Alfio Quarteroni

Numerical Models for Differential Problems

Springer

Contents

3

Elliptic equations

This chapter is devoted to the introduction of elliptic problems and to their weak formulation. Although our introduction is quite basic, the complete novice to functional analysis is invited to consult Chap. 2 before reading it.

For the sake of simplicity, most of our derivation will be given for one-dimensional and two-dimensional problems. However, the generalization to three-dimensional problems is (almost always) straightforward.

3.1 An elliptic problem example: the Poisson equation

Consider a domain $\Omega \subset \mathbb{R}^2$, i.e. an open bounded and connected set, and let $\partial\Omega$ be its boundary. We denote by $\mathbf{x}$ the spatial variable pair (x_1, x_2). The problem under examination is

$$-\Delta u = f \quad \text{in } \Omega, \tag{3.1}$$

where $f = f(\mathbf{x})$ is a given function and the symbol Δ denotes the Laplacian operator (1.6) in two dimensions. (3.1) is an elliptic, linear, non-homogeneous (if $f \neq 0$) second-order equation. We call (3.1) the *strong formulation* of the Poisson equation. We also recall that, in the case where $f = 0$, equation (3.1) is known as the Laplace equation.

Physically, u can represent the vertical displacement of an elastic membrane due to the application of a force with intensity equal to f, or the electric potential distribution due to an electric charge with density f.

To obtain a unique solution, suitable boundary conditions must be added to (3.1), that is we need information about the behavior of the solution u at the domain boundary $\partial\Omega$. For instance, the value of the displacement u on the boundary can be assigned

$$u = g \quad \text{on } \partial\Omega, \tag{3.2}$$

where g is a given function, and in such case we will talk about a *Dirichlet problem*. The case where $g = 0$ is said to be *homogeneous*.

A. Quarteroni: Numerical Models for Differential Problems.

Alternatively, the value of the *normal derivative* of u can be imposed

$$\nabla u \cdot \mathbf{n} = \frac{\partial u}{\partial n} = h \quad \text{on } \partial\Omega,$$

$\mathbf{n}$ being the outward unit normal vector on $\partial\Omega$ and h an assigned function. The associated problem is called a *Neumann problem* and corresponds, in the case of the membrane problem, to imposing the traction at the boundary of the membrane itself. Once again, the case $h = 0$ is said to be *homogeneous*.

Finally, different types of conditions can be assigned to different portions of the boundary of the computational domain Ω. For instance, supposing that $\partial\Omega = \Gamma_D \cup \Gamma_N$ with $\overset{\circ}{\Gamma}_D \cap \overset{\circ}{\Gamma}_N = \emptyset$, the following conditions can be imposed:

$$\begin{cases} u = g & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial n} = h & \text{on } \Gamma_N. \end{cases}$$

The notation $\overset{\circ}{\Gamma}$ has been used to indicate the interior of Γ. In such a case, the associated problem is said to be *mixed*.

Also in the case of homogeneous Dirichlet problems where f is a continuous function in $\overline{\Omega}$ (the closure of Ω), it is not guaranteed that problem (3.1), (3.2) admits a regular solution. For instance, if $\Omega = (0,1) \times (0,1)$ and $f = 1$, u could not belong to the space $C^2(\overline{\Omega})$. Indeed, if it were so, we would have

$$-\Delta u(0,0) = -\frac{\partial^2 u}{\partial x_1^2}(0,0) - \frac{\partial^2 u}{\partial x_2^2}(0,0) = 0$$

as the boundary conditions would imply that $u(x_1, 0) = u(0, x_2) = 0$ for all x_1, x_2 belonging to $[0,1]$. Hence u could not satisfy equation (3.1), that is

$$-\Delta u = 1 \quad \text{in } (0,1) \times (0,1).$$

What can be learned from this counterexample is that, even if $f \in C^0(\overline{\Omega})$, it makes no sense in general to look for a solution $u \in C^2(\overline{\Omega})$ to problem (3.1), (3.2), while one has greater probabilities to find a solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ (a larger space than $C^2(\overline{\Omega})$!).

We are therefore interested in finding an alternative formulation to the strong one, also because, as we will see in the following section, the latter does not allow the treatment of some physically significant cases. For instance, it is not guaranteed that, in the presence of non-smooth data, the physical solution lies in the space $C^2(\Omega) \cap C^0(\overline{\Omega})$, and not even that it lies in $C^1(\Omega) \cap C^0(\overline{\Omega})$.

3.2 The Poisson problem in the one-dimensional case

Our first step is the introduction of the weak formulation of a simple boundary-value problem in one dimension.

3.2.1 Homogeneous Dirichlet problem

Let us consider the homogeneous Dirichlet problem in the one-dimensional interval $\Omega = (0,1)$

$$\begin{cases} -u''(x) = f(x), & 0 < x < 1, \\ u(0) = 0, & u(1) = 0. \end{cases} \tag{3.3}$$

This problem governs, for instance, the equilibrium configuration of an elastic string with tension equal to one, fixed at the extrema, in a small displacement configuration and subject to a transversal force with intensity f. The overall force acting on section $(0,x)$ of the string is

$$F(x) = \int_0^x f(t)dt.$$

The function u describes the vertical displacement of the string relative to the resting position $u = 0$.

The strong formulation (3.3) is in general inadequate. If we consider, for instance, the case where the elastic string is subject to a charge concentrated in one or more points (in such case f can be represented via Dirac "deltas"), the physical solution exists and is continuous, but not differentiable. See the graphs of Fig. 3.1, where the case of a unit charge concentrated only in the point $x = 0.5$ is considered (left) and in the two points $x = 0.4$ and $x = 0.6$ (right). These functions cannot be solutions of (3.3), as the latter would require the solution to have a continuous second derivative. Similar considerations hold in the case where f is a piecewise constant function. For instance, in the case represented in Fig. 3.2 of a null load except for the interval $[0.4, 0.6]$ where it is equal to -1, the analytical solution is only of class $C^1([0,1])$, since it is given by

$$u(x) = \begin{cases} -\dfrac{1}{10}x & \text{for } x \in [0, 0.4], \\ \dfrac{1}{2}x^2 - \dfrac{1}{2}x + \dfrac{2}{25} & \text{for } x \in [0.4, 0.6], \\ -\dfrac{1}{10}(1-x) & \text{for } x \in [0.6, 1]. \end{cases}$$

A formulation of the problem alternative to the strong one is therefore necessary to allow reducing the order of the derivation required for the unknown solution u. We move from a second order differential problem to a first-order one in integral form, which is called the *weak formulation* of the differential problem.

To this end, we operate a sequence of formal transformations of (3.3), without worrying at this stage whether all the operations appearing in it are allowed. We start by multiplying equation (3.3) by a (so far arbitrary) *test function* v and integrating on the interval $(0,1)$,

$$-u''v = fv \quad \Rightarrow \quad -\int_0^1 u''v \, dx = \int_0^1 fv \, dx.$$

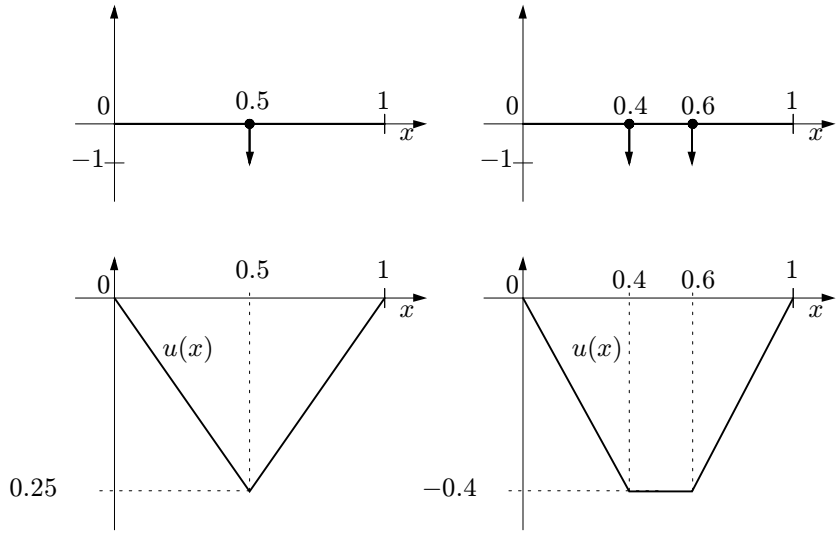


Fig. 3.1. We display on the left the equilibrium configuration of the string corresponding to the unit charge concentrated in $x = 0.5$, represented in the upper part of the figure. On the right we display the one corresponding to two unit charges concentrated in $x = 0.4$ and $x = 0.6$, also represented in the upper part of the figure

We apply the integration by parts formula to the first integral, with the purpose of eliminating the second derivative, in order to impose a lower regularity on the solution.

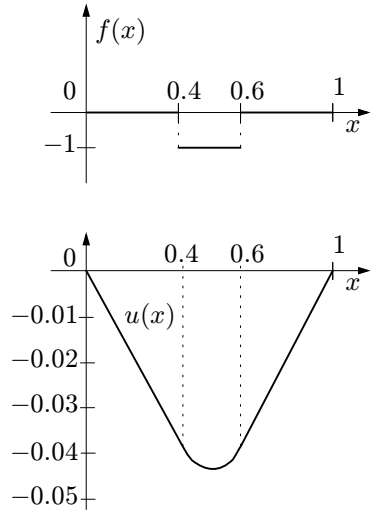


Fig. 3.2. Displacement relative to the discontinuous charge represented in the upper part of the figure

We find

$$-\int_0^1 u''v\ dx = \int_0^1 u'v'\ dx - [u'v]_0^1.$$

Since u is known at the boundary, we can consider only test functions which vanish at the extrema of the interval, hence the contribution of the boundary terms vanishes. In this way, the equation becomes

$$\int_0^1 u'v'\ dx = \int_0^1 fv\ dx. \tag{3.4}$$

The test function space V must therefore be such that if $v \in V$ then $v(0) = v(1) = 0$. Note that the solution u, being null at the boundary and having the same requirements of regularity as the test functions, will also be sought in the same space V.

It is now left to specify the regularity requirements which must be satisfied by the space V, so that all the operations introduced make sense. Evidently, if u and v belonged to $C^1([0,1])$, we would have $u', v' \in C^0([0,1])$ and therefore the integral appearing in the left-hand side of (3.4) would make sense. However, the examples in Fig. 3.1 tell us that the physical solutions might not be continuously differentiable: we must therefore require a lower regularity. Moreover, even when $f \in C^0([0,1])$, there is no certainty that the problem admits solutions in the space

$$V = \{v \in C^1([0,1]) : v(0) = v(1) = 0\}. \tag{3.5}$$

This may be attributed to the fact that such vector space, when provided with a scalar product

$$[u, v]_1 = \int_0^1 u'v'dx, \tag{3.6}$$

is not a complete space, that is, not all of the Cauchy sequences with values in V converge to an element of V. (Verify as an exercise that (3.6) is indeed a scalar product.)

Let us then proceed as follows. We recall the definition of the spaces L^p of the functions which are p-th power Lebesgue integrable. For $1 \le p < \infty$, these are defined as follows (see Sec. 2.5):

$$\mathrm{L}^p(0,1) = \{v : (0,1) \mapsto \mathbb{R} \text{ s.t. } \quad \|v\|_{\mathrm{L}^p(0,1)} = \left(\int_0^1 |v(x)|^p\ dx\right)^{1/p} < +\infty\}.$$

Since we want the integral $\int_0^1 u'v'\ dx$ to be well defined, the minimum requirement on u' and v' is that the product $u'v'$ lies in $\mathrm{L}^1(0,1)$. To this purpose, the following property holds:

Property 3.1 *Given two functions φ, $\psi : (0,1) \to \mathbb{R}$, if*

$$\varphi^2,\ \psi^2 \text{ are integrable then } \varphi\psi \text{ is integrable,}$$

that is, equivalently,

$$\varphi,\ \psi \in \mathrm{L}^2(0,1) \implies \varphi\psi \in \mathrm{L}^1(0,1).$$

This result is a direct consequence of the *Cauchy-Schwarz inequality*:

$$\left| \int_0^1 \varphi(x)\psi(x)\, dx \right| \leq \|\varphi\|_{\mathrm{L}^2(0,1)} \|\psi\|_{\mathrm{L}^2(0,1)}, \tag{3.7}$$

where

$$\|\varphi\|_{\mathrm{L}^2(0,1)} = \sqrt{\int_\Omega |\varphi(x)|^2 dx} \tag{3.8}$$

is the norm of φ in $\mathrm{L}^2(0,1)$. Since $\|\varphi\|_{\mathrm{L}^2(0,1)}, \|\psi\|_{\mathrm{L}^2(0,1)} < \infty$ by hypothesis, this proves that there also exists a (finite) integral of $\varphi\psi$.

In order for to the integrals appearing in (3.4) to make sense, functions, as well as their derivatives, must be square integrable. We therefore define the *Sobolev space*

$$\mathrm{H}^1(0,1) = \{v \in \mathrm{L}^2(0,1) : v' \in \mathrm{L}^2(0,1)\}.$$

The derivative must be interpreted in the sense of distributions (see Sec. 2.3). Let us hence choose as V the following subspace of $\mathrm{H}^1(0,1)$,

$$\mathrm{H}^1_0(0,1) = \{v \in \mathrm{H}^1(0,1) : v(0) = v(1) = 0\},$$

constituted by the functions of $\mathrm{H}^1(0,1)$ that are null at the extrema of the interval. If we suppose $f \in \mathrm{L}^2(0,1)$, the integral on the right-hand side of (3.4) also makes sense. Problem (3.3) is then reduced to the following integral problem,

$$\text{find } u \in V : \quad \int_0^1 u'v'\, dx = \int_0^1 fv\, dx \quad \forall v \in V, \tag{3.9}$$

with $V = \mathrm{H}^1_0(0,1)$.

Remark 3.1 In fact, the space $\mathrm{H}^1_0(0,1)$ is the closure, with respect to the scalar product (3.6), of the space defined in (3.5).
The functions of $\mathrm{H}^1(0,1)$ are not necessarily differentiable in a traditional sense, that is $\mathrm{H}^1(0,1) \not\subset C^1([0,1])$. For instance, functions that are piecewise continuous on a partition of the interval $(0,1)$ with derivatives that do not match at all endpoints of the partition belong to $\mathrm{H}^1(0,1)$ but not to $C^1([0,1])$. Hence, also continuous but not differentiable solutions of the previous examples are considered. •

The weak problem (3.9) turns out to be equivalent to a *variational problem*, thanks to the following result:

Theorem 3.1 *The problem*

$$\text{find } u \in V : \begin{cases} J(u) = \min_{v \in V} J(v) \quad \text{with} \\ J(v) = \dfrac{1}{2}\displaystyle\int_0^1 (v')^2 \, dx - \int_0^1 fv \, dx, \end{cases} \tag{3.10}$$

is equivalent to problem (3.9)*, in the sense that* u *is a solution of* (3.9) *if and only if* u *is a solution of* (3.10)*.*

Proof. Suppose that u is a solution of the variational problem (3.10). Then, setting $v = u + \delta w$, with $\delta \in \mathbb{R}$, we have that

$$J(u) \leq J(u + \delta w) \quad \forall w \in V.$$

The function $\psi(\delta) = J(u + \delta w)$ is a quadratic function in δ with minimum reached for $\delta = 0$. Thus,

$$\psi'(\delta)\Big|_{\delta=0} = \frac{\partial J(u + \delta w)}{\partial \delta}\Big|_{\delta=0} = 0.$$

From the definition of derivative we have

$$\frac{\partial J(u + \delta w)}{\partial \delta} = \lim_{\delta \to 0} \frac{J(u + \delta w) - J(u)}{\delta} \quad \forall w \in V.$$

Let us consider the term $J(u + \delta w)$:

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

Henceforth,

$$\frac{J(u + \delta w) - J(u)}{\delta} = \frac{1}{2}\int_0^1 [\delta w'^2 + 2u' w'] \, dx - \int_0^1 fw \, dx.$$

Passing to the limit for $\delta \to 0$ and imposing that it vanishes, we obtain

$$\int_0^1 u'w' \, dx - \int_0^1 fw \, dx = 0 \quad \forall w \in V,$$

that is, u satisfies the weak problem (3.9).

Conversely, if u is a solution of (3.9), by setting $v = \delta w$, we have in particular that

$$\int_0^1 u'\delta w' \, dx - \int_0^1 f\delta w \, dx = 0,$$

and therefore

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

Since

$$\frac{1}{2}\int_0^1 \delta^2 w'^2 \, dx \geq 0 \quad \forall \, w \in V, \; \forall \delta \in \mathbb{R},$$

we deduce that

$$J(u) \leq J(v) \quad \forall v \in V,$$

that is u also satisfies the variational problem (3.10). ◇

Remark 3.2 (Principle of virtual work) Let us consider again the problem of studying the configuration assumed by a unit tension string, fixed at the extrema and subject to a forcing term f, described by equation (3.3). We indicate with v an admissible displacement of the string (that is a null displacement at the extrema) from the equilibrium position u. Equation (3.9), expressing the equality between the work performed by the internal forces and by the external forces in correspondence to the displacement v, is nothing but the *principle of virtual work* of mechanics. Moreover, as in our case there exists a potential (indeed, $J(w)$ defined in (3.10) expresses the potential global energy corresponding to the configuration w of the system), the principle of virtual works establishes that any displacement allowed by the equilibrium configuration causes an increment of the system's potential energy. In this sense, Theorem 3.1 states that the weak solution is also the one minimizing the potential energy. •

3.2.2 Non-homogeneous Dirichlet problem

In the non-homogeneous case the boundary conditions in (3.3) are replaced by

$$u(0) = g_0, \quad u(1) = g_1,$$

g_0 and g_1 being two assigned values.

We can reconduct to the homogeneous case by noticing that if u is a solution of the non-homogeneous problem, then the function $\overset{\circ}{u} = u - [(1-x)g_0 + xg_1]$ is a solution of the corresponding homogeneous problem (3.3). The function $R_g = (1-x)g_0 + xg_1$ is said *lifting* (or *extension*, or *prolongation*) of the boundary data.

3.2.3 Neumann Problem

Let us now consider the following Neumann problem

$$\begin{cases} -u'' + \sigma u = f, & 0 < x < 1, \\ u'(0) = h_0, & u'(1) = h_1, \end{cases}$$

σ being a positive function and h_0, h_1 two real numbers. We observe that in the case where $\sigma = 0$ the solution of this problem would not be unique, being defined up to an additive constant. By applying the same procedure followed in the case of the Dirichlet problem, that is by multiplying the equation by a test function v, integrating on the interval $(0,1)$ and applying the formula of integration by parts, we get the equation

$$\int_0^1 u'v'\,dx + \int_0^1 \sigma uv\,dx - [u'v]_0^1 = \int_0^1 fv\,dx.$$

Let us suppose $f \in \mathrm{L}^2(0,1)$ and $\sigma \in \mathrm{L}^\infty(0,1)$ that is that σ be a bounded function almost everywhere (a.e.) on $(0,1)$ (see (2.14)). The boundary term is known thanks to the Neumann conditions. On the other hand, the unknown u is not known at the boundary in this case, hence it must not be required that v be null at the boundary. The weak formulation of the Neumann problem is therefore: *find* $u \in \mathrm{H}^1(0,1)$ *such that*

$$\int_0^1 u'v'\,dx + \int_0^1 \sigma uv\,dx = \int_0^1 fv\,dx + h_1 v(1) - h_0 v(0) \quad \forall v \in \mathrm{H}^1(0,1). \tag{3.11}$$

In the homogeneous case $h_0 = h_1 = 0$, the weak problem is characterized by the same equation as the Dirichlet case, but the space V of test functions is now $\mathrm{H}^1(0,1)$ instead of $\mathrm{H}_0^1(0,1)$.

3.2.4 Mixed homogeneous problem

Analogous considerations hold for the mixed homogeneous problem, that is when we have a homogeneous Dirichlet condition in one extreme and a homogeneous Neumann condition in the other,

$$\begin{cases} -u'' + \sigma u = f, & 0 < x < 1, \\ u(0) = 0, & u'(1) = 0. \end{cases} \tag{3.12}$$

In such case it must be required that the test functions be null in $x = 0$. Setting $\Gamma_D = \{0\}$ and defining

$$\mathrm{H}^1_{\Gamma_D}(0,1) = \{v \in \mathrm{H}^1(0,1) : \ v(0) = 0\},$$

the weak formulation of problem (3.12) is: *find* $u \in \mathrm{H}^1_{\Gamma_D}(0,1)$ *such that*

$$\int_0^1 u'v' \, dx + \int_0^1 \sigma uv \, dx = \int_0^1 fv \, dx \quad \forall v \in \mathrm{H}^1_{\Gamma_D}(0,1),$$

with $f \in \mathrm{L}^2(0,1)$ and $\sigma \in \mathrm{L}^\infty(0,1)$. The formulation is once again the same as in the homogeneous Dirichlet problem, however the space where to find the solution changes.

3.2.5 Mixed (or Robin) boundary conditions

Finally, consider the following problem

$$\begin{cases} -u'' + \sigma u = f, & 0 < x < 1, \\ u(0) = 0, & u'(1) + \gamma u(1) = r, \end{cases}$$

where $\gamma > 0$ and r are two assigned constants.
Also in this case, we will use test functions that are null at $x = 0$, the value of u being known thereby. As opposed to the Neumann case, the boundary term for $x = 1$, deriving from the integration by parts, no longer provides a known quantity, but a term proportional to the unknown u. As a matter of fact, we have

$$-[u'v]_0^1 = -rv(1) + \gamma u(1)v(1).$$

The weak formulation is therefore: *find* $u \in \mathrm{H}^1_{\Gamma_D}(0,1)$ *such that*

$$\int_0^1 u'v' \, dx + \int_0^1 \sigma uv \, dx + \gamma u(1)v(1) = \int_0^1 fv \, dx + rv(1) \quad \forall v \in \mathrm{H}^1_{\Gamma_D}(0,1).$$

A boundary condition that is a linear combination between the value of u and the value of its first derivative is called *Robin* (or *Newton*, or *third type*) *condition*.

3.3 The Poisson problem in the two-dimensional case

In this section, we consider the problems at the limits associated to the Poisson equation in the two-dimensional case.

3.3.1 The homogeneous Dirichlet problem

The problem consists in finding u such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.13}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with boundary $\partial\Omega$. We proceed in a similar way as for the one-dimensional case. By multiplying the differential equation in (3.13) by an arbitrary function v and integrating on Ω, we find

$$-\int_\Omega \Delta u v \, d\Omega = \int_\Omega f v \, d\Omega.$$

At this point, it is necessary to apply the multi-dimensional analogous of the one-dimensional formula of integration by parts. This can be obtained by applying the divergence (Gauss) theorem by which

$$\int_\Omega \mathrm{div}(\mathbf{a}) \, d\Omega = \int_{\partial\Omega} \mathbf{a} \cdot \mathbf{n} \, d\gamma, \tag{3.14}$$

$\mathbf{a}(\mathbf{x}) = (a_1(\mathbf{x}), a_2(\mathbf{x}))^T$ being a sufficiently regular vector function and $\mathbf{n}(\mathbf{x}) = (n_1(\mathbf{x}), n_2(\mathbf{x}))^T$ the outward unit normal vector on $\partial\Omega$. If we apply (3.14) first to the function $\mathbf{a} = (\varphi\psi, 0)^T$ and then to $\mathbf{a} = (0, \varphi\psi)^T$, we get the relations

$$\int_\Omega \frac{\partial \varphi}{\partial x_i} \psi \, d\Omega = -\int_\Omega \varphi \frac{\partial \psi}{\partial x_i} \, d\Omega + \int_{\partial\Omega} \varphi\psi n_i \, d\gamma, \quad i = 1, 2. \tag{3.15}$$

Note also that if we take $\mathbf{a} = \mathbf{b}\varphi$, where $\mathbf{b}$ and φ are respectively a vector and a scalar field, then (3.14) yields

$$\int_\Omega \varphi \mathrm{div}\mathbf{b} \, d\Omega = -\int_\Omega \mathbf{b} \cdot \nabla\varphi \, d\Omega + \int_{\partial\Omega} \mathbf{b} \cdot \mathbf{n}\varphi \, d\gamma \tag{3.16}$$

which is called *Green formula* for the divergence operator.

We exploit (3.15) by keeping into account the fact that $\Delta u = \mathrm{div}\nabla u = \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\Big(\frac{\partial u}{\partial x_i}\Big)$. Supposing that all the integrals that appear are meaningful, we find

$$\begin{aligned} -\int_\Omega \Delta u v \, d\Omega &= -\sum_{i=1}^{2} \int_\Omega \frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) v \, d\Omega \\ &= \sum_{i=1}^{2} \int_\Omega \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} \, d\Omega - \sum_{i=1}^{2} \int_{\partial\Omega} \frac{\partial u}{\partial x_i} v n_i \, d\gamma \\ &= \int_\Omega \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} \, d\Omega - \int_{\partial\Omega} \left(\sum_{i=1}^{2} \frac{\partial u}{\partial x_i} n_i\right) v \, d\gamma. \end{aligned}$$

We obtain the following relation, called *Green formula* for the Laplacian

$$-\int_\Omega \Delta u v \, d\Omega = \int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma. \tag{3.17}$$

Similarly to the one-dimensional case, the homogeneous Dirichlet problem will lead us to choosing test functions that vanish at the boundary, and, consequently, the boundary term that appears in (3.17) will in turn vanish.

Taking this into account, we get the following weak formulation for problem (3.13)

$$\text{find u } \in \mathrm{H}_0^1(\Omega): \quad \int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \quad \forall v \in \mathrm{H}_0^1(\Omega), \tag{3.18}$$

f being a fuction of $\mathrm{L}^2(\Omega)$ and having set

$$\begin{aligned} &\mathrm{H}^1(\Omega) = \{v : \Omega \to \mathbb{R} \text{ s.t. } v \in \mathrm{L}^2(\Omega), \frac{\partial v}{\partial x_i} \in \mathrm{L}^2(\Omega),\ i = 1, 2\}, \\ &\mathrm{H}_0^1(\Omega) = \{v \in \mathrm{H}^1(\Omega) :\ v = 0 \text{ on } \partial\Omega\}. \end{aligned}$$

The derivatives must be understood in the sense of distributions and the condition $v = 0$ on $\partial\Omega$ in the sense of the traces (see Chap. 2).
In particular, we observe that if $u, v \in \mathrm{H}_0^1(\Omega)$, then $\nabla u, \nabla v \in [\mathrm{L}^2(\Omega)]^2$ and therefore $\nabla u \cdot \nabla v \in \mathrm{L}^1(\Omega)$. The latter property is obtained by applying the following inequality

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

a direct consequence of the *Cauchy-Schwarz inequality* (2.16).

Hence, the integral appearing at the left of (3.18) is perfectly meaningful and so is the one appearing at the right.

Similarly to the one-dimensional case, it can be shown also in the two-dimensional case that problem (3.18) is equivalent to the following *variational problem*

$$\text{find } u \in V : \begin{cases} J(u) = \inf_{v \in V} J(v), \text{ with} \\ J(v) = \dfrac{1}{2}\displaystyle\int_{\Omega} |\nabla v|^2 \, d\Omega - \int_{\Omega} fv \, d\Omega, \end{cases}$$

having set $V = \mathrm{H}_0^1(\Omega)$.

We can rewrite the weak formulation (3.18) in a more compact way by introducing the following form

$$a : V \times V \to \mathbb{R}, \quad a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, d\Omega, \tag{3.19}$$

and the following functional

$$F : V \to \mathbb{R}, \quad F(v) = \int_{\Omega} fv \, d\Omega$$

(functionals and forms are introduced in Chap. 2).
Problem (3.18) therefore becomes:

$$\text{find } u \in V : \quad a(u, v) = F(v) \quad \forall v \in V.$$

We notice that $a(\cdot, \cdot)$ is a bilinear form (that is, linear with respect to both its arguments), while F is a linear functional. Then

$$|F(v)| \leq \|f\|_{\mathrm{L}^2(\Omega)} \|v\|_{\mathrm{L}^2(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)} \|v\|_{\mathrm{H}^1(\Omega)}.$$

Consequently, F is also bounded. Following the definition (2.2), its norm is bounded by $\|F\|_{V'} \leq \|f\|_{\mathrm{L}^2(\Omega)}$. Consequently, F belongs to V', the dual space of V, that is the set of linear and continuous functionals defined on V (see Sec. 2.1).

3.3.2 Equivalence, in the sense of distributions, between weak and strong form of the Dirichlet problem

We want to prove that the equations of problem (3.13) are actually satisfied by the weak solution, albeit only in the sense of distributions.
To this end, we consider the weak formulation (3.18). Let $\mathcal{D}(\Omega)$ now be the space of functions that are infinitely differentiable and with compact support in Ω (see Chap. 2). We recall that $\mathcal{D}(\Omega) \subset \mathrm{H}_0^1(\Omega)$. Hence, by choosing $v = \varphi \in \mathcal{D}(\Omega)$ in (3.18), we have

$$\int_{\Omega} \nabla u \cdot \nabla \varphi \, d\Omega = \int_{\Omega} f\varphi \, d\Omega \quad \forall \varphi \in \mathcal{D}(\Omega). \tag{3.20}$$

By applying Green's formula (3.17) to the left-hand side of (3.20), we find

$$-\int_\Omega \Delta u\varphi \, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n}\varphi \, d\gamma = \int_\Omega f\varphi \, d\Omega \quad \forall \varphi \in \mathcal{D}(\Omega),$$

where the integrals are to be understood in the sense of duality, that is:

$$\begin{aligned} -\int_\Omega \Delta u\varphi \, d\Omega &= {}_{\mathcal{D}'(\Omega)}\langle -\Delta u, \varphi\rangle_{\mathcal{D}(\Omega)}, \\ \int_{\partial\Omega} \frac{\partial u}{\partial n}\varphi \, d\gamma &= {}_{\mathcal{D}'(\partial\Omega)}\langle \frac{\partial u}{\partial n}, \varphi\rangle_{\mathcal{D}(\partial\Omega)}. \end{aligned}$$

Since $\varphi \in \mathcal{D}(\Omega)$, the boundary integral is null, so that

$${}_{\mathcal{D}'(\Omega)}\langle -\Delta u - f, \varphi\rangle_{\mathcal{D}(\Omega)} = 0 \quad \forall \varphi \in \mathcal{D}(\Omega),$$

which corresponds to saying that $-\Delta u - f$ is the null distribution, that is

$$-\Delta u = f \quad \text{in } \mathcal{D}'(\Omega).$$

The differential equation (3.13) is therefore verified, as long as we intend the derivatives in the sense of distributions and we interpret the equality between $-\Delta u$ and f not in a pointwise sense, but in the sense of distributions (and thus almost everywhere in Ω). Finally, the fact that u vanishes on the boundary (in the sense of traces) is a direct consequence of u being in $\mathrm{H}^1_0(\Omega)$.

3.3.3 The problem with mixed, non homogeneous conditions

The problem we want to solve is now the following

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial n} = \phi & \text{on } \Gamma_N, \end{cases} \tag{3.21}$$

where Γ_D and Γ_N yield a partition of $\partial\Omega$, that is $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\mathring{\Gamma}_D \cap \mathring{\Gamma}_N = \emptyset$ (see Fig. 3.3).

In the case of the Neumann problem, where $\Gamma_D = \emptyset$, the data f and ϕ must verify the following *compatibility condition*

$$-\int_{\partial\Omega} \phi \, d\gamma = \int_\Omega f \, d\Omega \tag{3.22}$$

so that the problem can have a solution. Condition (3.22) is deduced by integrating the differential equation in (3.21) and applying the divergence theorem (3.14)

$$-\int_\Omega \Delta u \, d\Omega = -\int_\Omega \mathrm{div}(\nabla u) \, d\Omega = -\int_{\partial\Omega} \frac{\partial u}{\partial n} \, d\gamma.$$

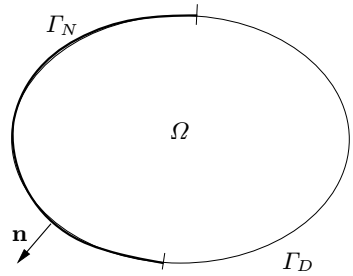


Fig. 3.3. The computational domain Ω

Moreover, we observe that also in the case of the Neumann problem, the solution is defined only up to an additive constant. In order to have uniqueness it would be sufficient, for example, to find a function with null average in Ω.

Let us now suppose that $\Gamma_D \neq \emptyset$ in order to ensure the uniqueness of the solution to the strong problem without conditions of compatibility on the data. Let us also suppose that $f \in \mathrm{L}^2(\Omega)$, $g \in \mathrm{H}^{1/2}(\Gamma_D)$ and $\phi \in \mathrm{L}^2(\Gamma_N)$, having denoted by $\mathrm{H}^{1/2}(\Gamma_D)$ the space of functions of $\mathrm{L}^2(\Gamma_D)$ that are traces of functions of $\mathrm{H}^1(\Omega)$ (see Sec. 2.4.3).

Thanks to Green's formula (3.17), we obtain from (3.21)

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma = \int_\Omega f v \, d\Omega. \tag{3.23}$$

We recall that $\partial u/\partial n = \phi$ on Γ_N and by exploiting the additivity of integrals, (3.23) becomes

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_{\Gamma_D} \frac{\partial u}{\partial n} v \, d\gamma - \int_{\Gamma_N} \phi \, v \, d\gamma = \int_\Omega f v \, d\Omega. \tag{3.24}$$

By imposing that the test function v vanish on Γ_D, the first boundary integral appearing in (3.24) vanishes. The mixed problem therefore admits the following weak formulation

$$\text{find } u \in V_g : \quad \int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\Gamma_N} \phi \, v \, d\gamma \quad \forall v \in V, \tag{3.25}$$

having denoted by V the space

$$V = \mathrm{H}^1_{\Gamma_D}(\Omega) = \{v \in \mathrm{H}^1(\Omega) : \, v|_{\Gamma_D} = 0\}, \tag{3.26}$$

and having set

$$V_g = \{v \in \mathrm{H}^1(\Omega) : v|_{\Gamma_D} = g\}.$$

The formulation (3.25) is not satisfactory, not only because the choice of spaces is "asymmetrical" ($v \in V$, while $u \in V_g$), but mainly because V_g is an affine manifold, but not a subspace of $\mathrm{H}^1(\Omega)$ (indeed, it is not true that linear combinations of elements of V_g are still elements of V_g).

We then proceed in a similar way as seen in Sec. 3.2.2. We suppose to know a function R_g, called *lifting of the boundary data*, such that

$$R_g \in \mathrm{H}^1(\Omega), \quad R_g|_{\Gamma_D} = g.$$

Furthermore, we suppose that such lifting be continuous, i.e. that

$$\exists C > 0 : \|R_g\|_{\mathrm{H}^1(\Omega)} \leq C\|g\|_{\mathrm{H}^{1/2}(\Gamma_D)} \forall g \in \mathrm{H}^{1/2}(\Gamma_D).$$

We set $\mathring{u} = u - R_g$ and we begin by observing that $\mathring{u}|_{\Gamma_D} = u|_{\Gamma_D} - R_g|_{\Gamma_D}$=0, that is $\mathring{u} \in \mathrm{H}^1_{\Gamma_D}(\Omega)$. Moreover, since $\nabla u = \nabla \mathring{u} + \nabla R_g$, problem (3.25) becomes

$$\text{find } \mathring{u} \in \mathrm{H}^1_{\Gamma_D}(\Omega) : \quad a(\mathring{u}, v) = F(v) \quad \forall v \in \mathrm{H}^1_{\Gamma_D}(\Omega), \tag{3.27}$$

having defined the bilinear form $a(\cdot,\cdot)$ as in (3.19), while the linear functional F now takes the form

$$F(v) = \int_\Omega fv \; d\Omega + \int_{\Gamma_N} \phi \; v \; d\gamma - \int_\Omega \nabla R_g \cdot \nabla v \; d\Omega.$$

The problem is now symmetric since the space where the (new) unknown solution is sought coincides with the test function space.

The Dirichlet conditions are said to be *essential* as they are imposed explicitly in the functional space in which the problem is set.

The Neumann conditions are instead said to be *natural*, as they are satisfied implicitly by the solution of the problem (to this end, see Sec. 3.3.4). This difference in treatment has important repercussions on the approximate problems.

Remark 3.3 The reduction of the problem to a "symmetric" form allows to obtain a linear system with a symmetric matrix when solving the problem numerically (for instance via the finite elements method). •

Remark 3.4 Building a lifting R_g of a boundary function with an arbitrary form can turn out to be problematic. Such task is simpler in the context of a numerical approximation, where one generally builds a lifting of an approximation of the g function (see Chap. 4). •

3.3.4 Equivalence, in the sense of distributions, between weak and strong form of the Neumann problem

Let us consider the non homogeneous Neumann problem

$$\begin{cases} -\Delta u + \sigma u = f & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = \phi & \text{on } \partial\Omega, \end{cases} \tag{3.28}$$

where σ is a positive constant or, more generally, a function $\sigma \in \mathrm{L}^\infty(\Omega)$ such that $\sigma(\mathbf{x}) \geq \alpha_0$ a.e. in Ω, for a well-chosen constant $\alpha_0 > 0$. Let us also suppose that $f \in \mathrm{L}^2(\Omega)$ and that $\phi \in \mathrm{L}^2(\partial\Omega)$. By proceeding as in Sec. 3.3.3, the following weak formulation can be derived

$$\text{find } u \in \mathrm{H}^1(\Omega) :$$
$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \int_\Omega \sigma u v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\partial\Omega} \phi v \, d\gamma \qquad \forall v \in \mathrm{H}^1(\Omega). \tag{3.29}$$

By taking $v = \varphi \in \mathcal{D}(\Omega)$ and counterintegrating by parts, we obtain

$$_{\mathcal{D}'(\Omega)}\langle -\Delta u + \sigma u - f, \varphi \rangle_{\mathcal{D}(\Omega)} = 0 \quad \forall \varphi \in \mathcal{D}(\Omega).$$

Hence

$$-\Delta u + \sigma u = f \quad \text{in } \mathcal{D}'(\Omega)$$

i.e.

$$-\Delta u + \sigma u - f = 0 \quad \text{a.e. in } \Omega. \tag{3.30}$$

In the case where $u \in C^2(\Omega)$ the application of Green's formula (3.17) in (3.29) leads to

$$\int_\Omega (-\Delta u + \sigma u - f) v \, d\Omega + \int_{\partial\Omega} (\frac{\partial u}{\partial n} - \phi) v = 0 \quad \forall v \in \mathrm{H}^1(\Omega),$$

and therefore, thanks to (3.30),

$$\frac{\partial u}{\partial n} = \phi \quad \text{on } \partial\Omega.$$

In the case where the solution u of (3.29) is only in $\mathrm{H}^1(\Omega)$ the generalized Green formula can be used, which states that there exists a unique linear and continuous functional $g \in (\mathrm{H}^{1/2}(\partial\Omega))'$ (called generalized normal derivative), which operates on the space $\mathrm{H}^{1/2}(\partial\Omega)$ satisfying

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \langle -\Delta u, v \rangle + \ll g, v \gg \quad \forall v \in \mathrm{H}^1(\Omega).$$

We have denoted by $< \cdot, \cdot >$ the duality between $\mathrm{H}^1(\Omega)$ and its dual, and by $\ll \cdot, \cdot \gg$ the duality between $\mathrm{H}^{1/2}(\partial\Omega)$ and its dual. Clearly g coincides with the classical normal derivative of u if u has sufficient regularity. For the sake of simplicity we use the notation $\partial u/\partial n$ for the generalized normal derivative in the remainder of this chapter. We therefore obtain that for $v \in \mathrm{H}^1(\Omega)$

$$\langle -\Delta u + \sigma u - f, v \rangle + \ll \partial u/\partial n - \phi, v \gg = 0;$$

thanks to (3.30), we finally conclude that

$$\ll \partial u/\partial n - \phi, v \gg = 0 \quad \forall v \in \mathrm{H}^1(\Omega),$$

and thus that $\partial u/\partial n = \phi$ a.e. on $\partial\Omega$.

