consistency and the fact that the LMM has $m - step$ requires that $a_m \neq 0$ and that a_0 and b_0 be both nonzero at the same time.

We have the following theorem:

Theorem C.20 (LMM root condition)**.** *Given the LMM of* (C.30)*, is zero-stable for the Cauchy Problem C.1 (with Lipschitz f) is and only if the first characteristic polynomial has roots that are all within the unit disk, and any root that is on the unit circle is simple.*

Proof. The proof of the necessary condition of the theorem is easy. Not so the sufficiency. Thus we prove only the former, and refer to the literature (e.g., [12]) for the proof.

$\Rightarrow$ For zero stability we look at the solution of the equation $y' = 0$. In this case the m-step LMM scheme is:

$$a_m y_{n+m} + \ldots + a_0 y_n = 0.$$

Looking for the solution of the form $y_n = z^n$, $z \in \mathbb{C}$, we have:

$$y_n = \sum_s \rho_s(n) z_s^n,$$

where z_s is a root of the first characteristic polynomial $\rho(z)$ of the LMM and $\rho_s(n)$ is a polynomial of degree $r - 1$, where r is the multiplicity of z_s. Obviously, if $\mid z_s \mid > 1$ there

exist some starting points for which the solution explodes as $\mid z_s \mid^n$. If $\mid z_s \mid = 1$ and $r > 1$, again there exist some starting points for which the solution grows as n^{h-1}. In both cases, the $y_n \longrightarrow \infty$ as $n \longrightarrow \infty$ and $h \longrightarrow 0$ with fixed nh, and the scheme is not zero-stable. □

Example C.20. Simple examples.

1. **(Backward and Forward) Euler and Crank-Nicolson.** All these schemes are zero stable. Since they differ in the way the function on the RHS of Problem C.1, their first characteristic polynomial is the same:

   $$\rho(z) = (z - 1).$$

   It has obviously a root $z = 1$ which is on the unit circle but is simple.

2. **Adams-Bashforth and Adams-Moulton.** Again Adams-Bashforth and Adams-Moulton differ for the right-hand-side. Their first characteristic polynomial is:

   $$\rho(z) = z^3(z - 1)$$

   and are both zero-stable

The following (order $p = 6$) LMM is not zero stable:

$$11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n = 3h\left(f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n\right).$$

Its first characteristic polynomial $\rho(z) = 11z^3 + 27z^2 - 27z - 11$ has roots given by: $z_1 = 1$; $z_2 \simeq -0.3189$; $z_e \simeq -3.1356$.

To verify the consistency of any LMM scheme we can use Taylor series (assuming enough regularity) to expand $y(t_{n+m})$ and $y'(t_{n+m})$ and evaluate the residual of the truncation error τ_n. Lengthy (but easy) calculations show that:

$$\begin{aligned}\tau_n =& \frac{\sum_{k=0}^m a_k y(t_{n+k}) - h\sum_{k=0}^m b_k f(t_{n+k}, y(t_{n+k}))}{h\sum_{k=0}^m b_k} \\ =& \frac{1}{h\sigma(1)}\left[c_0 y(t_n) + c_1 h y'(t_n) + c_2 h^2 y''(t_n) + c_3 h^3 y'''(t_n) + \ldots\right]\end{aligned} \tag{C.31}$$

where the constants are given by:

$$\begin{aligned} c_0 &= \sum_{k=0}^m a_k, & c_1 &= \sum_{k=1}^m k a_k - \sum_{k=0}^m b_k = \rho'(1) - \sigma(1) \\ c_2 &= \sum_{k=1}^m \frac{k^2}{2!} a_k - \sum_{k=1}^m k b_k & c_s &= \sum_{k=1}^m \frac{k^2}{2!} a_k - \sum_{k=1}^m \frac{k^{s-1}}{(s-1)!} b_k. \end{aligned} \tag{C.32}$$

For the consistency of the scheme we require that $c_0 = c_1 = 0$. This results into $\rho(1) = 0$ and $\rho'(1) = \sigma(1) \neq 0$. Hence $z_s = 1$ is a root of the first characteristic polynomials that lies on the unit circle, but it is simple, and thus it does not violate the root condition.
We are now in a position to define the order of accuracy, exactly as done in the case of one-step methods.