3.4 More general elliptic problems

Let us now consider the problem

$$\begin{cases} -\mathrm{div}(\mu\nabla u) + \sigma u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_D, \\ \mu\dfrac{\partial u}{\partial n} = \phi & \text{on } \Gamma_N, \end{cases} \tag{3.31}$$

where $\Gamma_D \cup \Gamma_N = \partial\Omega$ with $\overset{\circ}{\Gamma}_D \cap \overset{\circ}{\Gamma}_N = \emptyset$. We will suppose that $f \in \mathrm{L}^2(\Omega)$, $\mu, \sigma \in \mathrm{L}^\infty(\Omega)$. Furthermore, we suppose that $\exists \mu_0 > 0$ such that $\mu(\mathbf{x}) \geq \mu_0$ and $\sigma(\mathbf{x}) \geq 0$ a.e. in Ω. Only in the case where $\sigma = 0$ we will require that Γ_D be non-empty in order to prevent the solution from losing uniqueness. Finally, we will suppose that g and ϕ are sufficiently regular functions on $\partial\Omega$, for instance $g \in \mathrm{H}^{1/2}(\Gamma_D)$ and $\phi \in \mathrm{L}^2(\Gamma_N)$.

Also in this case, we proceed by multiplying the equation by a test function v and by integrating (once again formally) on the domain Ω:

$$\int_\Omega [-\mathrm{div}(\mu\nabla u) + \sigma u]\, v \, d\Omega = \int_\Omega f v \, d\Omega.$$

By applying Green's formula we obtain

$$\int_\Omega \mu\nabla u \cdot \nabla v \, d\Omega + \int_\Omega \sigma\, u\, v \, d\Omega - \int_{\partial\Omega} \mu\frac{\partial u}{\partial n} v \, d\gamma = \int_\Omega f v \, d\Omega,$$

which can also be rewritten as

$$\int_\Omega \mu\nabla u \cdot \nabla v \, d\Omega + \int_\Omega \sigma\, u\, v \, d\Omega - \int_{\Gamma_D} \mu\frac{\partial u}{\partial n} v \, d\gamma = \int_\Omega f v \, d\Omega + \int_{\Gamma_N} \mu\frac{\partial u}{\partial n} v \, d\gamma.$$

The function $\mu\partial u/\partial n$ is called *conormal derivative* of u associated to the operator $-\mathrm{div}(\mu\nabla u)$. On Γ_D we impose that the test function v is null, while on Γ_N we impose that the conormal derivative is equal to ϕ. We obtain

$$\int_\Omega \mu\nabla u \cdot \nabla v \, d\Omega + \int_\Omega \sigma\, u\, v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\Gamma_N} \phi v \, d\gamma.$$

Having denoted by R_g a lifting of g, we set $\overset{\circ}{u} = u - R_g$. The weak formulation of problem (3.31) is therefore

$$\begin{aligned} &\text{find } \overset{\circ}{u} \in \mathrm{H}^1_{\Gamma_D}(\Omega) : \\ &\int_\Omega \mu\nabla\overset{\circ}{u} \cdot \nabla v \, d\Omega + \int_\Omega \sigma\overset{\circ}{u} v \, d\Omega = \int_\Omega f v \, d\Omega \\ &-\int_\Omega \mu\nabla R_g \cdot \nabla v \, d\Omega - \int_\Omega \sigma R_g v \, d\Omega + \int_{\Gamma_N} \phi v \, d\gamma \qquad \forall v \in \mathrm{H}^1_{\Gamma_D}(\Omega). \end{aligned}$$

We define the bilinear form

$$a : V \times V \to \mathbb{R}, \quad a(u,v) = \int_\Omega \mu \nabla u \cdot \nabla v \, d\Omega + \int_\Omega \sigma \, u \, v \, d\Omega,$$

and the linear and continuous functional

$$F : V \to \mathbb{R}, \;\; F(v) = -a(R_g, v) + \int_\Omega f v \, d\Omega + \int_{\Gamma_N} \phi v \, d\gamma. \tag{3.32}$$

The previous problem can then be rewritten as

$$\text{find } \overset{\circ}{u} \in \mathrm{H}^1_{\Gamma_D}(\Omega) : \quad a(\overset{\circ}{u}, v) = F(v) \quad \forall v \in \mathrm{H}^1_{\Gamma_D}(\Omega). \tag{3.33}$$

A yet more general problem than (3.31) is the following

$$\begin{cases} Lu = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial n_L} = \phi & \text{on } \Gamma_N, \end{cases}$$

where, as usual, $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\overset{\circ}{\Gamma}_D \cap \overset{\circ}{\Gamma}_N = \emptyset$, and having defined

$$Lu = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) + \sigma u.$$

The a_{ij} coefficients are functions defined on Ω. The derivative

$$\frac{\partial u}{\partial n_L} = \sum_{i,j=1}^{2} a_{ij} \frac{\partial u}{\partial x_j} n_i \tag{3.34}$$

is called *conormal derivative* of u associated to the operator L (it coincides with the normal derivative when $Lu = -\Delta u$).
Let us suppose that $\sigma(\mathbf{x}) \in \mathrm{L}^\infty(\Omega)$ and that $\exists \alpha_0 > 0$ such that $\sigma(\mathbf{x}) \geq \alpha_0$ a.e. in Ω. Furthermore, let us suppose that the coefficients $a_{ij} : \bar{\Omega} \to \mathbb{R}$ are continuous functions $\forall i, j = 1, 2$ and that there exists a positive constant α such that

$$\forall \boldsymbol{\xi} = (\xi_1, \xi_2)^T \in \mathbb{R}^2 \quad \sum_{i,j=1}^{2} a_{ij}(\mathbf{x}) \xi_i \xi_j \geq \alpha \sum_{i=1}^{2} \xi_i^2 \quad \text{a.e. in } \Omega. \tag{3.35}$$

In such case, the weak formulation is still the same as (3.33), the functional F is still the one introduced in (3.32), while

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^{2} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sigma u v \right) d\Omega. \tag{3.36}$$

It can be shown (see Exercise 2) that under the ellipticity hypothesis on the coefficients (3.35), this bilinear form is continuous and coercive, in the sense of definitions (2.6) and (2.9). These properties will be exploited in the analysis of well-posedness of problem (3.33) (see Sec. 3.5).
Elliptic problems for fourth-order operators are proposed in Exercises 4 and 6, while an elliptic problem deriving from the linear elasticity theory is analyzed in Exercise 7.

Remark 3.5 (Robin conditions) The case where Robin boundary conditions are enforced on the whole boundary, say

$$\mu\frac{\partial u}{\partial n} + \gamma u \;=\; 0 \quad \text{on } \partial\Omega,$$

requires more care. The weak form of the problem reads

$$\text{find } u \in H^1(\Omega) \;:\; a(u,v) = \int_\Omega f v d\Omega \quad \forall v \in H^1(\Omega),$$

where the bilinear form $a(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v d\Omega + \int_\Omega \gamma uv d\Omega$ this time is not coercive if $\gamma < 0$. The analysis of this problem can be carried out by means of the Peetre-Tartar lemma, see [EG04].

•

3.5 Existence and uniqueness theorem

The following fundamental result holds (refer to Sec. 2.1 for definitions):

Lemma 3.1 (Lax-Milgram) *Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ a continuous and coercive bilinear form, $F(\cdot) : V \to \mathbb{R}$ a linear and continuous functional. Then, there exists one unique solution to the problem*

$$\text{find } u \in V : \quad a(u,v) = F(v) \quad \forall v \in V. \tag{3.37}$$

Proof. This is based on two classical results of Functional Analysis: the Riesz representation theorem (see Theorem 2.1, Chap. 2), and the Banach closed range theorem. The interested reader can refer to, e.g., [QV94, Chap. 5]. ◇

The Lax-Milgram Lemma thus ensures that the weak formulation of an elliptic problem is well posed, as long as the hypotheses on the form $a(\cdot,\cdot)$ and on the functional $F(\cdot)$ are verified. Several consequences derive from this Lemma. We report one of the most important in the following Corollary.

Corollary 3.1 *The solution of* (3.37) *is bounded by the data, that is*

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{V'},$$

where α *is the coercivity constant of the bilinear form* $a(\cdot,\cdot)$*, while* $\|F\|_{V'}$ *is the norm of the functional* F*, see (2.2).*

Proof. It is sufficient to choose $v = u$ in (3.37) and then to use the coercivity of the bilinear form $a(\cdot,\cdot)$. Indeed, we have

$$\alpha\|u\|_V^2 \leq a(u,u) = F(u).$$

On the other hand, since F is linear and continuous, it is also bounded and the upper bound

$$|F(u)| \leq \|F\|_{V'}\|u\|_V$$

holds, hence the thesis follows. ◇

Remark 3.6 If the bilinear form $a(\cdot,\cdot)$ is also *symmetric*, that is

$$a(u,v) = a(v,u) \quad \forall u, v \in V,$$

then (3.37) is equivalent to the following variational problem (see Exercise 1)

$$\begin{cases} \text{find } u \in V: \quad J(u) = \min_{v\in V} J(v), \\ \text{with } J(v) = \dfrac{1}{2}a(v,v) - F(v). \end{cases} \tag{3.38}$$

•

3.6 Adjoint operator and adjoint problem

In this section we will introduce the concept of *adjoint* of a given operator in Hilbert spaces, as well as the *adjoint* (or *dual*) problem of a given boundary value problem. Then we will show how to obtain dual problems, with associated boundary conditions. The adjoint problem of a given differential problem plays a fundamental role, for instance, in the context of derivations of error estimates for Galerkin methods, both a priori and a posteriori (see Sections 4.5.4 and 4.6.4-4.6.5, respectively), but also for the solution of optimal control problems, as we will see in Chap. 16.

Let V be a Hilbert space with scalar product $(\cdot,\cdot)_V$ and norm $\|\cdot\|_V$, and let V' be its dual space. Let $a: V \times V \to \mathbb{R}$ be a continuous and coercive bilinear form and let $A: V \to V'$ be its associated elliptic operator, that is $A \in \mathcal{L}(V, V')$,

$$_{V'}\langle Av, w\rangle_V \; = \; a(v,w) \quad \forall v, w \in V. \tag{3.39}$$

Let $a^* : V \times V \to \mathbb{R}$ be the bilinear form defined by

$$a^*(w, v) = a(v, w) \quad \forall v, w \in V, \tag{3.40}$$

and consider the operator $A^* : V \to V'$ associated to the form $a^*(\cdot, \cdot)$, that is

$$_{V'}\langle A^* w, v\rangle_V = a^*(w, v) \quad \forall v, w \in V. \tag{3.41}$$

Thanks to (3.40) we have the following relation, known as the *Lagrange identity*

$$_{V'}\langle A^* w, v\rangle_V = {}_{V'}\langle Av, w\rangle_V \quad \forall v, w \in V. \tag{3.42}$$

Note that this is precisely the equation that stands at the base of the definition (2.19) of the adjoint of a given operator A acting between a Hilbert space and its dual. For coherence with (2.19), we should have noted this operator A'. However, we prefer to denote it A^* because the latter notation is more customarily used in the context of elliptic boundary value problems.

If $a(\cdot, \cdot)$ is a symmetric form, $a^*(\cdot, \cdot)$ coincides with $a(\cdot, \cdot)$ and A^* with A. In such case A is said to be *self-adjoint*; A is said to be *normal* if $AA^* = A^*A$.
Naturally, the identity operator I is self-adjoint ($I = I^*$), while if an operator is self-adjoint, then it is also normal.
Some properties of the adjoint operators which are a consequence of the previous definition, are listed below:

- A being linear and continuous, then also A^* is, that is $A^* \in \mathcal{L}(V, V')$;
- $\|A^*\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} = \|A\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')}$ (these norms are defined in (2.20));
- $(A + B)^* = A^* + B^*$;
- $(AB)^* = B^* A^*$;
- $(A^*)^* = A$;
- $(A^{-1})^* = (A^*)^{-1}$ (if A is invertible);
- $(\alpha A)^* = \alpha A^* \qquad \forall \alpha \in \mathbb{R}$.

When we need to find the adjoint (or dual) problem of a given (primal) problem, we will use the Lagrange identity to characterize the differential equation of the dual problem, as well as its boundary conditions.

We provide an example of such a procedure, starting from a simple one-dimensional diffusion transport equation, completed by homogeneous Robin-Dirichlet boundary conditions

$$\begin{cases} Av = -v'' + v' = f, & x \in I = (0, 1), \\ v'(0) + \beta v(0) = 0, \qquad v(1) = 0, \end{cases} \tag{3.43}$$

assuming β constant. Note that the weak form of this problem is

$$\text{find } u \in V \text{ s.t. } a(u, v) = \int_0^1 f v dx \quad \forall v \in V, \tag{3.44}$$

being $V = \{v \in H^1(0,1) : v(1) = 0\}$ and

$$a : V \times V \to \mathbb{R}, \quad a(u,v) = \int_0^1 (u' - u)v'dx - (\beta + 1)u(0)v(0).$$

Thanks to (3.40) we obtain, $\forall v, w \in V$,

$$\begin{aligned} a^*(w,v) = a(v,w) &= \int_0^1 (v' - v)w'dx - (\beta + 1)v(0)w(0) \\ &= -\int_0^1 v(w'' + w')dx + [vw']_0^1 - (\beta + 1)v(0)w(0) \\ &= \int_0^1 (-w'' - w')vdx - [w'(0) + (\beta + 1)w(0)]v(0). \end{aligned}$$

Since definition (3.41) must hold, we will have

$$A^*w = -w'' - w' \quad in \ \mathcal{D}'(0,1).$$

Moreover, as $v(0)$ is arbitrary, w will need to satisfy the boundary conditions

$$[w' + (\beta + 1)w](0) = 0, \quad w(1) = 0.$$

We observe that the transport field of the dual problem has an opposite direction with respect to that of the primal problem. Moreover, to homogeneous Robin-Dirichlet boundary conditions for the primal problem (3.43) correspond conditions of exactly the same nature for the dual problem.

The procedure illustrated for problem (3.43) can clearly be extended to the multidimensional case. In Table 3.3 we provide a list of several differential operators with boundary conditions, and their corresponding adjoint operators with associated boundary conditions. (On the functions appearing in the table, assume all the necessary regularity for the considered differential operators to be well-defined). We note, in particular, that to a given type of primal conditions do not necessarily correspond dual conditions of the same type, and that, for an operator that is not self-adjoint, to a conservative (resp. non-conservative) formulation of the primal problem there does correspond a non-conservative (resp. conservative) formulation of the dual one.

The extension of the analysis in the previous section to the non-linear case is not so immediate. For simplicity, we consider the one-dimensional problem

$$\begin{cases} A(v)v = -v'' + vv' = f, & x \in I = (0,1), \\ v(0) = v(1) = 0, \end{cases} \tag{3.45}$$

having denoted by $A(v)$ the operator

$$A(v)\cdot = -\frac{d^2\cdot}{dx^2} + v\frac{d\cdot}{dx}. \tag{3.46}$$

Table 3.3. Differential operators and boundary conditions (B.C.) for the primal problem and corresponding dual (adjoint) operators (with associated boundary conditions)

Primal operator	*Primal B.C.*	*Dual (adjoint) operator*	*Dual B.C.*
$-\Delta u$	$u = 0$ on Γ $\dfrac{\partial u}{\partial n} = 0$ on $\partial\Omega\backslash\Gamma$	$-\Delta w$	$w = 0$ on Γ, $\dfrac{\partial w}{\partial n} = 0$ on $\partial\Omega\backslash\Gamma$
$-\Delta u + \sigma u$	$u = 0$ on Γ, $\dfrac{\partial u}{\partial n} + \gamma u = 0$ on $\partial\Omega\backslash\Gamma$	$-\Delta w + \sigma w$	$w = 0$ on Γ, $\dfrac{\partial w}{\partial n} + \gamma w = 0$ on $\partial\Omega\backslash\Gamma$
$-\Delta u + \mathbf{b}\cdot\nabla u + \sigma u$, $\operatorname{div}\mathbf{b} = 0$	$u = 0$ on Γ, $\dfrac{\partial u}{\partial n} + \gamma u = 0$ on $\partial\Omega\backslash\Gamma$	$-\Delta w - \mathbf{b}\cdot\nabla w + \sigma w$, $\operatorname{div}\mathbf{b} = 0$	$w = 0$ on Γ, $\dfrac{\partial w}{\partial n} + (\mathbf{b}\cdot\mathbf{n} + \gamma)w = 0$ on $\partial\Omega\backslash\Gamma$
$-\Delta u + \mathbf{b}\cdot\nabla u + \sigma u$, $\operatorname{div}\mathbf{b} = 0$	$u = 0$ on Γ, $\dfrac{\partial u}{\partial n} = 0$ on $\partial\Omega\backslash\Gamma$	$-\Delta w - \mathbf{b}\cdot\nabla w + \sigma w$, $\operatorname{div}\mathbf{b} = 0$	$w = 0$ on Γ, $\dfrac{\partial w}{\partial n} + \mathbf{b}\cdot\mathbf{n}\, w = 0$ on $\partial\Omega\backslash\Gamma$
$-\operatorname{div}(\mu\nabla u) + \operatorname{div}(\mathbf{b}u) + \sigma u$	$u = 0$ on Γ, $\mu\dfrac{\partial u}{\partial n} - \mathbf{b}\cdot\mathbf{n}u = 0$ on $\partial\Omega\backslash\Gamma$	$-\operatorname{div}(\mu\nabla w) - \mathbf{b}\cdot\nabla w + \sigma w$	$w = 0$ on Γ, $\mu\dfrac{\partial w}{\partial n} = 0$ on $\partial\Omega\backslash\Gamma$
$-\operatorname{div}(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u$, $\operatorname{div}\mathbf{b} = 0$	$u = 0$ on Γ, $\mu\dfrac{\partial u}{\partial n} = 0$ on $\partial\Omega\backslash\Gamma$	$-\operatorname{div}(\mu\nabla w) - \operatorname{div}(\mathbf{b}w) + \sigma w$, $\operatorname{div}\mathbf{b} = 0$	$w = 0$ on Γ, $\mu\dfrac{\partial w}{\partial n} + \mathbf{b}\cdot\mathbf{n}w = 0$ on $\partial\Omega\backslash\Gamma$

The Lagrange identity (3.42) is now generalized as

$$ {}_{V'}\langle A(v)u, w\rangle_V = {}_V\langle u, A^*(v)w\rangle_{V'} \tag{3.47} $$

for each $u \in D(A)$ and $w \in D(A^*)$, $D(A)$ being the set of functions of class C^2 that are null at $x = 0$ and $x = 1$, and $D(A^*)$ the domain of the adjoint (or dual) operator A^* whose properties will be identified by imposing the fulfillment of (3.47). Starting from such identity, let us see which adjoint operator A^* and which dual boundary conditions we get for problem (3.45). By integrating by parts the diffusion term twice and the transport term of order one once, we obtain

$$ \begin{aligned} {}_{V'}\langle A(v)u, w\rangle_V &= -\int_0^1 u''\, w\, dx + \int_0^1 v u'\, w\, dx \\ &= \int_0^1 u'\, w'\, dx - u'\, w\Big|_0^1 - \int_0^1 (v\, w)' u\, dx + v\, u\, w\Big|_0^1 \\ &= -\int_0^1 u\, w''\, dx + u\, w'\Big|_0^1 - u'\, w\Big|_0^1 - \int_0^1 (v\, w)' u\, dx + v\, u\, w\Big|_0^1 . \end{aligned} \tag{3.48} $$

Let us analyze the boundary terms separately, by expliciting the contributions at both extrema. In order to guarantee (3.47), it must be

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

for each u and $v \in D(A)$. We observe that the fact that u belongs to $D(A)$ allows us to immediately vanish the two first and two last terms, so that we end up having

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

Since such relation must hold for each $u \in D(A)$, we must choose homogeneous Dirichlet conditions for the dual operator, i.e.

$$ w(0) = w(1) = 0. \tag{3.49} $$

Reverting to (3.48), we then have

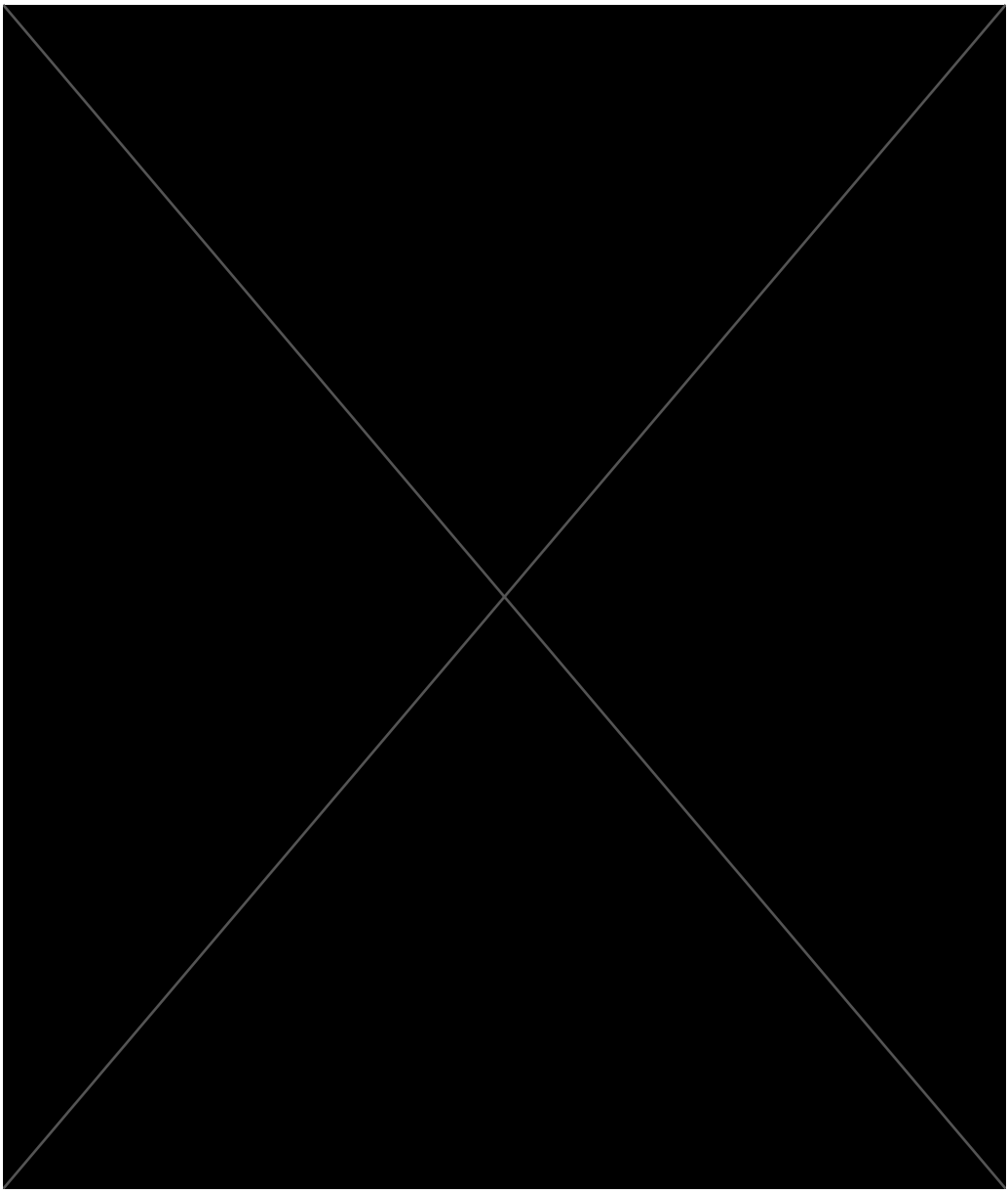
$$ \begin{aligned} {}_{V'}\langle A(v)u, w\rangle_V &= -\int_0^1 u''\, w\, dx + \int_0^1 v u'\, w\, dx \\ &= -\int_0^1 u\, w''\, dx - \int_0^1 (v\, w)' u\, dx = {}_V\langle u, A^*(v)w\rangle_{V'}. \end{aligned} $$

The adjoint operator A^* of the primal operator A defined in (3.46) therefore results to be

$$ A^*(v)\cdot = -\frac{d^2\cdot}{dx^2} + \frac{d}{dx} v\cdot $$

while the dual boundary conditions are provided by (3.49). To conclude, we note that the dual problem is always linear, even though we started from a non-linear primal problem.

For more details on the differentiation and on the analysis of the adjoint problems, we refer the reader to, e.g., [Mar95].



4

The Galerkin finite element method for elliptic problems

In this chapter, we describe the numerical solution of the elliptic boundary-value problems considered in Chap. 3 by introducing the Galerkin method. We then illustrate the finite element method as a particular case. The latter will be further developed in the following chapters.

4.1 Approximation via the Galerkin method

As seen in Chap. 3.2, the weak formulation of a generic elliptic problem set on a domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, can be written in the following way

$$\text{find } u \in V: \quad a(u, v) = F(v) \quad \forall v \in V, \tag{4.1}$$

V being an appropriate Hilbert space, subspace of $\mathrm{H}^1(\Omega)$, $a(\cdot, \cdot)$ being a continuous and coercive bilinear form from $V \times V$ in $\mathbb{R}$, $F(\cdot)$ being a continuous linear functional from V in $\mathbb{R}$. Under such hypotheses, the Lax-Milgram Lemma of Sec. 3.5 ensures existence and uniqueness of the solution.

Let V_h be a family of spaces that depends on a positive parameter h, such that

$$V_h \subset V, \quad \dim V_h = N_h < \infty \quad \forall h > 0.$$

The approximate problem takes the form

$$\text{find } u_h \in V_h: \quad a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h, \tag{4.2}$$

and is called *Galerkin problem*. Denoting with $\{\varphi_j, j = 1, 2, \ldots, N_h\}$ a basis of V_h, it suffices that (4.2) be verified for each function of the basis, as all the functions in the space V_h are a linear combination of the φ_j. We will then require that

$$a(u_h, \varphi_i) = F(\varphi_i), \quad i = 1, 2, \ldots, N_h. \tag{4.3}$$

Obviously, since $u_h \in V_h$,

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

A. Quarteroni: Numerical Models for Differential Problems.

where the $u_j, j = 1, \ldots, N_h$, are unknown coefficients. The equations (4.3) then become

$$\sum_{j=1}^{N_h} u_j \, a(\varphi_j, \varphi_i) = F(\varphi_i), \quad i = 1, 2, \ldots, N_h. \tag{4.4}$$

We denote by A the matrix (called *stiffness* matrix) with elements

$$a_{ij}{=}a(\varphi_j, \varphi_i),$$

and by $\mathbf{f}$ the vector with components $f_i = F(\varphi_i)$. If we denote by $\mathbf{u}$ the vector having as components the unknown coefficients u_j, (4.4) is equivalent to the linear system

$$\mathrm{A}\mathbf{u} = \mathbf{f}. \tag{4.5}$$

We point out some characteristics of the stiffness matrix that are independent of the basis chosen for V_h, but exclusively depend on the properties of the weak problem that is being approximated. Others instead, such as the condition number or the sparsity structure, depend on the basis under exam and are therefore reported in the sections dedicated to the specific numerical methods. For instance, bases formed by functions with small support are appealing as all the elements a_{ij} relating to basis functions having supports with null intersections will result to be null. More in general, from a computational viewpoint, the most convenient choices of V_h will be the ones requiring a modest computational effort for the computation of the matrix elements as well as the known term $\mathbf{f}$.

Theorem 4.1 *The matrix* A *associated to the discretization of an elliptic problem with the Galerkin method is positive definite.*

Proof. We recall that a matrix $\mathrm{B} \in \mathbb{R}^{n\times n}$ is said to be positive definite if

$$\mathbf{v}^T \mathrm{B}\mathbf{v} \geq 0 \quad \forall \mathbf{v} \in \mathbb{R}^n \quad \text{and also } \mathbf{v}^T \mathrm{B}\mathbf{v} = 0 \;\Leftrightarrow\; \mathbf{v} = \mathbf{0}. \tag{4.6}$$

The correspondence

$$\mathbf{v} = (v_i) \in \mathbb{R}^{N_h} \leftrightarrow \quad v_h(x) = \sum_{j=1}^{N_h} v_j \phi_j \in V_h \tag{4.7}$$

defines a bijection between the spaces $\mathbb{R}^{N_h}$ and V_h . Given a generic vector $\mathbf{v} = (v_i)$ di $\mathbb{R}^{N_h}$, thanks to the bilinearity and coercivity of the form $a(\cdot, \cdot)$, we obtain

$$\begin{aligned}
\mathbf{v}^T \mathrm{A}\mathbf{v} &= \sum_{j=1}^{N_h}\sum_{i=1}^{N_h} v_i a_{ij} v_j = \sum_{j=1}^{N_h}\sum_{i=1}^{N_h} v_i a(\varphi_j, \varphi_i) v_j \\
&= \sum_{j=1}^{N_h}\sum_{i=1}^{N_h} a(v_j\varphi_j, v_i\varphi_i) = a\left(\sum_{j=1}^{N_h} v_j\varphi_j, \sum_{i=1}^{N_h} v_i\varphi_i\right) \\
&= a(v_h, v_h) \geq \alpha \|v_h\|_V^2 \geq 0.
\end{aligned}$$

Moreover, if $\mathbf{v}^T \mathrm{A}\mathbf{v} = 0$, then, by what we have just obtained, $\|v_h\|_V^2 = 0$ too, i.e. $v_h = 0$ and so $\mathbf{v} = \mathbf{0}$. Consequently, the thesis is proven as the two conditions in (4.6) are fulfilled. ◇

Furthermore, the following property can be proven (see Exercise 4):

Property 4.1 *The matrix* A *is simmetric if and only if the bilinear form* $a(\cdot,\cdot)$ *is symmetric.*

For instance, in the case of the Poisson problem with either Dirichlet (3.18) or mixed (3.27) boundary conditions, the matrix A is symmetric and positive definite. The numerical solution of such a system can be efficiently performed both using direct methods such as the Cholesky factorization, and iterative methods such as the conjugate gradient method (see Chap. 7 and, e.g., [QSS07, Chap. 4]).

4.2 Analysis of the Galerkin method

In this section, we aim at studying the Galerkin method, and in particular at verifying three of its fundamental properties:

- *existence* and *uniqueness* of the discrete solution u_h;
- *stability* of the discrete solution u_h;
- *convergence* of u_h to the exact solution u of problem (4.1), for $h \to 0$.

4.2.1 Existence and uniqueness

The Lax-Milgram Lemma, stated in Sec. 3.5, holds for any Hilbert space, hence, in particular, for the space V_h, as the latter is a closed subspace of the Hilbert space V. Furthermore, the bilinear form $a(\cdot,\cdot)$ and the functional $F(\cdot)$ are the same as in the variational problem (4.1). The hypotheses required by the Lemma are therefore fulfilled. The following result then derives:

Corollary 4.1 *The solution of the Galerkin problem* (4.2) *exists and is unique.*

It is nonetheless instructive to provide a constructive proof of this Corollary without using the Lax-Milgram Lemma. As we have seen, indeed, the Galerkin problem (4.2) is equivalent to the linear system (4.5). Proving the existence and uniqueness for one means to automatically prove the existence and uniqueness of the other. We therefore focus our attention on the linear system (4.5).

The matrix A is invertible as the unique solution of system $\mathrm{A}\mathbf{u} = \mathbf{0}$ is the identically null solution. This immediately descends from the fact that A is positive definite. Consequently, the linear system (4.5) admits a unique solution, hence also its corresponding Galerkin problem admits a unique solution.

4.2.2 Stability

Corollary 3.1 allows us to provide the following stability result.

Corollary 4.2 *The Galerkin method is stable, uniformly with respect to h, as the following upper bound holds for the solution*

$$\|u_h\|_V \leq \frac{1}{\alpha}\|F\|_{V'}.$$

The stability of the method guarantees that the norm $\|u_h\|_V$ of the discrete solution remains bounded for h tending to zero, uniformly with respect to h. Equivalently, it guarantees that $\|u_h - w_h\|_V \leq \frac{1}{\alpha}\|F - G\|_{V'}$, u_h and w_h being numerical solutions corresponding to two different data F and G.

4.2.3 Convergence

We now want to prove that the weak solution of the Galerkin problem converges to the solution of the weak problem (4.1) when h tends to zero. Consequently, by taking a sufficiently small h, it will be possible to approximate the exact solution u as accurately as desired by the Galerkin solution u_h.

Let us first prove the following consistency property.

Lemma 4.1 (Céa) *The Galerkin method is strongly consistent, that is*

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{4.8}$$

Proof. Since $V_h \subset V$, the exact solution u satisfies the weak problem (4.1) for each element $v = v_h \in V_h$, hence we have

$$a(u, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{4.9}$$

By subtracting side to side (4.2) from (4.9), we obtain

$$a(u, v_h) - a(u_h, v_h) = 0 \quad \forall v_h \in V_h,$$

from which, thanks to the bilinearity of the form $a(\cdot, \cdot)$, the thesis follows. ◇

Let us point out that (4.9) coincides with the definition of strong consistency given in (1.10).

Property (4.8) is known as Galerkin orthogonality. The reason is that, if $a(\cdot, \cdot)$ is symmetric, it defines a scalar product in V. Then, the consistency property is interpreted as the orthogonality with respect to the scalar product $a(\cdot, \cdot)$, between the

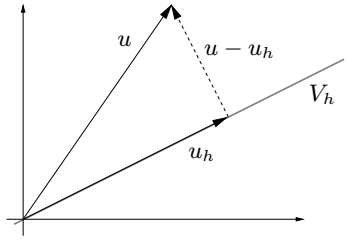


Fig. 4.1. Geometric interpretation of the Céa lemma

approximation error, $u - u_h$, and the subspace V_h. In this sense, analogously to the euclidian case, the solution u_h of the Galerkin method is said to be the *orthogonal projection* on V_h of the exact solution u. Among all elements of V_h, v_h is the one minimizing the distance to the exact solution u in the *energy norm*, i.e. in the following norm induced by the scalar product $a(\cdot, \cdot)$:

$$\|u - u_h\|_a = \sqrt{a(u - u_h, u - u_h)}.$$

Remark 4.1 The geometric interpretation of the Galerkin method makes sense only in the case where the form $a(\cdot, \cdot)$ is symmetric. However, this does not impair the generality of the method or its consistency property in the case where the bilinear form is not symmetric. •

Let us now consider the value taken by the bilinear form when both its arguments are equal to $u - u_h$. If v_h is an arbitrary element of V_h we obtain

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h).$$

The last term is null thanks to (4.8), as $v_h - u_h \in V_h$. Moreover

$$|a(u - u_h, u - v_h)| \leq M\|u - u_h\|_V \|u - v_h\|_V,$$

having exploited the continuity of the bilinear form. On the other hand, by the coercivity of $a(\cdot, \cdot)$ it follows

$$a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|_V^2$$

hence we have

$$\|u - u_h\|_V \leq \frac{M}{\alpha}\|u - v_h\|_V \quad \forall v_h \in V_h.$$

Such inequality holds for all functions $v_h \in V_h$ and therefore we find

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|_V. \tag{4.10}$$

It is then evident that in order for the method to converge, it will be sufficient to require that, for h tending to zero, the space V_h tends to "fill" the entire space V. Precisely, it must turn out that

$$\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0 \quad \forall v \in V. \tag{4.11}$$

In that case, the Galerkin method is convergent and it can be written that

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

The space V_h must therefore be carefully chosen in order to guarantee the density property (4.11). Once this requirement is satisfied, convergence will be verified in any case, independently of how u is made; conversely, the speed with which the discrete solution converges to the exact solution, i.e. the order of decay of the error with respect to h, will depend, in general, on both the choice of V_h and the regularity of u (see Theorem 4.3).

Remark 4.2 Obviously, $\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V$. Consequently, by (4.10), if $\frac{M}{\alpha}$ is of the order of unity, the error due to the Galerkin method can be identified with the best approximation error for u in V_h. In any case, both errors have the same infinitesimal order with respect to h. •

Remark 4.3 In the case where $a(\cdot,\cdot)$ is a symmetric bilinear form, and also continuous and coercive, then (4.10) can be improved as follows (see Exercise 5)

$$\|u - u_h\|_V \leq \sqrt{\frac{M}{\alpha}} \inf_{w_h \in V_h} \|u - w_h\|_V. \tag{4.12}$$

•

4.3 The finite element method in the one-dimensional case

Let us suppose that Ω be an interval (a, b). The goal of this section is to create approximations of the space $\mathrm{H}^1(a, b)$, that depend on a parameter h. To this end, we introduce a partition $\mathcal{T}_h$ of (a, b) in $N + 1$ subintervals $K_j = (x_{j-1}, x_j)$, also called *elements*, having width $h_j = x_j - x_{j-1}$ with

$$a = x_0 < x_1 < \ldots < x_N < x_{N+1} = b, \tag{4.13}$$

and set $h = \max_j h_j$.
Since the functions of $\mathrm{H}^1(a, b)$ are continuous functions on $[a, b]$, we can construct the following family of spaces

$$X_h^r = \left\{ v_h \in C^0\left(\overline{\Omega}\right) : \, v_h|_{K_j} \in \mathbb{P}_r \ \forall K_j \in \mathcal{T}_h \right\}, \quad r = 1, 2, \ldots \tag{4.14}$$

having denoted by $\mathbb{P}_r$ the space of polynomials with degree lower than or equal to r in the variable x. The spaces X_h^r are all subspaces of $\mathrm{H}^1(a,b)$ as they are constituted by differentiable functions except for at most a finite number of points (the vertices x_i of the partition $\mathcal{T}_h$). They represent possible choices for the space V_h, provided that the boundary conditions are properly incorporated. The fact that the functions of X_h^r are locally (elementwise) polynomials will make the stiffness matrix easy to compute.

We must now choose a basis $\{\varphi_i\}$ for the X_h^r space. It is convenient, by what exposed in Sec. 4.1, that the support of the generic basis function φ_i have non-empty intersection only with that of a negligible number of other functions of the basis. In such way, many elements of the stiffness matrix will be null. It is also convenient that the basis be *Lagrangian*: in that case, the coefficients of the expansion of a generic function $v_h \in X_h^r$ on the basis itself will be the values taken by v_h in carefully chosen points, which we call *nodes* and which, as we will see, generally form a superset of the vertices of $\mathcal{T}_h$. This does not prevent the use of non-lagrangian bases, especially in their hierarchical version (as we will see later). We now provide some examples of bases for the spaces X_h^1 and X_h^2.