Definition C.21 (Order of accuracy of LMM)**.** An LMM (eq. (C.30)) is said to posses *order of accuracy* p if the truncation error is $\mathcal{O}(h^p)$, i.e., if there exists a $h_0 > 0$ such that

$$| \tau_n | \leq K h^p \qquad 0 \leq h \leq h_0$$

for all $m + 1$ starting points $(t_n, y(t_n))$, $(t_{n+1}, y(t_{n+1}))$, ..., $(t_{n+m}, y(t_{n+m}))$.

Looking at eq. (C.31), we note that to have order of accuracy p we need $c_0 = c_1 = \ldots = c_p = 0$ and $c_{p+1} \neq 0$. The constant c_{p+1} is called the *error constant* and we have:

$$\tau_n = \frac{c_{p+1}}{\sigma(1)} h^p y^{(p+1)}(t_n) + \mathcal{O}\left(h^{p+1}\right).$$

Example C.21. We want to build a $m = 2$-step LMM with maximum order and one free parameter. A two-step LMM is written as:

$$a_0 y_{n+1} + a_1 y_{n+1} + a_2 y_n = h\left(b_2 y'_{n+2} + b_1 y'_{n+1} + b_0 y'_n\right)$$

where we need to impose the requirements $a_0 \neq 0$ and $a_0^2 + b_0^2 \neq 0$. Expanding in Taylor series and substituting, or using directly the expressions of the constants c_i given by (C.32), we have:

$$\begin{aligned}(a_2 + a_1 + a_0) y(t_n) + (2a_2 + a_1) y'(t_n) + (2a_2 + \frac{1}{2} a_1) y''(t_n) + \ldots \\ = h\left[(b_2 + b_1 + b_0) y'(t_n) + (2b_2 + b_1) y''(t_n) + \ldots\right].\end{aligned}$$

From the first consistency condition we obtain $c_0 = a_2 + a_1 + a_0 = 0$. Since we want one free parameter, we take $a_2 = 1$ and we let $a_0 = \alpha$ be the free parameter. From the last condition we have thus $a_1 = -1 - \alpha$. The remaining conditions needed for consistency are then:

$$c_1 = 2 + a_1 - (b_2 + b_1 + b_0) = 0 \quad c_2 = \frac{1}{2}(a_1 + 4) - (b_1 + 2b_2) = 0 \quad c_3 = \frac{1}{3!}(a_1 + 8) - \frac{1}{2}(b_1 + 4b_2) = 0$$

leading to:

$$b_0 = -\frac{1}{12}(1 + 5\alpha) \quad b_1 = \frac{2}{3}(1 - \alpha) \quad b_2 = \frac{1}{12}(5 + \alpha).$$

The final scheme is then:

$$y_{n+2} - (1+\alpha)y_{n+1} + \alpha y_n = h\left[\frac{1}{12}(5+\alpha)y'_{n+2} + \frac{2}{3}(1-\alpha)y'_{n+1} - \frac{1}{12}(1+5\alpha)y'_{n+2}\right].$$

The error constants c_4 and c_5 are:

$$c_4 = -\frac{1}{4!}(1+\alpha) \qquad c_5 = -\frac{1}{3\cdot 5!}(17+13\alpha).$$

We can see that we cannot have both $c_4 = c_5 = 0$. Hence, to have maximum order we choose $\alpha = -1$ so that $c_4 = 0$ and we have order $p = 4$ and error constant $c_5 = -4/3\cdot 5!$. The wanted scheme is:

$$y_{n+1} = y_n + \frac{h}{3}\left[y'_{n+2} + 4y'_{n+1} + y'_n\right]$$

which is Simpson rule.

We now look at the convergence of LMMs.

Definition C.22 (Convergence of LMM)**.** An m-step LMM for the Cauchy Problem C.1 is *convergent* for any initial value y_0 if

$$\lim_{\substack{h\longrightarrow 0 \\ nh\longrightarrow t-t_0}} y_n = y(t).$$

for all $t \in (t_0, T]$ and for all solutions of the LMM difference equations with consistent starting values, i.e., $y_s = \eta_s(h)$, $s = 0, \ldots, m-1$, and $\lim_{\substack{h\longrightarrow 0 \\ nh\longrightarrow t-t_0}} \eta_s(h) = y_0$.

We state here without proof two important theorems due to Dahlquist.

Theorem C.23 (Dahlquist (Lax-Richtmeyer))**.** *A consistent m-step LM scheme with consistent starting values converges if and only if it is zero-stable. If $y(t) \in \mathcal{C}^{p+1}([t_0, T])$ and $\tau_n = \mathcal{O}(h^p)$ then the scheme has order of accuracy p and the global error is $\mathcal{O}(h^p)$.*

Theorem C.24 (Dahlquist I barrier)**.** *There are no zero-stable m-step LM methods with order of accuracy larger than $m+2$ if m is even or $m+1$ is m is odd.*

Absolute stability. We now look at the properties of absolute stability, i.e., how the scheme behave when applied to the test equation:

$$\begin{aligned} y' &= \lambda y \qquad \lambda \in \mathbb{C} \quad \mathrm{Re}(\lambda) < 0 \\ y(t_0) &= y_0 \end{aligned} \tag{C.33}$$

at a fixed h and as $n \longrightarrow \infty$, i.e., $t \longrightarrow \infty$. We note that since $\mid y(t) \mid \longrightarrow 0$ as $t \longrightarrow \infty$, we must have that $y_n \longrightarrow 0$ as $n \longrightarrow \infty$.
Applying the LMM (C.30) to (C.33) we obtain:

$$\sum_{k=0}^{m} (a_k - h\lambda b_k)\, y_{n+l} = 0. \tag{C.34}$$

The polynomial

$$\pi_m(z; h\lambda) = \rho(z) - h\lambda\sigma(z)$$

is called the stability polynomial. Since the solution of the difference equation (C.34) is given by $y_n = \sum_{k=0}^{m} \gamma_k z_k^n$, we have that $\mid y_n \mid \longrightarrow 0$ if all the roots of the stability polynomials are internal to the unit disk, $\mid z_k \mid < 1$.

Definition C.25. The m-step LMM is *absolutely stable* if all roots of $\pi_m(z; h\lambda)$ are such that $\mid z_k \mid < 1$, $k = 1, \ldots, m$. Otherwise, it is *absolutely unstable.*
The region $\mathcal{A}$ of the complex plane $h\lambda$ for which LMM is absolutely stable is called the *absolute stability region.*

We have the following.

Theorem C.26. *For* $\mathrm{Re}(h\lambda) > 0$*, all linear multistep methods are absolutely unstable.*

Example C.22. We report here some examples of stability regions.

Forward Euler The FE method for the test equation is an LMM with $a_1 = 1$, $a_0 = -1$, $b_1 = 0$, and $b_0 = 1$:

$$y_{n+1} - y_n = h y'_n.$$

The stability polynomial is thus given by:

$$\pi_m(z, h\lambda) = -1 + z - h\lambda.$$

Its only root is given by $z_1 = 1 + h\lambda$. The region of absolute stability is thus:

$$\mathcal{A}_{FE} = \{h\lambda \in \mathbb{C} : \mid 1 + h\lambda \mid < 1, \mathrm{Re}(\lambda) < 0\}$$

which is the unit disk centered in -1 (Figure C.6, left).