4.3.1 The space X_h^1

It is constituted by the piecewise continuous and linear functions on a partition $\mathcal{T}_h$ of (a,b) of the form (4.13). Since only one straight line can pass by two different points and the functions of X_h^1 are continuous, the *degrees of freedom* of the functions of this space, i.e. the values that must be assigned to univoquely define the functions themselves, will be equal to the number $N+2$ of vertices of the partition itself. In this case, therefore, nodes and vertices coincide. Consequently, having assigned $N+2$ basis functions φ_i, $i=0,\ldots,N+1$, the whole space X_h^1 will be completely defined. The characteristic Lagrangian basis functions are characterized by the following property

$$\varphi_i \in X_h^1 \quad \text{such that} \quad \varphi_i(x_j) = \delta_{ij}, \quad i,j = 0,1,\ldots,N+1,$$

δ_{ij} being the Kronecker delta. The function φ_i is therefore piecewise linear and equal to one at x_i and zero at the remaining nodes of the partition (see Fig. 4.2). Its expression is given by

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{for } x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{for } x_i \le x \le x_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.15}$$

Obviously φ_i has the union of the only intervals $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$ as support, if $i \neq 0$ or $i \neq N+1$ (for $i = 0$ or $i = N+1$ the support will be limited to the

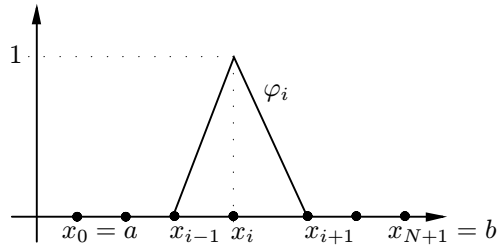


Fig. 4.2. The basis function of X_h^1 associated to node x_i

interval $[x_0, x_1]$ or $[x_N, x_{N+1}]$, respectively). Consequently, only the basis functions φ_{i-1} and φ_{i+1} have a support with non-empty intersection with that of φ_i, henceforth the stiffness matrix is tridiagonal as $a_{ij} = 0$ if $j \notin \{i-1, i, i+1\}$.

As visible in expression (4.15) the two basis functions φ_i and φ_{i+1} defined on each interval $[x_i, x_{i+1}]$, basically repeat themselves with no changes, up to a scaling factor linked to the length of the interval itself. In practice, the two basis functions φ_i and φ_{i+1} can be obtained by transforming two basis functions $\widehat{\varphi}_0$ and $\widehat{\varphi}_1$ built once and for all on a reference interval, typically the $[0, 1]$ interval.

To this end, it is sufficient to exploit the fact that the generic interval (x_i, x_{i+1}) of the partition of (a, b) can be obtained starting from the interval $(0, 1)$ via the linear transformation $\phi : [0, 1] \to [x_i, x_{i+1}]$ defined as

$$x = \phi(\xi) = x_i + \xi(x_{i+1} - x_i). \tag{4.16}$$

If we define the two basis functions $\widehat{\varphi}_0$ and $\widehat{\varphi}_1$ on $[0, 1]$ as

$$\widehat{\varphi}_0(\xi) = 1 - \xi, \quad \widehat{\varphi}_1(\xi) = \xi,$$

the basis functions φ_i and φ_{i+1} on $[x_i, x_{i+1}]$ will simply be given by

$$\varphi_i(x) = \widehat{\varphi}_0(\xi(x)), \quad \varphi_{i+1}(x) = \widehat{\varphi}_1(\xi(x))$$

since $\xi(x) = (x - x_i)/(x_{i+1} - x_i)$ (see Fig. 4.3 and 4.4).

This way of proceeding (defining the basis on a reference element and then transforming it on a specific element) will be of fundamental importance when considering problems in several dimensions.

4.3.2 The space X_h^2

The functions of X_h^2 are piecewise polynomials of degree 2 on each interval of $\mathcal{T}_h$ and, consequently, are univocally set once the values they take in three distinct points of each interval K_j are assigned. To guarantee the continuity of the functions of X_h^2 two of these points will be the extrema of the generic interval of $\mathcal{T}_h$, the third will be the midpoint of the latter. The degrees of freedom of the space X_h^2 are therefore the

values of v_h taken at the extrema of the intervals composing the partition $\mathcal{T}_h$ and at their midpoints. We order the nodes starting from $x_0 = a$ to $x_{2N+2} = b$; in such way the midpoints correspond to the nodes with odd indices, and the extrema to the nodes with even indices (refer to Exercise 6 for alternative numberings).

Exactly as in the previous case the Lagrangian basis for X_h^2 is the one formed by the functions

$$\varphi_i \in X_h^2 \quad \text{such that} \quad \varphi_i(x_j) = \delta_{ij}, \quad i, j = 0, 1, \ldots, 2N+2.$$

These are therefore piecewise quadratic functions that are equal to 1 at the node to which they are associated and are null at the remaining nodes. We report the explicit expression of the generic basis function associated to the extrema of the intervals in the partition:

$$(i \text{ even}) \quad \varphi_i(x) = \begin{cases} \dfrac{(x - x_{i-1})(x - x_{i-2})}{(x_i - x_{i-1})(x_i - x_{i-2})} & \text{if } x_{i-2} \le x \le x_i, \\ \dfrac{(x_{i+1} - x)(x_{i+2} - x)}{(x_{i+1} - x_i)(x_{i+2} - x_i)} & \text{if } x_i \le x \le x_{i+2}, \\ 0 & \text{otherwise.} \end{cases}$$

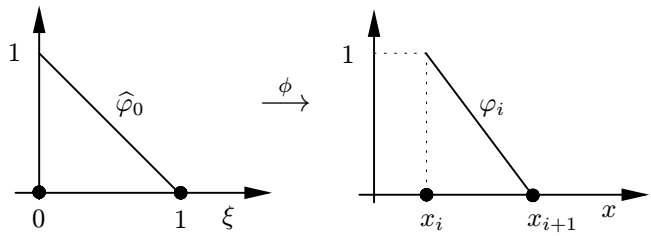


Fig. 4.3. The basis function φ_i in $[x_i, x_{i+1}]$ and the corresponding basis function $\widehat{\varphi}_0$ on the reference element

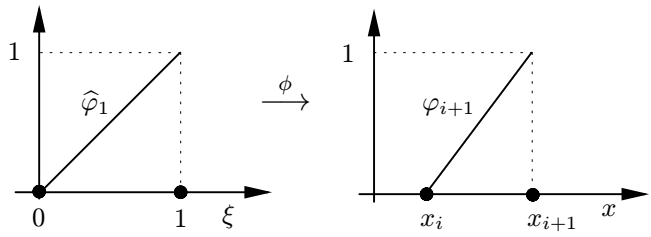


Fig. 4.4. The basis function φ_{i+1} in $[x_i, x_{i+1}]$ and the corresponding basis function $\widehat{\varphi}_1$ on the reference element

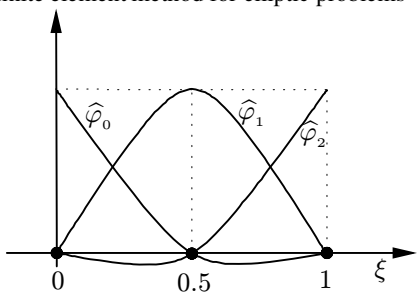


Fig. 4.5. The basis functions X_h^2 on the reference interval

For the midpoints of the intervals, we have

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

As in the case of linear finite elements, in order to describe the basis it is sufficient to provide the expression of the basis functions on the reference interval $[0,1]$ and then to transform the latter via (4.16). We have

$$\widehat{\varphi}_0(\xi) = (1-\xi)(1-2\xi), \quad \widehat{\varphi}_1(\xi) = 4(1-\xi)\xi, \quad \widehat{\varphi}_2(\xi) = \xi(2\xi-1).$$

We report a representation of these functions in Fig. 4.5. Note that the generic basis function φ_{2i+1} relative to node x_{2i+1} has a support coinciding with the element to which the midpoint belongs. For its peculiar form it is known as *bubble function*.
As previously anticipated, we can also introduce other non-lagrangian bases. A particularly interesting one is the one constructed (locally) by the three functions

$$\widehat{\psi}_0(\xi) = 1-\xi, \quad \widehat{\psi}_1(\xi) = \xi, \quad \widehat{\psi}_2(\xi) = (1-\xi)\xi.$$

A basis of this kind is said to be *hierarchical* as, to construct the basis for X_h^2, it exploits the basis functions of the immediately lower-dimension space, X_h^1. It is convenient from a computational viewpoint if one decides, during the approximation of a problem, to increase only locally, i.e. only for such elements, the degree of interpolation (that is if one intends to perform the so-called adaptivity in the degree, or *adaptivity of type p*).

The Lagrange polynomials are linearly independent by construction. In general however, such property must be verified to ensure that the set of chosen polynomials is effectively a basis. In the case of functions $\widehat{\psi}_0$, $\widehat{\psi}_1$ and $\widehat{\psi}_2$ we must verify that

$$\text{if } \alpha_0\widehat{\psi}_0(\xi) + \alpha_1\widehat{\psi}_1(\xi) + \alpha_2\widehat{\psi}_2(\xi) = 0 \ \forall \xi, \quad \text{then} \quad \alpha_0 = \alpha_1 = \alpha_2 = 0.$$

Indeed, the equation

$$\alpha_0\widehat{\psi}_0(\xi) + \alpha_1\widehat{\psi}_1(\xi) + \alpha_2\widehat{\psi}_2(\xi) = \alpha_0 + \xi(\alpha_1 - \alpha_0 + \alpha_2) - \alpha_2\xi^2 = 0$$

implies $\alpha_0 = 0$, $\alpha_2 = 0$ and therefore $\alpha_1 = 0$. We notice that the stiffness matrix in the case of finite elements of degree 2 will be pentadiagonal.

By proceeding in the same way it will be possible to generate bases for X_h^r with an arbitrary positive integer r: we point out however that as the polynomial degree increases, the number of degrees of freedom increases and so does the computational cost of solving the linear system (4.5). Moreover, a well known fact from the polynomial interpolation theory, the use of high degrees combined with equispaced node distributions, leads to less and less stable approximations, in spite of the theoretical increase in accuracy. A successful remedy is provided by the spectral element approximation that, using well-chosen nodes (the ones from the Gaussian quadrature), allows to generate approximations with arbitrarily high accuracy. To this purpose see Chap. 10.

4.3.3 The approximation with linear finite elements

We now examine how to approximate the following problem

$$\begin{cases} -u'' + \sigma u = f, & a < x < b, \\ u(a) = 0, & u(b) = 0, \end{cases}$$

whose weak formulation, as we have seen in the previous chapter, is

$$\text{find } u \in \mathrm{H}_0^1(a,b): \quad \int_a^b u'v'\,dx + \int_a^b \sigma uv\,dx = \int_a^b fv\,dx \quad \forall v \in \mathrm{H}_0^1(a,b).$$

As we did in (4.13), we introduce a decomposition $\mathcal{T}_h$ of $(0,1)$ in $N+1$ subintervals K_j and use linear finite elements. We therefore introduce the space

$$V_h = \{v_h \in X_h^1 : \ v_h(a) = v_h(b) = 0\} \tag{4.17}$$

that is the space of piecewise linear functions that vanish at the boundary (a function of such space has been introduced in Fig. 4.6). This is a subspace of $\mathrm{H}_0^1(a,b)$.

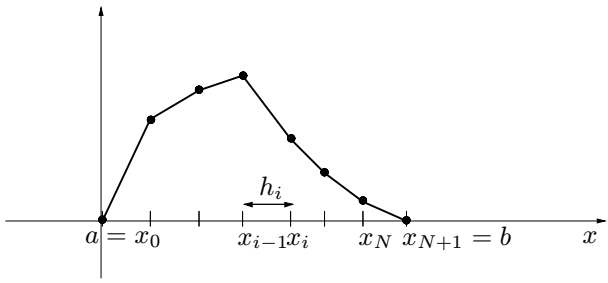


Fig. 4.6. Example of a function of V_h

The corresponding finite element problem is therefore given by

$$\text{find } u_h \in V_h : \int_a^b u_h' v_h' \, dx + \int_a^b \sigma u_h v_h \, dx = \int_a^b f v_h \, dx \quad \forall v_h \in V_h. \tag{4.18}$$

We use as a basis of X_h^1 the set of hat functions defined in (4.15) by caring to only consider the indices $1 \leq i \leq N$. By expressing u_h as a linear combination of such functions $u_h(x) = \sum_{i=1}^{N} u_i \varphi_i(x)$, and imposing that (4.18) be satisfied for each element of the basis of V_h, we obtain a system of N equations

$$\mathrm{A}\mathbf{u} = \mathbf{f}, \tag{4.19}$$

where

$$\mathrm{A} = [a_{ij}], \quad a_{ij} = \int_a^b \varphi_j' \varphi_i' \, dx + \int_a^b \sigma \varphi_j \varphi_i \, dx \, ;$$

$$\mathbf{u} = [u_i]; \quad \mathbf{f} = [f_i], \quad f_i = \int_a^b f \varphi_i \, dx.$$

Note that $u_i = u_h(x_i),\ 1 \leq i \leq N$, that is the finite element unknowns are the nodal values of the finite element solution u_h.
To find the numerical solution u_h it is now sufficient to solve the linear system (4.19).

In the case of linear finite elements, the stiffness matrix A is not only sparse, but also results to be tridiagonal. To compute its elements, we proceed as follows. As we have seen it is not necessary to directly operate on the basis functions on the single intervals, but it is sufficient to refer to the ones defined on the reference interval: it will then be enough to appropriately transform the integrals that appear in the definition of the coefficients of A.

A generic non-null element of the stiffness matrix is given by

$$a_{ij} = \int_a^b (\varphi_i' \varphi_j' + \sigma \varphi_i \varphi_j) dx = \int_{x_{i-1}}^{x_i} (\varphi_i' \varphi_j' + \sigma \varphi_i \varphi_j) dx + \int_{x_i}^{x_{i+1}} (\varphi_i' \varphi_j' + \sigma \varphi_i \varphi_j) dx.$$

Let us consider the first addendum by supposing $j = i - 1$. Evidently, via the coordinate transformation (4.16), we can re-write it as

$$\int_{x_{i-1}}^{x_i} (\varphi_i' \varphi_{i-1}' + \sigma \varphi_i \varphi_{i-1}) dx =$$

$$\int_0^1 [\varphi_i'(x(\xi)) \varphi_{i-1}'(x(\xi)) + \sigma(x(\xi)) \varphi_i(x(\xi)) \varphi_{i-1}(x(\xi))] h_i \, d\xi,$$

having noted that $dx = d(x_{i-1} + \xi h_i) = h_i d\xi$. On the other hand $\varphi_i(x(\xi)) = \widehat{\varphi}_1(\xi)$ and $\varphi_{i-1}(x(\xi)) = \widehat{\varphi}_0(\xi)$. We also note that

$$\frac{d}{dx}\varphi_i(x(\xi)) = \frac{d\xi}{dx}\widehat{\varphi}_1'(\xi) = \frac{1}{h_i}\widehat{\varphi}_1'(\xi).$$

Similarly, we find that $\varphi_{i-1}'(x(\xi)) = (1/h_i)\widehat{\varphi}_0'(\xi)$. Hence, the element $a_{i,i-1}$ becomes

$$a_{i,i-1} = \int_0^1 \left(\frac{1}{h_i}\widehat{\varphi}_1'(\xi)\widehat{\varphi}_0'(\xi) + \sigma\widehat{\varphi}_1(\xi)\widehat{\varphi}_0(\xi)h_i\right)\, d\xi.$$

The advantage of this expression lies in the fact that in the case of constant coefficients, all the integrals appearing within matrix A can be computed once and for all. We will see in the multi-dimensional case that this way of proceeding maintains its importance also in the case of variable coefficients.

4.3.4 Interpolation operator and interpolation error

Let us set $I = (a, b)$. For each $v \in C^0(\overline{I})$, we define *interpolant* of v in the space of X_h^1, determined by the partition $\mathcal{T}_h$, the function $\Pi_h^1 v$ such that

$$\Pi_h^1 v(x_i) = v(x_i) \quad \forall x_i, \text{ node of the partition}, i = 0, \ldots, N+1.$$

By using the Lagrangian basis $\{\varphi_i\}$ of the space X_h^1, the interpolant can be expressed in the following way

$$\Pi_h^1 v(x) = \sum_{i=0}^{N+1} v(x_i)\varphi_i(x).$$

Hence, when v and a basis of X_h^1 are known, the interpolant of v is easy to compute. The operator $\Pi_h^1 : C^0(\overline{I}) \mapsto X_h^1$ mapping a function v to its interpolant $\Pi_h^1 v$ is called *interpolation operator*.

Analogously, we can define the operators $\Pi_h^r : C^0(\overline{I}) \mapsto X_h^r$, for all $r \geq 1$. Having denoted by $\Pi_{K_j}^r$ the local interpolation operator mapping a function v to the polynomial $\Pi_{K_j}^r v \in \mathbb{P}_r(K_j)$, interpolating v at the $r+1$ nodes of the element $K_j \in \mathcal{T}_h$, we define $\Pi_h^r v$ as

$$\Pi_h^r v \in X_h^r \ : \quad \Pi_h^r v\big|_{K_j} = \Pi_K^r(v\big|_{K_j}) \qquad \forall K_j \in \mathcal{T}_h. \tag{4.20}$$

Theorem 4.2 *Let $v \in \mathrm{H}^{r+1}(I)$, for $r \geq 1$, and let $\Pi_h^r v \in X_h^r$ be its interpolating function defined in* (4.20)*. The following estimate of the interpolation error holds*

$$|v - \Pi_h^r v|_{\mathrm{H}^k(I)} \leq C_{k,r} h^{r+1-k} |v|_{\mathrm{H}^{r+1}(I)} \qquad \text{for } k = 0, 1. \tag{4.21}$$

The constants $C_{k,r}$ are independent of v and h. We recall that $\mathrm{H}^0(I) = \mathrm{L}^2(I)$ and that $|\cdot|_{\mathrm{H}^0(I)} = \|\cdot\|_{\mathrm{L}^2(I)}$.

Proof. We prove (4.21) for the case $r = 1$, and refer to [QV94, Chap. 3] or [Cia78] for the more general case. We start by observing that if $v \in H^{r+1}(I)$ then $v \in C^r(I)$. In particular, for $r = 1$, $v \in C^1(I)$. Let us set $e = v - \Pi_h^1 v$. Since $e(x_j) = 0$ for each node x_j, the Rolle theorem allows to conclude that there exist some $\xi_j \in K_j = (x_{j-1}, x_j)$, with $j = 1, \ldots, N+1$, for which we have $e'(\xi_j) = 0$.

$\Pi_h^1 v$ being a linear function in each interval K_j, we obtain that for $x \in K_j$

$$e'(x) = \int_{\xi_j}^{x} e''(s)ds = \int_{\xi_j}^{x} v''(s)ds,$$

from which we deduce that

$$|e'(x)| \leq \int_{x_{j-1}}^{x_j} |v''(s)|ds \qquad \text{for } x \in K_j.$$

Now, by using the Cauchy-Schwarz inequality we obtain

$$|e'(x)| \leq \left(\int_{x_{j-1}}^{x_j} 1^2 ds \right)^{1/2} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds \right)^{1/2} \leq h^{1/2} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds \right)^{1/2}. \tag{4.22}$$

Hence,

$$\int_{x_{j-1}}^{x_j} |e'(x)|^2 dx \leq h^2 \int_{x_{j-1}}^{x_j} |v''(s)|^2 ds. \tag{4.23}$$

An upper bound for $e(x)$ can be obtained by noting that, for each $x \in K_j$, $e(x) = \int_{x_{j-1}}^{x} e'(s)ds$, and therefore, by applying inequality (4.22),

$$|e(x)| \leq \int_{x_{j-1}}^{x_j} |e'(s)|ds \leq h^{3/2} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds \right)^{1/2}.$$

Hence,

$$\int_{x_{j-1}}^{x_j} |e(x)|^2 dx \leq h^4 \int_{x_{j-1}}^{x_j} |v''(s)|^2 ds. \tag{4.24}$$

By summing over the indices j from 1 to $N+1$ in (4.23) and (4.24) we obtain the inequalities

$$\left(\int_a^b |e'(x)|^2 dx\right)^{1/2} \leq h \left(\int_a^b |v''(x)|^2 dx\right)^{1/2}$$

and

$$\left(\int_a^b |e(x)|^2 dx\right)^{1/2} \leq h^2 \left(\int_a^b |v''(x)|^2 dx\right)^{1/2}$$

respectively, that correspond to the desired estimates (4.21) for $r = 1$, with $C_{k,1} = 1$ and $k = 0, 1$. ◇

4.3.5 Estimate of the finite element error in the H^1 norm

Thanks to the result (4.21) we can obtain an estimate of the approximation error of the finite element method.

Theorem 4.3 *Let $u \in V$ be the exact solution of the variational problem* (4.1) *(in our case $\Omega = I = (a, b)$) and u_h its approximate solution via the finite element method of degree r, i.e. the solution of problem* (4.2) *where $V_h = X_h^r \cap V$. Moreover, let $u \in \mathrm{H}^{p+1}(I)$, for a suitable p such that $r \leq p$. Then, the following inequality, also called a priori error estimate, holds*

$$\|u - u_h\|_V \leq \frac{M}{\alpha} C h^r |u|_{\mathrm{H}^{r+1}(I)}, \tag{4.25}$$

C being a constant independent of u and h.

Proof. From (4.10), by setting $w_h = \Pi_h^r u$, the interpolant of degree r of u in the space V_h, we obtain

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

The right-hand side can now be bounded from above via the interpolation error estimate (4.21) for $k = 1$, from which the thesis follows. ◇

It follows from the latter theorem that, in order to increase the accuracy, two different strategies can be followed: reducing h, i.e. refining the grid, or increasing r, that is using finite elements of higher degree. However, the latter strategy makes sense only if the solution u is regular enough: as a matter of fact, from (4.25) we immediately infer that, if $u \in V \cap \mathrm{H}^{p+1}(I)$, the maximum value of r that it makes sense to take is $r = p$. Values higher than r do not ensure a better rate of convergence: therefore

if the solution is not very regular it is not convenient to use finite elements of high degree, as the greater computational cost is not compensated by an improvement of the convergence. An interesting case is when the solution only has the minimum regularity ($p = 0$). From the relations (4.10) and (4.11) we obtain that there is convergence anyhow, but the estimate (4.25) is no longer valid. It is then impossible to say how the norm V of the error tends to zero when h decreases. We summarize these situations in Table 4.1.

Table 4.1. Order of convergence with respect to h for the finite element method for varying regularity of the solution and degree r of the finite elements. We have highlighted on each column the result corresponding to the "optimal" choice of the polynomial degree

r		$u \in \mathrm{H}^1(I)$	$u \in \mathrm{H}^2(I)$	$u \in \mathrm{H}^3(I)$	$u \in \mathrm{H}^4(I)$	$u \in \mathrm{H}^5(I)$
1	converge	$\boxed{h^1}$	h^1	h^1	h^1	
2	converge	h^1	$\boxed{h^2}$	h^2	h^2	
3	converge	h^1	h^2	$\boxed{h^3}$	h^3	
4	converge	h^1	h^2	h^3	$\boxed{h^4}$	

In general, we can state that: if $u \in \mathrm{H}^{p+1}(I)$, for a given $p > 0$, then there exists a constant C independent of u and h, such that

$$\|u - u_h\|_{\mathrm{H}^1(I)} \leq Ch^s |u|_{\mathrm{H}^{s+1}(I)}, \quad s = \min\{r, p\}. \tag{4.26}$$

4.4 Finite elements, simplices and barycentric coordinates

Before introducing finite element spaces in 2D and 3D domains we can attempt to provide a formal definition of *finite element*.

4.4.1 An abstract definition of finite element in the Lagrangian case

From the examples we considered we can deduce that there are three ingredients allowing to characterize univocally a finite element in the general case, i.e. independently of the dimension:

- the domain of definition K of the element. In the one-dimensional case it is an interval, in the two-dimensional case it is generally a triangle but it can also be a quadrilateral; in the three-dimensional case it can be a tetrahedron, a prism or a hexahedron;
- a space of polynomials Π_r of dimension N_r defined on K and a basis $\{\varphi_j\}_{j=1}^{N_r}$ of Π_r;

- a set of functionals on Π_r, $\Sigma = \{\gamma_i : \Pi_r \to \mathbb{R}\}_{i=1}^{N_r}$ satisfying $\gamma_i(\varphi_j) = \delta_{ij}$, δ_{ij} being the Kronecker delta. These allow to univocally identify the coefficients $\{\alpha_j\}_{j=1}^{N_r}$ of the expansion of a polynomial $p \in \Pi_r$ with respect to the chosen basis, $p(x) = \sum_{j=1}^{N_r} \alpha_j \varphi_j(x)$. As a matter of fact, we have $\alpha_i = \gamma_i(p)$, $i = 1, \ldots, N_r$. These coefficients are called *degrees of freedom* of the finite element.

In the case of *Lagrange finite elements* the chosen basis is provided by the Lagrange polynomials and the degree of freedom α_i is equal to the value taken by the polynomial p at a point $\mathbf{a}_i$ of K, called *node*, that is we have $\alpha_i = p(\mathbf{a}_i)$, $i = 1, \ldots, N_r$. We can then set, with a slight notation abuse, $\Sigma = \{\mathbf{a}_j\}_{j=1}^{N_r}$, as knowing the position of the nodes allows us to find the degrees of freedom (notice however that this is not true in general, think only of the case of the hierarchical basis introduced previously). In the remainder, we will exclusively refer to the case of Lagrange finite elements.

In the construction of a Lagrange finite element, the choice of nodes is not arbitrary. Indeed, the problem of interpolation on a given set K may be ill-posed. For this reason the following definition proves useful:

Definition 4.1 *A set $\Sigma = \{\mathbf{a}_j\}_{j=1}^{N_r}$ of points of K is called unisolvent on Π_r if, given N_r arbitrary scalars α_j, $j = 1, \ldots, N_r$, there exists a unique function $p \in \Pi_r$ such that*

$$p(\mathbf{a}_j) = \alpha_j, \quad j = 1, \ldots, N_r.$$

In such case, the triple (K, Σ, Π_r) is called *Lagrangian finite element*. In the case of Lagrangian finite elements, the element is generally recalled by citing the sole polynomial space: hence the linear finite elements introduced previously are called $\mathbb{P}_1$, the quadratic ones $\mathbb{P}_2$, and so forth.

As we have seen in the 1D case, for the finite elements $\mathbb{P}_1$ and $\mathbb{P}_2$ it is convenient to define the finite element starting from a reference element $\widehat{K}$; typically this is the interval $(0, 1)$. It will commonly be the right triangle with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$ in the two-dimensional case (when using triangular elements). (See Sec. 4.4.2 for the case in arbitrary dimensions.) Hence, via a transformation ϕ, we move to the finite element defined on K. The transformation therefore concerns the finite element as a whole. More precisely, we observe that if $(\widehat{K}, \widehat{\Sigma}, \widehat{\Pi}_r)$ is a Lagrangian finite element and $\phi : \widehat{K} \to \mathbb{R}^d$ a continuous and injective application, and we define

$$K = \phi(\widehat{K}), \quad P_r = \{p : K \to \mathbb{R} : \ p \circ \phi \in \widehat{\Pi}_r\}, \quad \Sigma = \phi(\widehat{\Sigma}),$$

then (K, Σ, P_r) is still said to be a Lagrangian finite element. The space of polynomials defined on triangles and tetrahedra can be introduced as follows.

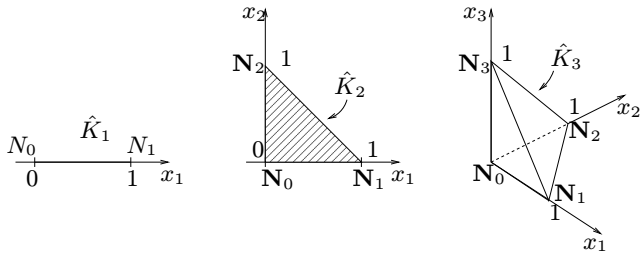


Fig. 4.7. The unitary simplex in $\mathbb{R}^d$, $d = 1, 2, 3$

4.4.2 Simplices

If $\{\mathbf{N}_0, \ldots, \mathbf{N}_d\}$ are $d+1$ points in $\mathbb{R}^d$, $d \geq 1$, and the vectors $\{\mathbf{N}_1 - \mathbf{N}_0, \ldots, \mathbf{N}_d - \mathbf{N}_0\}$ are linearly independent, then the convex hull of $\{\mathbf{N}_0, \ldots, \mathbf{N}_d\}$ is called a *simplex*, and $\{\mathbf{N}_0, \ldots, \mathbf{N}_d\}$ area called the *vertices* of the simplex. The *unitary simplex* of $\mathbb{R}^d$ is the set

$$\hat{K}_d = \{\mathbf{x} \in \mathbb{R}^d : x_i \geq 0,\ 1 \leq i \leq d,\ \sum_{i=1}^{d} x_i \leq 1\}, \tag{4.27}$$

and it is a unitary interval in $\mathbb{R}^1$, a unitary triangle in $\mathbb{R}^d$, a unitary tetrahedron in $\mathbb{R}^d$ (see Fig. 4.7). Its vertices are ordered in such a way that the cartesian coordinates of $\mathbf{N}_i$ are all null unless the i-th one that is equal to 1. On a d-dimensional simplex, the space of polynomials $\mathbb{P}_r$ is defined as follows

$$\mathbb{P}_r = \{p(\mathbf{x}) = \sum_{\substack{0 \leq i_1, \ldots, i_d \\ i_1 + \cdots + i_d \leq d}} a_{i_1 \ldots i_d} x_1^{i_1} \ldots x_d^{i_d},\quad a_{i_1 \ldots i_d} \in \mathbb{R}\}. \tag{4.28}$$

Then

$$N_r = \dim \mathbb{P}_r = \binom{r+d}{r} = \frac{1}{d!} \prod_{k=1}^{d} (r+k). \tag{4.29}$$

4.4.3 Barycentric coordinates

For a given simplex K in $\mathbb{R}^d$ (see Sect. 4.5.1) it is sometimes convenient to consider a coordinate frame alternative to the cartesian one, that of the *barycentric coordinates*. The latter are $d+1$ functions, $\{\lambda_0, \ldots, \lambda_d\}$, defined as follows

$$\lambda_i : \mathbb{R}^d \to \mathbb{R}\,,\ \lambda_i(\mathbf{x}) = 1 - \frac{(\mathbf{x} - \mathbf{N}_i) \cdot \mathbf{n}_i}{(\mathbf{N}_j - \mathbf{N}_i) \cdot \mathbf{n}_i},\quad 0 \leq i \leq d. \tag{4.30}$$

For every $i = 0, \ldots, d$ let F_i denote the *face* of K opposite to $\mathbf{N}_i$; F_i is in fact a vertex if $d = 1$, an edge if $d = 2$, a triangle if $d = 3$. In (4.30), $\mathbf{n}_i$ denotes the outward

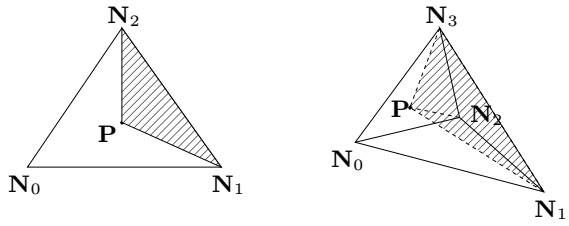


Fig. 4.8. The barycentric coordinate λ_i of the point $\mathbf{P}$ is the ratio $\frac{|K_i|}{|K|}$ between the measure of simplex K_i (whose vertices are $\mathbf{P}$ and $\{\mathbf{P}_j, j \neq i\}$) and that of the given simplex K (a triangle on the left, a tetrahedron on the right). The shadowed simplex is K_0

normal to F_i, while N_j is an arbitrary vertex belonging to F_i. The definition of λ_i is however independent of which vertex of F_i is chosen.

Barycentric coordinates have a geometrical meaning. Indeed, for every point $\mathbf{P}$ belonging to K, its barycentric coordinate λ_i, $0 \leq i \leq d$, represents the ratio between the measure of the simplex K_i whose vertices are $\mathbf{P}$ and the vertices of K sitting on the face F_i opposite to the vertex $\mathbf{P}_i$, and the measure of K. See Fig. 4.8.

Remark 4.4 Let us consider the unitary simplex $\hat{K}_d$, whose vertices $\{\hat{N}_0, \ldots, \hat{N}_d\}$ are ordered in such a way that all the cartesian coordinates of $\mathbf{N}_i$ are null unless x_i which is equal to one. Then

$$\lambda_i = x_i, \quad 1 \leq i \leq d, \quad \lambda_0 = 1 - \sum_{i=1}^{d} \lambda_i \,. \tag{4.31}$$

The barycentric coordinate λ_i is therefore an affine function that is equal to 1 at $\mathbf{N}_i$ and vanishes on the face F_i opposite to $\mathbf{N}_i$.

On a general simplex K in $\mathbb{R}^d$, the following *partition of unity* property is satisfied

$$0 \leq \lambda_i(x) \leq 1\,, \quad \sum_{i=0}^{d} \lambda_i(\mathbf{x}) = 1 \quad \forall \mathbf{x} \in K. \tag{4.32}$$

•

A point $\mathbf{P}$ belonging to the interior of K has therefore all its barycentric coordinates positive. This property is useful whenever one has to check which triangle in 2D or tetrahedron in 3D a given point belongs to, a situation that occurs when using Lagrangian derivatives (see Sect. 15.7.2) or computing suitable quantities (fluxes, streamlines, etc.) as a post-processing of finite element computations.

A noticeable property is that the center of gravity of K has all its barycentric coordinates equal to $(d+1)^{-1}$. Another remarkable property is that

$$\varphi_i = \lambda_i, \quad 0 \leq i \leq d, \tag{4.33}$$

where $\{\varphi_i,\ 0 \le i \le d\}$ are the characteristic Lagrangian functions on the simplex K of degree $r = 1$, that is

$$\varphi_i \in \mathbb{P}_1(K_d), \quad \varphi_i(\mathbf{N}_j) = \delta_{ij}, \quad 0 \le j \le d. \tag{4.34}$$

(See Fig. 4.10, left, for the nodes.)
For $r = 2$ the above identity (4.33) does not hold anymore, however the characteristic Lagrangiang functions $\{\varphi_i\}$ can still be expressed in terms of the barycentric coordinates $\{\lambda_i\}$ as follows:

$$\begin{cases} \varphi_i = \lambda_i(2\lambda_i - 1), & 0 \le i \le d, \\ \varphi_{d+i+j} = 4\lambda_i\lambda_j, & 0 \le i < j \le d. \end{cases} \tag{4.35}$$

For $0 \le i \le d$, φ_i is the characteristic Lagrangian function associated to the vertex $\mathbf{N}_i$, while for $0 \le i < j \le d$, φ_{d+i+j} is the characteristic Lagrangian function associated to the midpoint of the edge whose endpoints are the vertices $\mathbf{N}_i$ and $\mathbf{N}_j$ (see Fig. 4.10, center).

The previous identities justify the name of "coordinates" that is used for the λ_i's. Indeed, if $\mathbf{P}$ is a generic point of the simplex K, its cartesian coordinates $\{x_j^{(P)}, 1 \le j \le d\}$ can be expressed in terms of the barycentric coordinates $\{\lambda_i^{(P)}, 0 \le i \le d\}$ as follows

$$x_j^{(P)} = \sum_{i=0}^{d} \lambda_i^{(P)} x_j^{(i)},\ 1 \le j \le d, \tag{4.36}$$

where $\{x_j^{(i)}, 1 \le j \le d\}$ denote the cartesian coordinates of the i-th vertex $\mathbf{N}_i$ of the simplex K.

4.5 The finite element method in the multi-dimensional case

In this section we extend the finite element method introduced previously for one-dimensional problems to the case of boundary-value problems in multi-dimensional regions. We will also specifically refer to the case of simplices. Many of the presented results are in any case immediately extensible to more general finite elements (see, for instance, [QV94]).

For the sake of simplicity, most often we will consider domains $\Omega \subset \mathbb{R}^2$ with polygonal shape and meshes (or grids) $\mathcal{T}_h$ which represent their coverage with non-overlapping triangles. For this reason, $\mathcal{T}_h$ is also called a triangulation. We refer to Chap. 6 for a more detailed description of the essential features of a generic grid $\mathcal{T}_h$. This way, the discretized domain

$$\Omega_h = int\Big(\bigcup_{K \in \mathcal{T}_h} K\Big)$$

represented by the internal part of the union of the triangles of $\mathcal{T}_h$ perfectly coincides with Ω. We recall that we denote by $int(A)$ the internal part of the set A, that is the

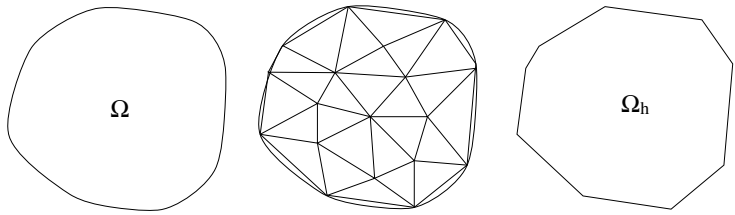


Fig. 4.9. Example of the grid of a non-polygonal domain. The grid induces an approximation Ω_h of the domain Ω such that $\lim_{h\to 0} \text{meas}(\Omega - \Omega_h) = 0$. This issue is not addressed in the present text. The interested reader may consult, for instance, [Cia78] or [SF73]

region obtained by excluding the boundary from A. In fact, we will not discuss the issue relating to the approximation of a non-polygonal domain with a finite element grid (see Fig. 4.9). Hence, from now on we will adopt the symbol Ω to denote without distinction both the computational domain and its (optional) approximation.

Also in the multidimensional case, the h parameter is related to the spacing of the grid. Having set $h_K = \text{diam}(K)$, for each $K \in \mathcal{T}_h$, where $\text{diam}(K) = \max_{x,y\in K} |x-y|$ is the *diameter* of element K, we define $h = \max_{K\in\mathcal{T}_h} h_K$. Moreover, we will impose that the grid satisfy the following *regularity* condition. Let ρ_K be the diameter of the circle inscribed in the triangle K (also called *sphericity* of K); a family of grids $\{\mathcal{T}_h, h > 0\}$ is said to be *regular* if, for a suitable $\delta > 0$, the condition

$$\frac{h_K}{\rho_K} \leq \delta \quad \forall K \in \mathcal{T}_h \tag{4.37}$$

is verified. We observe that condition (4.37) instantly excludes very deformed (i.e. stretched) triangles, and hence the option of using *anisotropic* computational grids.

On the other hand, anisotropic grids are often used in the context of fluid dynamics problems in the presence of boundary layers. See Remark 4.6, and especially references [AFG+00, DV02, FMP04]. Additional details on the generation of grids on two-dimensional domains are provided in Chap. 6.
We denote by $\mathbb{P}_r$ the space of polynomials of global degree less than or equal to r, for $r = 1, 2, \ldots$. According to the general formula (4.28) we find

$$\begin{aligned}
\mathbb{P}_1 &= \{p(x_1,x_2) = a + bx_1 + cx_2, \text{ with } a,b,c \in \mathbb{R}\},\\
\mathbb{P}_2 &= \{p(x_1,x_2) = a + bx_1 + cx_2 + dx_1x_2 + ex_1^2 + fx_2^2, \text{ with } a,b,c,d,e,f \in \mathbb{R}\},\\
&\vdots\\
\mathbb{P}_r &= \{p(x_1,x_2) = \sum_{i,j\geq 0, i+j\leq r} a_{ij}x_1^i x_2^j, \text{ with } a_{ij} \in \mathbb{R}\}.
\end{aligned}$$

According to (4.29), the spaces $\mathbb{P}_r$ have dimension $(r+1)(r+2)/2$. For instance, it results that $\dim\ \mathbb{P}_1 = 3$, $\dim\ \mathbb{P}_2 = 6$ and $\dim\ \mathbb{P}_3 = 10$, hence on every element of the grid $\mathcal{T}_h$ the generic function v_h is well defined whenever its value at 3, 6 resp. 10 suitably chosen nodes, is known (see Fig. 4.10). The nodes for linear $(r = 1)$,

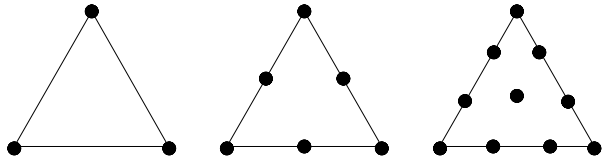


Fig. 4.10. Nodes for linear ($r = 1$, left), quadratic ($r = 2$, center) and cubic ($r = 3$, right) polynomials on a triangle. Such sets of nodes are unisolvent

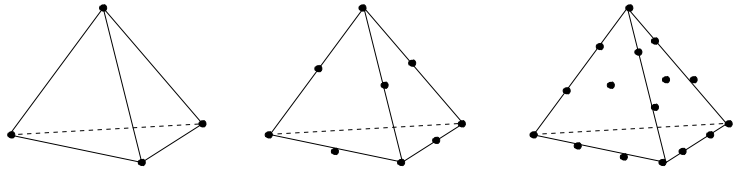


Fig. 4.11. Nodes for linear ($r = 1$, left), quadratic ($r = 2$, center) and cubic ($r = 3$, right) polynomials on a tetrahedron (only those on visible faces are shown)

quadratic $(r = 2)$, and cubic $(r = 3)$ polynomials on a three dimensional simplex are shown in Fig. 4.11.