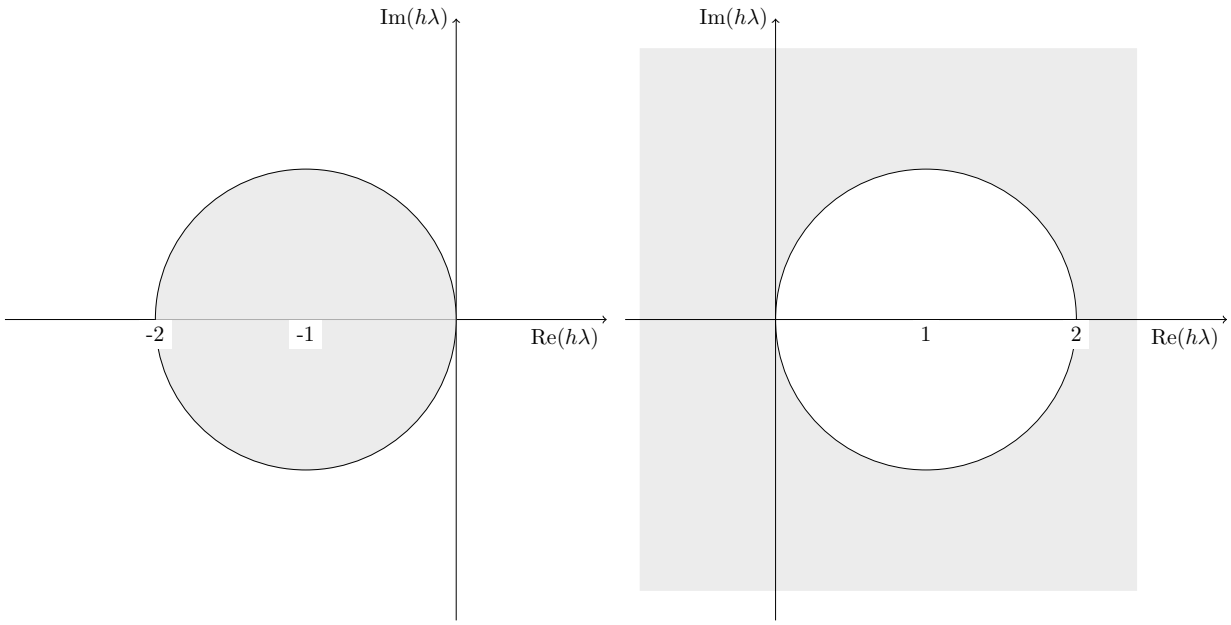


FIGURE C.6: *Regions of absolute stability for Forward Euler (left) and Backward Euler (right)*

Backward Euler The BE method for the test equation is an LMM with $a_1 = 1$, $a_0 = -1$, $b_1 = 1$, and $b_0 = 0$:

$$y_{n+1} - y_n = hy'_{n+1}.$$

The stability polynomial is thus given by:

$$\pi_m(z, h\lambda) = -1 + z(1 - h\lambda).$$

Its only root is given by $z_1 = 1/(1 - h\lambda)$ whose absolute value is always smaller than unity since $\mathrm{Re}(\lambda) < 0$. The region of absolute stability is thus:

$$\mathcal{A}_{BE} = \{h\lambda \in \mathbb{C} : \mid 1 - h\lambda \mid > 1, \mathrm{Re}(\lambda) < 0\}$$

which is the region outside the unit disk centered in $+1$ (Figure C.6, right).

C.5 Systems of ODEs

We extend here the above results to the case of systems of ODEs of the first order. Hence, we assume that $y \in \mathbb{R}^d$ is an d-dimensional vector i.e., $y(t) = (y_1(t), \ldots, y_d(t))$ and the function $f : \mathbb{R} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is also an n-dimensional vector. The vector function y satisfies the following system of ODEs of the first order:

$$\begin{cases} P(t)y' = f(t, y), & t \in [t_0, T] \\ y(t_0) = y_0 & \in \mathbb{R}^d, \end{cases} \tag{C.35}$$

where matrix P is d-dimensional and non-singular, with coefficients that depend possibly on t, y' is the derivative of y, and y_0 is the vector of given initial conditions. The numerical methods presented in section C can be applied also to Eq. (C.35). Note that, when the function f is nonlinear, the use of implicit schemes implies the solution of a d-dimensional nonlinear system of equations at each step of the scheme.
If f is a linear function of y, i.e., it can be written as $f(t, y) = H(t)y + q(t)$, where H and q are d-dimensional matrix and vector, respectively. Eq. (C.35) rewrites:

$$P(t)y' + H(t)y + q(t) = 0. \tag{C.36}$$

Assuming for simplicity that both P and H ar not dependent upon t and that $q = 0$, the solution can be written explicitely as:

$$y(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t} u_i, \tag{C.37}$$

where $\lambda_1, \ldots, \lambda_n$ and $u_1, \ldots, u_n$ are the eigenvalues and the eigenvectors of the matrix $-P^{-1}H$. The coefficients $c_1, \ldots c_n$ satisfy the initial condition:

$$y_0 = c_1 u_1 + \ldots + c_d u_d.$$

Actually, solution (C.37) can be extended to the time-dependent case and for $q \neq 0$, using standard techniques. However, this approach at the solution is complicated as the computation of the eigenvalues and eigenvectors of $-P^{-1}H$ is an expensive procedure, and is seldom done. One typically resorts to numerical schemes also for linear equations.
Before applying the LMM to the system, we transform it in a more common form, by multiplying the above system by P^{-1} to yield:

$$y' = Ay + g \qquad A = -P^{-1}H, \quad g = -P^{-1}q.$$

Applying the LMM of eq (C.30) we get:

$$\sum_{k=0}^{m} \left(a_k I - h b_k A\right) y_{n+k} = h\sigma(1)g.$$

We need the hypothesis that A is diagonalizable, so that there exists a matrix U such that $UAU^{-1} = \Lambda$, where Λ is the $d \times d$ diagonal matrix containing the eigenvalues of A:

$$UAU^{-1} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}.$$

Defining $z = Uy$ and substituting we have immediately:

$$\sum_{k=0}^{m} \left(a_k U - h b_k U A\right) U^{-1} y_{n+k} = hc = h\sigma(1) U g,$$

from which: Assuming for simplicity that P and H do not depend on t and $q = 0$, then Eq. (C.36) admits an explicit solution:

$$\sum_{k=0}^{m} \left(a_k I - h b_k \Lambda\right) z_{n+k} = hc,$$

or, for the component i:

$$\sum_{k=0}^{m} \left(a_k - h b_k \lambda_i\right) z_{n+k,i} = hc_i. \tag{C.38}$$

We see that the stability polynomial can be defined in terms of the eigenvalues λ_i of A, and we can state the following:

Definition C.27. An m-step LMM is *absolutely stable* in an open set $\mathcal{A}$ of the complex plane if for all $h\lambda_i \in \mathcal{A}$ the roots of the stability polynomial $\pi_m(z; h\lambda_i)$ of (C.38) are such that $\mid z_k \mid < 1$, $k = 1, \ldots, m$ for all eigenvectors λ_i of A. The set $\mathcal{A}$ is called the *region of absolute stability* of the LMM.

Example C.23. Consider the second order constant coefficient equation:

$$y'' + (\lambda + 1)y' + \lambda y = 0 \qquad y(0) = 1; \quad y'(0) = \lambda - 2.$$

This equation can be easily solved by standard procedures to obtain:

$$y(x) = 2e^{-t} - e^{-\lambda t}.$$

We transform this into a system of first order equations by setting $z = (u, v)^T$, where $u = y$ and $v = y'$. Then we have:

$$\begin{array}{l} z' = Az \\ z(0) = z_0 \end{array} \qquad A = \begin{pmatrix} 0 & -1 \\ -\lambda & -(\lambda+1) \end{pmatrix} \quad z_0 = \begin{pmatrix} \lambda - 2 \\ 2 \end{pmatrix}$$

The eigenvalues of matrix A are $\mu_1 = -1$ and $\mu_2 = -\lambda$. The solution of the system is given by:

$$u(t) = 2e^{-t} - e^{-\lambda t} \qquad v(t) = -2e^{-t} + \lambda e^{-\lambda t}. \tag{C.39}$$