4.5.1 Finite element solution of the Poisson problem

We introduce the space of finite elements

$$X_h^r = \left\{ v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathbb{P}_r \, \forall K \in \mathcal{T}_h \right\}, \qquad r = 1, 2, \ldots \tag{4.38}$$

that is the space of globally continuous functions that are polynomials of degree r on the single triangles (elements) of the triangulation $\mathcal{T}_h$.
Moreover, we define

$$\overset{\circ}{X}{}_h^r = \{ v_h \in X_h^r : v_h|_{\partial\Omega} = 0 \}. \tag{4.39}$$

The spaces X_h^r e $\overset{\circ}{X}{}_h^r$ are suitable for the approximation of $\mathrm{H}^1(\Omega)$, resp. $\mathrm{H}_0^1(\Omega)$, thanks to the following property (for its proof see, e.g., [QV94]):

Property 4.2 *A sufficient condition for a function v to belong to $\mathrm{H}^1(\Omega)$ is that $v \in C^0(\overline{\Omega})$ and moreover that v belong to $\mathrm{H}^1(K)$ $\forall K \in \mathcal{T}_h$.*

Having set $V_h = \overset{\circ}{X}{}_h^r$, we can introduce the following finite element problem for the approximation of the Poisson problem (3.1) with Dirichlet boundary condition (3.2), in the homogeneous case (that is with $g = 0$)

$$\text{find } u_h \in V_h : \int_\Omega \nabla u_h \cdot \nabla v_h \, d\Omega = \int_\Omega f v_h \, d\Omega \quad \forall \, v_h \in V_h. \tag{4.40}$$

As in the one-dimensional case, each function $v_h \in V_h$ is characterized, univocally, by the values it takes at the nodes $\mathbf{N}_i$, with $i = 1, \ldots, N_h$, of the grid $\mathcal{T}_h$ (excluding the boundary nodes where $v_h = 0$); consequently, a basis in the space V_h can be the set of the characteristic Lagrangian functions $\varphi_j \in V_h$, $j = 1, \ldots, N_h$, such that

$$\varphi_j(\mathbf{N}_i) = \delta_{ij} = \begin{cases} 0 & i \neq j, \\ 1 & i = j, \end{cases} \quad i, j = 1, \ldots, N_h. \tag{4.41}$$

In particular, if $r = 1$, the nodes are vertices of the elements, with the exception of those vertices belonging to the boundary of Ω, while the generic function φ_j is linear on each triangle and is equal to 1 at the node $\mathbf{N}_j$ and 0 at all the other nodes of the triangulation (see Fig. 4.12).

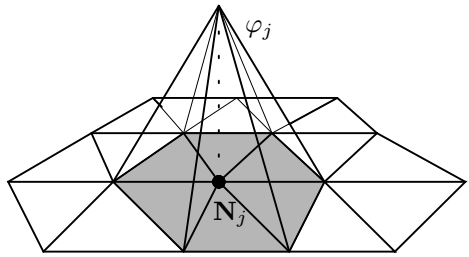


Fig. 4.12. The basis function φ_j of the space X_h^1 and its support

A generic function $v_h \in V_h$ can be expressed through a linear combination of the basis functions of V_h in the following way

$$v_h(\mathbf{x}) = \sum_{i=1}^{N_h} v_i \varphi_i(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega, \text{ with } v_i = v_h(\mathbf{N}_i). \tag{4.42}$$

By expressing the discrete solution u_h in terms of the basis $\{\varphi_j\}$ via (4.42), $u_h(\mathbf{x}) = \sum_{j=1}^{N_h} u_j \varphi_j(\mathbf{x})$, with $u_j = u_h(\mathbf{N}_j)$, and imposing that it verifies (4.40) for each function of the basis itself, we find the following linear system of N_h equations in the N_h unknowns u_j, equivalent to problem (4.40),

$$\sum_{j=1}^{N_h} u_j \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, d\Omega = \int_\Omega f\varphi_i \, d\Omega, \quad i = 1, \ldots, N_h. \tag{4.43}$$

The stiffness matrix has dimensions $N_h \times N_h$ and is defined as

$$\mathrm{A} = [a_{ij}] \quad \text{with} \quad a_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, d\Omega. \tag{4.44}$$

Moreover, we introduce the vectors

$$\mathbf{u} = [u_j] \quad \text{with} \quad u_j = u_h(\mathbf{N}_j), \qquad \mathbf{f} = [f_i] \quad \text{with} \quad f_i = \int_\Omega f\varphi_i \, d\Omega. \quad (4.45)$$

The linear system (4.43) can then be written as

$$\mathrm{A}\mathbf{u} = \mathbf{f}. \quad (4.46)$$

As in the one-dimensional case, the unknowns are the nodal values of the finite element solution. It is evident that, since the *support* of the generic function with basis φ_i is only formed by the triangles having node $\mathbf{N}_i$ in common, A is a sparse matrix. In particular, the number of non-null elements of A is of the order of N_h as a_{ij} is different from zero only if $\mathbf{N}_j$ and $\mathbf{N}_i$ are nodes of the same triangle. It is not guaranteed instead that A has a definite structure (e.g. banded), as that will depend on how the nodes are numbered.

Let us consider now the case of a *non-homogeneous* Dirichlet problem represented by equations (3.1)-(3.2). We have seen in the previous chapter that we can in any case resort to the homogeneous case through a lifting (also called extension, or prolongation) of the boundary datum. In the corresponding discrete problem we build a lifting of a well-chosen approximation of the boundary datum, by proceeding in the following way.
We denote by N_h the internal nodes of the grid $\mathcal{T}_h$ and by N_h^t the total number, thus including the boundary nodes, that for the sake of simplicity we will suppose to be numbered last. The set of boundary nodes will then be formed by $\{\mathbf{N}_i,\ i = N_h + 1, \ldots, N_h^t\}$. A possible approximation g_h of the boundary datum g can be obtained by interpolating g on the space formed by the trace functions on $\partial\Omega$ of functions of X_h^r. This can be written as a linear combination of the traces of the basis functions of X_h^r associated to the boundary nodes

$$g_h(\mathbf{x}) = \sum_{i=N_h+1}^{N_h^t} g(\mathbf{N}_i)\varphi_i\big|_{\partial\Omega}(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Omega. \quad (4.47)$$

Its lifting $R_{g_h} \in X_h^r$ is constructed as follows

$$R_{g_h}(\mathbf{x}) = \sum_{i=N_h+1}^{N_h^t} g(\mathbf{N}_i)\varphi_i(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega. \quad (4.48)$$

In Fig. 4.13, we provide an example of a possible lifting of a non-homogeneous Dirichlet boundary datum (3.2), in the case where g has a non-constant value. The finite element formulation of the Poisson problem then becomes

find $\overset{\circ}{u}_h \in V_h$:

$$\int_\Omega \nabla \overset{\circ}{u}_h \cdot \nabla v_h \, d\Omega = \int_\Omega f v_h \, d\Omega - \int_\Omega \nabla R_{g_h} \cdot \nabla v_h \, d\Omega \quad \forall\, v_h \in V_h. \quad (4.49)$$

The approximate solution will then be provided by $u_h = \overset{\circ}{u}_h + R_{g_h}$.

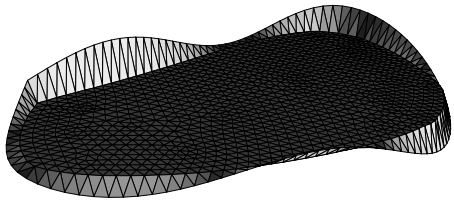


Fig. 4.13. Example of a lifting of a non-homogeneous Dirichlet boundary datum $u = g$, g being variable

Notice that, thanks to the particular lifting we adopted, we can give the following algebraic interpretation to (4.49)

$$\mathrm{A}\mathbf{u} = \mathbf{f} - \mathrm{B}\mathbf{g}$$

where A and $\mathbf{f}$ are defined as in (4.44) and (4.45), now with $u_j = \overset{\circ}{u}_h\ (\mathbf{N}_j)$. Having set $N_h^b = N_h^t - N_h$ (this is the number of boundary nodes), the vector $\mathbf{g} \in \mathbb{R}^{N_h^b}$ and the matrix $\mathrm{B} \in \mathbb{R}^{N_h \times N_h^b}$ have respectively the components

$$g_i = g(\mathbf{N}_{i+N_h}), \quad i = 1, \ldots, N_h^b,$$

$$b_{ij} = \int_\Omega \nabla\varphi_{j+N_h} \cdot \nabla\varphi_i \, d\Omega, \quad i = 1, \ldots, N_h,\ j = 1, \ldots, N_h^b.$$

Remark 4.5 Matrices A and B are both sparse. An efficient program will store exclusively their non-null elements. (See, e.g., [Saa96] for a description of possible storage formats for sparse matrices, and also Chap. 8). In particular, thanks to the special lifting we have adopted, in the B matrix, all the lines corresponding to non-adjacent nodes to a boundary node will be null. (Two grid nodes are said to be adjacent if there exists an element $K \in \mathcal{T}_h$ to which they both belong.) •

4.5.2 Conditioning of the stiffness matrix

We have seen that the stiffness matrix $\mathrm{A} = [a(\varphi_j, \varphi_i)]$ associated to the Galerkin problem and therefore, in particular, to the finite element method, is positive definite; moreover A is symmetric if the bilinear form $a(\cdot, \cdot)$ is symmetric.

For a symmetric and positive definite matrix, its condition number with respect to the norm 2 is given by

$$K_2(\mathrm{A}) = \frac{\lambda_{max}(\mathrm{A})}{\lambda_{min}(\mathrm{A})},$$

$\lambda_{max}(\mathrm{A})$ and $\lambda_{min}(\mathrm{A})$ being the maximum and minimum eigenvalues, respectively, of A.

It can be proven that, both in the one-dimensional and the multi-dimensional case, the following relation holds for the stiffness matrix

$$K_2(\mathrm{A}) = Ch^{-2}, \tag{4.50}$$

where C is a constant independent of the h parameter, but dependent on the degree of the finite elements being used.

To prove (4.50), we recall that the eigenvalues of the matrix A verify the relation

$$\mathrm{A}\mathbf{v} = \lambda_h \mathbf{v},$$

$\mathbf{v}$ being the eigenvector corresponding to the eigenvalue λ_h. Let v_h be the function of the space V_h whose nodal values are the components v_i of $\mathbf{v}$, see (4.7). We suppose $a(\cdot,\cdot)$ to be symmetric, thus A is symmetric and its eigenvalues are real and positive. We then have

$$\lambda_h = \frac{(\mathrm{A}\mathbf{v},\mathbf{v})}{|\mathbf{v}|^2} = \frac{a(v_h,v_h)}{|\mathbf{v}|^2} \tag{4.51}$$

where $|\cdot|$ is the Euclidean vector norm. We suppose that the grid family $\{\mathcal{T}_h, h>0\}$ is regular (i.e. satisfies (4.37)) and moreover is *quasi-uniform*, i.e. such that there exists a constant $\tau > 0$:

$$\min_{K\in\mathcal{T}_h} h_K \geq \tau\, h \qquad \forall h > 0.$$

We now observe that, under the hypotheses made on $\mathcal{T}_h$, the following *inverse inequality* holds (for the proof, refer to [QV94])

$$\exists\, C_I > 0 \quad : \quad \forall v_h \in V_h, \quad \|\nabla v_h\|_{\mathrm{L}^2(\Omega)} \leq C_I\, h^{-1}\, \|v_h\|_{\mathrm{L}^2(\Omega)}, \tag{4.52}$$

the constant C_I being independent of h. We can now prove that there exist two constants C_1, $C_2 > 0$ such that, for each $v_h \in V_h$ as in (4.7), we have

$$C_1\, h^d\, |\mathbf{v}|^2 \leq \|v_h\|^2_{\mathrm{L}^2(\Omega)} \leq C_2\, h^d\, |\mathbf{v}|^2 \tag{4.53}$$

d being the spatial dimension, with $d = 1, 2, 3$. For the proof in the general case we refer to [QV94], Proposition 6.3.1. We here limit ourselves to proving the second inequality in the one-dimensional case ($d = 1$) and for linear finite elements. Indeed, on each element $K_i = [x_{i-1}, x_i]$, we have

$$\int_{K_i} v_h^2(x)\, dx = \int_{K_i} \big(v_{i-1}\varphi_{i-1}(x) + v_i\varphi_i(x)\big)^2\, dx,$$

with φ_{i-1} e φ_i defined according to (4.15). Then, a direct computation shows that

$$\int_{K_i} v_h^2(x)\, dx \leq 2\,\big(v_{i-1}^2 \int_{K_i} \varphi_{i-1}^2(x)\, dx + v_i^2 \int_{K_i} \varphi_i^2(x)\, dx\big) = \frac{2}{3}\, h_i\, \big(v_{i-1}^2 + v_i^2\big)$$

with $h_i = x_i - x_{i-1}$. The inequality

$$\|v_h\|^2_{\mathrm{L}^2(\Omega)} \leq C\, h\, |\mathbf{v}|^2$$

with $C = 4/3$, can be found by simply summing the intervals K and observing that each nodal contribution v_i is counted twice.
On the other hand, from (4.51), we obtain, thanks to the continuity and coercivity of the bilinear form $a(\cdot,\cdot)$,

$$\alpha \frac{\|v_h\|^2_{\mathrm{H}^1(\Omega)}}{|\mathbf{v}|^2} \leq \lambda_h \leq M \frac{\|v_h\|^2_{\mathrm{H}^1(\Omega)}}{|\mathbf{v}|^2},$$

M and α being the continuity and coercivity constant, respectively. Now, $\|v_h\|^2_{\mathrm{H}^1(\Omega)} \geq \|v_h\|^2_{\mathrm{L}^2(\Omega)}$ by the definition of the norm in $\mathrm{H}^1(\Omega)$, while $\|v_h\|_{\mathrm{H}^1(\Omega)} \leq C_3\, h^{-1}\, \|v_h\|_{\mathrm{L}^2(\Omega)}$ (for a well-chosen constant $C_3 > 0$) thanks to (4.52). Thus, by using inequalities (4.53), we obtain

$$\alpha\, C_1\, h^d \leq \lambda_h \leq M\, C_3^2\, C_2\, h^{-2}\, h^d.$$

We therefore have

$$\frac{\lambda_{max}(\mathrm{A})}{\lambda_{min}(\mathrm{A})} \leq \frac{M\, C_3^2\, C_2}{\alpha\, C_1}\, h^{-2}$$

that is (4.50).

When the grid-size h decreases, the condition number of the stiffness matrix increases, and therefore the associated system becomes more and more ill-conditioned. In particular, if the datum $\mathbf{f}$ of the linear system (4.46) is subject to a perturbation $\boldsymbol{\delta}\mathbf{f}$ (i.e. it is affected by error), the latter in turn affects the solution with a perturbation $\boldsymbol{\delta}\mathbf{u}$; it can then be proven that, if there are no perturbations on the matrix A, then

$$\frac{|\boldsymbol{\delta}\mathbf{u}|}{|\mathbf{u}|} \leq K_2(\mathrm{A}) \frac{|\boldsymbol{\delta}\mathbf{f}|}{|\mathbf{f}|}.$$

It is evident that the higher is the conditioning number, the more the solution resents from the perturbation on the data. (On the other hand, notice that the latter is always affected by perturbations on the data caused by the inevitable roundoff errors introduced by the computer.)

As a further example we can study how conditioning affects the solution method. Consider, for instance, solving the linear system (4.46) using the conjugate gradient method (see Chap. 7). Then a sequence $\mathbf{u}^{(k)}$ of approximate solutions is iteratively constructed, converging to the exact solution $\mathbf{u}$. In particular, we have

$$\|\mathbf{u}^{(k)} - \mathbf{u}\|_{\mathrm{A}} \leq 2 \left(\frac{\sqrt{K_2(\mathrm{A})} - 1}{\sqrt{K_2(\mathrm{A})} + 1} \right)^k \|\mathbf{u}^{(0)} - \mathbf{u}\|_{\mathrm{A}},$$

having denoted by $\|\mathbf{v}\|_{\mathrm{A}} = \sqrt{\mathbf{v}^T \mathrm{A} \mathbf{v}}$ the so-called " norm A" of a generic vector $\mathbf{v} \in \mathbb{R}^{N_h}$. If we define

$$\rho = \frac{\sqrt{K_2(\mathrm{A})} - 1}{\sqrt{K_2(\mathrm{A})} + 1},$$

such quantity gives an idea of the convergence rate of the method: the closer ρ is to 0, the faster the method converges, the closer ρ is to 1, the slower the convergence

will be. Indeed, following (4.50), the more accurate one wants to be, by decreasing h, the more ill-conditioned the system will be, and therefore the more "problematic" its solution will turn out to be.

This calls for the system to be preconditioned, i.e. it is necessary to find an invertible matrix P, called *preconditioner*, such that

$$K_2(\mathrm{P}^{-1}\mathrm{A}) \ll K_2(\mathrm{A}),$$

and then to apply the iterative method to the system preconditioned with P (see Chap. 7).

4.5.3 Estimate of the approximation error in the energy norm

Analogously to the one-dimensional case, for each $v \in C^0(\overline{\Omega})$ we define *interpolant* of v in the space of X_h^1, determined by the grid $\mathcal{T}_h$, the function $\Pi_h^1 v$ such that

$$\Pi_h^1 v(\mathbf{N}_i) = v(\mathbf{N}_i) \quad \text{for each node } \mathbf{N}_i \text{ of } \mathcal{T}_h, \text{ for } i = 1, \ldots, N_h.$$

If $\{\varphi_i\}$ is the Lagrangian basis of the space X_h^1, then

$$\Pi_h^1 v(\mathbf{x}) = \sum_{i=1}^{N_h} v(\mathbf{N}_i)\varphi_i(\mathbf{x}).$$

The operator $\Pi_h^1 : C^0(\overline{\Omega}) \to X_h^1$, associating to a continuous function v its interpolant $\Pi_h^1 v$ is called *interpolation operator*.

Analogously, we can define an operator $\Pi_h^r : C^0(\overline{\Omega}) \to X_h^r$, for each integer $r \geq 1$. Having denoted by Π_K^r the local interpolation operator associated to a continuous function v the polynomial $\Pi_K^r v \in \mathbb{P}_r(K)$, interpolating v in the degrees of freedom of the element $K \in \mathcal{T}_h$, we define

$$\Pi_h^r v \in X_h^r \ : \quad \Pi_h^r v\big|_K = \Pi_K^r(v\big|_K) \qquad \forall K \in \mathcal{T}_h. \tag{4.54}$$

We will suppose that $\mathcal{T}_h$ belongs to a family of regular grids of Ω.

In order to obtain an estimate for the approximation error $\|u - u_h\|_V$ we follow a similar procedure to the one used in Theorem 4.3 for the one-dimensional case. The first step is to derive a suitable estimate for the interpolation error. To this end, we will obtain useful information starting from the geometric parameters of each triangle K, i.e. its diameter h_K and sphericity ρ_K. Moreover, we will exploit the affine and invertible transformation $F_K : \widehat{K} \to K$ between the reference triangle $\widehat{K}$ and the generic triangle K (see Fig. 4.14). Such map is defined by $F_K(\hat{\mathbf{x}}) = B_K\hat{\mathbf{x}} + \mathbf{b}_K$, with $B_K \in \mathbb{R}^{2\times 2}$ and $\mathbf{b}_K \in \mathbb{R}^2$, and satisfies the relation $F_K(\widehat{K}) = K$. We recall that the choice of the reference triangle $\widehat{K}$ is not univoque.

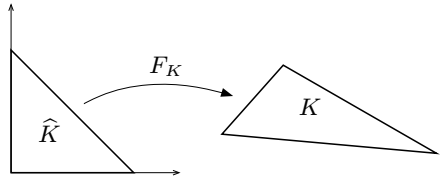


Fig. 4.14. The map F_K between the reference triangle $\widehat{K}$ and the generic triangle K

We will need some preliminary results.

Lemma 4.2 (Transformation of the seminorms) *For each integer $m \geq 0$ and each $v \in \mathrm{H}^m(K)$, let $\hat{v} : \widehat{K} \to \mathbb{R}$ be the function defined by $\hat{v} = v \circ F_K$. Then $\hat{v} \in \mathrm{H}^m(\widehat{K})$. Moreover, there exists a constant $C = C(m) > 0$ such that:*

$$|\hat{v}|_{\mathrm{H}^m(\widehat{K})} \leq C \, \|B_K\|^m \, |\det B_K|^{-\frac{1}{2}} \, |v|_{\mathrm{H}^m(K)}, \tag{4.55}$$

$$|v|_{\mathrm{H}^m(K)} \leq C \, \|B_K^{-1}\|^m \, |\det B_K|^{\frac{1}{2}} \, |\hat{v}|_{\mathrm{H}^m(\widehat{K})}, \tag{4.56}$$

$\|\cdot\|$ being the matrix norm associated to the euclidean vector norm $|\cdot|$, i.e.

$$\|B_K\| = \sup_{\boldsymbol{\xi} \in \mathbb{R}^2, \boldsymbol{\xi} \neq \mathbf{0}} \frac{|B_K \boldsymbol{\xi}|}{|\boldsymbol{\xi}|}. \tag{4.57}$$

Proof. Since $C^m(K) \subset \mathrm{H}^m(K)$ with dense inclusion, for each $m \geq 0$, we can limit ourselves to proving the previous two inequalities for the functions of $C^m(K)$, then extending by density the result to the functions of $\mathrm{H}^m(K)$. The derivatives in the remainders will therefore have to be intended in the classical sense. We recall that

$$|\hat{v}|_{\mathrm{H}^m(\widehat{K})} = \Big(\sum_{|\boldsymbol{\alpha}|=m} \int_{\widehat{K}} |D^{\boldsymbol{\alpha}} \hat{v}|^2 \, d\hat{\mathbf{x}} \Big)^{1/2},$$

by referring to Chap. 2.3 for the definition of the derivative $D^{\boldsymbol{\alpha}}$. By using the chain rule for the differentiation of composite functions, we obtain

$$\|D^{\boldsymbol{\alpha}} \hat{v}\|_{\mathrm{L}^2(\widehat{K})} \leq C \|B_K\|^m \sum_{|\boldsymbol{\beta}|=m} \|(D^{\boldsymbol{\beta}} v) \circ F_K\|_{\mathrm{L}^2(\widehat{K})}.$$

Then

$$\|D^{\boldsymbol{\alpha}} \hat{v}\|_{\mathrm{L}^2(\widehat{K})} \leq C \|B_K\|^m |\det B_K|^{-\frac{1}{2}} \, \|D^{\boldsymbol{\alpha}} v\|_{\mathrm{L}^2(K)}.$$

Inequality (4.55) follows after summing on the multi-index $\boldsymbol{\alpha}$, for $|\boldsymbol{\alpha}| = m$. The result (4.56) can be proven by proceeding in a similar way. ◇

Lemma 4.3 (Estimates for the norms $\|B_K\|$ and $\|B_K^{-1}\|$) *We have the following upper bounds:*

$$\|B_K\| \leq \frac{h_K}{\hat{\rho}}, \tag{4.58}$$

$$\|B_K^{-1}\| \leq \frac{\hat{h}}{\rho_K}, \tag{4.59}$$

$\hat{h}$ and $\hat{\rho}$ being the diameter and the sphericity of the reference triangle $\widehat{K}$.

Proof. Thanks to (4.57) we have

$$\|B_K\| = \frac{1}{\hat{\rho}} \sup_{\boldsymbol{\xi}\in\mathbb{R}^2, |\boldsymbol{\xi}|=\hat{\rho}} |B_K\boldsymbol{\xi}|.$$

For each $\boldsymbol{\xi}$, with $|\boldsymbol{\xi}| = \hat{\rho}$, we can find two points $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}} \in \widehat{K}$ such that $\hat{\mathbf{x}} - \hat{\mathbf{y}} = \boldsymbol{\xi}$. Since $B_K\boldsymbol{\xi} = F_K(\hat{\mathbf{x}}) - F_K(\hat{\mathbf{y}})$, we have $|B_K\boldsymbol{\xi}| \leq h_K$, that is (4.58).
An analogous procedure leads to the result (4.59). ◇

What we now need is an estimate in $\mathrm{H}^m(\widehat{K})$ of the seminorm of $(v - \Pi_K^r v)\circ F_K$, for each function v of $\mathrm{H}^m(K)$. In the remainder, we denote the interpolant $\Pi_K^r v \circ F_K$ with $[\Pi_K^r v]\widehat{\ }$. The nodes of K are $\mathbf{N}_i^K = F_K(\widehat{\mathbf{N}}_i)$, $\widehat{\mathbf{N}}_i$ being the nodes of $\widehat{K}$, and, analogously, the basis functions $\hat{\varphi}_i$ defined on $\widehat{K}$ are identified by the relation $\hat{\varphi}_i = \varphi_i^K \circ F_K$, having denoted by φ_i^K the basis functions associated to the element K. Thus,

$$[\Pi_K^r v]\widehat{\ } = \Pi_K^r v \circ F_K = \sum_{i=1}^{M_K} v(\mathbf{N}_i^K)\varphi_i^K \circ F_K = \sum_{i=1}^{M_K} v(F_K(\widehat{\mathbf{N}}_i))\hat{\varphi}_i = \Pi_{\widehat{K}}^r \hat{v},$$

M_K being the number of nodes on K determined by the choice made for the degree r. It then follows that

$$|(v - \Pi_K^r v)\circ F_K|_{\mathrm{H}^m(\widehat{K})} = |\hat{v} - \Pi_{\widehat{K}}^r \hat{v}|_{\mathrm{H}^m(\widehat{K})}. \tag{4.60}$$

In order to estimate the second member of the previous equality, we start by proving the following result:

Lemma 4.4 (Bramble-Hilbert Lemma) *Let $\widehat{L} : \mathrm{H}^{r+1}(\widehat{K}) \to \mathrm{H}^m(\widehat{K})$, with $m \geq 0$ and $r \geq 0$, be a linear and continuous transformation such that*

$$\widehat{L}(\hat{p}) = 0 \qquad \forall \hat{p} \in \mathbb{P}_r(\widehat{K}). \tag{4.61}$$

Then, for each $\hat{v} \in \mathrm{H}^{r+1}(\widehat{K})$, we have

$$|\widehat{L}(\hat{v})|_{\mathrm{H}^m(\widehat{K})} \leq \|\widehat{L}\|_{\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}),\mathrm{H}^m(\widehat{K}))} \inf_{\hat{p}\in\mathbb{P}_r(\widehat{K})} \|\hat{v} + \hat{p}\|_{\mathrm{H}^{r+1}(\widehat{K})}, \tag{4.62}$$

where $\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}), \mathrm{H}^m(\widehat{K}))$ denotes the space of linear and continuous transformations $l : \mathrm{H}^{r+1}(\widehat{K}) \to \mathrm{H}^m(\widehat{K})$ the norm of which is

$$\|l\|_{\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}),\mathrm{H}^m(\widehat{K}))} = \sup_{v\in\mathrm{H}^{r+1}(\widehat{K}),v\neq 0} \frac{\|l(v)\|_{\mathrm{H}^m(\widehat{K}))}}{\|v\|_{\mathrm{H}^{r+1}(\widehat{K})}}. \tag{4.63}$$

Proof. Let $\hat{v} \in \mathrm{H}^{r+1}(\widehat{K})$. For each $\hat{p} \in \mathbb{P}_r(\widehat{K})$, thanks to (4.61) and to the norm definition (4.63), we obtain

$$|\widehat{L}(\hat{v})|_{\mathrm{H}^m(\widehat{K})} = |\widehat{L}(\hat{v} + \hat{p})|_{\mathrm{H}^m(\widehat{K})} \leq \|\widehat{L}\|_{\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}),\mathrm{H}^m(\widehat{K}))} \|\hat{v} + \hat{p}\|_{\mathrm{H}^{r+1}(\widehat{K})}.$$

The result (4.62) can be deduced thanks to the fact that $\hat{p}$ is arbitrary. ◇

The following result (whose proof is given, e.g., in [QV94, Chap. 3]) provides the last necessary tool to obtain the estimate for the interpolation error that we are seeking.

Lemma 4.5 (Deny-Lions Lemma) *For each $r \geq 0$, there exists a constant $C = C(r, \widehat{K})$ such that*

$$\inf_{\hat{p}\in\mathbb{P}_r} \|\hat{v} + \hat{p}\|_{\mathrm{H}^{r+1}(\widehat{K})} \leq C\, |\hat{v}|_{\mathrm{H}^{r+1}(\widehat{K})} \qquad \forall \hat{v} \in \mathrm{H}^{r+1}(\widehat{K}). \tag{4.64}$$

As a consequence of the two previous lemmas, we can provide the following

Corollary 4.3 *Let $\widehat{L} : \mathrm{H}^{r+1}(\widehat{K}) \to \mathrm{H}^m(\widehat{K})$, with $m \geq 0$ and $r \geq 0$, be a linear and continuous transformation such that $\widehat{L}(\hat{p}) = 0\ \forall \hat{p} \in \mathbb{P}_r(\widehat{K})$. Then, there exists a constant $C = C(r, \widehat{K})$ such that, for each $\hat{v} \in \mathrm{H}^{r+1}(\widehat{K})$, we have*

$$|\widehat{L}(\hat{v})|_{\mathrm{H}^m(\widehat{K})} \leq C\, \|\widehat{L}\|_{\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}),\mathrm{H}^m(\widehat{K}))}\, |\hat{v}|_{\mathrm{H}^{r+1}(\widehat{K})}. \tag{4.65}$$

We are now able to prove the sought interpolation error estimate.

Theorem 4.4 (Local estimate of the interpolation error) *Let $r \geq 1$ and $0 \leq m \leq r+1$. Then, there exists a constant $C = C(r, m, \widehat{K}) > 0$ such that*

$$|v - \Pi_K^r v|_{\mathrm{H}^m(K)} \leq C \frac{h_K^{r+1}}{\rho_K^m} |v|_{\mathrm{H}^{r+1}(K)} \qquad \forall v \in \mathrm{H}^{r+1}(K). \tag{4.66}$$

Proof. From Property 2.3 we derive first of all that $\mathrm{H}^{r+1}(K) \subset C^0(K)$, for $r \geq 1$. The interpolation operator Π_K^r thus results to be well defined in $\mathrm{H}^{r+1}(K)$. By using, in order, the results (4.56), (4.60), (4.59) and (4.65), we have

$$\begin{aligned}
|v - \Pi_K^r v|_{\mathrm{H}^m(K)} &\leq C_1 \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} |\hat{v} - \Pi_{\widehat{K}}^r \hat{v}|_{\mathrm{H}^m(\widehat{K})} \\
&\leq C_1 \frac{\hat{h}^m}{\rho_K^m} |\det B_K|^{\frac{1}{2}} |\underbrace{\hat{v} - \Pi_{\widehat{K}}^r \hat{v}}_{\widehat{L}(\hat{v})}|_{\mathrm{H}^m(\widehat{K})} \\
&\leq C_2 \frac{\hat{h}^m}{\rho_K^m} |\det B_K|^{\frac{1}{2}} \|\widehat{L}\|_{\mathcal{L}(\mathrm{H}^{r+1}(\widehat{K}), \mathrm{H}^m(\widehat{K}))} |\hat{v}|_{\mathrm{H}^{r+1}(\widehat{K})} \\
&= C_3 \frac{1}{\rho_K^m} |\det B_K|^{\frac{1}{2}} |\hat{v}|_{\mathrm{H}^{r+1}(\widehat{K})},
\end{aligned}$$

$C_1 = C_1(m)$, $C_2 = C_2(r, m, \widehat{K})$ and $C_3 = C_3(r, m, \widehat{K})$ being suitably chosen constants. We note that the result (4.65) has been applied by identifying $\widehat{L}$ with the operator $I - \Pi_{\widehat{K}}^r$, with $(I - \Pi_{\widehat{K}}^r)\hat{p} = 0$, for $\hat{p} \in \mathbb{P}_r(\widehat{K})$. Moreover the quantity $\hat{h}^m$ and the norm of the operator $\widehat{L}$ have been included in the constant C_3.
At this point, by applying (4.55) and (4.58) we obtain the result (4.66), that is

$$|v - \Pi_K^r v|_{\mathrm{H}^m(K)} \leq C_4 \frac{1}{\rho_K^m} \|B_K\|^{r+1} |v|_{\mathrm{H}^{r+1}(K)} \leq C_5 \frac{h_K^{r+1}}{\rho_K^m} |v|_{\mathrm{H}^{r+1}(K)}, \tag{4.67}$$

$C_4 = C_4(r, m, \widehat{K})$ and $C_5 = C_5(r, m, \widehat{K})$ being two well-chosen constants. The quantity $\hat{\rho}^{r+1}$ generated by (4.58) and relating to the sphericity of the reference element has been directly included in the constant C_5. ◇

Finally, we can prove the global estimate for the interpolation error:

Theorem 4.5 (Global estimate for the interpolation error) *Let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular grids of the domain Ω and let $m = 0, 1$ and $r \geq 1$. Then, there exists a constant $C = C(r, m, \widehat{K}) > 0$ such that*

$$|v - \Pi_h^r v|_{\mathrm{H}^m(\Omega)} \leq C \Big(\sum_{K \in \mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{\mathrm{H}^{r+1}(K)} \Big)^{1/2} \qquad \forall v \in \mathrm{H}^{r+1}(\Omega). \tag{4.68}$$

In particular, we obtain

$$|v - \Pi_h^r v|_{\mathrm{H}^m(\Omega)} \leq C\, h^{r+1-m} |v|_{\mathrm{H}^{r+1}(\Omega)} \qquad \forall v \in \mathrm{H}^{r+1}(\Omega). \tag{4.69}$$

Proof. Thanks to (4.66) and to the regularity condition (4.37), we have

$$\begin{aligned} |v - \Pi_h^r v|^2_{\mathrm{H}^m(\Omega)} &= \sum_{K \in \mathcal{T}_h} |v - \Pi_K^r v|^2_{\mathrm{H}^m(K)} \\ &\leq C_1 \sum_{K \in \mathcal{T}_h} \left(\frac{h_K^{r+1}}{\rho_K^m} \right)^2 |v|^2_{\mathrm{H}^{r+1}(K)} \\ &= C_1 \sum_{K \in \mathcal{T}_h} \left(\frac{h_K}{\rho_K} \right)^{2m} h_K^{2(r+1-m)} |v|^2_{\mathrm{H}^{r+1}(K)} \\ &\leq C_1\, \delta^{2m} \sum_{K \in \mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{\mathrm{H}^{r+1}(K)}, \end{aligned}$$

i.e. (4.68), with $C_1 = C_1(r, m, \widehat{K})$ and $C = C_1\, \delta^{2m}$. (4.69) follows thanks to the fact that $h_K \leq h$, for each $K \in \mathcal{T}_h$, and that

$$|v|_{\mathrm{H}^p(\Omega)} = \Big(\sum_{K \in \mathcal{T}_h} |v|^2_{\mathrm{H}^p(K)} \Big)^{1/2},$$

for each integer $p \geq 0$. ◇

In the $m = 0$ case, regularity of the grid is not necessary to obtain the estimate (4.69). This is no longer true for $m = 1$. As a matter of fact, given a triangle K and a function $v \in \mathrm{H}^{r+1}(K)$, with $r \geq 1$, it can be proven that the following inequality holds [QV94],

$$|v - \Pi_h^r v|_{\mathrm{H}^m(K)} \leq \tilde{C} \frac{h_K^{r+1}}{\rho_K^m} |v|_{\mathrm{H}^{r+1}(K)}, \quad m = 0, 1,$$

with $\tilde{C}$ independent of v and $\mathcal{T}_h$. Hence, in the $m = 1$ case for a family of regular grids we obtain (4.69) by setting $C = \delta \tilde{C}$, δ being the constant appearing in (4.37).

On the other hand, the need for a regularity condition can be proven by considering the particular case where, for each $C > 0$, a (non-regular) grid can be constructed for which inequality (4.69) is not true, as we are about to prove in the following example which relates to the case $r = 1$.

Example 4.1 Consider the triangle K_l illustrated in Fig. 4.15, with vertices $(0, 0)$, $(1, 0)$, $(0.5, l)$, with $l \leq \frac{\sqrt{3}}{2}$, and the function $v(x_1, x_2) = x_1^2$. Clearly $v \in \mathrm{H}^2(K_l)$ and its interpolant on K_l is given by $\Pi_h^1 v(x_1, x_2) = x_1 - (4l)^{-1} x_2$. Since in this case $h_{K_l} = 1$, the inequality (4.69), applied to the single triangle K_l, would yield

$$|v - \Pi_h^1 v|_{\mathrm{H}^1(K_l)} \leq C |v|_{\mathrm{H}^2(K_l)}. \tag{4.70}$$

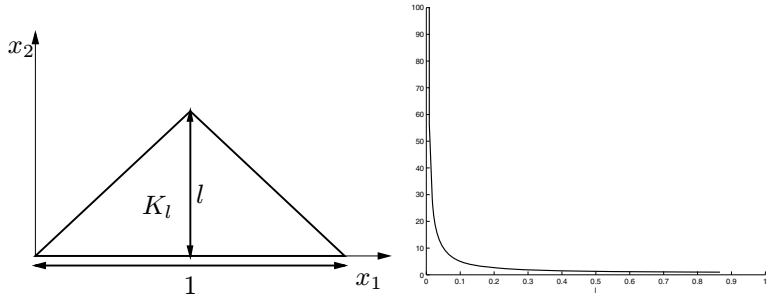


Fig. 4.15. The triangle K_l (left) and the behavior of the relation $|v - \Pi_h^1 v|_{\mathrm{H}^1(K_l)} / |v|_{\mathrm{H}^2(K_l)}$ as a function of l (right)

Let us now consider the behavior of the relation

$$\eta_l = \frac{|v - \Pi_h^1 v|_{\mathrm{H}^1(K_l)}}{|v|_{\mathrm{H}^2(K_l)}}$$

when l tends to zero, that is when the triangle is squeezed. We note that allowing l to tend to zero is equivalent to violating the regularity condition (4.37) as, for small enough values of l, $h_{K_l} = 1$, while, denoting by p_{K_l} the perimeter of K_l and by $|K_l|$ we denote the surface of the element K_l, the sphericity of K_l

$$\rho_{K_l} = \frac{4|K_l|}{p_{K_l}} = \frac{2l}{1 + \sqrt{1 + 4l^2}}$$

tends to zero. We have

$$\eta_l \geq \frac{\|\partial_{x_2}(v - \Pi_h^1 v)\|_{\mathrm{L}^2(K_l)}}{|v|_{\mathrm{H}^2(K_l)}} = \left(\frac{\int_{K_l} \left(\frac{1}{4l}\right)^2 \, d\mathbf{x}}{2l} \right)^{\frac{1}{2}} = \frac{1}{8l}.$$

Hence $\lim_{l \to 0} \eta_l = +\infty$ (see Fig. 4.15). Consequently, there cannot exist a constant C, independent of $\mathcal{T}_h$, for which (4.70) holds.