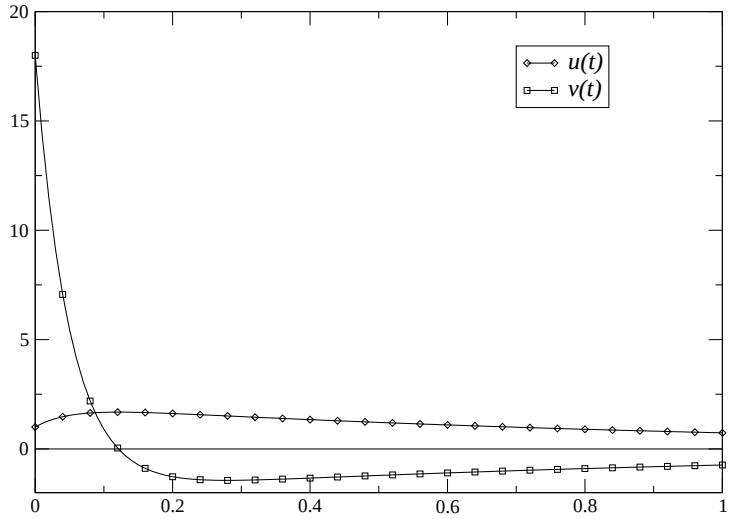


FIGURE C.7: *Functions $u(t)$ and $v(t)$ of eq.* (C.39) *of Example C.23 for* $\lambda = 20$

The stability polynomial of the Forward Euler method is:

$$\pi_{FE,m}(z; h\mu) = -1 - z - h\mu, \qquad \text{for } \mu = \mu_1 = -1 \text{ and } \mu = \mu_2 = -\lambda.$$

The root condition for stability explicitates into the following conditions that should be satisfied simultaneously:

$$\begin{aligned} \mu = \mu_1 \qquad & | \, 1 - h \, | < 1 && 0 < h < 2 \\ \mu = \mu_2 \qquad & | \, 1 - h\lambda \, | < 1 && 0 < h < \frac{2}{\lambda} \end{aligned}$$

Obviously, the second equation is more stringent if $\lambda > 1$.
Let us look at a concrete case, and take $\lambda = 20$ ($v(0) = 18$). The solutions are shown in Figure C.7. In this case, absolute stability of Forward Euler requires $h < 1/10$.

It is clear from the Example above that the two functions that solve the system behave differently. In particular, $u(t)$ is smooth and varies slowly in time, while $v(t)$ displays a fast initial transient, i.e., at small times $v(t)$ shows a strong variations. This fast transient is responsible for the strongest stability constraint of the time step size of Forward Euler. This is a prototypical *stiff* system of equations.

C.5.1 Stability of LMMs for stiff systems

If we apply Backward Euler to the problem of Example C.23, we find immediately that its stability polynomial is:

$$\pi_{BE,m}(z; h\mu) = z(1 - h\mu) - 1, \qquad \text{for } \mu = \mu_1 = -1 \text{ and } \mu = \mu_2 = -\lambda.$$

and the region of absolute stability is given by:

$$\mathcal{A} = \{h\mu : \mid 1 - h\mu \mid > 1\}.$$

In particular, $\mathcal{A}$ containes the entire negative complex half plane. We conclude that BE is always absolutely stable independently of h. This motivates the following:

Definition C.28 (A-stability)**.** An m-step LMM is *A-stable* if its region of absolute stability $\mathcal{A}$ containes the negative complex half plane $\text{Re}(h\lambda) < 0$.

Then, Backward Euler is A-stable. Unfortunately, there is Dahlquist's II barrier to the order of accuracy of A-stable LMMs, that we state without proof.

Theorem C.29 (Dahlquist II barrier)**.** *1. There are no explicit LMMs that are A-stable.*

2. The maximum order of accuracy of an A-stable implicit LMM is 2.

3. The A-stable implicit LMM with smallest error constant is Crank-Nicolson.

In light of this theorem, we conclude that it is inconvenient to use explicit schemes for stiff problems. But we are restricted with the order of accuracy that we can use.
In practice, the A-stability property is often too strong, and there are implicit schemes that are not A-stable but that work well for most stiff problems. Thus we define the following stability classes.