■

The theorem on interpolation error estimate immediately provides us with an estimate of the approximation error of the Galerkin method. The proof is analogous to that of Theorem 4.3 for the one-dimensional case. Indeed, it is sufficient to apply (4.10) and Theorem 4.5 (for $m = 1$) to obtain the following error estimate:

Theorem 4.6 *Let $u \in V$ be the exact solution of the variational problem* (4.1) *and u_h its approximate solution using the finite element method of degree r. If $u \in \mathrm{H}^{r+1}(\Omega)$, then the following* a priori *error estimates hold:*

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq \frac{M}{\alpha} C \Big(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|^2_{\mathrm{H}^{r+1}(K)} \Big)^{1/2}, \tag{4.71}$$

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq \frac{M}{\alpha} C h^r |u|_{\mathrm{H}^{r+1}(\Omega)}, \tag{4.72}$$

C being a constant independent of h and u.

Also in the multi-dimensional case, in order to increase the accuracy two different strategies can therefore be followed:

1. decreasing h, i.e. refine the grid;
2. increasing r, i.e. use finite elements of higher degree.

However, the latter approach can only be pursued if the solution u is regular enough. In general, we can say that, if $u \in \mathrm{H}^{p+1}(\Omega)$ for some $p > 0$, then

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq C h^s |u|_{\mathrm{H}^{s+1}(\Omega)}, \quad s = \min\{r, p\}, \tag{4.73}$$

as already observed in the one-dimensional case (see (4.26)). Moreover, it is possible to prove an error estimate in the maximum norm. For instance, if $r = 1$, one has

$$\|u - u_h\|_{L^\infty(\Omega)} \leq C h^2 |\log h| |u|_{W^{2,\infty}(\Omega)}$$

where C is a positive constant independent of h and the last term on the right hand side is the seminorm of u in the Sobolev space $W^{2,\infty}(\Omega)$ (see Sect. 2.5). For the proof of this and other error estimates in $W^{k,\infty}(\Omega)$-norms see, e.g., [Cia78] and [BS94].

Remark 4.6 (Case of anisotropic grids) The interpolation error estimate (4.66) (and the consequent discretization error estimate) can be generalized in the case of *anisotropic grids*. In such case however, the left term of (4.66) takes a more complex expression: these estimates, in fact, because of their *directional* nature, must take into account information coming from characteristic directions associated to the single triangles which replace the "global" information concentrated in the seminorm $|v|_{\mathrm{H}^{r+1}(K)}$. The interested reader can consult [Ape99, FP01]. Moreover, we refer to Fig. 4.18 and 11.14 for examples of anisotropic grids. •

4.5.4 Estimate of the approximation error in the L^2 norm

The inequality (4.72) provides an estimate of the approximation error in the energy norm. Analogously, it is possible to obtain an error estimate in the L^2 norm. Since the latter norm is weaker than the previous one, one must expect a higher convergence rate with respect to h.

Lemma 4.6 (Elliptic regularity) *Consider the homogeneous Dirichlet problem*

$$\begin{cases} -\Delta w = g & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega, \end{cases}$$

with $g \in \mathrm{L}^2(\Omega)$. If $\partial\Omega$ is sufficiently regular (for instance, if $\partial\Omega$ is a curve of class C^2, or else if Ω is a convex polygon), then $w \in \mathrm{H}^2(\Omega)$ and moreover there exists a constant $C > 0$ such that

$$\|w\|_{\mathrm{H}^2(\Omega)} \leq C\|g\|_{\mathrm{L}^2(\Omega)}. \tag{4.74}$$

For the proof see, e.g., [Bre86, Gri76].

Theorem 4.7 *Let $u \in V$ be the exact solution of the variational problem* (4.1) *and u_h its approximate solution obtained with the finite element method of degree r. Moreover, let $u \in \mathrm{H}^{p+1}(\Omega)$ for a given $p > 0$. Then, the following* a priori *error estimate in the norm of $\mathrm{L}^2(\Omega)$ holds*

$$\|u - u_h\|_{\mathrm{L}^2(\Omega)} \leq Ch^{s+1}|u|_{\mathrm{H}^{s+1}(\Omega)}, \quad s = \min\{r, p\}, \tag{4.75}$$

C being a constant independent of h and u.

Proof. We will limit ourselves to proving this result for the Poisson problem (3.13), the weak formulation of which is given in (3.18). Let $e_h = u - u_h$ be the approximation error and consider the following auxiliary Poisson problem (called *adjoint problem*, see Sec. 3.6) with known term given by the error function e_h

$$\begin{cases} -\Delta\phi = e_h & \text{in } \Omega, \\ \phi = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.76}$$

whose weak formulation is

$$\text{find } \phi \in V: \quad a(\phi, v) = \int_\Omega e_h v \, d\Omega \qquad \forall\, v \in V, \tag{4.77}$$

with $V = \mathrm{H}_0^1(\Omega)$. Taking $v = e_h$ $(\in V)$, we have

$$\|e_h\|^2_{\mathrm{L}^2(\Omega)} = a(\phi, e_h).$$

Since the bilinear form is symmetric, by the Galerkin orthogonality (4.8) we have

$$a(e_h, \phi_h) = a(\phi_h, e_h) = 0 \qquad \forall\, \phi_h \in V_h.$$

It follows that

$$\|e_h\|^2_{\mathrm{L}^2(\Omega)} = a(\phi, e_h) = a(\phi - \phi_h, e_h). \tag{4.78}$$

Now, taking $\phi_h = \Pi^1_h \phi$, applying the Cauchy-Schwarz inequality to the bilinear form $a(\cdot, \cdot)$ and using the interpolation error estimate (4.69) we obtain

$$\|e_h\|^2_{\mathrm{L}^2(\Omega)} \leq |e_h|_{\mathrm{H}^1(\Omega)} |\phi - \phi_h|_{\mathrm{H}^1(\Omega)} \leq |e_h|_{\mathrm{H}^1(\Omega)} Ch |\phi|_{\mathrm{H}^2(\Omega)}. \tag{4.79}$$

Notice that the interpolation operator Π^1_h can be applied to ϕ as, thanks to Lemma 4.6, $\phi \in \mathrm{H}^2(\Omega)$ and thus, in particular, $\phi \in C^0(\overline{\Omega})$, thanks to property 2.3 in Chap. 2.
By applying Lemma 4.6 to the adjoint problem (4.76) we obtain the inequality

$$|\phi|_{\mathrm{H}^2(\Omega)} \leq C \|e_h\|_{\mathrm{L}^2(\Omega)}, \tag{4.80}$$

which, applied to (4.79), eventually provides

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

where C accounts for all the constants which appeared so far. By now exploiting the error estimate in the energy norm (4.72), we obtain (4.75). ◇

Let us generalize the result we have just proven for the Poisson problem to the case of a generic elliptic boundary-value problem approximated with finite elements and for which an estimate of the approximation error in the energy norm such as (4.72) holds, and so does an elliptic regularity property analogous to the one expressed in Lemma 4.6.
In particular, let us consider the case where the bilinear form $a(\cdot, \cdot)$ is not necessarily symmetric. Let u be the exact solution of the problem

$$\text{find } u \in V: \quad a(u, v) = (f, v) \quad \forall v \in V, \tag{4.81}$$

and u_h the solution of the Galerkin problem

$$\text{find } u_h \in V_h: \quad a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h.$$

Finally, suppose that the error estimate (4.72) holds and let us consider the following problem, which we will call *adjoint problem* of (4.81): for each $g \in \mathrm{L}^2(\Omega)$,

$$\text{find } \phi = \phi(g) \in V: \quad a^*(\phi, v) = (g, v) \quad \forall v \in V, \tag{4.82}$$

where we have defined (see (3.40))

$$a^* : V \times V \to R,\ a^*(w, v) = a(v, w)\ \forall w, v \in V. \tag{4.83}$$

Obviously, should a be symmetric, the two problems coincide, as seen for instance in the case of problem (4.77).

Notice that the unknown is now the second argument of $a(\cdot,\cdot)$, while in the primal problem (4.81) the unknown is the first argument of $a(\cdot,\cdot)$. Let us suppose that for the solution u of the primal problem (4.81) an elliptic regularity result holds; it can then be verified that the same result is valid for the adjoint problem (4.82), that is

$$\exists\, C > 0 \ : \qquad \|\phi(g)\|_{\mathrm{H}^2(\Omega)} \leq C\, \|g\|_{\mathrm{L}^2(\Omega)} \quad \forall g \in \mathrm{L}^2(\Omega).$$

In particular, this is true for a generic elliptic problem with Dirichlet or Neumann (but not mixed) data on a polygonal and convex domain Ω [Gri76]. We now choose $g = e_h$ and denote, for simplicity, $\phi = \phi(e_h)$. Furthermore, having chosen $v = e_h$, we have

$$\|e_h\|^2_{\mathrm{L}^2(\Omega)} = a(e_h, \phi).$$

Since, by the elliptic regularity of the adjoint problem, $\phi \in \mathrm{H}^2(\Omega)$ and $\|\phi\|_{\mathrm{H}^2(\Omega)} \leq C\|e_h\|_{\mathrm{L}^2(\Omega)}$ thanks to the Galerkin orthogonality, we have that

$$\begin{aligned}
\|e_h\|^2_{\mathrm{L}^2(\Omega)} &= a(e_h, \phi) = a(e_h, \phi - \Pi^1_h\phi) \\
&\leq C_1\|e_h\|_{\mathrm{H}^1(\Omega)}\, \|\phi - \Pi^1_h\phi\|_{\mathrm{H}^1(\Omega)} \\
&\leq C_2\|e_h\|_{\mathrm{H}^1(\Omega)}\, h\, \|\phi\|_{\mathrm{H}^2(\Omega)} \\
&\leq C_3\|e_h\|_{\mathrm{H}^1(\Omega)}\, h\, \|e_h\|_{\mathrm{L}^2(\Omega)},
\end{aligned}$$

where we have exploited the continuity of the form $a(\cdot,\cdot)$ and the estimate (4.72). Thus

$$\|e_h\|_{\mathrm{L}^2(\Omega)} \leq C_3 h \|e_h\|_{\mathrm{H}^1(\Omega)},$$

from which (4.75) follows, using the estimate (4.73) of the error in $\mathrm{H}^1(\Omega)$.

Remark 4.7 The technique illustrated above, depending upon the use of the adjoint problem for the estimate of the L^2-norm of the discretization error, is known in the literature as *Aubin-Nitsche trick* [Aub67, Nit68]. Several examples of the determination of the adjoint of a given problem will be presented in Sec. 3.6. •

Example 4.2 We consider the model problem $-\Delta u + u = f$ in $\Omega = (0,1)^2$ with $u = g$ on $\partial\Omega$. Suppose to choose the known term f and the function g so that the exact solution of the problem is $u(x,y) = \sin(2\pi x)\cos(2\pi y)$. We solve such a problem with the Galerkin method with finite elements of degree 1 and 2 on a uniform grid with stepsize h. The graph of Fig. 4.16 shows the behavior of the error when the grid-size h decreases, both in the norm $\mathrm{L}^2(\Omega)$ and in that of $\mathrm{H}^1(\Omega)$. As shown by inspecting the slope of the lines in the figure, the error decrease when using L^2 norm (crossed lines) is quadratic if linear finite elements are used (solid line), and cubic when quadratic finite elements are used (etched line).

With respect to the H^1 norm (lines without crosses) instead, there is a linear reduction of the erorr with respect to the linear finite elements (solid line), and quadratic when quadratic finite elements are used (etched line). Fig. 4.17 shows the solution on the grid with grid-size 1/8 obtained with linear (left) and quadratic (right) finite elements. ■

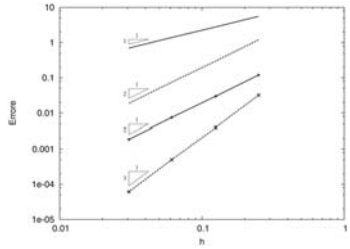


Fig. 4.16. Behavior with respect to h of the error in $\mathrm{H}^1(\Omega)$ norm (lines without crosses) and in $\mathrm{L}^2(\Omega)$ norm (lines with crosses) for linear (solid lines) and quadratic (etched lines) finite elements for the solution of the problem reported in Example 4.2

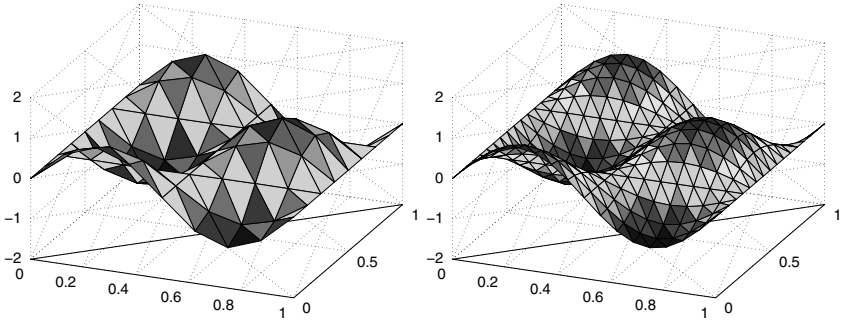
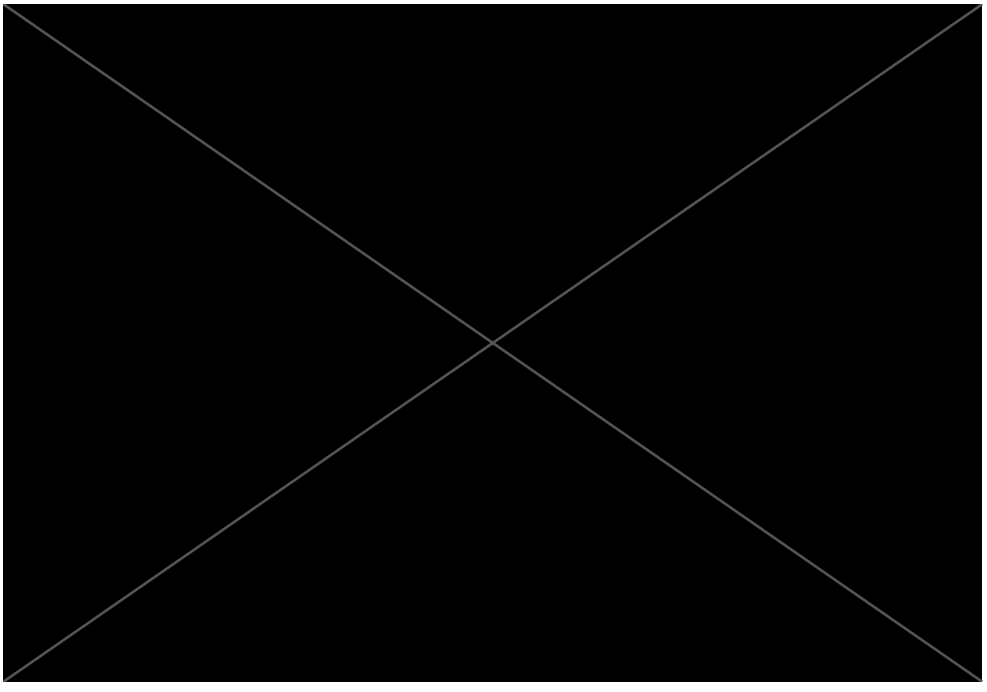


Fig. 4.17. Solutions computed using piecewise linear (left) and piecewise quadratic (right) finite elements on a uniform grid with grid-size 1/8



5

Parabolic equations

In this chapter, we consider parabolic equations of the form

$$\frac{\partial u}{\partial t} + Lu = f, \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{5.1}$$

where Ω is a domain of $\mathbb{R}^d$, $d = 1, 2, 3$, $f = f(\mathbf{x}, t)$ is a given function, $L = L(\mathbf{x})$ is a generic elliptic operator acting on the unknown $u = u(\mathbf{x}, t)$. When solved only for a bounded temporal interval, say for $0 < t < T$, the region $Q_T = \Omega \times (0, T)$ is called *cylinder* in the space $\mathbb{R}^d \times \mathbb{R}^+$ (see Fig. 5.1). In the case where $T = +\infty$, $Q = \{(\mathbf{x}, t) : \mathbf{x} \in \Omega, t > 0\}$ will be an infinite cylinder.
Equation (5.1) must be completed by assigning an initial condition

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{5.2}$$

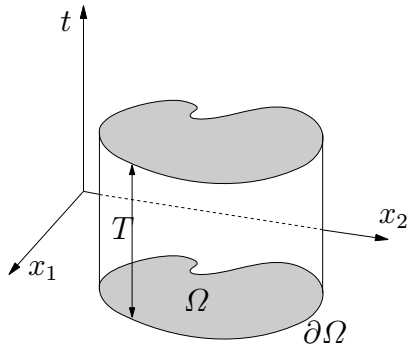


Fig. 5.1. The cylinder $Q_T = \Omega \times (0, T)$, $\Omega \subset \mathbb{R}^2$

A. Quarteroni: Numerical Models for Differential Problems.

together with boundary conditions, which can take the following form

$$\begin{aligned} u(\mathbf{x},t) &= \varphi(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_D \text{ and } t>0, \\ \frac{\partial u(\mathbf{x},t)}{\partial n} &= \psi(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_N \text{ and } t>0, \end{aligned} \tag{5.3}$$

where u_0, φ and ψ are given functions and $\{\Gamma_D, \Gamma_N\}$ provides a boundary partition, that is $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\overset{\circ}{\Gamma}_D \cap \overset{\circ}{\Gamma}_N = \emptyset$. For obvious reasons, Γ_D is called Dirichlet boundary and Γ_N Neumann boundary.
In the one-dimensional case, the problem

$$\begin{aligned} &\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = f, \quad 0<x<d, \quad t>0, \\ &u(x,0) = u_0(x), \quad 0<x<d, \\ &u(0,t) = u(d,t) = 0, \quad t>0, \end{aligned} \tag{5.4}$$

describes the evolution of the temperature $u(x,t)$ at point x and time t of a metallic bar of length d occupying the interval $[0,d]$, whose thermal conductivity is ν and whose extrema are kept at a constant temperature of zero degrees. The function u_0 describes the initial temperature, while f represents the caloric production (per unit length) provided by the bar. For this reason, (5.4) is called *heat equation*. For a particular case, see Example 1.5 of Chap. 1.

5.1 Weak formulation and its approximation

In order to numerically solve problem (5.1)-(5.3), we will introduce a weak formulation, as we did to handle elliptic problems.
We proceed formally, by multiplying for each $t>0$ the differential equation by a test function $v = v(\mathbf{x})$ and integrating on Ω. We set $V = \mathrm{H}^1_{\Gamma_D}(\Omega)$ (see (3.26)) and for each $t>0$ we seek $u(t) \in V$ s.t.

$$\int_\Omega \frac{\partial u(t)}{\partial t} v \, d\Omega + a(u(t),v) = \int_0^t f(t) v d\Omega \qquad \forall v \in V, \tag{5.5}$$

with $u(0) = u_0$, where $a(\cdot,\cdot)$ is the bilinear form associated to the elliptic operator L, and where we have supposed for simplicity $\varphi = 0$ and $\psi = 0$. The modification of (5.5) in the case where $\varphi \neq 0$ and $\psi \neq 0$ is left to the reader.
A sufficient condition for the existence and uniqueness of the solution to problem (5.5) is that the following hypotheses hold:
the bilinear form $a(\cdot,\cdot)$ is continuous and *weakly coercive*, that is

$$\exists \lambda \geq 0, \ \exists \alpha > 0 : \quad a(v,v) + \lambda \|v\|^2_{\mathrm{L}^2(\Omega)} \geq \alpha \|v\|^2_V \quad \forall v \in V,$$

so we find again for $\lambda = 0$ the standard definition of coercivity. Moreover, we require $u_0 \in \mathrm{L}^2(\Omega)$ and $f \in \mathrm{L}^2(Q)$. Then, problem (5.5) admits a unique solution $u \in \mathrm{L}^2(\mathbb{R}^+; V) \cap C^0(\mathbb{R}^+; \mathrm{L}^2(\Omega))$, with $V = \mathrm{H}^1_{\Gamma_D}(\Omega)$.
For the definition of these functional spaces, see Sec. 2.7. For the proof, see [QV94, Sec. 11.1.1].
Some a priori estimates of the solution u will be provided in the following section.

We now consider the Galerkin approximation of problem (5.5):
for each $t > 0$, find $u_h(t) \in V_h$ s.t.

$$\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + a(u_h(t), v_h) = \int_\Omega f(t) v_h \, d\Omega \quad \forall v_h \in V_h \tag{5.6}$$

with $u_h(0) = u_{0h}$, where $V_h \subset V$ is a suitable space of finite dimension and u_{0h} is a convenient approximation of u_0 in the space V_h. Such problem is called *semi-discretization* of (5.5), as the temporal variable has not yet been discretized.
To provide an algebraic interpretation of (5.6) we introduce a basis $\{\varphi_j\}$ for V_h (as we did in the previous chapters), and we observe that it suffices that (5.6) is verified for the basis functions in order to be satisfied by all the functions of the subspace. Moreover, as for each $t > 0$ the solution to the Galerkin problem belongs to the subspace as well, we will have

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

where the $\{u_j(t)\}$ coefficients represent the unknowns of problem (5.6).

Denoting by $\dot{u}_j(t)$ the derivatives of the function $u_j(t)$ with respect to time, (5.6) becomes

$$\int_\Omega \sum_{j=1}^{N_h} \dot{u}_j(t) \varphi_j \varphi_i \, d\Omega + a\left(\sum_{j=1}^{N_h} u_j(t) \varphi_j, \varphi_i \right) = \int_\Omega f(t) \phi_i \, d\Omega, \qquad i = 1, 2, \ldots, N_h,$$

that is

$$\sum_{j=1}^{N_h} \dot{u}_j(t) \underbrace{\int_\Omega \varphi_j \varphi_i \, d\Omega}_{m_{ij}} + \sum_{j=1}^{N_h} u_j(t) \underbrace{a(\varphi_j, \varphi_i)}_{a_{ij}} = \underbrace{\int_\Omega f(t) \phi_i \, d\Omega}_{f_i(t)}, \qquad i = 1, 2, \ldots, N_h. \tag{5.7}$$

If we define the vector of unknowns $\mathbf{u} = (u_1(t),\ u_2(t),\ \ldots, u_{N_h}(t))^T$, the *mass matrix* $\mathrm{M} = [m_{ij}]$, the stiffness matrix $\mathrm{A} = [a_{ij}]$ and the right-hand side vector $\mathbf{f} = (f_1(t),\ f_2(t), \ldots, f_{N_h}(t))^T$, the system (5.7) can be rewritten in matrix form as

$$\mathrm{M}\dot{\mathbf{u}}(t) + \mathrm{A}\mathbf{u}(t) = \mathbf{f}(t).$$

For the numerical solution of this ODE system, many finite difference methods are available. See, e.g., [QSS07, Chap. 11]. Here we limit ourselves to considering the so-called θ-method. The latter discretizes the temporal derivative by a simple incremental ratio and replaces the other terms via a linear combination of the value at time t^k and of the value at time t^{k+1}, depending on the real parameter θ ($0 \leq \theta \leq 1$),

$$\mathrm{M}\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + \mathrm{A}[\theta\mathbf{u}^{k+1} + (1-\theta)\mathbf{u}^k] = \theta\mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k. \tag{5.8}$$

As usual, the real positive parameter $\Delta t = t^{k+1} - t^k$, $k = 0, 1, ...$, denotes the discretization step (here assumed to be constant), while the super-index k indicates that the quantity under consideration refers to the time t^k. Let us see some particular cases of (5.8):

– for $\theta = 0$ we obtain the *forward Euler* (or *explicit* Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^k = \mathbf{f}^k$$

which is first-order accurate with respect to Δt;

– for $\theta = 1$ we have the *backward Euler* (or *implicit* Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^{k+1} = \mathbf{f}^{k+1}$$

which is itself first-order with respect to Δt;

– for $\theta = 1/2$ we have the *Crank-Nicolson* (or *trapezoidal*) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\Delta t} + \frac{1}{2}\mathrm{A}\left(\mathbf{u}^{k+1} + \mathbf{u}^k\right) = \frac{1}{2}\left(\mathbf{f}^{k+1} + \mathbf{f}^k\right)$$

which is second-order accurate with respect to Δt. (More precisely, $\theta = 1/2$ is the only value for which we obtain a second-order method.)

Let us consider the two extreme cases, $\theta = 0$ and $\theta = 1$. For both, we obtain a system of linear equations: if $\theta = 0$, the system to solve has matrix $\frac{\mathrm{M}}{\Delta t}$, in the second case it has matrix $\frac{\mathrm{M}}{\Delta t} + A$. We observe that the M matrix is invertible, being positive definite (see Exercise 1).

In the $\theta = 0$ case, if we make matrix M diagonal, we actually decouple the system equations. This operation is performed by executing the so-called *lumping* of the mass matrix (see Sec. 11.5). However, this scheme is not unconditionally stable (see Sec. 5.4) and, in the case where V_h is a subspace of finite elements, we have the following stability condition (see Sec. 5.4)

$$\exists c > 0 \quad : \quad \Delta t \leq ch^2 \qquad \forall h > 0,$$

that does not allow an arbitrary choice of Δt with respect to h.

In the case $\theta > 0$, the system will have the form $K\mathbf{u}^{k+1} = \mathbf{g}$, where $\mathbf{g}$ is the known term and $\mathrm{K} = \frac{\mathrm{M}}{\Delta t} + \theta\mathrm{A}$. Such matrix is however invariant in time (the operator

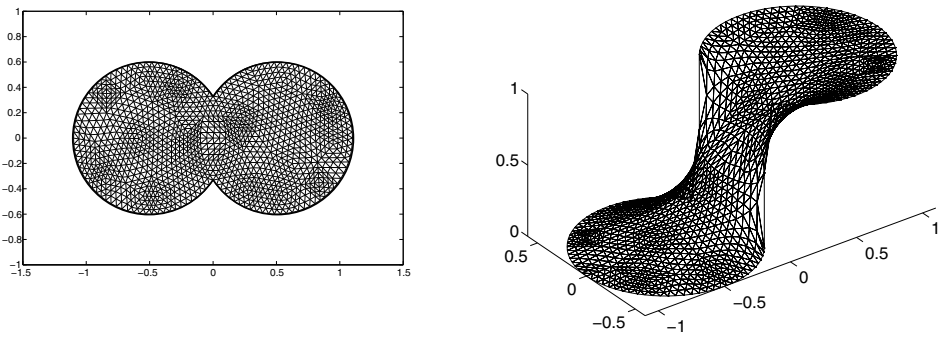


Fig. 5.2. Solution of the heat equation for the problem of Example 5.1

L, and therefore the matrix A, being independent of time); if the space mesh does not change, it can then be factorized once and for all at the beginning of the process. Since M is symmetric, if A is symmetric too, the K matrix associated to the system will also be symmetric. Hence, we can use, for instance, the Cholesky factorization, K=H H^T, H being lower triangular. At each time step, we will therefore have to solve two triangular systems in N_h unknowns:

$$\begin{aligned} &\mathrm{H}\mathbf{y} = \mathbf{g}, \\ &\mathrm{H}^T\mathbf{u}^{k+1} = \mathbf{y} \end{aligned}$$

(see Chap. 7 and also [QSS07, Chap. 3]).

Example 5.1 Let us suppose to solve the heat equation $\frac{\partial u}{\partial t} - 0.1\Delta u = 0$ on the domain $\Omega \subset \mathbb{R}^2$ of Fig. 5.2 (left) which is the union of two circles of radius 0.5 and center $(-0.5, 0)$ resp. $(0.5, 0)$). We assign Dirichlet conditions on the whole boundary taking $u(\mathbf{x}, t) = 1$ for the points on $\partial\Omega$ for which $x_1 \geq 0$ and $u(\mathbf{x}, t) = 0$ if $x_1 < 0$. The initial condition is $u(\mathbf{x}, 0) = 1$ for $x_1 \geq 0$ and null elsewhere. In Fig. 5.2, we report the solution obtained at time $t = 1$. We have used linear finite elements in space and the implicit Euler method in time with $\Delta t = 0.01$. As it can be seen, the initial discontinuity has been regularized, in accordance with the boundary conditions. ■

5.2 A priori estimates

Let us consider problem (5.5); since the corresponding equations must hold for each $v \in V$, it will be legitimate to pose $v = u(t)$ (t being given), solution of the problem itself, yielding

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega + a(u(t), u(t)) = \int_\Omega f(t)u(t)\, d\Omega \qquad \forall t > 0. \tag{5.9}$$

Considering the individual terms, we have

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\int_\Omega |u(t)|^2 d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|^2_{L^2(\Omega)}. \tag{5.10}$$

If we assume for simplicity that the bilinear form is coercive (with coercivity constant equal to α), we obtain

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

while thanks to the Cauchy-Schwarz inequality, we find

$$(f(t), u(t)) \leq \|f(t)\|_{\mathrm{L}^2(\Omega)} \|u(t)\|_{\mathrm{L}^2(\Omega)}. \tag{5.11}$$

In the remainder, we will often use the following *Young inequality*

$$\forall a, b \in \mathbb{R}, \qquad ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \qquad \forall \varepsilon > 0, \tag{5.12}$$

that derives from the elementary inequality

$$\left(\sqrt{\varepsilon}\, a - \frac{1}{2\sqrt{\varepsilon}}\, b\right)^2 \geq 0.$$

Using first the Poincaré inequality (2.13) and next the Young inequality, we obtain

$$\begin{aligned} \tfrac{1}{2}\tfrac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha \|\nabla u(t)\|^2_{L^2(\Omega)} &\leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \\ &\leq \tfrac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}. \end{aligned} \tag{5.13}$$

Then, by integrating in time we obtain, for all $t > 0$,

$$\begin{aligned} &\|u(t)\|^2_{L^2(\Omega)} + \alpha \textstyle\int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds \\ &\leq \|u_0\|^2_{L^2(\Omega)} + \tfrac{C_\Omega^2}{\alpha} \textstyle\int_0^t \|f(s)\|^2_{L^2(\Omega)} ds. \end{aligned} \tag{5.14}$$

This is an a priori energy estimate. Different kinds of a priori estimates can be obtained as follows. Note that

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = \|u(t)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}.$$

Then from (5.9), using (5.10) and (5.11) we obtain (still using the Poincaré inequality)

$$\begin{aligned} \|u(t)\|_{L^2(\Omega)} \tfrac{d}{dt}\|u(t)\|_{L^2(\Omega)} \quad &+ \quad \tfrac{\alpha}{C_\Omega}\|u(t)\|_{L^2(\Omega)} \|\nabla u(t)\|_{L^2(\Omega)} \\ &\leq \quad \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)}, \qquad t > 0. \end{aligned}$$

If $\|u(t)\|_{L^2(\Omega)} \neq 0$ (otherwise we should proceed differently, however the final result is still true) we can divide by $\|u(t)\|_{L^2(\Omega)}$ and integrate in time to yield

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \quad t > 0. \tag{5.15}$$

This is a further a priori estimate.

Let us now use the first inequality in (5.13), and integrate in time to yield

$$\begin{array}{l} \|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2 ds \\ \leq \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u(s)\|_{L^2(\Omega)} ds \\ \leq \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \cdot (\|u_0\|^2_{L^2(\Omega)} + \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau) ds \\ \text{(using (5.15))} \\ = \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u_0\|_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau \\ = (\|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\| ds)^2. \end{array} \tag{5.16}$$

The latter equality follows upon noticing that

$$\|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau = \frac{d}{ds} (\int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau)^2.$$

We therefore conclude with the additional a priori estimate

$$\begin{array}{l} (\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds)^{\frac{1}{2}} \\ \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \qquad t > 0. \end{array} \tag{5.17}$$

We have seen that we can formulate the Galerkin problem (5.6) for problem (5.5) and that the latter, under suitable hypotheses, admits a unique solution. Similarly to what we did for problem (5.5) we can prove the following a priori (stability) estimates for the solution to problem (5.6):

$$\begin{array}{rcl} \|u_h(t)\|^2_{L^2(\Omega)} & + & \alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)} ds \\ & \leq & \|u_{0h}(t)\|^2_{L^2(\Omega)} + \frac{C^2_\Omega}{\alpha} \int_0^t \|f(s)\|^2_{L^2(\Omega)} ds, \quad t > 0. \end{array} \tag{5.18}$$

For its proof we can take, for every $t > 0$, $v_h = u_h(t)$ and proceed as we did to obtain (5.13). Then, by recalling that the initial data is $u_h(0) = u_{0h}$, we can deduce the following discrete counterparts of (5.15) and (5.17):

$$\|u_h(t)\|_{L^2(\Omega)} \quad \leq \quad \|u_{0h}(t)\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \quad t > 0, \tag{5.19}$$

and

$$\begin{array}{l} (\|u_h(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)} ds)^{\frac{1}{2}} \\ \leq \|u_{0h}(t)\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \qquad t > 0. \end{array} \tag{5.20}$$

5.3 Convergence analysis of the semi-discrete problem

Let us consider the problem (5.5) and its approximation (5.6). We want to prove the convergence of u_h to u in suitable norms.
By the coercivity hypotheses we can write

$$\begin{aligned}\alpha\|u-u_h\|^2_{\mathrm{H}^1(\Omega)} &\leq a(u-u_h,u-u_h)\\ &= a(u-u_h,u-v_h)+a(u-u_h,v_h-u_h) \quad \forall v_h\in V_h.\end{aligned}$$

By subtracting equation (5.6) from equation (5.5) and setting $w_h = v_h - u_h$ we have

$$\left(\frac{\partial(u-u_h)}{\partial t},w_h\right)+a(u-u_h,w_h)=0,$$

where $(v,w)=\int_\Omega vw\,d\Omega$ is the scalar product of $L^2(\Omega)$. Then

$$\alpha\|u-u_h\|^2_{\mathrm{H}^1(\Omega)} \leq a(u-u_h,u-v_h)-\left(\frac{\partial(u-u_h)}{\partial t},w_h\right). \tag{5.21}$$

We analyze the two right-hand side terms separately:

– using the continuity of the form $a(\cdot,\cdot)$ and the Young inequality, we obtain

$$\begin{aligned}a(u-u_h,u-v_h) &\leq M\|u-u_h\|_{\mathrm{H}^1(\Omega)}\|u-v_h\|_{\mathrm{H}^1(\Omega)}\\ &\leq \frac{\alpha}{2}\|u-u_h\|^2_{\mathrm{H}^1(\Omega)}+\frac{M^2}{2\alpha}\|u-v_h\|^2_{\mathrm{H}^1(\Omega)};\end{aligned}$$

– writing w_h in the form $w_h=(v_h-u)+(u-u_h)$ we obtain

$$-\left(\frac{\partial(u-u_h)}{\partial t},w_h\right)=\left(\frac{\partial(u-u_h)}{\partial t},u-v_h\right)-\frac{1}{2}\frac{d}{dt}\|u-u_h\|^2_{\mathrm{L}^2(\Omega)}. \tag{5.22}$$

Replacing these two results in (5.21), we obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|u-u_h\|^2_{\mathrm{L}^2(\Omega)}+\frac{\alpha}{2}\|u-u_h\|^2_{\mathrm{H}^1(\Omega)}\\ &\leq \frac{M^2}{2\alpha}\|u-v_h\|^2_{\mathrm{H}^1(\Omega)}+(\frac{\partial(u-u_h)}{\partial t},u-v_h).\end{aligned}$$

Multiplying both sides by 2 and integrating in time between 0 and t we find

$$\begin{aligned}&\|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)}+\alpha\int_0^t\|(u-u_h)(s)\|^2_{\mathrm{H}^1(\Omega)}\,ds \leq \|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)}\\ &+\quad\frac{M^2}{\alpha}\int_0^t\|u(s)-v_h\|^2_{\mathrm{H}^1(\Omega)}\,ds+2\int_0^t\left(\frac{\partial}{\partial t}(u-u_h)(s),u(s)-v_h\right)\,ds.\end{aligned} \tag{5.23}$$

Integrating by parts and using the Young inequality, we obtain

$$\begin{aligned}
&\int_0^t \left(\frac{\partial}{\partial t}(u-u_h)(s), u(s)-v_h\right) ds = -\int_0^t \left((u-u_h)(s), \frac{\partial}{\partial t}(u(s)-v_h)\right) ds \\
&\qquad +((u-u_h)(t),(u-v_h)(t)) - ((u-u_h)(0),(u-v_h)(0)) \\
&\leq \frac{1}{4}\int_0^t \|(u-u_h)(s)\|^2_{\mathrm{L}^2(\Omega)}\, ds + \int_0^t \|\frac{\partial(u(s)-v_h)}{\partial t}\|^2_{\mathrm{L}^2(\Omega)}\, ds + \frac{1}{4}\|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} \\
&\qquad +\|(u-v_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \|(u-u_h)(0)\|_{\mathrm{L}^2(\Omega)}\|(u-v_h)(0)\|_{\mathrm{L}^2(\Omega)}.
\end{aligned}$$

From (5.23) we thus obtain

$$\begin{aligned}
&\frac{1}{2}\|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \alpha\int_0^t \|(u-u_h)(s)\|^2_{\mathrm{H}^1(\Omega)}\, ds \\
&\leq \frac{M^2}{\alpha}\int_0^t \|u(s)-v_h\|^2_{\mathrm{H}^1(\Omega)}\, ds + 2\int_0^t \|\frac{\partial(u(s)-v_h)}{\partial t}\|^2_{\mathrm{L}^2(\Omega)}\, ds \\
&+2\|(u-v_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)} \\
&+2\|(u-u_h)(0)\|_{\mathrm{L}^2(\Omega)}\|(u-v_h)(0)\|_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\int_0^t \|(u-u_h)(s)\|^2_{\mathrm{L}^2(\Omega)}\, ds.
\end{aligned} \tag{5.24}$$

Let us now suppose that V_h is the space of finite elements of degree r, more precisely $V_h = \{v_h \in X_h^r : v_h|_{\Gamma_D} = 0\}$, and let us choose, at each t, $v_h = \Pi_h^r u(t)$, the interpolant of $u(t)$ in V_h (see (4.20)). Thanks to (4.69) we have, assuming that u is sufficiently regular,

$$h\|u-\Pi_h^r u\|_{\mathrm{H}^1(\Omega)} + \|u-\Pi_h^r u\|_{\mathrm{L}^2(\Omega)} \leq C_2 h^{r+1}|u|_{\mathrm{H}^{r+1}(\Omega)}.$$

Hence, the addenda of the right-hand side of inequality (5.24) are bounded as follows:

$$\begin{aligned}
E_1 &= \frac{M^2}{\alpha}\int_0^t \|u(s)-v_h\|^2_{\mathrm{H}^1(\Omega)}\, ds \leq C_1 h^{2r}\int_0^t |u(s)|^2_{\mathrm{H}^{r+1}(\Omega)}\, ds, \\
E_2 &= 2\int_0^t \left\|\frac{\partial(u-v_h)}{\partial t}(s)\right\|^2_{\mathrm{L}^2(\Omega)} ds \leq C_2 h^{2r}\int_0^t \left|\frac{\partial u}{\partial t}(s)\right|^2_{\mathrm{H}^r(\Omega)} ds, \\
E_3 &= 2\|(u-v_h)(t)\|^2_{\mathrm{L}^2(\Omega)} \leq C_3 h^{2r}|u|^2_{\mathrm{H}^r(\Omega)}, \\
E_4 &= \|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)} + 2\|(u-u_h)(0)\|_{\mathrm{L}^2(\Omega)}\|(u-v_h)(0)\|_{\mathrm{L}^2(\Omega)} \\
&\leq C_4 h^{2r}|u(0)|^2_{\mathrm{H}^r(\Omega)}.
\end{aligned}$$

Consequently,

$$E_1 + E_2 + E_3 + E_4 \leq Ch^{2r}N(u),$$

where $N(u)$ is a suitable function depending on u and on $\dfrac{\partial u}{\partial t}$. This way, we obtain the inequality

$$\frac{1}{2}\|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \alpha\int_0^t \|(u-u_h)(s)\|^2_{\mathrm{H}^1(\Omega)}\,ds$$
$$\leq Ch^{2r}N(u) + \frac{1}{2}\int_0^t \|(u-u_h)(s)\|^2_{\mathrm{L}^2(\Omega)}$$

and finally, applying the Gronwall lemma (see Sec. 2.7), we obtain the a priori error estimate for all $t > 0$

$$\|u(t)-u_h(t)\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha\int_0^t \|u(s)-u_h(s)\|^2_{\mathrm{H}^1(\Omega)}ds \leq Ch^{2r}N(u)\,e^t. \quad (5.25)$$

By a different proof technique that does not make use of the Gronwall lemma we can prove an error inequality similar to (5.25), however without the exponential factor e^t at the right-hand side.
We proceed as follows. If we subtract (5.6) from (5.5) and set $E_h = u - u_h$, we obtain that

$$(\frac{\partial E_h}{\partial t}, v_h) + a(E_h, v_h) = 0 \quad \forall v_h \in V_h, \quad \forall t > 0.$$

If for the sake of simplicity, we suppose that $a(\cdot,\cdot)$ is symmetric, we can define the orthogonal projection operator

$$\Pi^r_{1,h} : V \to V_h \ : \ \forall w \in V,\ a(\Pi^r_{1,h}w - w, v_h) = 0 \qquad \forall v_h \in V_h. \quad (5.26)$$

Using the results seen in Chap. 3, we can prove (see [QV94, Sec. 3.5]) that there exists a constant $C > 0$ s.t., $\forall w \in V \cap H^{r+1}(\Omega)$,

$$\|\Pi^r_{1,h}w - w\|_{\mathrm{H}^1(\Omega)} + h^{-1}\|\Pi^r_{1,h}w - w\|_{\mathrm{L}^2(\Omega)} \leq Ch^p|w|_{\mathrm{H}^{p+1}(\Omega)}, 0 \leq p \leq r. \quad (5.27)$$

Then we set

$$E_h = \sigma_h + e_h = (u - \Pi^r_{1,h}u) + (\Pi^r_{1,h}u - u_h). \quad (5.28)$$

Note that the orthogonal projection error σ_h can be bounded by inequality (5.27) and that e_h is an element of the subspace V_h. Then

$$(\frac{\partial e_h}{\partial t}, v_h) + a(e_h, v_h) = -(\frac{\partial \sigma_h}{\partial t}, v_h) - a(\sigma_h, v_h) \quad \forall v_h \in V_h \quad \forall t > 0.$$

If we take at every $t > 0$, $v_h = e_h(t)$, and proceed as done in Sec. 5.2 to deduce the a priori estimates on the semi-discrete solution u_h, we obtain

$$\begin{aligned} \tfrac{1}{2}\tfrac{d}{dt}\|e_h(t)\|^2_{L^2(\Omega)} \quad & +\alpha\|\nabla e_h(t)\|^2_{L^2(\Omega)} \\ & \leq |a(\sigma_h(t), e_h(t))| + |(\tfrac{\partial}{\partial t}\sigma_h(t), e_h(t))|. \end{aligned} \tag{5.29}$$

Using the continuity of the bilinear form $a(\cdot,\cdot)$ (M being the continuity constant) and the Young inequality, we obtain

$$|a(\sigma_h(t), e_h(t))| \leq \frac{\alpha}{4}\|\nabla e_h(t)\|^2_{L^2(\Omega)} + \frac{M^2}{\alpha}\|\nabla \sigma_h(t)\|^2_{L^2(\Omega)}.$$

Moreover, using the Poincaré inequality and once more the Young inequality it follows that

$$\begin{aligned} |(\tfrac{\partial}{\partial t}\sigma_h(t), e_h(t))| \quad & \leq \|\tfrac{\partial}{\partial t}\sigma_h(t)\|_{L^2(\Omega)} C_\Omega \|\nabla e_h(t)\|_{L^2(\Omega)} \\ & \leq \tfrac{\alpha}{4}\|\nabla e_h(t)\|^2_{L^2(\Omega)} + \tfrac{C^2_\Omega}{\alpha}\|\tfrac{\partial}{\partial t}\sigma_h(t)\|^2_{L^2(\Omega)}. \end{aligned}$$

Using these bounds in (5.29) we obtain, after integrating w.r.t. t:

$$\begin{aligned} & \|e_h(t)\|^2_{L^2(\Omega)} + \alpha \textstyle\int_0^t \|\nabla e_h(t)\|^2_{L^2(\Omega)} ds \\ & \leq \|e_h(0)\|^2_{L^2(\Omega)} + \tfrac{2M^2}{\alpha}\textstyle\int_0^t \|\nabla \sigma_h(s)\|^2_{L^2(\Omega)} ds \\ & + \tfrac{2C^2_\Omega}{\alpha}\textstyle\int_0^t \|\tfrac{\partial}{\partial t}\sigma_h(s)\|^2_{L^2(\Omega)} ds, \quad t > 0. \end{aligned}$$

At this point we can use (5.27) to bound the errors on the right-hand side:

$$\begin{aligned} & \|\nabla \sigma_h(t)\|_{L^2(\Omega)} \leq Ch^r |u(t)|_{H^{r+1}(\Omega)}, \\ & \|\tfrac{\partial}{\partial t}\sigma_h(t)\|_{L^2(\Omega)} = \|(\tfrac{\partial u}{\partial t} - \Pi^r_{1,h}\tfrac{\partial u}{\partial t})(t)\|_{L^2(\Omega)} \leq Ch^r |\tfrac{\partial u(t)}{\partial t}|_{H^{r+1}(\Omega)}. \end{aligned}$$

Finally, note that $\|e_h(0)\|_{L^2(\Omega)} \leq Ch^r |u_0|_{H^r(\Omega)}$, still using (5.27).
Since, for any norm $\|\cdot\|$,

$$\|u - u_h\| \leq \|\sigma_h\| + \|e_h\|$$

(owing to 5.28), using the previous estimates we can conclude that there exists a constant $C > 0$ independent of both t and h s.t.

$$\begin{aligned} & \|u(t) - u_h(t)\|^2_{L^2(\Omega)} + \alpha \textstyle\int_0^t \|\nabla u(s) - \nabla u_h(s)\|^2_{L^2(\Omega)} ds \\ & \leq Ch^{2r}\{|u_0|^2_{H^r(\Omega)} + \textstyle\int_0^t |u(s)|^2_{H^{r+1}(\Omega)} ds + \int_0^t |\tfrac{\partial u(s)}{\partial t}|^2_{H^{r+1}(\Omega)} ds\}. \end{aligned}$$

This error inequality has a form similar to (5.25), however, as anticipated, the right-hand side does not include the exponential factor e^t any longer. Further error estimates are proven, e.g. in [QV94, Chap. 11].

5.4 Stability analysis of the θ-method

We now analyze the stability of the fully discretized problem.
Applying the θ-method to the Galerkin problem (5.6) we obtain

$$\begin{aligned}\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h\right) \\ = \theta F^{k+1}(v_h) + (1-\theta)F^k(v_h) \qquad \forall v_h \in V_h,\end{aligned} \tag{5.30}$$

for each $k \geq 0$, with $u_h^0 = u_{0h}$; F^k indicates that the functional is evaluated at time t^k.
We will limit ourselves to the case where $F = 0$ and start to consider the case of the implicit Euler method where $\theta = 1$, that is

$$\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(u_h^{k+1}, v_h\right) = 0 \qquad \forall v_h \in V_h.$$

By choosing $v_h = u_h^{k+1}$, we obtain

$$(u_h^{k+1}, u_h^{k+1}) + \Delta t\, a\left(u_h^{k+1}, u_h^{k+1}\right) = (u_h^k, u_h^{k+1}).$$

By now exploiting the following inequalities

$$a(u_h^{k+1}, u_h^{k+1}) \geq \alpha \|u_h^{k+1}\|_V^2, \qquad (u_h^k, u_h^{k+1}) \leq \frac{1}{2}\|u_h^k\|_{\mathrm{L}^2(\Omega)}^2 + \frac{1}{2}\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}^2,$$

the former deriving from the coercivity of the bilinear form $a(\cdot,\cdot)$ and the latter from the Cauchy-Schwarz and Young inequalities, we obtain

$$\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t\|u_h^{k+1}\|_V^2 \leq \|u_h^k\|_{\mathrm{L}^2(\Omega)}^2. \tag{5.31}$$

By summing over the k index from 0 to $n-1$ we deduce that

$$\|u_h^n\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t\sum_{k=0}^{n-1}\|u_h^{k+1}\|_V^2 \leq \|u_{0h}\|_{\mathrm{L}^2(\Omega)}^2.$$

In the case where $f \neq 0$, using the discrete Gronwall lemma (see Sec. 2.7) it can be proven in a similar way that

$$\|u_h^n\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t\sum_{k=1}^{n}\|u_h^k\|_V^2 \leq C(t^n)\left(\|u_{0h}\|_{\mathrm{L}^2(\Omega)}^2 + \sum_{k=1}^{n}\Delta t\|f^k\|_{\mathrm{L}^2(\Omega)}^2\right). \tag{5.32}$$

Such relation is similar to (5.20), provided that the integrals $\int_0^t \cdot\, ds$ are approximated by a composite numerical integration formula with step Δt.

Finally, observing that $\|u_h^{k+1}\|_V \geq \|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}$, we deduce from (5.31) that, for each given $\Delta t > 0$,

$$\lim_{k\to\infty} \|u_h^k\|_{\mathrm{L}^2(\Omega)} = 0,$$

that is the backward Euler method is absolutely stable without any restriction on the time step Δt.

Before analyzing the general case where θ is an arbitrary parameter ranging between 0 and 1, we introduce the following definition.
We say that the scalar λ is an *eigenvalue of the bilinear form* $a(\cdot,\cdot) : V \times V \mapsto \mathbb{R}$ and that $w \in V$ is its corresponding *eigenfunction* if it turns out that

$$a(w, v) = \lambda(w, v) \qquad \forall v \in V.$$

If the bilinear form $a(\cdot,\cdot)$ is symmetric and coercive, it has infinite eigenvalues, all positive real, which form an infinite sequence; moreover, its eigenfunctions form a basis of the space V.

The eigenvalues and eigenfunctions of $a(\cdot,\cdot)$ can be approximated by finding the pairs $\lambda_h \in \mathbb{R}$ and $w_h \in V_h$ which satisfy

$$a(w_h, v_h) = \lambda_h(w_h, v_h) \quad \forall v_h \in V_h. \tag{5.33}$$

From an algebraic viewpoint, problem (5.33) can be formulated as follows

$$\mathrm{A}\mathbf{w} = \lambda_h \mathrm{M}\mathbf{w},$$

where A is the stiffness matrix and M the mass matrix. We are therefore dealing with a *generalized eigenvalue problem*.

Such eigenvalues are all positive and as many as N_h (N_h being as usual the dimension of the subspace V_h); after ordering them in ascending order, $\lambda_h^1 \leq \lambda_h^2 \leq \ldots \leq \lambda_h^{N_h}$, we have

$$\lambda_h^{N_h} \to \infty \qquad \text{for } N_h \to \infty.$$

Moreover, the corresponding eigenfunctions form a basis for the subspace V_h and can be chosen in order to be *orthonormal* with respect to the scalar product of $\mathrm{L}^2(\Omega)$. This means that, denoting by w_h^i the eigenfunction corresponding to the eigenvalue λ_h^i, we have $(w_h^i, w_h^j) = \delta_{ij} \quad \forall i, j = 1, \ldots, N_h$. Thus, each function $v_h \in V_h$ can be represented as follows

$$v_h(\mathbf{x}) = \sum_{j=1}^{N_h} v_j w_h^j(\mathbf{x})$$

and, thanks to the eigenfunction orthonormality,

$$\|v_h\|^2_{\mathrm{L}^2(\Omega)} = \sum_{j=1}^{N_h} v_j^2. \tag{5.34}$$

Let us now consider an arbitrary $\theta \in [0,1]$ and let us limit ourselves to the case where the bilinear form $a(\cdot,\cdot)$ is symmetric (otherwise, although the final stability result holds in general, the following proof would not be applicable, as the eigenfunctions would not necessarily form a basis). Let $\{w_h^i\}$ still denote the discrete (orthonormal) eigenfunctions of $a(\cdot,\cdot)$. Since $u_h^k \in V_h$, we can write

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

We observe that in this modal expansion, the u_j^k no longer represent the nodal values of u_h^k. If we now set $F = 0$ in (5.30) and take $v_h = w_h^i$, we find

$$\frac{1}{\Delta t}\sum_{j=1}^{N_h}[u_j^{k+1} - u_j^k]\left(w_h^j, w_h^i\right) + \sum_{j=1}^{N_h}[\theta u_j^{k+1} + (1-\theta)u_j^k]a(w_h^j, w_h^i) = 0,$$

for each $i = 1, \ldots, N_h$. For each pair $i, j = 1, \ldots, N_h$ we have

$$a(w_h^j, w_h^i) = \lambda_h^j (w_h^j, w_h^i) = \lambda_h^j \delta_{ij} = \lambda_h^i,$$

and thus, for each $i = 1, \ldots, N_h$,

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} + [\theta u_i^{k+1} + (1-\theta)u_i^k]\lambda_h^i = 0.$$

Solving now with respect to u_i^{k+1}, we find

$$u_i^{k+1} = u_i^k \frac{1 - (1-\theta)\lambda_h^i \Delta t}{1 + \theta \lambda_h^i \Delta t}.$$

Recalling (5.34), we can conclude that for the method to be absolutely stable, we must satisfy the inequality

$$\left| \frac{1 - (1-\theta)\lambda_h^i \Delta t}{1 + \theta \lambda_h^i \Delta t} \right| < 1,$$

that is

$$-1 - \theta\lambda_h^i \Delta t < 1 - (1-\theta)\lambda_h^i \Delta t < 1 + \theta \lambda_h^i \Delta t.$$

Hence,

$$-\frac{2}{\lambda_h^i \Delta t} - \theta < \theta - 1 < \theta.$$

The second inequality is always verified, while the first one can be rewritten as

$$2\theta - 1 > -\frac{2}{\lambda_h^i \Delta t}.$$

If $\theta \geq 1/2$, the left-hand side is non-negative, while the right-hand side is negative, so the inequality holds for each Δt. Instead, if $\theta < 1/2$, the inequality is satisfied (hence the method is stable) only if

$$\Delta t < \frac{2}{(1-2\theta)\lambda_h^i}. \tag{5.35}$$

As such relation must hold for all the eigenvalues λ_h^i of the bilinear form, it will suffice to require that it holds for the maximum among them, which we have supposed to be $\lambda_h^{N_h}$. To summarize, we have:

– *if $\theta \geq 1/2$, the θ-method is unconditionally stable, i.e. it is stable for each Δt;*

– *if $\theta < 1/2$, the θ-method is stable only for*

$$\Delta t \leq \frac{2}{(1-2\theta)\lambda_h^{N_h}}.$$

Thanks to the definition of eigenvalue (5.33) and to the continuity property of $a(\cdot,\cdot)$, we deduce

$$\lambda_h^{N_h} = \frac{a(w_{N_h}, w_{N_h})}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq \frac{M\|w_{N_h}\|^2_V}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq M(1 + C^2 h^{-2}).$$

The constant $C > 0$ which appears in the latter step derives from the following *inverse inequality*

$$\exists C > 0 \ : \ \|\nabla v_h\|_{\mathrm{L}^2(\Omega)} \leq C h^{-1} \|v_h\|_{\mathrm{L}^2(\Omega)} \qquad \forall v_h \in V_h,$$

for whose proof we refer to [QV94, Chap. 3].
Hence, for h small enough, $\lambda_h^{N_h} \leq Ch^{-2}$. In fact, we can prove that $\lambda_h^{N_h}$ is indeed of the order of h^{-2}, that is

$$\lambda_h^{N_h} = max_i \lambda_h^i \simeq ch^{-2}.$$

Keeping this into account, we obtain that for $\theta < 1/2$ the method is absolutely stable only if

$$\Delta t \leq C(\theta) h^2, \tag{5.36}$$

where $C(\theta)$ denotes a positive constant depending on θ.

The latter relation implies that, for $\theta < 1/2$, Δt cannot be chosen arbitrarily but is bound to the choice of h.

5.5 Convergence analysis of the θ-method

We can prove the following convergence theorem

Theorem 5.1 *Under the hypothesis that u_0, f and the exact solution are sufficiently regular, the following* a priori error estimate *holds:* $\forall n \geq 1$,

$$\|u(t^n) - u_h^n\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha\Delta t \sum_{k=1}^{n} \|u(t^k) - u_h^k\|^2_V \leq C(u_0, f, u)(\Delta t^{p(\theta)} + h^{2r}),$$

where $p(\theta) = 2$ if $\theta \neq 1/2$, $p(1/2) = 4$ and C depends on its arguments but not on h and Δt.

Proof. The proof is carried out by comparing the solution of the fully discretized problem (5.30) with that of the semi-discrete problem (5.6), using the stability result (5.32) as well as the decay rate of the truncation error of the time discretization. For simplicity, we will limit ourselves to considering the backward Euler method (corresponding to $\theta = 1$)

$$\frac{1}{\Delta t}(u_h^{k+1} - u_h^k, v_h) + a(u_h^{k+1}, v_h) = (f^{k+1}, v_h) \qquad \forall v_h \in V_h. \tag{5.37}$$

We refer the reader to [QV94], Sec. 11.3.1, for the proof in the general case.

Let $\Pi^r_{1,h}$ be the orthogonal projector operator introduced in (5.26). Then

$$\|u(t^k) - u_h^k\|_{\mathrm{L}^2(\Omega)} \leq \|u(t^k) - \Pi^r_{1,h} u(t^k)\|_{\mathrm{L}^2(\Omega)} + \|\Pi^r_{1,h} u(t^k) - u_h^k\|_{\mathrm{L}^2(\Omega)}. \tag{5.38}$$

The first term can be estimated by referring to (5.27). To analyze the second term, having set $\varepsilon_h^k = u_h^k - \Pi^r_{1,h} u(t^k)$, we obtain

$$\frac{1}{\Delta t}(\varepsilon_h^{k+1} - \varepsilon_h^k, v_h) + a(\varepsilon_h^{k+1}, v_h) = (\delta^{k+1}, v_h) \qquad \forall v_h \in V_h, \tag{5.39}$$

having set, $\forall v_h \in V_h$,

$$(\delta^{k+1}, v_h) = (f^{k+1}, v_h) - \frac{1}{\Delta t}(\Pi^r_{1,h}(u(t^{k+1}) - u(t^k)), v_h) - a(u(t^{k+1}), v_h) \tag{5.40}$$

and having exploited on the last addendum the orthogonality (5.26) of the operator $\Pi^r_{1,h}$. The sequence $\{\varepsilon_h^k,\ k = 0, 1 \ldots\}$ satisfies problem (5.39), which is similar in all to (5.37) (provided that we take δ^{k+1} instead of f^{k+1}). By adapting the stability estimate (5.32), we obtain, for each $n \geq 1$,

$$\|\varepsilon_h^n\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha\Delta t \sum_{k=1}^{n} \|\varepsilon_h^k\|^2_V \leq C(t^n) \left(\|\varepsilon_h^0\|^2_{\mathrm{L}^2(\Omega)} + \sum_{k=1}^{n} \Delta t \|\delta^k\|^2_{\mathrm{L}^2(\Omega)} \right). \tag{5.41}$$

The norm associated to the initial time-level can easily be estimated: for instance, if $u_{0h} = \Pi_h^r u_0$ is the finite element interpolant of u_0, by suitably using the estimates (4.69) and (5.27) we obtain

$$\begin{aligned} \|\varepsilon_h^0\|_{\mathrm{L}^2(\Omega)} &= \|u_{0h} - \Pi_{1,h}^r u_0\|_{\mathrm{L}^2(\Omega)} \\ &\leq \|\Pi_h^r u_0 - u_0\|_{\mathrm{L}^2(\Omega)} + \|u_0 - \Pi_{1,h}^r u_0\|_{\mathrm{L}^2(\Omega)} \leq C\, h^r\, |u_0|_{\mathrm{H}^r(\Omega)}. \end{aligned} \tag{5.42}$$

Let us now focus on estimating the norm $\|\delta^k\|_{\mathrm{L}^2(\Omega)}$. We note that, thanks to (5.5),

$$(f^{k+1}, v_h) - a(u(t^{k+1}), v_h) = \left(\frac{\partial u(t^{k+1})}{\partial t}, v_h\right).$$

This allows us to rewrite (5.40) as

$$\begin{aligned} (\delta^{k+1}, v_h) &= \left(\frac{\partial u(t^{k+1})}{\partial t}, v_h\right) - \frac{1}{\Delta t}(\Pi_{1,h}^r(u(t^{k+1}) - u(t^k)), v_h) \\ &= \left(\frac{\partial u(t^{k+1})}{\partial t} - \frac{u(t^{k+1}) - u(t^k)}{\Delta t}, v_h\right) + \left(\left(I - \Pi_{1,h}^r\right)\left(\frac{u(t^{k+1}) - u(t^k)}{\Delta t}\right), v_h\right). \end{aligned} \tag{5.43}$$

Using the Taylor formula with the remainder in integral form, we have

$$\frac{\partial u(t^{k+1})}{\partial t} - \frac{u(t^{k+1}) - u(t^k)}{\Delta t} = \frac{1}{\Delta t}\int_{t^k}^{t^{k+1}} (s - t^k)\frac{\partial^2 u}{\partial t^2}(s)\, ds, \tag{5.44}$$

having made the suitable regularity requirements on the u function with respect to the temporal variable. By now using the fundamental theorem of integration and exploiting the commutativity between the projection operator $\Pi_{1,h}^r$ and the temporal derivative, we obtain

$$\left(I - \Pi_{1,h}^r\right)\left(u(t^{k+1}) - u(t^k)\right) = \int_{t^k}^{t^{k+1}} \left(I - \Pi_{1,h}^r\right)\left(\frac{\partial u}{\partial t}\right)(s)\, ds. \tag{5.45}$$

By choosing $v_h = \delta^{k+1}$ in (5.43), thanks to (5.44) and (5.45), we can deduce the following upper bound

$$\begin{aligned} &\|\delta^{k+1}\|_{\mathrm{L}^2(\Omega)} \\ &\leq \left\|\frac{1}{\Delta t}\int_{t^k}^{t^{k+1}} (s - t^k)\frac{\partial^2 u}{\partial t^2}(s)\, ds\right\|_{\mathrm{L}^2(\Omega)} + \left\|\frac{1}{\Delta t}\int_{t^k}^{t^{k+1}} \left(I - \Pi_{1,h}^r\right)\left(\frac{\partial u}{\partial t}\right)(s)\, ds\right\|_{\mathrm{L}^2(\Omega)} \\ &\leq \int_{t^k}^{t^{k+1}} \left\|\frac{\partial^2 u}{\partial t^2}(s)\right\|_{\mathrm{L}^2(\Omega)} ds + \frac{1}{\Delta t}\int_{t^k}^{t^{k+1}} \left\|\left(I - \Pi_{1,h}^r\right)\left(\frac{\partial u}{\partial t}\right)(s)\right\|_{\mathrm{L}^2(\Omega)} ds. \end{aligned} \tag{5.46}$$

By reverting to the stability estimate (5.41) and exploiting (5.42) and the estimate (5.46) with suitably scaled indexes, we have

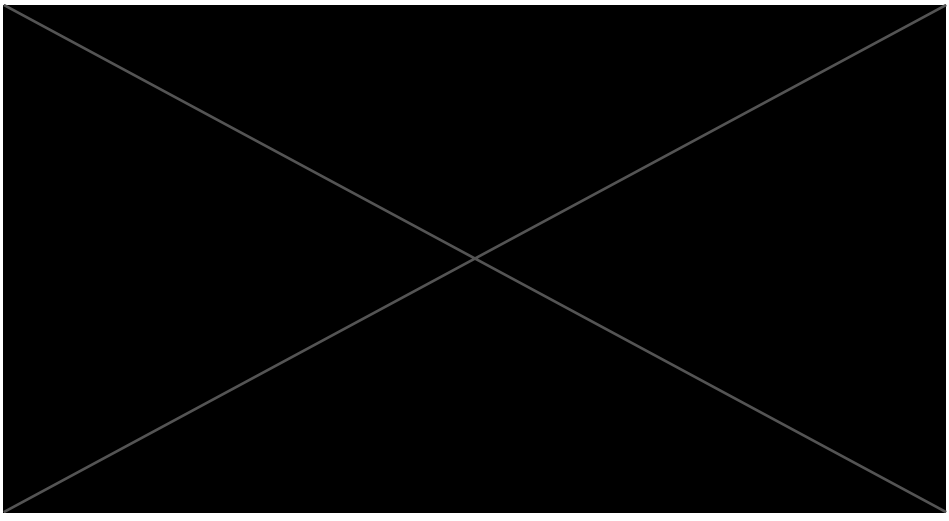
$$\begin{aligned}\|\varepsilon_h^n\|^2_{\mathrm{L}^2(\Omega)} &\leq C(t^n)\Bigg(h^{2r}\,|u_0|^2_{\mathrm{H}^r(\Omega)} + \sum_{k=1}^{n}\Delta t\Bigg[\Bigg(\int_{t^{k-1}}^{t^k}\left\|\frac{\partial^2 u}{\partial t^2}(s)\right\|_{\mathrm{L}^2(\Omega)} ds\Bigg)^2 \\ &+ \frac{1}{\Delta t^2}\Bigg(\int_{t^{k-1}}^{t^k}\left\|\Big(I-\Pi^r_{1,h}\Big)\Big(\frac{\partial u}{\partial t}\Big)(s)\right\|_{\mathrm{L}^2(\Omega)} ds\Bigg)^2\Bigg]\Bigg).\end{aligned}$$

Then, using the Cauchy-Schwarz inequality and the estimate (5.27) for the projection operator $\Pi^r_{1,h}$, we obtain

$$\begin{aligned}\|\varepsilon_h^n\|^2_{\mathrm{L}^2(\Omega)} &\leq C(t^n)\Bigg(h^{2r}\,|u_0|^2_{\mathrm{H}^r(\Omega)} + \sum_{k=1}^{n}\Delta t\Bigg[\Delta t\int_{t^{k-1}}^{t^k}\left\|\frac{\partial^2 u}{\partial t^2}(s)\right\|^2_{\mathrm{L}^2(\Omega)} ds \\ &+ \frac{1}{\Delta t^2}\Bigg(\int_{t^{k-1}}^{t^k} h^r\left|\frac{\partial u}{\partial t}(s)\right|_{\mathrm{H}^r(\Omega)} ds\Bigg)^2\Bigg]\Bigg) \\ &\leq C(t^n)\Bigg(h^{2r}\,|u_0|^2_{\mathrm{H}^r(\Omega)} + \Delta t^2\sum_{k=1}^{n}\int_{t^{k-1}}^{t^k}\left\|\frac{\partial^2 u}{\partial t^2}(s)\right\|^2_{\mathrm{L}^2(\Omega)} ds \\ &+ \frac{1}{\Delta t}\,h^{2r}\sum_{k=1}^{n}\Delta t\int_{t^{k-1}}^{t^k}\left|\frac{\partial u}{\partial t}(s)\right|^2_{\mathrm{H}^r(\Omega)} ds\Bigg).\end{aligned}$$

The result now follows using (5.38) and the estimate (5.27). ◇

More stability and convergence estimates can be found in [Tho84].



7

Algorithms for the solution of linear systems

In this chapter we make a quick and elementary introduction of some of the basic algorithms that are used to solve a system of linear algebraic equations. For a more thorough presentation we advise the reader to refer to, e.g., [QSS07], Chap. 3 and 4, [Saa96] and [vdV03].

A system of m linear equations in n unknowns is a set of algebraic relations of the form

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, m \tag{7.1}$$

x_j being the unknowns, a_{ij} the system's coefficients and b_i the known terms. The system (7.1) will more commonly be written in matrix form

$$\mathrm{A}\mathbf{x} = \mathbf{b}, \tag{7.2}$$

having denoted by $\mathrm{A} = (a_{ij}) \in \mathbb{R}^{m \times n}$ the coefficient matrix, $\mathbf{b}=(b_i) \in \mathbb{R}^m$ being the known term vector (also named as the right-hand side) and $\mathbf{x}=(x_i) \in \mathbb{R}^n$ the unknown vector. We call *solution* of (7.2) any n-uple of values x_i verifying (7.1).

In the following sections, we recall some numerical techniques for the solution of (7.2) in the case where $m = n$; we will obviously suppose that A is non-singular, i.e. that $\det(\mathrm{A}) \neq 0$. Numerical methods are called *direct* if they lead to the solution of the system in a finite number of operations, or *iterative* if they require a (theoretically) infinite number.

7.1 Direct methods

The solution of a linear system can be performed through the Gauss elimination method (GEM), where the initial system A$\mathbf{x}$=$\mathbf{b}$ is reconducted in n steps to an equivalent system (i.e. having the same solution) of the form $\mathrm{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ where $\mathrm{A}^{(n)} = \mathrm{U}$ is a nonsingular upper triangular matrix and $\mathbf{b}^{(n)}$ is a new known term. It will be possible to solve the latter system with a computational cost of the order of n^2 operations,

A. Quarteroni: Numerical Models for Differential Problems.

through the following backward substitution algorithm:

$$\begin{aligned} x_n &= \frac{b_n^{(n)}}{u_{nn}}, \\ x_i &= \frac{1}{u_{ii}}\left(b_i^{(n)} - \sum_{j=i+1}^{n} u_{ij}x_j\right), \quad i = n-1, \ldots, 1. \end{aligned} \tag{7.3}$$

Denoting by $\mathrm{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$ the original system, the k-th step of GEM is achieved via the following formulae:

$$\begin{aligned} m_{ik} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, && i = k+1, \ldots, n, \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, && i, j = k+1, \ldots, n \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik}b_k^{(k)}, && i = k+1, \ldots, n. \end{aligned} \tag{7.4}$$

We note that in this way, the elements $a_{ij}^{(k+1)}$ with $j = k$ and $i = k+1, \ldots, n$ result to be null. The m_{ik} elements are called *multipliers*, while the denominators $a_{kk}^{(k)}$ are named *pivotal elements*. The GEM can obviously be achieved only if all the pivotal elements result to be non null. This happens, for instance, for symmetric positive definite matrices and for strict diagonal dominant ones. In general, it will be necessary to resort to the *pivoting* method, i.e. to the swapping of rows (and/or columns) of $\mathrm{A}^{(k)}$, in order to ensure that the element $a_{kk}^{(k)}$ be non-null.
To complete the Gauss eliminations, we need $2(n-1)n(n+1)/3 + n(n-1)$ flops, to which we must add n^2 flops to solve the upper triangular system $\mathrm{U}\mathbf{x} = \mathbf{b}^{(n)}$ via the backward substitution method. Hence, about $(2n^3/3 + 2n^2)$ flops are needed to solve the linear system via the GEM. More simply, by neglecting the lower order terms with respect to n, it can be said that the Gaussian elimination process requires $2n^3/3$ flops.

It can be verified that the GEM is equivalent to factorizing the A matrix, i.e. to rewriting A as the product LU of two matrices. The U matrix, upper triangular, coincides with the matrix $\mathrm{A}^{(n)}$ obtained at the end of the elimination process. The L matrix is lower triangular, its diagonal elements are equal to 1 while the ones located in the remaining lower triangular portion are equal to the multipliers.

Once the matrices L and U are known, the solution of the initial linear system simply involves the solution (in a sequence) of the two triangular systems

$$\mathrm{L}\mathbf{y} = \mathbf{b}, \quad \mathrm{U}\mathbf{x} = \mathbf{y}.$$

Obviously, the computational cost of the factorization process is the same as the one required by the GEM. The advantages of such a reinterpretation are evident: as L and U depend on A only and not on the known term, the same factorization can be used to solve different linear systems having the same matrix A, but a variable known term **b** (think for instance of the discretization of a linear parabolic problem by an implicit method where at each time step it is necessary to solve a system with the same matrix

all the time, but with a different known term). Consequently, as the computational cost is concentrated in the elimination procedure (about $2n^3/3$ flops), we have in this way a considerable reduction in the number of operations when we want to solve several linear systems having the same matrix.

If A is a symmetric positive definite matrix, the LU factorization can be conveniently specialized. Indeed, there exists only one upper triangular matrix H with positive elements on the diagonal such that

$$\mathrm{A} = \mathrm{H}^{\mathrm{T}}\mathrm{H}. \tag{7.5}$$

Equation (7.5) is the so-called Cholesky factorization. The h_{ij} elements of H^T are given by the following formulae: $h_{11} = \sqrt{a_{11}}$ and, for $i = 2, \ldots, n$,

$$h_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} h_{ik} h_{jk} \right) / h_{jj}, \quad j = 1, \ldots, i-1,$$
$$h_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} h_{ik}^2 \right)^{1/2}.$$

This algorithm only requires about $n^3/3$ flops, i.e. saves about twice the computing time of the LU factorization and about half the memory.

Let us now consider the particular case of a linear system with non-singular *tridiagonal* matrix A of the form

$$\mathrm{A} = \begin{bmatrix} a_1 & c_1 & & 0 \\ b_2 & a_2 & \ddots & \\ & \ddots & & c_{n-1} \\ 0 & & b_n & a_n \end{bmatrix}.$$

In this case, the L and U matrices of the LU factorization of A are two bidiagonal matrices of the type

$$\mathrm{L} = \begin{bmatrix} 1 & & & 0 \\ \beta_2 & 1 & & \\ & \ddots & \ddots & \\ 0 & & \beta_n & 1 \end{bmatrix}, \quad \mathrm{U} = \begin{bmatrix} \alpha_1 & c_1 & & 0 \\ & \alpha_2 & \ddots & \\ & & \ddots & c_{n-1} \\ 0 & & & \alpha_n \end{bmatrix}.$$

The α_i and β_i unknown coefficients can be easily computed through the following equations:

$$\alpha_1 = a_1, \quad \beta_i = \frac{b_i}{\alpha_{i-1}}, \quad \alpha_i = a_i - \beta_i c_{i-1}, \; i = 2, \ldots, n.$$

This algorithm is named *Thomas algorithm* and can be seen as a particular kind of LU factorization without pivoting.

7.2 Iterative methods

Iterative methods aim at constructing the solution $\mathbf{x}$ of a linear system as the limit of a sequence $\{\mathbf{x}^{(n)}\}$ of vectors. To obtain the single elements of the sequence, computing the residue $\mathbf{r}^{(n)} = \mathbf{b} - A\mathbf{x}^{(n)}$ of the system is required. In the case where the matrix is full and of order n, the computational cost of an iterative method is therefore of the order of n^2 operations per iteration. Such cost must be compared with the approximately $2n^3/3$ operations required by a direct method. Consequently, iterative methods are competitive with direct methods only if the number of necessary iterations to reach convergence (within a given tolerance) is independent of n or depends on n in a sub-linear way.

Other considerations in the choice between an iterative method and a direct one intervene as soon as the matrix is sparse.

7.2.1 Classical iterative methods

A general strategy to construct iterative methods is based on an additive decomposition, called splitting, starting from matrix A of the form A=P−N, where P and N are two suitable matrices and P is non-singular. For reasons which will become evident in the remainder, P is also called *preconditioning matrix* or *preconditioner*.

Precisely, given $\mathbf{x}^{(0)}$, we obtain $\mathbf{x}^{(k)}$ for $k \geq 1$ by solving the new systems

$$\mathrm{P}\mathbf{x}^{(k+1)} = \mathrm{N}\mathbf{x}^{(k)} + \mathbf{b}, \quad k \geq 0 \tag{7.6}$$

or, equivalently,

$$\mathbf{x}^{(k+1)} = \mathrm{B}\mathbf{x}^{(k)} + \mathrm{P}^{-1}\mathbf{b}, \quad k \geq 0 \tag{7.7}$$

having denoted by $\mathrm{B} = \mathrm{P}^{-1}\mathrm{N}$ the *iteration matrix*.

We are interested in *convergent* iterative methods, i.e. such that $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}$ for each choice of the *initial vector* $\mathbf{x}^{(0)}$, having denoted by $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ the error. As with a recursive argument we find

$$\mathbf{e}^{(k)} = \mathrm{B}^k\mathbf{e}^{(0)}, \qquad \forall k = 0, 1, \ldots \tag{7.8}$$

We can conclude that an iterative method of the form (7.6) is convergent if and only if $\rho(\mathrm{B}) < 1$, $\rho(\mathrm{B})$ being the spectral radius of the iteration matrix B, i.e. the maximum modulus of the eigenvalues of B.

Equation (7.6) can also be formulated in the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathrm{P}^{-1}\mathbf{r}^{(k)}, \tag{7.9}$$

having denoted by

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(k)} \tag{7.10}$$

the *residue* at step k. Equation (7.9) thus expresses the fact that to update the solution at step $k + 1$, it is necessary to solve a linear system with matrix P. Hence, besides

being non-singular, P must be invertible at a low computational cost if we want to prevent the overall cost of the scheme from increasing excessively (obviously, in the limit case where P is equal to A and N=0, method (7.9) would converge in only one iteration, but at the cost of a direct method).

Let us now see how to accelerate the convergence of the iterative methods (7.6) by exploiting the latter form. We denote by

$$\mathrm{R_P} = \mathrm{I} - \mathrm{P}^{-1}\mathrm{A}$$

the iteration matrix associated to method (7.9). Matrix (7.9) can be generalized by introducing a suitable relaxation (or acceleration) parameter α. This way, we obtain the *stationary Richardson methods* (more simply called *Richardson* methods), of the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathrm{P}^{-1}\mathbf{r}^{(k)}, \qquad k \geq 0. \tag{7.11}$$

More generally, supposing α to be dependent on the iteration index, we obtain the *non-stationary Richardson methods* given by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathrm{P}^{-1}\mathbf{r}^{(k)}, \qquad k \geq 0. \tag{7.12}$$

If we set $\alpha = 1$, we can recover two classical iterative methods: the Jacobi method if $\mathrm{P} = \mathrm{D(A)}$ (the diagonal part of A), the Gauss-Seidel method if $\mathrm{P} = \mathrm{L(A)}$ (the lower triangular part of A).
The iteration matrix at step k for such methods is given by

$$\mathrm{R}(\alpha_k) = \mathrm{I} - \alpha_k \mathrm{P}^{-1}\mathrm{A},$$

(note that the latter depends on k). In the case where P=I, the methods under exam will be called *non preconditioned*.