Definition C.30 ($A(\alpha)$-stability)**.** An m-step LMM is *$A(\alpha)$-stable*, for $\alpha \in (0, \pi/2)$, if its region of absolute stability $\mathcal{A}$ contains the wedge:

$$\mathcal{W}_\alpha = \{h\lambda : \pi - \alpha < \arg(h\lambda) < \pi + \alpha\}$$

Definition C.31 ($A(0)$-stability)**.** An m-step LMM is *$A(0)$-stable* if there exists $\alpha \in (0, \pi/2)$ for which it is $A(\alpha)$-stable.

Figure C.8 gives a geometrical interpretation of the above definitions. If $h\lambda$ belongs to the shaded region, then the scheme is absolutely stable.

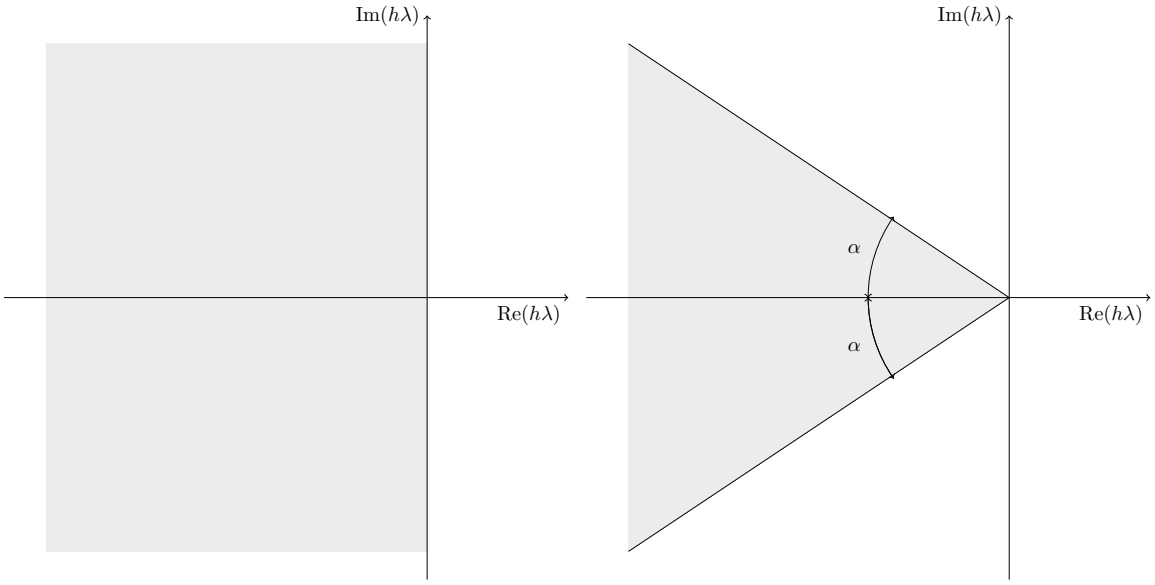


FIGURE C.8: *A-stable region (left). $A(\alpha)$-stable region (right)*

C.5.2 Forward Euler

We apply the forward Euler scheme (explicit) to Eq. (C.36) with step $h = \Delta t$:

$$\frac{1}{\Delta t} P_t y_{t+\Delta t} = \left(\frac{1}{\Delta t} P_t - H_t \right) y_t - q_t.$$

As seen in the previous section, this scheme is consistent and of the first order. Also in this case the scheme is conditionally stable, in fact the error can be written as:

$$\frac{1}{\Delta t} P e_{t+\Delta t} = \left(\frac{1}{\Delta t} P_t - H_t \right) e_t$$

i.e.,

$$e_k = E^k e_0,$$

where $t = k\Delta t$ and $E = \Delta t P^{-1} \left(\frac{1}{\Delta t} P - H \right)$. Stability of the method requires:

$$\lim_{k \to \infty} E^k = 0$$

which is satisfied if and only if the spectral radius of E, $\rho(E)$ is smaller than one. Note that $E = I - \Delta t P^{-1} H$, where I is the n-dimensional identity matrix. Assuming that $P^{-1}H$ has positive real eigenvalues, the condition for the stability of the scheme is

$$\Delta t < \frac{2}{\rho(P^{-1}H)}.$$

C.5.3 Backward Euler

The backward Euler scheme is:

$$\left(\frac{1}{\Delta t}P_{t+\Delta t} + H_{t+\Delta t}\right) y_{t+\Delta t} = \frac{1}{\Delta t}P_{t+\Delta t}y_t - q_{t+\Delta t}.$$