We can rewrite (7.12) (and therefore also (7.11)) in a form of greater computational interest. Indeed, having set $\mathbf{z}^{(k)} = \mathrm{P}^{-1}\mathbf{r}^{(k)}$ (the so-called *preconditioned residue*), we have that $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}$ and $\mathbf{r}^{(k+1)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathrm{A}\mathbf{z}^{(k)}$. To summarize, a non-stationary Richardson method at step $k+1$ requires the following operations:

$$\begin{aligned}
&\text{solving the linear system } \mathrm{P}\mathbf{z}^{(k)} = \mathbf{r}^{(k)},\\
&\text{computing the acceleration parameter } \alpha_k,\\
&\text{updating the solution } \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)},\\
&\text{updating the residue } \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathrm{A}\mathbf{z}^{(k)}.
\end{aligned} \tag{7.13}$$

As far as convergence is concerned for the stationary Richardson method (for which $\alpha_k = \alpha$, for each $k \geq 0$) the following result holds:

Property 7.1 *If* P *is a non-singular matrix, the stationary Richardson method* (7.11) *is convergent if and only if*

$$\frac{2\mathrm{Re}\lambda_i}{\alpha|\lambda_i|^2} > 1 \quad \forall i = 1, \ldots, n, \tag{7.14}$$

λ_i *being the eigenvalues of* $\mathrm{P}^{-1}\mathrm{A}$.
Moreover, if we suppose that $\mathrm{P}^{-1}\mathrm{A}$ *has positive real eigenvalues, ordered in such a way that* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$*, then, the stationary Richardson method* (7.11) *converges if and only if* $0 < \alpha < 2/\lambda_1$*. Having set*

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n} \tag{7.15}$$

the spectral radius of the iteration matrix R_α *is minimal if* $\alpha = \alpha_{opt}$*, with*

$$\rho_{opt} = \min_\alpha [\rho(\mathrm{R}_\alpha)] = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}. \tag{7.16}$$

If P and A are both symmetric and positive definite, it can be proven that the Richardson method converges monotonically with respect to the vector norms $\|\cdot\|_2$ and $\|\cdot\|_\mathrm{A}$. We recall that $\|\mathbf{v}\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$ and $\|\mathbf{v}\|_\mathrm{A} = (\sum_{i,j=1}^n v_i a_{ij} v_j)^{1/2}$.
In this case, thanks to (7.16), we can relate ρ_{opt} with the condition number introduced in Sec. 4.5.2 in the following way:

$$\rho_{opt} = \frac{K_2(\mathrm{P}^{-1}\mathrm{A}) - 1}{K_2(\mathrm{P}^{-1}\mathrm{A}) + 1}, \quad \alpha_{opt} = \frac{2\|\mathrm{A}^{-1}\mathrm{P}\|_2}{K_2(\mathrm{P}^{-1}\mathrm{A}) + 1}. \tag{7.17}$$

The importance of the choice of the P preconditioner in a Richardson method can therefore be clearly understood. We refer to Chap. 4 of [QSS07] for some examples of preconditioners.

7.2.2 Gradient and conjugate gradient methods

The optimal expression of the acceleration parameter α, indicated in (7.15), results to be of little practical utility, as it requires knowing the maximal and minimal eigenvalues of the matrix $\mathrm{P}^{-1}\mathrm{A}$. In the particular case of positive definite symmetric matrices, it is however possible to evaluate the optimal acceleration parameter in a *dynamic* way, that is as a function of quantities computed by the method itself at step k, as we show below.

First of all, we observe that in the case where A is a symmetric positive definite matrix, solving system (7.2) is equivalent to finding the minimum $\mathbf{x} \in \mathbb{R}^n$ of the quadratic form

$$\Phi(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T\mathrm{A}\mathbf{y} - \mathbf{y}^T\mathbf{b},$$

called *energy of the system* (7.2).
The problem is thus reconducted to determining the minimum point $\mathbf{x}$ of Φ starting from a point $\mathbf{x}^{(0)} \in \mathbb{R}^n$ and, consequently, choosing suitable directions along which to move to approach the solution $\mathbf{x}$ as quickly as possible. The optimal direction, joining $\mathbf{x}^{(0)}$ and $\mathbf{x}$, is obviously unknown a priori: we will therefore have to move from $\mathbf{x}^{(0)}$ along another direction $\mathbf{d}^{(0)}$ and fix a new point $\mathbf{x}^{(1)}$ on the latter, from which to repeat the procedure until convergence.

At the generic step k we will then determine $\mathbf{x}^{(k+1)}$ as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}, \tag{7.18}$$

α_k being the value fixing the length of the step along $\mathbf{d}^{(k)}$. The most natural idea, consisting in taking as downhill direction the maximal incline direction for Φ, given by $\mathbf{r}^{(k)} = -\nabla\Phi(\mathbf{x}^{(k)})$, leads to the *gradient* or *steepest descent method*.
The latter yields to the following algorithm: given $\mathbf{x}^{(0)} \in \mathbb{R}^n$, having set $\mathbf{r}^{(0)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(0)}$, for $k = 0, 1, \ldots$ until convergence, we compute

$$\alpha_k = \frac{{\mathbf{r}^{(k)}}^T \mathbf{r}^{(k)}}{{\mathbf{r}^{(k)}}^T \mathrm{A}\mathbf{r}^{(k)}}, \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)},$$
$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathrm{A}\mathbf{r}^{(k)}.$$

Its preconditioned version takes the following form: given $\mathbf{x}^{(0)} \in \mathbb{R}^n$, having set $\mathbf{r}^{(0)} = b - \mathrm{A}\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} = P^{-1}\mathbf{r}^{(0)}$, for $k = 0, 1, \ldots$ until convergence we compute

$$\alpha_k = \frac{{\mathbf{z}^{(k)}}^T \mathbf{r}^{(k)}}{{\mathbf{z}^{(k)}}^T \mathrm{A}\mathbf{z}^{(k)}}, \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}.$$
$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathrm{A}\mathbf{z}^{(k)}, \quad P\mathbf{z}^{(k+1)} = \mathbf{r}^{(k+1)}.$$

As far as the convergence properties of the descent method are concerned, the following result holds

Theorem 7.1 *Let* A *be symmetric and positive definite, then the gradient method converges for each value of the initial datum* $\mathbf{x}^{(0)}$ *and*

$$\|\mathbf{e}^{(k+1)}\|_{\mathrm{A}} \leq \frac{K_2(\mathrm{A}) - 1}{K_2(\mathrm{A}) + 1} \|\mathbf{e}^{(k)}\|_{\mathrm{A}}, \qquad k = 0, 1, \ldots \tag{7.19}$$

where $\|\cdot\|_{\mathrm{A}}$ *is the previously defined energy norm.*

A similar result, with $K_2(\mathrm{A})$ replaced by $K_2(\mathrm{P}^{-1}\mathrm{A})$, holds also in the case of the preconditioned gradient method, as long as we assume that P is also symmetric and positive definite.

An even more effective alternative consists in using the *conjugate gradient method* where the descent directions no longer coincide with that of the residue. In particular,

having set $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, we seek directions of the form

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta_k \mathbf{p}^{(k)}, \quad k = 0, 1, \ldots \tag{7.20}$$

where the parameters $\beta_k \in \mathbb{R}$ are to be determined so that

$$(\mathrm{A}\mathbf{p}^{(j)})^T \mathbf{p}^{(k+1)} = 0, \quad j = 0, 1, \ldots, k. \tag{7.21}$$

Directions of this type are called A-orthogonal. The method in the preconditioned case then takes the form: given $\mathbf{x}^{(0)} \in \mathbb{R}^n$, having set $\mathbf{r}^{(0)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} = \mathrm{P}^{-1}\mathbf{r}^{(0)}$ and $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$, the k-th iteration, with $k = 0, 1 \ldots$, is

$$\begin{aligned}
&\alpha_k = \frac{\mathbf{p}^{(k)^T}\mathbf{r}^{(k)}}{(\mathrm{A}\mathbf{p}^{(k)})^T\mathbf{p}^{(k)}},\\
&\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)},\\
&\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathrm{A}\mathbf{p}^{(k)},\\
&\mathrm{P}\mathbf{z}^{(k+1)} = \mathbf{r}^{(k+1)},\\
&\beta_k = \frac{(\mathrm{A}\mathbf{p}^{(k)})^T\mathbf{z}^{(k+1)}}{\mathbf{p}^{(k)^T}\mathrm{A}\mathbf{p}^{(k)}},\\
&\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} - \beta_k \mathbf{p}^{(k)}.
\end{aligned}$$

The parameter α_k is chosen in order to guarantee that the error $\|(e)^{(k+1)}\|_{\mathrm{A}}$ be minimized along the descent direction $\mathbf{p}^{(k)}$. The β_k parameter, instead, is chosen so that the new direction $\mathbf{p}^{(k+1)}$ be A-conjugate with $\mathbf{p}^{(k)}$, or $(\mathrm{A}\mathbf{p}^{(k)})^T\mathbf{p}^{(k+1)} = 0$. Indeed, it can be proven (thanks to the induction principle) that if the latter relation is verified, then so are all the ones in (7.21) relating to $j = 0, ..., k-1$. For a complete derivation of the method, see e.g. [QSS07, Chap. 4] or [Saa96].
It can be proven that the conjugate gradient method converges in exact arithmetics in at most n steps and that

$$\|\mathbf{e}^{(k)}\|_{\mathrm{A}} \leq \frac{2c^k}{1 + c^{2k}} \|\mathbf{e}^{(0)}\|_{\mathrm{A}}, \tag{7.22}$$

with

$$c = \frac{\sqrt{K_2(\mathrm{P}^{-1}\mathrm{A})} - 1}{\sqrt{K_2(\mathrm{P}^{-1}\mathrm{A})} + 1}. \tag{7.23}$$

Consequently, in the absence of roundoff errors, the CG method can be seen as a direct method as it terminates after a finite number of operations.
On the other hand, for matrices of large dimension, it is usually applied as an iterative method and is arrested when an error estimator (as for instance the relative residue) is less than a given tolerance.

Thanks to (7.23), the dependence on the reduction factor of the error on the matrix condition number is more favorable than the one of the gradient method (due to the presence of the square root of $K_2(P^{-1}\mathrm{A})$).

It can be noted that the number of iterations required for convergence (up to a prescribed tolerance) is proportional to $\frac{1}{2}\sqrt{K_2(P^{-1}A)}$ for the preconditioned conjugate gradient method, a decided improvement w.r.t $\frac{1}{2}K_2(P^{-1}A)$ for the preconditioned gradient method. Of course, the PCG method is more costly per iteration, both in CPU time and storage.

7.2.3 Krylov subspace methods

Generalizations of the gradient method in the case where matrix A is not symmetric lead to the so-called Krylov methods. Notable examples are the GMRES method and the conjugate bigradient method BiCG, as well as its stabilized version, the BiCGSTAB method. We refer the interested reader to [QSS07, Chap. 4], [Saa96] and [vdV03].

Here we briefly review the GMRES (generalized minimal residual) method. We start by a revisitation of the Richardson method (7.13) with $\mathrm{P}=\mathrm{I}$; the residual at the k-th step can be related to the initial residual as

$$\mathbf{r}^{(k)} = \prod_{j=0}^{k-1}(\mathrm{I}-\alpha_j\mathrm{A})\mathbf{r}^{(0)} = p_k(\mathrm{A})\mathbf{r}^{(0)}, \tag{7.24}$$

where $p_k(\mathrm{A})$ is a polynomial in A of degree k. If we introduce the space

$$K_m(\mathrm{A};\mathbf{v}) = \mathrm{span}\{\mathbf{v}, \mathrm{A}\mathbf{v}, \ldots, \mathrm{A}^{m-1}\mathbf{v}\}, \tag{7.25}$$

it immediately appears from (7.24) that $\mathbf{r}^{(k)} \in K_{k+1}(\mathrm{A};\mathbf{r}^{(0)})$. The space defined in (7.25) is called the *Krylov subspace* of order m associated with matrix A and vector $\mathbf{v}$. It is a subspace of the set spanned by all the vectors $\mathbf{u} \in \mathbb{R}^n$ that can be written as $\mathbf{u} = p_{m-1}(\mathrm{A})\mathbf{v}$, where p_{m-1} is a polynomial in A of degree $\leq m-1$.

Similarly, the iterate $\mathbf{x}^{(k)}$ of the Richardson method can be represented as follows

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + \sum_{j=0}^{k-1}\alpha_j\mathbf{r}^{(j)},$$

whence $\mathbf{x}^{(k)}$ belongs to the space

$$W_k = \{\mathbf{v} = \mathbf{x}^{(0)} + \mathbf{y},\ \mathbf{y} \in K_k(\mathrm{A};\mathbf{r}^{(0)})\}. \tag{7.26}$$

Notice also that $\sum_{j=0}^{k-1}\alpha_j\mathbf{r}^{(j)}$ is a polynomial in A of degree less than $k-1$. In the nonpreconditioned Richardson method we are thus looking for an approximate solution to $\mathbf{x}$ in the space W_k. More generally, one can devise methods that search for approximate solutions of the form

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + q_{k-1}(\mathrm{A})\mathbf{r}^{(0)}, \tag{7.27}$$

where q_{k-1} is a polynomial selected in such a way that $\mathbf{x}^{(k)}$ be, in a sense that must be made precise, the best approximation of $\mathbf{x}$ in W_k. A method that looks for a solution of the form (7.27) with W_k defined as in (7.26) is called a *Krylov method*.

A first question concerning Krylov subspace iterations is whether the dimension of $K_m(\mathrm{A};\mathbf{v})$ increases as the order m grows. A partial answer is provided by the following result.

Property 7.2 *Let $\mathrm{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{v} \in \mathbb{R}^n$. The Krylov subspace $K_m(\mathrm{A};\mathbf{v})$ has dimension equal to m iff the degree of $\mathbf{v}$ with respect to A, denoted by $\deg_{\mathrm{A}}(\mathbf{v})$, is not less than m; the degree of $\mathbf{v}$ is defined as the minimum degree of a monic nonnull polynomial p in A, for which $p(\mathrm{A})\mathbf{v} = \mathbf{0}$.*

The dimension of $K_m(\mathrm{A};\mathbf{v})$ is thus equal to the minimum between m and the degree of $\mathbf{v}$ with respect to A and, as a consequence, the dimension of the Krylov subspaces is certainly a nondecreasing function of m. The degree of $\mathbf{v}$ cannot be greater than n due to the Cayley-Hamilton theorem (see [QSS07, Sec. 1.7]).

Example 7.1 Consider the 4×4 matrix $\mathrm{A} = \mathrm{tridiag}_4(-1, 2, -1)$. The vector $\mathbf{v} = [1, 1, 1, 1]^T$ has degree 2 with respect to A since $p_2(\mathrm{A})\mathbf{v} = \mathbf{0}$ with $p_2(\mathrm{A}) = \mathrm{I}_4 - 3\mathrm{A} + \mathrm{A}^2$ (I_4 is the 4×4 identity matrix), while there is no monic polynomial p_1 of degree 1 for which $p_1(\mathrm{A})\mathbf{v} = \mathbf{0}$. As a consequence, all Krylov subspaces from $K_2(\mathrm{A};\mathbf{v})$ on, have dimension equal to 2. The vector $\mathbf{w} = [1, 1, -1, 1]^T$ has, instead, degree 4 with respect to A. ■

For a fixed m, it is possible to compute an orthonormal basis for $K_m(\mathrm{A};\mathbf{v})$ using the so-called *Arnoldi algorithm*.

Setting $\mathbf{v}_1 = \mathbf{v}/\|\mathbf{v}\|_2$, this method generates an orthonormal basis $\{\mathbf{v}_i\}$ for $K_m(\mathrm{A};\mathbf{v}_1)$ using the Gram-Schmidt procedure (see [QSS07, Sec. 3.4.3]). For $k = 1,\ldots,m$, the Arnoldi algorithm computes

$$\begin{aligned} &h_{ik} = \mathbf{v}_i^T \mathrm{A}\mathbf{v}_k, && i = 1, 2, \ldots, k, \\ &\mathbf{w}_k = \mathrm{A}\mathbf{v}_k - \sum_{i=1}^{k} h_{ik}\mathbf{v}_i, && h_{k+1,k} = \|\mathbf{w}_k\|_2. \end{aligned} \tag{7.28}$$

If $\mathbf{w}_k = \mathbf{0}$ the process terminates and in such a case we say that a *breakdown* of the algorithm has occurred; otherwise, we set $\mathbf{v}_{k+1} = \mathbf{w}_k/\|\mathbf{w}_k\|_2$ and the algorithm restarts, incrementing k by 1.

It can be shown that if the method terminates at the step m then the vectors $\mathbf{v}_1,\ldots,\mathbf{v}_m$ form a basis for $K_m(\mathrm{A};\mathbf{v})$. In such a case, if we denote by $\mathrm{V}_m \in \mathbb{R}^{n\times m}$ the matrix whose columns are the vectors $\mathbf{v}_i$, we have

$$\mathrm{V}_m^T \mathrm{A}\mathrm{V}_m = \mathrm{H}_m, \quad \mathrm{V}_{m+1}^T \mathrm{A}\mathrm{V}_m = \widehat{\mathrm{H}}_m, \tag{7.29}$$

where $\widehat{\mathrm{H}}_m \in \mathbb{R}^{(m+1)\times m}$ is the upper Hessenberg matrix whose entries h_{ij} are given by (7.28) and $\mathrm{H}_m \in \mathbb{R}^{m\times m}$ is the restriction of $\widehat{\mathrm{H}}_m$ to the first m rows and m columns.

The algorithm terminates at an intermediate step $k < m$ iff $\deg_{\mathrm{A}}(\mathbf{v}_1) = k$. As for the stability of the procedure, all the considerations valid for the Gram-Schmidt method hold. For more efficient and stable computational variants of (7.28), we refer to [Saa96].

We are now ready to solve the linear system (7.2) by a Krylov method. We look for the iterate $\mathbf{x}^{(k)}$ under the form (7.27); for a given $\mathbf{r}^{(0)}$, $\mathbf{x}^{(k)}$ is selected as being the unique element in W_k which satisfies a criterion of minimal distance from $\mathbf{x}$. The criterion for selecting $\mathbf{x}^{(k)}$ is precisely the distinguishing feature of a Krylov method.

The most natural idea consists of searching for $\mathbf{x}^{(k)} \in W_k$ as the vector which minimizes the Euclidean norm of the error. This approach, however, does not work in practice since $\mathbf{x}^{(k)}$ would depend on the (unknown) solution $\mathbf{x}$. Two alternative strategies can be pursued:

1. compute $\mathbf{x}^{(k)} \in W_k$ enforcing that the residual $\mathbf{r}^{(k)}$ is orthogonal to any vector in $K_k(\mathrm{A}; \mathbf{r}^{(0)})$, i.e., we look for $\mathbf{x}^{(k)} \in W_k$ such that
$$\mathbf{v}^T(\mathbf{b} - \mathrm{A}\mathbf{x}^{(k)}) = 0 \qquad \forall \mathbf{v} \in K_k(\mathrm{A}; \mathbf{r}^{(0)}); \tag{7.30}$$
2. compute $\mathbf{x}^{(k)} \in W_k$ minimizing the Euclidean norm of the residual $\|\mathbf{r}^{(k)}\|_2$, i.e.
$$\|\mathbf{b} - \mathrm{A}\mathbf{x}^{(k)}\|_2 = \min_{\mathbf{v} \in W_k} \|\mathbf{b} - \mathrm{A}\mathbf{v}\|_2. \tag{7.31}$$

Satisfying (7.30) leads to the Arnoldi method for linear systems (more commonly known as FOM, *full orthogonalization method*), while satisfying (7.31) yields the GMRES (*generalized minimum residual*) method.

We shall assume that k steps of the Arnoldi algorithm have been carried out, in such a way that an orthonormal basis for $K_k(\mathrm{A}; \mathbf{r}^{(0)})$ has been generated and stored into the column vectors of the matrix V_k with $\mathbf{v}_1 = \mathbf{r}^{(0)}/\|\mathbf{r}^{(0)}\|_2$. In such a case the new iterate $\mathbf{x}^{(k)}$ can always be written as

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + \mathrm{V}_k \mathbf{z}^{(k)}, \tag{7.32}$$

where $\mathbf{z}^{(k)}$ must be selected according to a suitable criterion that we are going to specify. Consequently we have

$$\mathbf{r}^{(k)} = \mathbf{r}^{(0)} - \mathrm{A}\mathrm{V}_k \mathbf{z}^{(k)}. \tag{7.33}$$

Since $\mathbf{r}^{(0)} = \mathbf{v}_1 \|\mathbf{r}^{(0)}\|_2$, using (7.29), relation (7.33) becomes

$$\mathbf{r}^{(k)} = \mathrm{V}_{k+1}(\|\mathbf{r}^{(0)}\|_2 \mathbf{e}_1 - \widehat{\mathrm{H}}_k \mathbf{z}^{(k)}), \tag{7.34}$$

where $\mathbf{e}_1$ is the first unit vector of $\mathbb{R}^{k+1}$. Therefore, in the GMRES method the solution at step k can be computed through (7.32), provided

$$\mathbf{z}^{(k)} \text{ minimizes } \| \, \|\mathbf{r}^{(0)}\|_2 \mathbf{e}_1 - \widehat{\mathrm{H}}_k \mathbf{z}^{(k)}\|_2 \tag{7.35}$$

(we note that the matrix V_{k+1} appearing in (7.34) does not alter the value of $\|\cdot\|_2$ since it is orthogonal). Having to solve at each step a least-squares problem of size k, the GMRES method will be the more effective the smaller is the number of iterations.

Similarly to the CG method, the GMRES method has the finite termination property, that is it terminates at most after n iterations, yielding the exact solution (in exact arithmetic). Indeed, the $k-th$ iterate minimizes the residual in the Krylov subspace K_k. Since every subspace is contained in the next subspace, the residual decreases monotonically. After n iterations, where n is the size of the matrix A, the Krylov space K_n is the whole of $\mathbb{R}^n$ and hence the GMRES method arrives at the exact solution. Premature stops are due to a breakdown in the orthonormalization Arnoldi algorithm. More precisely, we have the following result.

Property 7.3 *A breakdown occurs for the GMRES method at a step m (with $m < n$) if and only if the computed solution $\mathbf{x}^{(m)}$ coincides with the exact solution to the system.*

However, the idea is that after a small number of iterations (relative to n), the vector $\mathbf{x_k}$ is already a good approximation to the exact solution. This is confirmed by the convergence results that we report later in this section.

To improve the efficiency of the GMRES algorithm it is necessary to devise a stopping criterion which does not require the explicit evaluation of the residual at each step. This is possible, provided that the linear system with upper Hessenberg matrix $\widehat{\mathrm{H}}_k$ is appropriately solved.

In practice, matrix $\widehat{\mathrm{H}}_k$ in (7.29) is transformed into an upper triangular matrix $\mathrm{R}_k \in \mathbb{R}^{(k+1)\times k}$ with $r_{k+1,k} = 0$ such that $\mathrm{Q}_k^T \mathrm{R}_k = \widehat{\mathrm{H}}_k$, where Q_k is a matrix obtained as the product of k Givens rotations. Then, since Q_k is orthogonal, it can be seen that minimizing $\|\|\mathbf{r}^{(0)}\|_2 \mathbf{e}_1 - \widehat{\mathrm{H}}_k \mathbf{z}^{(k)}\|_2$ is equivalent to minimize $\|\mathbf{f}_k - \mathrm{R}_k \mathbf{z}^{(k)}\|_2$, with $\mathbf{f}_k = \mathrm{Q}_k \|\mathbf{r}^{(0)}\|_2 \mathbf{e}_1$. It can also be shown that the $k+1$-th component of $\mathbf{f}_k$ is, in absolute value, the Euclidean norm of the residual at the k-th step.

As FOM, the GMRES method entails a high computational effort and a large amount of memory, unless convergence occurs after few iterations. For this reason, two variants of the algorithm are available, one named GMRES(m) and based on the *restart* after m steps, with $\mathbf{x}(m)$ as initial guess, the other named Quasi-GMRES or QGMRES and based on stopping the Arnoldi orthogonalization process. It is worth noting that these two methods do not enjoy Property 7.3.

The convergence analysis of GMRES is not trivial, and we report just some of the more elementary results here. If A is positive definite, i.e., its symmetric part A_S has positive eigenvalues, then the k-th residual decreases according to the following bound

$$\|\mathbf{r}^{(k)}\|_2 \leq \sin^k(\beta) \|\mathbf{r}^{(0)}\|_2 \,, \tag{7.36}$$

where $\cos(\beta) = \lambda_{\min}(L_S)/\|L\|$ with $\beta \in [0, \pi/2)$. Moreover, GMRES(m) converges for all $m \geq 1$. In order to obtain a bound on the residual at a step $k \geq 1$, let us assume that the matrix A is diagonalizable

$$A = T \Lambda T^{-1} \,,$$

where Λ is the diagonal matrix of eigenvalues, $\{\lambda_j\}_{j=1,\dots,n}$, and $T = [\boldsymbol{\omega}^1, \dots, \boldsymbol{\omega}^n]$ is the matrix whose columns are the right eigenvectors of A. Under these assumptions, the residual norm after k steps of GMRES satisfies

$$\|\mathbf{r}^{(k)}\| \leq K_2(T)\delta\|\mathbf{r}^{(0)}\| \,,$$

where $K_2(T) = \|T\|_2\|T^{-1}\|_2$ is the condition number of T and

$$\delta = \min_{p\in\mathbb{P}_k, p(0)=1} \max_{1\leq i\leq k} |p(\lambda_i)| \,.$$

Moreover, suppose that the initial residual is dominated by m eigenvectors, i.e., $\mathbf{r}^0 = \sum_{j=1}^m \alpha_j\boldsymbol{\omega}^j + \mathbf{e}$, with $\|\mathbf{e}\|$ small in comparison to $\|\sum_{j=1}^m \alpha_j\boldsymbol{\omega}^j\|$, and assume that if some complex $\boldsymbol{\omega}^j$ appears in the previous sum, then its conjugate $\overline{\boldsymbol{\omega}}^j$ appears as well. Then

$$\|\mathbf{r}^{(k)}\| \leq K_2(T)c_k\|\mathbf{e}\| \,,$$

$$c_k = \max_{p>k} \prod_{j=1}^{k} \left|\frac{\lambda_p - \lambda_j}{\lambda_j}\right| \,.$$

Very often, c_k is of order one; hence, k steps of GMRES reduce the residual norm to the order of $\|\mathbf{e}\|$ provided that $\kappa_2(T)$ is not too large.

In general, as highlighted from the previous estimate, the eigenvalue information alone is not enough, and information on the eigensystem is also needed. If the eigensystem is orthogonal, as for normal matrices, then $K_2(T) = 1$, and the eigenvalues are descriptive for convergence. Otherwise, upper bounds for $\|\mathbf{r}^{(k)}\|$ can be provided in terms of both spectral and pseudospectral information, as well as the so-called *field of values* of A

$$\mathcal{F}(A) = \{\mathbf{v}^* A\mathbf{v} \mid \|\mathbf{v}\| = 1\}.$$

If $0 \notin \mathcal{F}(A)$, then the estimate (7.36) can be improved by replacing $\lambda_{\min}(A_S)$ with $\text{dist}(0, \mathcal{F}(A))$.

An extensive discussion of convergence of GMRES and GMRES(m) can be found in [Saa96], [Emb99], [Emb03], [TE05], and [vdV03].

The GMRES method can of course be implemented for a preconditioned system. We provide here an implementation of the preconditioned GMRES method with a left preconditioner P.

Preconditioned GMRES (PGMRES) Method. Initialize

$$\mathbf{x}^{(0)}, P\mathbf{r}^{(0)} = \mathbf{f} - A\mathbf{x}^{(0)}, \beta = \|\mathbf{r}^{(0)}\|_2, \mathbf{x}^{(1)} = \mathbf{r}^{(0)}/\beta.$$

Iterate

$$
\begin{aligned}
&For\ j = 1, \ldots, k\ Do \\
&\quad \text{Compute } P\mathbf{w}^{(j)} = A\mathbf{x}^{(j)} \\
&\quad For\ i = 1, \ldots, j\ Do \\
&\qquad g_{ij} = (\mathbf{x}^{(i)})^T \mathbf{w}^{(j)} \\
&\qquad \mathbf{w}^{(j)} = \mathbf{w}^{(j)} - g_{ij}\mathbf{x}_i \\
&\quad End\ Do \\
&\quad g_{j+1,j} = \|\mathbf{w}^{(j)}\|_2 \\
&\quad (if\ g_{j+1,j} = 0 \text{ set } k = j \text{ and } Goto\ (1)) \\
&\quad \mathbf{x}^{(j+1)} = \mathbf{w}^{(j)}/g_{j+1,j} \\
&End\ Do \\
&V_k = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}],\ \hat{H}_k = \{g_{ij}\},\ 1 \le j \le k,\ 1 \le i \le j+1; \\
&(1) \text{ Compute } \mathbf{z}^{(k)}\ , \text{ the minimizer of } \|\beta \mathbf{e}_1 - \hat{H}_k \mathbf{z}\| \\
&\text{Set } \mathbf{x}^{(k)} = \mathbf{x}^{(0)} + V_k \mathbf{z}^{(k)}
\end{aligned}
\tag{7.37}
$$

More generally, as proposed by Saad (1996), a variable preconditioner P_k can be used at the k-th iteration, yielding the so-called *flexible GMRES* method. The use of a variable preconditioner is especially interesting in those situations where the preconditioner is not explicitly given, but implicitly defined, for instance, as an approximate Jacobian in a Newton iteration or by a few steps of an inner iteration process (see Chap. 15). Another meaningful case is the one of domain decomposition preconditioners (of either Schwarz or Schur type) where the preconditioning step involves one or several substeps of local solves in the subdomains (see Chap. 17).

Several considerations for the practical implementation of GMRES, its relation with FOM, how to restart GMRES, and the Householder version of GMRES can be found in [Saa96].

Remark 7.1 (Projection methods) Denoting by Y_k and L_k two generic m-dimensional subspaces of $\mathbb{R}^n$, we call *projection method* a process which generates an approximate solution $\mathbf{x}^{(k)}$ at step k, enforcing that $\mathbf{x}^{(k)} \in Y_k$ and that the residual $\mathbf{r}^{(k)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(k)}$ be orthogonal to L_k. If $Y_k = L_k$, the projection process is said to be *orthogonal*, *oblique* otherwise (see [Saa96]).

The Krylov subspace iterations can be regarded as being projection methods. For instance, the Arnoldi method (see [Saa96]) is an orthogonal projection method where $L_k = Y_k = K_k(\mathrm{A}; \mathbf{r}^{(0)})$, while the GMRES method is an oblique projection method with $Y_k = K_k(\mathrm{A}; \mathbf{r}^{(0)})$ and $L_k = \mathrm{A}Y_k$. It is worth noticing that some classical methods introduced in previous sections fall into this category. For example, the Gauss-Seidel method is an orthogonal projection method where at the k-th step $K_k(\mathrm{A}; \mathbf{r}^{(0)}) = \mathrm{span}\{\mathbf{e}_k\}$, with $k = 1, \ldots, n$. The projection steps are carried out cyclically from 1 to n until convergence. •

12

Finite differences for hyperbolic equations

In this chapter, we will deal with time-dependent problems of hyperbolic type. For their derivation and for an in-depth analysis see e.g. [Sal08], Chap. 4. We will limit ourselves to considering the numerical approximation using the finite difference method, which was historically the first one to be applied to this type of equations. To introduce in a simple way the basic concepts of the theory, most of our presentation will concern problems depending on a single space variable. Finite element approximations will be addressed in Chap. 13, the extension to nonlinear problems in Chap. 14.

12.1 A scalar transport problem

Let us consider the following scalar hyperbolic problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0, \quad x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{cases} \tag{12.1}$$

where $a \in \mathbb{R} \setminus \{0\}$. The solution of such problem is a wave travelling at velocity a, given by

$$u(x,t) = u_0(x - at), \quad t \geq 0.$$

We consider the curves $x(t)$ in the plane (x,t), solutions of the following ordinary differential equations

$$\begin{cases} \dfrac{dx}{dt} = a, \qquad t > 0, \\ x(0) = x_0, \end{cases}$$

for varying values of $x_0 \in \mathbb{R}$.

A. Quarteroni: Numerical Models for Differential Problems.

Such curves are called *characteristic lines* (often simply characteristics) and the solution along these lines remains constant as

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0.$$

In the case of the more general problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} + a_0 u = f, & x \in \mathbb{R}, t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{12.2}$$

where a, a_0, and f are given functions of the (x,t) variables, the characteristic lines $x(t)$ are the solutions of the Cauchy problem

$$\begin{cases} \dfrac{dx}{dt} = a(x,t), & t > 0, \\ x(0) = x_0. \end{cases}$$

In such case, the solutions of (12.2) satisfy the following relation

$$\frac{d}{dt}u(x(t),t) = f(x(t),t) - a_0(x(t),t)u(x(t),t).$$

It is therefore possible to extract the solution u by solving an ordinary differential equation on each characteristic curve (this approach leads to the so-called *characteristic method*).
Let us now consider problem (12.1) in a bounded interval. For instance, let us suppose $x \in [0,1]$ and $a > 0$. As u is constant on the characteristics, from Fig. 12.1 we deduce that the value of the solution at point P coincides with the value of u_0 at the foot P_0 of the characteristic outgoing from P. Instead, the characteristic outgoing from point Q, intersects the straight line $x = 0$ for $t > 0$. The point $x = 0$ is therefore an inflow point and must necessarily be assigned the value of u. Note that if $a < 0$, the inflow point would be $x = 1$.

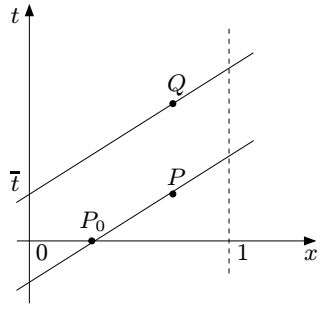


Fig. 12.1. Examples of characteristic lines (of straight lines in this case) issuing from points P and Q

By referring to problem (12.1) it is useful to observe that if u_0 were a discontinuous function at x_0, then such discontinuity would propagate along the characteristic outgoing from x_0 (this process can be rigorously formalized from a mathematical viewpoint, by introducing the concept of *weak solution* for hyperbolic problems). In order to regularize the discontinuity, one could approximate the initial datum u_0 with a sequence of regular functions $u_0^\varepsilon(x), \varepsilon > 0$. However, this procedure is only effective if the hyperbolic problem is linear. The solutions of non-linear hyperbolic problems can indeed develop discontinuities also for regular initial data (as we will see in Chap. 14). In this case, the strategy (which also inspires numerical methods) is to regularize the differential equation itself rather than the initial datum. In such perspective, we can consider the following diffusion-transport equation

$$\frac{\partial u^\varepsilon}{\partial t} + a\frac{\partial u^\varepsilon}{\partial x} = \varepsilon\frac{\partial^2 u^\varepsilon}{\partial x^2}, \quad x \in \mathbb{R},\ t > 0,$$

for small values of $\varepsilon > 0$, which can be regarded as a parabolic regularization of equation (12.1). If we set $u^\varepsilon(x, 0) = u_0(x)$, we can prove that

$$\lim_{\varepsilon \to 0^+} u^\varepsilon(x, t) = u_0(x - at),\ t > 0,\ x \in \mathbb{R}.$$

12.1.1 An a priori estimate

Let us now return to the transport-reaction problem (12.2) on a bounded interval

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} + a_0 u = f, & x \in (\alpha, \beta),\ t > 0, \\ u(x, 0) = u_0(x), & x \in [\alpha, \beta], \\ u(\alpha, t) = \varphi(t), & t > 0, \end{cases} \tag{12.3}$$

where $a(x)$, $f(x, t)$ and $\varphi(t)$ are assigned functions; we have made the assumption that $a(x) > 0$, so that $x = \alpha$ is the inflow point (where to impose the boundary condition), while $x = \beta$ is the outflow point.

By multiplying the first equation of (12.3) by u, integrating with respect to x and using the formula of integration by parts, we obtain for each $t > 0$

$$\frac{1}{2}\frac{d}{dt}\int_\alpha^\beta u^2\, dx + \int_\alpha^\beta (a_0 - \frac{1}{2}a_x)u^2\, dx + \frac{1}{2}(au^2)(\beta) - \frac{1}{2}(au^2)(\alpha) = \int_\alpha^\beta fu\, dx.$$

By supposing that there exists a $\mu_0 \geq 0$ s.t.

$$a_0 - \tfrac{1}{2}a_x \geq \mu_0 \quad \forall x \in [\alpha, \beta],$$

we find

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{\mathrm{L}^2(\alpha,\beta)} + \mu_0\|u(t)\|^2_{\mathrm{L}^2(\alpha,\beta)} + \frac{1}{2}(au^2)(\beta) \leq \int_\alpha^\beta fu\, dx + \frac{1}{2}a(\alpha)\varphi^2(t).$$

If f and φ are identically null, then

$$\|u(t)\|_{\mathrm{L}^2(\alpha,\beta)} \leq \|u_0\|_{\mathrm{L}^2(\alpha,\beta)} \quad \forall t > 0.$$

In the case of the more general problem (12.2), if we suppose that $\mu_0 > 0$, thanks to the Cauchy-Schwarz and Young inequalities we have

$$\int_\alpha^\beta fu\, dx \leq \|f\|_{\mathrm{L}^2(\alpha,\beta)}\|u\|_{\mathrm{L}^2(\alpha,\beta)} \leq \frac{\mu_0}{2}\|u\|^2_{\mathrm{L}^2(\alpha,\beta)} + \frac{1}{2\mu_0}\|f\|^2_{\mathrm{L}^2(\alpha,\beta)}.$$

Integrating over time, we get the following a priori estimate

$$\begin{aligned}\|u(t)\|^2_{\mathrm{L}^2(\alpha,\beta)} \quad &+ \quad \mu_0\int_0^t \|u(s)\|^2_{\mathrm{L}^2(\alpha,\beta)}\, ds + a(\beta)\int_0^t u^2(\beta,s)\, ds \\ &\leq \quad \|u_0\|^2_{\mathrm{L}^2(\alpha,\beta)} + a(\alpha)\int_0^t \varphi^2(s)\, ds + \frac{1}{\mu_0}\int_0^t \|f\|^2_{\mathrm{L}^2(\alpha,\beta)}\, ds.\end{aligned}$$

An alternative estimate that does not require differentiability of $a(x)$ but uses the hypothesis that $a_0 \leq a(x) \leq a_1$ for two suitable positive constants a_0 and a_1 can be obtained by multiplying the equation by a^{-1},

$$a^{-1}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = a^{-1} f.$$

By now multiplying by u and integrating between α and β we obtain, after a few simple steps,

$$\frac{1}{2}\frac{d}{dt}\int_\alpha^\beta a^{-1}(x)u^2(x,t)\, dx + \frac{1}{2}u^2(\beta,t) = \int_\alpha^\beta a^{-1}(x)f(x,t)u(x,t)dx + \frac{1}{2}\varphi^2(t).$$

If $f = 0$ we immediately obtain

$$\|u(t)\|^2_a + \int_0^t u^2(\beta,s)ds = \|u_0\|^2_a + \int_0^t \varphi^2(s)ds,\ t > 0.$$

We have defined

$$\|v\|_a = \left(\int_\alpha^\beta a^{-1}(x)v^2(x)dx\right)^{\frac{1}{2}}.$$

Thanks to the lower and upper bound of a^{-1}, the latter is an equivalent norm to that of $\mathrm{L}^2(\alpha,\beta)$. On the other hand, if $f \neq 0$, we can proceed as follows

$$\|u(t)\|^2_a + \int_0^t u^2(\beta,s)\, ds \leq \|u_0\|^2_a + \int_0^t \varphi^2(s)\, ds + \int_0^t \|f\|^2_a\, ds + \int_0^t \|u(s)\|^2_a ds,$$

having used the Cauchy-Schwarz inequality.

By now applying the Gronwall lemma (see Lemma 2.2) we obtain, for each $t > 0$,

$$\|u(t)\|_a^2 + \int_0^t u^2(\beta, s)\, ds \leq e^t \left(\|u_0\|_a^2 + \int_0^t \varphi^2(s) ds + \int_0^t \|f\|_a^2\, ds \right). \tag{12.4}$$

12.2 Systems of linear hyperbolic equations

Let us consider a linear system of the form

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + \mathrm{A}\dfrac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & x \in \mathbb{R},\ t > 0, \\ \mathbf{u}(0, x) = \mathbf{u}_0(x), & x \in \mathbb{R}, \end{cases} \tag{12.5}$$

where $\mathbf{u} : [0, \infty) \times \mathbb{R} \to \mathbb{R}^p$, $\mathrm{A} : \mathbb{R} \to \mathbb{R}^{p \times p}$ is a given matrix, and $\mathbf{u}_0 : \mathbb{R} \to \mathbb{R}^p$ is the initial datum.