The error matrix is:

$$E = (P + \Delta t H)^{-1} P = \left(I + \Delta t H P^{-1}\right)^{-1},$$

with spectral radius that is always smaller than one under the hypothesis that $P^{-1}H$ has positive real eigenvalues. For this reason, also in the multidimensional case the backward Euler scheme is unconditionally stable.

C.5.4 Crank-Nicolson

The Crank-Nicolson scheme is:

$$\left(\frac{1}{\Delta t}P_{t+\Delta t} + \frac{1}{2}H_{t+\Delta t}\right) y_{t+\Delta t} = \left(\frac{1}{\Delta t}P_{t+\Delta t} - H_t\right) y_t - q_{t+\Delta t}.$$

Also in this case it is easy to see that the spectral radius of the error matrix is always smaller than one under the hypothesis that $P^{-1}H$ has positive real eigenvalues.

References

[1] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1st edition, 1998. ISBN 0898714125.

[2] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[3] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext (Berlin. Print). Springer, 2010.

[4] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, Berlin, 1991.

[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods*. Scientific Computation. Springer, Berlin, 2007. ISBN 978-3-540-30727-3. Evolution to complex geometries and applications to fluid dynamics.

[6] H. S. Carslaw and J. C. Jaeger. *Conduction of Heat in Solids*. Oxford University Press, USA, 1986.

[7] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, Providence (R.I.), 2010.

[8] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In C. Ph. and L. J. L., editors, *Handbook of numerical analysis*, volume VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[9] G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-31716-0. Modern techniques and their applications, A Wiley-Interscience Publication.

[10] B. Fornberg. *A practical guide to pseudospectral methods*, volume 1 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 1996. ISBN 0-521-49582-2. doi: 10.1017/CBO9780511626357. URL `http://dx.doi.org/10.1017/CBO9780511626357`.

[11] B. Fornberg and D. M. Sloan. A review of pseudospectral methods for solving partial differential equations. In *Acta numerica, 1994*, Acta Numer., pages 203–267. Cambridge Univ. Press, Cambridge, 1994.

[12] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1971. ISBN 0136266061.

[13] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I (2Nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag New York, Inc., New York, NY, USA, 1993.

[14] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag New York, Inc., New York, NY, USA, 1993.

[15] R. Helmig. *Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems*. Environmental engineering. Springer, Berlin, New York, 1997. ISBN 3-540-62703-0.

[16] C. Hirsch. *Numerical computation of internal and external flows. volume 1. , fundamentals of computational fluid dynamics*. Butterworth-Heinemann, Amsterdam, Boston, London, 2007. ISBN 978-0-7506-6594-0.

[17] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, 1995.

[18] A. Logg. Efficient representation of computational meshes. *International Journal of Computational Science and Engineering*, 4(4):283, 2009.

[19] R. Peyret. *Spectral methods for incompressible viscous flow*, volume 148 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002. ISBN 0-387-95221-7. doi: 10.1007/978-1-4757-6557-1. URL `http://dx.doi.org/10.1007/978-1-4757-6557-1`.

[20] M. Putti and C. Cordes. Finite element approximation of the diffusion operator on tetrahedra. *SIAM J. Sci. Stat. Comp.*, 19:1154–1168, 1998.

[21] A. Quarteroni. *Numerical Models for Differential Problems*. Springer-Verlag Italia, Milano, 2009.

[22] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[23] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Springer, Berlin, Heidelberg, second edition, 2007.

[24] J. J. Risler. *Mathematical Methods for CAD*. Cambridge University Press, New York, NY, USA, 1993. ISBN 0521436915.

[25] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, New York, NY, 1980.