Let us first consider the case where the coefficients of A are constant (i.e. independent of both x and t). The system (12.5) is called *hyperbolic* if A can be diagonalized and has real eigenvalues. In such case, there exists a non-singular matrix $\mathrm{T} : \mathbb{R} \to \mathbb{R}^{p \times p}$ such that

$$\mathrm{A} = \mathrm{T} \Lambda \mathrm{T}^{-1},$$

where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$, with $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, p$, is the diagonal matrix of the eigenvalues of A while $\mathrm{T} = [\boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \ldots, \boldsymbol{\omega}^p]$ is the matrix whose column vectors are the right eigenvectors of A, that is

$$\mathrm{A}\boldsymbol{\omega}^k = \lambda_k \boldsymbol{\omega}^k, \quad k = 1, \ldots, p.$$

Through this similarity transformation, it is possible to rewrite the system (12.5) in the form

$$\frac{\partial \mathbf{w}}{\partial t} + \Lambda \frac{\partial \mathbf{w}}{\partial x} = \mathbf{0}, \tag{12.6}$$

where $\mathbf{w} = \mathrm{T}^{-1}\mathbf{u}$ are called *characteristic variables*. In this way, we obtain p independent equations of the form

$$\frac{\partial w_k}{\partial t} + \lambda_k \frac{\partial w_k}{\partial x} = 0, \quad k = 1, \ldots, p,$$

analogous in all to the one of problem (12.1) (provided that we suppose a_0 and f null). The solution w_k is therefore constant along each *characteristic curve* $x = x(t)$, solution of the Cauchy problem

$$\begin{cases} \dfrac{dx}{dt} = \lambda_k, & t > 0, \\ x(0) = x_0. \end{cases} \tag{12.7}$$

Since the λ_k are constant, the characteristic curves are in fact the lines $x(t) = x_0 + \lambda_k t$ and the solutions feature the form $w_k(x,t) = \psi_k(x - \lambda_k t)$, where ψ_k is a function of a single variable, determined by the initial conditions. In the case of problem (12.5), we have that $\psi_k(x) = w_k(x,0)$, thus the solution $\mathbf{u} = \mathrm{T}\mathbf{w}$ will be of the form

$$\mathbf{u}(x,t) = \sum_{k=1}^{p} w_k(x - \lambda_k t, 0)\boldsymbol{\omega}^k.$$

As we can see, the latter is composed by p travelling, non-interacting waves.
As in a strictly hyperbolic system p different characteristic lines exit each point $(\overline{x}, \overline{t})$ of the plane (x,t), $u(\overline{x}, \overline{t})$ will only depend on the initial datum at the points $\overline{x} - \lambda_k \overline{t}$, for $k = 1, \ldots, p$. For this reason, the set of the p points that form the feet of the characteristics outgoing point $(\overline{x}, \overline{t})$, that is

$$D(\overline{x}, \overline{t}) = \{x \in \mathbb{R} \,|\, x = \overline{x} - \lambda_k \overline{t}\ ,\ k = 1, ..., p\}, \tag{12.8}$$

is called *domain of dependence* at point $(\overline{x}, \overline{t})$ of the solution $\mathbf{u}$.

In case we consider a bounded interval (α, β) instead of the whole real line, the sign of $\lambda_k,\ k = 1, \ldots, p$, denotes the inflow point for each of the characteristic variables. The ψ_k function in the case of a problem set on a bounded interval will be determined not only by the initial conditions, but also by the boundary conditions provided to the inflow of each characteristic variable. Having considered a point $(\overline{x}, \overline{t})$ with $\overline{x} \in (\alpha, \beta)$ and $\overline{t} > 0$, if $\overline{x} - \lambda_k \overline{t} \in (\alpha, \beta)$ then $w_k(\overline{x}, \overline{t})$ is determined by the initial condition, in particular we have $w_k(\overline{x}, \overline{t}) = w_k(\overline{x} - \lambda_k \overline{t}, 0)$. Conversely, if $\overline{x} - \lambda_k \overline{t} \notin (\alpha, \beta)$ then the value of $w_k(\overline{x}, \overline{t})$ will depend on the boundary condition (see Fig. 12.2):

$$\text{if } \lambda_k > 0\ ,\ w_k(\overline{x}, \overline{t}) = w_k(\alpha, \frac{\overline{x} - \alpha}{\lambda_k}),$$
$$\text{if } \lambda_k < 0\ ,\ w_k(\overline{x}, \overline{t}) = w_k(\beta, \frac{\overline{x} - \beta}{\lambda_k}).$$

As a consequence, the number of positive eigenvalues determines the number of boundary conditions to be assigned at $x = \alpha$, while at $x = \beta$ we will need to assign as many conditions as is the number of negative eigenvalues.

In the case where the coefficients of the matrix A in (12.5) are functions of x and t, we denote respectively by

$$\mathrm{L} = \begin{bmatrix} \mathbf{l}_1^T \\ \vdots \\ \mathbf{l}_p^T \end{bmatrix} \qquad \text{and} \qquad \mathrm{R} = [\mathbf{r}_1 \ldots \mathbf{r}_p]\,,$$

the matrices containing the left resp. right eigenvectors of A, whose elements satisfy the relations

$$\mathrm{A}\mathbf{r}_k = \lambda_k \mathbf{r}_k, \qquad \mathbf{l}_k^T \mathrm{A} = \lambda_k \mathbf{l}_k^T\,,$$

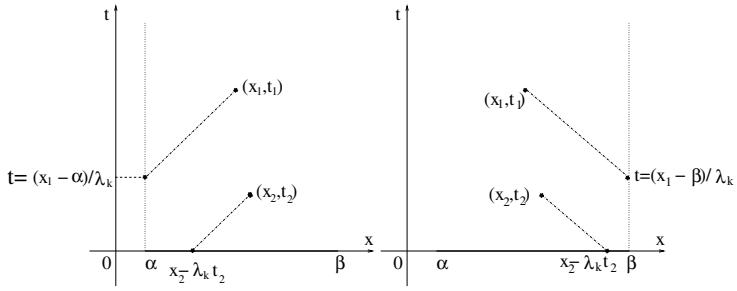


Fig. 12.2. The value of w_k at a point in the plane (x, t) depends either on the boundary condition or on the initial condition, based on the value of $x - \lambda_k t$. Both the positive (right) and negative (left) λ_k cases are reported

that is

$$\mathrm{AR} = \mathrm{R}\Lambda, \qquad \mathrm{LA} = \Lambda\mathrm{L}.$$

Without loss of generality, we can suppose that $\mathrm{LR} = \mathrm{I}$. Let us now suppose that there exists a vector function $\mathbf{w}$ satisfying the relations

$$\frac{\partial \mathbf{w}}{\partial \mathbf{u}} = \mathrm{R}^{-1}, \qquad \text{that is} \qquad \frac{\partial \mathbf{u}_k}{\partial \mathbf{w}} = \mathbf{r}_k, \qquad k = 1, \ldots, p.$$

Proceeding as we did initially, we obtain

$$\mathrm{R}^{-1}\frac{\partial \mathbf{u}}{\partial t} + \Lambda \mathrm{R}^{-1}\frac{\partial \mathbf{u}}{\partial x} = 0$$

hence the new diagonal system (12.6). By reintroducing the characteristic curves (12.7) (the latter will no longer be straight lines as the eigenvalues λ_k vary for different values of x and t), $\mathbf{w}$ is constant along them. The components of $\mathbf{w}$ are therefore still called characteristic variables. As $\mathrm{R}^{-1} = \mathrm{L}$ (thanks to the normalization relation) we obtain

$$\frac{\partial w_k}{\partial \mathbf{u}} \cdot \mathbf{r}_m = \mathbf{l}_k \cdot \mathbf{r}_m = \delta_{km}, \qquad k, m = 1, \ldots, p.$$

The functions w_k, $k = 1, \ldots, p$ are called *Riemann invariants* of the hyperbolic system.

12.2.1 The wave equation

Let us consider the following second order hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} - \gamma^2 \frac{\partial^2 u}{\partial x^2} = f, \quad x \in (\alpha, \beta), \quad t > 0. \tag{12.9}$$

Let

$$u(x, 0) = u_0(x) \quad \text{and} \quad \frac{\partial u}{\partial t}(x, 0) = v_0(x), \quad x \in (\alpha, \beta),$$

be the initial data and let us suppose, moreover, that u be identically null at the boundary

$$u(\alpha,t)=0 \quad \text{and} \quad u(\beta,t)=0, \quad t>0. \tag{12.10}$$

In this case, u can represent the vertical displacement of a vibrating elastic chord with lenght $\beta-\alpha$, fixed at the endpoints, and γ is a coefficient that depends on the specific mass of the chord and on its tension. The chord is subject to a vertical force whose density is f. The functions $u_0(x)$ and $v_0(x)$ describe the initial displacement resp. velocity of the chord.

For simplicity of notation, we denote by u_t the derivative $\frac{\partial u}{\partial t}$, by u_x the derivative $\frac{\partial u}{\partial x}$ and we use similar notations for the second derivatives.

Let us now suppose that f be null. From equation (12.9) we can deduce that in this case, the kinetic energy of the system is preserved, that is (see Exercise 1)

$$\|u_t(t)\|^2_{\mathrm{L}^2(\alpha,\beta)}+\gamma^2\,\|u_x(t)\|^2_{\mathrm{L}^2(\alpha,\beta)}=\|v_0\|^2_{\mathrm{L}^2(\alpha,\beta)}+\gamma^2\,\|u_{0x}\|^2_{\mathrm{L}^2(\alpha,\beta)}. \tag{12.11}$$

With the change of variables

$$\omega_1=u_x, \quad \omega_2=u_t,$$

the wave equation (12.9) becomes the following first-order system

$$\frac{\partial \boldsymbol{\omega}}{\partial t}+\mathrm{A}\frac{\partial \boldsymbol{\omega}}{\partial x}=\mathbf{f}, \quad x\in(\alpha,\beta), \quad t>0, \tag{12.12}$$

where

$$\boldsymbol{\omega}=\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}, \quad \mathrm{A}=\begin{bmatrix}0 & -1\\ -\gamma^2 & 0\end{bmatrix}, \quad \mathbf{f}=\begin{bmatrix}0\\ f\end{bmatrix},$$

whose initial conditions are $\omega_1(x,0)=u_0'(x)$ and $\omega_2(x,0)=v_0(x)$.
Since the eigenvalues of A are two distinct real numbers $\pm\gamma$ (representing the wave propagation rates), the system (12.12) is hyperbolic.
Note that, also in this case, to regular initial data correspond regular solutions, while discontinuities in the initial data will propagate along the characteristic lines $\dfrac{dx}{dt}=\pm\gamma$.

12.3 The finite difference method

Out of simplicity, we will now consider the case of problem (12.1). To numerically solve the latter, we can use spatio-temporal discretizations based on the finite difference method. In this case, the half-plane $\{t>0\}$ is discretized choosing a temporal step Δt, a spatial discretization step h and defining the gridpoints (x_j,t^n) in the following way

$$x_j=jh, \qquad j\in\mathbb{Z}, \qquad t^n=n\Delta t, \qquad n\in\mathbb{N}.$$

Let

$$\lambda = \Delta t/h,$$

and let us define

$$x_{j+1/2} = x_j + h/2.$$

We seek discrete solutions u_j^n which approximate $u(x_j, t^n)$ for each j and n.

The hyperbolic initial value problems are often discretized in time using explicit methods. Of course, this imposes restrictions on the values of λ that implicit methods generally don't have. For instance, let us consider problem (12.1). Any explicit finite difference method can be written in the form

$$u_j^{n+1} = u_j^n - \lambda(H_{j+1/2}^n - H_{j-1/2}^n), \tag{12.13}$$

where $H_{j+1/2}^n = H(u_j^n, u_{j+1}^n)$ for a suitable function $H(\cdot, \cdot)$ called *numerical flux*.

The numerical scheme (12.13) is basically the outcome of the following consideration. Suppose that a is constant and let us write equation (12.1) in conservation form

$$\frac{\partial u}{\partial t} + \frac{\partial (au)}{\partial x} = 0,$$

au being the *flux* associated to the equation. By integrating in space, we obtain

$$\int_{x_{j-1/2}}^{x_{j+1/2}} \frac{\partial u}{\partial t}\, dx + [au]_{x_{j-1/2}}^{x_{j+1/2}} = 0, \quad j \in \mathbb{Z},$$

that is

$$\frac{\partial}{\partial t} U_j + \frac{(au)(x_{j+\frac{1}{2}}) - (au)(x_{j-\frac{1}{2}})}{h} = 0, \quad \text{where} \quad U_j = h^{-1} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\, dx.$$

Equation (12.13) can now be interpreted as an approximation where the temporal derivative is discretized using the forward Euler finite difference scheme, U_j is replaced by u_j and $H_{j+1/2}$ is a suitable approximation of $(au)(x_{j+\frac{1}{2}})$.

12.3.1 Discretization of the scalar equation

In the context of explicit methods, the numerical methods are distinguished based on how the numerical flux H is chosen. In particular, we cite the following methods:

- **forward/centered Euler (FE/C)**

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2} a(u_{j+1}^n - u_{j-1}^n), \tag{12.14}$$

that takes the form (12.13) provided we define

$$H_{j+1/2} = \frac{1}{2}a(u_{j+1} + u_j). \tag{12.15}$$

– **Lax-Friedrichs (LF)**

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\lambda}{2}a(u_{j+1}^n - u_{j-1}^n), \tag{12.16}$$

also of the form (12.13) with

$$H_{j+1/2} = \frac{1}{2}[a(u_{j+1} + u_j) - \lambda^{-1}(u_{j+1} - u_j)]. \tag{12.17}$$

– **Lax-Wendroff (LW)**

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}a(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda^2}{2}a^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \tag{12.18}$$

that can be rewritten in the form (12.13) provided that we take

$$H_{j+1/2} = \frac{1}{2}[a(u_{j+1} + u_j) - \lambda a^2(u_{j+1} - u_j)]. \tag{12.19}$$

– **Upwind (or forward/decentered Euler) (U)**

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}a(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda}{2}|a|(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \tag{12.20}$$

corresponding to the form (12.13) provided that we choose

$$H_{j+1/2} = \frac{1}{2}[a(u_{j+1} + u_j) - |a|(u_{j+1} - u_j)]. \tag{12.21}$$

The LF method represents a modification of the FE/C method consisting in replacing the nodal value u_j^n in (12.14) with the average of the previous nodal value u_{j-1}^n and of the following one, u_{j+1}^n.
The LW method can be derived by first applying the Taylor development with respect to the temporal variable

$$u^{n+1} = u^n + (\partial_t u)^n \Delta t + (\partial_{tt} u)^n \frac{\Delta t^2}{2} + \mathcal{O}(\Delta t^3),$$

where $(\partial_t u)^n$ denotes the partial derivative of u at time t^n. Then, using equation (12.1), we replace $\partial_t u$ by $-a\partial_x u$, and $\partial_{tt} u$ by $a^2 \partial_{xx} u$. Neglecting the remainder $\mathcal{O}(\Delta t^3)$ and approximating the spatial derivatives with centered finite differences, we get to (12.18). Finally, the U method is obtained by discretizing the convective term $a\partial_x u$ of the equation with the upwind finite difference, as seen in Chap. 11, Sec. 11.6.

All of the previously introduced schemes are explicit. An example of implicit method is the following:

– **Backward/centered Euler (BE/C)**

$$u_j^{n+1} + \frac{\lambda}{2}a(u_{j+1}^{n+1} - u_{j-1}^{n+1}) = u_j^n. \tag{12.22}$$

Naturally, the implicit schemes can also be rewritten in a general form that is similar to (12.13) where H^n is replaced by H^{n+1}. In the specific case, the numerical flux will again be defined by (12.15).

The advantage of the (12.13) formulation is that it can easily be extended to the case of more general hyperbolic problems.
In particular, we will examine the case of linear systems in Chap. 12.3.2. The extension to the case of non-linear hyperbolic equations will instead be considered in Sec. 14.2.
Finally, we point out the following schemes for approximating the wave equation (12.9), again in the $f = 0$ case:

– **Leap-Frog**

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = (\gamma\lambda)^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \tag{12.23}$$

– **Newmark**

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = \frac{(\gamma\lambda)^2}{4}\left(w_j^{n-1} + 2w_j^n + w_j^{n+1}\right), \tag{12.24}$$

where $w_j^n = u_{j+1}^n - 2u_j^n + u_{j-1}^n$.

12.3.2 Discretization of linear hyperbolic systems

Let us consider the linear system (12.5). Generalizing (12.13), a numerical scheme for a finite difference approximation can be written in the form

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\mathbf{H}_{j+1/2}^n - \mathbf{H}_{j-1/2}^n),$$

where $\mathbf{u}_j^n$ is the vector approximating $\mathbf{u}(x_j, t^n)$. Now, $\mathbf{H}_{j+1/2}$ is a *vector numerical flux*. Its formal expression can easily be derived by generalizing with respect to the scalar case and replacing in (12.15), (12.17), (12.19), (12.21), a, a^2, and $|a|$ respectively with A, A^2, and $|\mathrm{A}|$, being

$$|\mathrm{A}| = \mathrm{T}|\Lambda|\mathrm{T}^{-1},$$

where $|\Lambda| = \mathrm{diag}(|\lambda_1|, ..., |\lambda_p|)$ and T is the matrix of eigenvectors of A.

For instance, transforming system (12.5) in p independent transport equations and approximating each of these with an upwind scheme for scalar equations, we obtain the following upwind numerical scheme for the initial system

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{\lambda}{2}\mathrm{A}(\mathbf{u}_{j+1}^n - \mathbf{u}_{j-1}^n) + \frac{\lambda}{2}|\mathrm{A}|(\mathbf{u}_{j+1}^n - 2\mathbf{u}_j^n + \mathbf{u}_{j-1}^n).$$

The numerical flux of such scheme is

$$\mathbf{H}_{j+\frac{1}{2}} = \frac{1}{2}[\mathrm{A}(\mathbf{u}_{j+1} + \mathbf{u}_j) - |\mathrm{A}|(\mathbf{u}_{j+1} - \mathbf{u}_j)].$$

The Lax-Wendroff method becomes

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{1}{2}\lambda \mathrm{A}(\mathbf{u}_{j+1}^n - \mathbf{u}_{j-1}^n) + \frac{1}{2}\lambda^2 \mathrm{A}^2(\mathbf{u}_{j+1}^n - 2\mathbf{u}_j^n + \mathbf{u}_{j-1}^n),$$

and its numerical flux is

$$\mathbf{H}_{j+\frac{1}{2}} = \frac{1}{2}[\mathrm{A}(\mathbf{u}_{j+1} - \mathbf{u}_j) - \lambda \mathrm{A}^2(\mathbf{u}_{j+1} - \mathbf{u}_j)].$$

12.3.3 Boundary treatment

In case we want to discretize the hyperbolic equation (12.3) on a bounded interval, we will obviously need to use the inflow node $x = \alpha$ to impose the boundary condition, say $u_0^{n+1} = \varphi(t^{n+1})$, while in all the other nodes x_j, $1 \leq j \leq m$ (including the outflow node $x_m = \beta$) we will write the finite difference scheme.

However, we must observe that the schemes using a centered discretization of the space derivative require a particular treatment at x_m. Indeed, they would require using the value u_{m+1}, the latter being unavailable as it relates to the point with coordinates $\beta + h$ which lies outside the integration interval. The problem can be solved in various ways. An option is to only use the upwind decentered discretization on the last node, as such discretization does not require knowing the datum in x_{m+1}; this approach however is only a first-order one. Alternatively, the value u_m^{n+1} can be obtained through an extrapolation from the values available at the internal nodes. An example could be an extrapolation along the characteristic lines applied to a scheme for which $\lambda a \leq 1$; this provides $u_m^{n+1} = u_{m-1}^n \lambda a + u_m^n(1 - \lambda a)$.

A further option consists in applying the centered finite difference scheme to the outflow node x_m as well and use, in place of u_{m+1}^n, an approximation based on a constant extrapolation ($u_{m+1}^n = u_m^n$), or on a linear one ($u_{m+1}^n = 2u_m^n - u_{m-1}^n$).

This matter becomes more problematic in the case of hyperbolic systems, where we must resort to compatibility equations. To gain a more in-depth view of these aspects and to analyze their possible instabilities deriving from the numerical boundary treatment, the reader can refer to Strickwerda [Str89], [QV94, Chap. 14] and [LeV07].

12.4 Analysis of the finite difference methods

We analyze the consistency, stability and convergence properties of the finite difference methods we introduced previously.

12.4.1 Consistency and convergence

For a given numerical scheme, the local truncation error is the error generated by expecting the exact solution to verify the numerical scheme itself.

For instance, in the case of scheme (12.14), having denoted by u the solution of the exact problem (12.1), we can define the truncation error at point (x_j, t^n) as follows

$$\tau_j^n = \frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\Delta t} + a\frac{u(x_{j+1}, t^n) - u(x_{j-1}, t^n)}{2h}.$$

If the *truncation error*

$$\tau(\Delta t, h) = \max_{j,n} |\tau_j^n|$$

tends to zero when Δt and h tend to zero, independently, then the numerical scheme will be said to be *consistent*.

Moreover, we will say that a numerical scheme is *accurate to the order p in time* and *to the order q in space* (for suitable integers p and q), if for a sufficiently regular solution of the exact problem, we have

$$\tau(\Delta t, h) = \mathcal{O}(\Delta t^p + h^q).$$

Using the Taylor developments suitably, we can then see that the truncation error of the previously introduced methods behaves as follows:

- **Euler (forward or backward) / centered**: $\mathcal{O}(\Delta t + h^2)$;
- **Upwind**: $\mathcal{O}(\Delta t + h)$;
- **Lax-Friedrichs** : $\mathcal{O}(\frac{h^2}{\Delta t} + \Delta t + h^2)$;
- **Lax-Wendroff** : $\mathcal{O}(\Delta t^2 + h^2 + h^2 \Delta t)$.

Finally, we will say that a scheme is *convergent* (in the maximum norm) if

$$\lim_{\Delta t, h \to 0} (\max_{j,n} |u(x_j, t^n) - u_j^n|) = 0.$$

Obviously, we can also consider weaker norms, such as $\|\cdot\|_{\Delta,1}$ and $\|\cdot\|_{\Delta,2}$ which we will introduce in (12.26).

12.4.2 Stability

We will say that a numerical method for a linear hyperbolic problem is *stable* if for each time T, there exists a constant $C_T > 0$ (possibly dependent on T) such that for each $h > 0$, there exists $\delta_0 > 0$ (possibly dependent on h) s.t. for each $0 < \Delta t < \delta_0$ we have

$$\|\mathbf{u}^n\|_\Delta \leq C_T \|\mathbf{u}^0\|_\Delta, \tag{12.25}$$

for each n such that $n\Delta t \leq T$, and for each initial datum $\mathbf{u}_0$. Note that C_T should not depend on Δt and h. Often (always, in the case of explicit methods) we will have stability only if the temporal step is sufficiently small with respect to the spatial one, that is for $\delta_0 = \delta_0(h)$.

The notation $\|\cdot\|_\Delta$ denotes a suitable discrete norm, for instance

$$\|\mathbf{v}\|_{\Delta,p} = \left(h \sum_{j=-\infty}^{\infty} |v_j|^p \right)^{\frac{1}{p}} \text{ for } p = 1, 2, \quad \|\mathbf{v}\|_{\Delta,\infty} = \sup_j |v_j|. \tag{12.26}$$

Note how $\|\mathbf{v}\|_{\Delta,p}$ represents an approximation of the $\mathrm{L}^p(\mathbb{R})$ norm, for $p = 1, 2$ or $+\infty$.

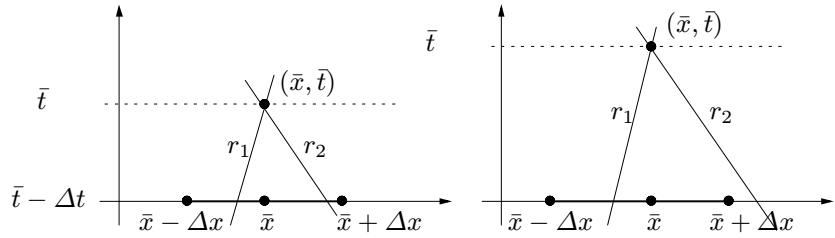


Fig. 12.3. Geometric interpretation of the CFL condition for a system with $p = 2$, where $r_i = \bar{x} - \lambda_i(t - \bar{t})$ $i = 1, 2$. The CFL condition is satisfied in the left case, and violated in the right case

The implicit backward/centered Euler scheme (12.22) is stable in the $\|\cdot\|_{\Delta,2}$ norm for any choice of the Δt and h parameters (see Exercise 2).

A scheme is called *strongly stable* with respect to the $\|\cdot\|_\Delta$ norm if

$$\|\mathbf{u}^n\|_\Delta \leq \|\mathbf{u}^{n-1}\|_\Delta, \tag{12.27}$$

for each n such that $n\Delta t \leq T$, and for each initial datum $\mathbf{u}_0$, which implies that (12.25) is verified with $C_T = 1$.

Remark 12.1 In the context of hyperbolic problems, solutions for long time intervals (i.e. for $T \gg 1$) are often sought . Such cases usually require a strongly stable scheme, as this guarantees that the numerical solution is limited for each value of T. •

As we will see, a necessary condition for the stability of an explicit numerical scheme of the form (12.13) is that the temporal and spatial discretization steps be linked by the following relation

$$|a\lambda| \leq 1, \text{ or } \Delta t \leq \frac{h}{|a|} \tag{12.28}$$

called *CFL condition* (from Courant, Friedrichs and Lewy). The number $a\lambda$ is commonly called *CFL number*; this is an a-dimensional quantity (a being a velocity).

The geometrical interpretation of the CFL stability condition is the following. In a finite difference scheme, the value of u_j^{n+1} generally depends on the values u_{j+i}^n of u^n at the three points x_{j+i}, $i = -1, 0, 1$. Proceeding backwards, we deduce that the solution u_j^{n+1} will only depend on the initial data at the points x_{j+i}, for $i = -(n+1), ..., (n+1)$ (see Fig. 12.3).

Denoting by *numerical domain of dependence* $D_{\Delta t}(x_j, t^n)$ the domain of dependency of u_j^n, which will therefore be called numerical dependency domain of u_j^n, the former will verify

$$D_{\Delta t}(x_j, t^n) \subset \{x \in \mathbb{R} : \ |x - x_j| \leq nh = \frac{t^n}{\lambda}\}.$$

Consequently, for each given point $(\overline{x}, \overline{t})$ we have

$$D_{\Delta t}(\overline{x}, \overline{t}) \subset \{x \in \mathbb{R} : \ |x - \overline{x}| \leq \frac{\overline{t}}{\lambda}\}.$$

In particular, taking the limit for $\Delta t \to 0$, and fixing λ, the numerical dependency domain becomes

$$D_0(\overline{x}, \overline{t}) = \{x \in \mathbb{R} : \ |x - \overline{x}| \leq \frac{\overline{t}}{\lambda}\}.$$

The condition (12.28) is then equivalent to the inclusion

$$D(\overline{x}, \overline{t}) \subset D_0(\overline{x}, \overline{t}), \tag{12.29}$$

where $D(\overline{x}, \overline{t})$ is the dependency domain of the exact solution defined in (12.8). Note that in the scalar case, $p = 1$ and $\lambda_1 = a$.

Remark 12.2 The CFL condition establishes, in particular, that there exist no explicit finite different schemes, unconditionally stable and consistent for hyperbolic initial value problems. Indeed, if the CFL condition were violated, there would exist at least a point x^* in the dependency domain that does not belong to the numerical dependency domain. Then, changing the initial datum to x^* will only modify the exact solution and not the numerical one. This implies a non-convergence of the method and therefore also its instability. Indeed, for a consistent method, the Lax-Richtmyer equivalence theorem states that stability is a necessary and sufficient condition for its convergence. •

Remark 12.3 In the case where $a = a(x, t)$ is no longer constant in (12.1), the CFL condition becomes

$$\Delta t \leq \frac{h}{\sup\limits_{x \in \mathbb{R},\ t>0} |a(x,t)|},$$

and even if the spatial discretization step varies, we have

$$\Delta t \leq \min_k \frac{h_k}{\sup\limits_{x \in (x_k, x_{k+1}),\ t>0} |a(x,t)|},$$

as $h_k = x_{k+1} - x_k$. •

Referring to the hyperbolic system (12.5), the stability condition CFL, analogous in all to (12.28), will be

$$\left|\lambda_k \frac{\Delta t}{h}\right| \leq 1, \quad k = 1, \ldots, p, \qquad \text{or, equivalently,} \qquad \Delta t \leq \frac{h}{\max_k |\lambda_k|},$$

where $\{\lambda_k,\ k = 1 \ldots, p\}$ are the eigenvalues of A.

This condition as well can be written in the form (12.29). The latter expresses the requirement that each line of the form $x = \overline{x} - \lambda_k(\overline{t} - t)$, $k = 1, \ldots, p$, must intersect the horizontal line $t = \overline{t} - \Delta t$ at points $x^{(k)}$ which lie within the numerical dependency domain.

Theorem 12.1 *If the CFL condition (12.28) is satisfied, the upwind, Lax-Friedrichs and Lax-Wendroff schemes are strongly stable in the norm* $\|\cdot\|_{\Delta,1}$.

Proof. To prove the stability of the upwind scheme (12.20) we rewrite it in the following form (having supposed $a > 0$)

$$u_j^{n+1} = u_j^n - \lambda a(u_j^n - u_{j-1}^n).$$

Then

$$\|\mathbf{u}^{n+1}\|_{\Delta,1} \leq h\sum_j |(1 - \lambda a)u_j^n| + h\sum_j |\lambda a u_{j-1}^n|.$$

Under the hypothesis (12.28) both values λa and $1 - \lambda a$ are non-negative. Hence,

$$\|\mathbf{u}^{n+1}\|_{\Delta,1} \leq h(1 - \lambda a)\sum_j |u_j^n| + h\lambda a\sum_j |u_{j-1}^n| = \|\mathbf{u}^n\|_{\Delta,1},$$

that is, inequality (12.25) holds with $C_T = 1$. The scheme is therefore strongly stable with respect to the norm $\|\cdot\|_\Delta = \|\cdot\|_{\Delta,1}$.

For the Lax-Friedrichs scheme, always under the CFL condition (12.28), we derive from (12.16) that

$$u_j^{n+1} = \frac{1}{2}(1 - \lambda a)u_{j+1}^n + \frac{1}{2}(1 + \lambda a)u_{j-1}^n,$$

so

$$\begin{aligned}\|\mathbf{u}^{n+1}\|_{\Delta,1} &\leq \frac{1}{2}h\left[\sum_j \left|(1 - \lambda a)u_{j+1}^n\right| + \sum_j \left|(1 + \lambda a)u_{j-1}^n\right|\right] \\ &\leq \frac{1}{2}(1 - \lambda a)\|\mathbf{u}^n\|_{\Delta,1} + \frac{1}{2}(1 + \lambda a)\|\mathbf{u}^n\|_{\Delta,1} = \|\mathbf{u}^n\|_{\Delta,1}.\end{aligned}$$

For the Lax-Wendroff scheme, the proof is analogous (see e.g. [QV94, Chap. 14] or [Str89]). ◇

Finally, we can prove that, if the CFL condition is verified, the upwind scheme satisfies

$$\|\mathbf{u}^n\|_{\Delta,\infty} \leq \|\mathbf{u}^0\|_{\Delta,\infty} \quad \forall n \geq 0, \tag{12.30}$$

i.e. it is strongly stable in the $\|\cdot\|_{\Delta,\infty}$ norm. The relation (12.30) is called *discrete maximum principle* (see Exercise 4).

Theorem 12.2 *The backward Euler scheme BE/C is strongly stable in the norm* $||\cdot||_{\Delta,2}$*, with no restriction on* Δt*. The forward Euler scheme FE/C, instead, is never strongly stable. However, it is stable with constant* $C_T = e^{T/2}$ *provided that we assume that* Δt *satisfies the following condition (more restrictive than the CFL condition)*

$$\Delta t \leq \left(\frac{h}{a}\right)^2. \tag{12.31}$$

Proof. We observe that

$$(B-A)B = \frac{1}{2}(B^2 - A^2 + (B-A)^2) \quad \forall A, B \in \mathbb{R}. \tag{12.32}$$

As a matter of fact

$$(B-A)B = (B-A)^2 + (B-A)A = \frac{1}{2}((B-A)^2 + (B-A)(B+A)).$$

Multiplying (12.22) by u_j^{n+1} we find

$$(u_j^{n+1})^2 + (u_j^{n+1} - u_j^n)^2 = (u_j^n)^2 - \lambda a(u_{j+1}^{n+1} - u_{j-1}^{n+1})u_j^{n+1}.$$

Observing that

$$\sum_{j\in\mathbb{Z}}(u_{j+1}^{n+1} - u_{j-1}^{n+1})u_j^{n+1} = 0 \tag{12.33}$$

(being a telescopic sum), we immediately obtain that $||\mathbf{u}^{n+1}||_{\Delta,2}^2 \leq ||\mathbf{u}^n||_{\Delta,2}^2$, which is the result sought for the BE/C scheme.

Let us now move to the FE/C scheme and multiply (12.14) by u_j^n. Observing that

$$(B-A)A = \frac{1}{2}(B^2 - A^2 - (B-A)^2) \quad \forall A, B \in \mathbb{R}, \tag{12.34}$$

we find

$$(u_j^{n+1})^2 = (u_j^n)^2 + (u_j^{n+1} - u_j^n)^2 - \lambda a(u_{j+1}^n - u_{j-1}^n)u_j^n.$$

On the other hand, we obtain once again from (12.14) that

$$u_j^{n+1} - u_j^n = -\frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n),$$

and therefore

$$(u_j^{n+1})^2 = (u_j^n)^2 + \left(\frac{\lambda a}{2}\right)^2 (u_{j+1}^n - u_{j-1}^n)^2 - \lambda a(u_{j+1}^n - u_{j-1}^n)u_j^n.$$

Now summing on j and observing that the last addendum yields a telescopic sum (hence it does not provide any contribution) we obtain, after multiplying by h,

$$\|\mathbf{u}^{n+1}\|^2_{\Delta,2} = \|\mathbf{u}^n\|^2_{\Delta,2} + \left(\frac{\lambda a}{2}\right)^2 h \sum_{j\in\mathbb{Z}} (u^n_{j+1} - u^n_{j-1})^2,$$

from which we infer that there is no value of Δt for which the method is strongly stable. However, as

$$(u^n_{j+1} - u^n_{j-1})^2 \leq 2\left[(u^n_{j+1})^2 + (u^n_{j-1})^2\right],$$

we find that, under the hypothesis (12.31),

$$\|\mathbf{u}^{n+1}\|^2_{\Delta,2} \leq (1 + \lambda^2 a^2)\|\mathbf{u}^n\|^2_{\Delta,2} \leq (1 + \Delta t)\|\mathbf{u}^n\|^2_{\Delta,2}.$$

By recursion, we find

$$\|\mathbf{u}^n\|^2_{\Delta,2} \leq (1 + \Delta t)^n \|\mathbf{u}^0\|^2_{\Delta,2} \leq e^T \|\mathbf{u}^0\|^2_{\Delta,2},$$

where we have used inequality

$$(1 + \Delta t)^n \leq e^{n\Delta t} \leq e^T \quad \forall n \text{ s.t. } t^n \leq T.$$

We conclude that

$$\|\mathbf{u}^n\|_{\Delta,2} \leq e^{T/2} \|\mathbf{u}^0\|_{\Delta,2},$$

which is the stability result sought for the FE/C scheme. ◇

12.4.3 Von Neumann analysis and amplification coefficients

The stability of a scheme in the norm $||\cdot||_{\Delta,2}$ can be also studied by the von Neumann analysis. To this purpose, we hypothesize that the function $u_0(x)$ is 2π-periodic and thus it can be written as a Fourier series as follows

$$u_0(x) = \sum_{k=-\infty}^{\infty} \alpha_k e^{ikx}, \tag{12.35}$$

where

$$\alpha_k = \frac{1}{2\pi} \int_0^{2\pi} u_0(x)\, e^{-ikx}\, dx$$

is the k-th Fourier coefficient. Hence,

$$u^0_j = u_0(x_j) = \sum_{k=-\infty}^{\infty} \alpha_k e^{ikjh}, \quad j = 0, \pm 1, \pm 2, \cdots$$

Table 12.1. Amplification coefficient for the different numerical schemes in Sec. 12.3.1. We recall that $\lambda = \Delta t/h$

Scheme	γ_k
Forward/centered Euler	$1 - ia\lambda \sin(kh)$
Backward/centered Euler	$(1 + ia\lambda \sin(kh))^{-1}$
Upwind	$1 - \lvert a\rvert\lambda(1 - e^{-ikh})$
Lax-Friedrichs	$\cos kh - ia\lambda \sin(kh)$
Lax-Wendroff	$1 - ia\lambda \sin(kh) - a^2\lambda^2(1 - cos(kh))$

It can be verified that applying any of the difference schemes seen in Sec. 12.3.1 we get the following relation

$$u_j^n = \sum_{k=-\infty}^{\infty} \alpha_k e^{ikjh} \gamma_k^n, \quad j = 0, \pm 1, \pm 2, \ldots, \quad n \geq 1. \tag{12.36}$$

The number $\gamma_k \in \mathbb{C}$ is called *amplification coefficient* of the k-th frequency (or harmonic), and characterizes the scheme under exam. For instance, in the case of the forward centered Euler method (FE/C) we find

$$\begin{aligned} u_j^1 &= \sum_{k=-\infty}^{\infty} \alpha_k e^{ikjh} \left(1 - \frac{a\Delta t}{2h}(e^{ikh} - e^{-ikh})\right) \\ &= \sum_{k=-\infty}^{\infty} \alpha_k e^{ikjh} \left(1 - \frac{a\Delta t}{h} i \sin(kh)\right). \end{aligned}$$

Hence,

$$\gamma_k = 1 - \frac{a\Delta t}{h} i \sin(kh) \qquad \text{and thus} \qquad |\gamma_k| = \left\{1 + \left(\frac{a\Delta t}{h} \sin(kh)\right)^2\right\}^{\frac{1}{2}}.$$

As there exist values of k for which $|\gamma_k| > 1$, there is no value of Δt and h for which the scheme is strongly stable.

Proceeding in a similar way for the other schemes, we find the coefficients reported in Table 12.1.

We will now see how the von Neumann analysis can be applied to study the stability of a numerical scheme with respect to the $\|\cdot\|_{\Delta,2}$ norm and to ascertain its dissipation and dispersion characteristics.

To this purpose, we prove the following result:

Theorem 12.3 *If there exists a number $\beta \geq 0$, and a positive integer m such that, for suitable choices of Δt and h, we have $|\gamma_k| \leq (1 + \beta\Delta t)^{\frac{1}{m}}$ for each k, then the scheme is stable with respect to the norm $\|\cdot\|_{\Delta,2}$ with a stability constant $C_T = e^{\beta T/m}$. In particular, if we can take $\beta = 0$ (and therefore $|\gamma_k| \leq 1\ \forall k$) then the scheme is strongly stable with respect to the same norm.*

Proof. We will suppose that problem (12.1) is formulated on the interval $[0, 2\pi]$. In such interval, let us consider $N+1$ equidistant nodes,

$$x_j = jh, \quad j = 0, \ldots, N, \quad \text{with} \quad h = \frac{2\pi}{N},$$

(N being an even positive integer) where to satisfy the numerical scheme (12.13). Moreover, we will suppose for simplicity that the initial datum u_0 be periodic. As the numerical scheme only depends on the values of u_0 at the x_j nodes, we can replace u_0 by the Fourier polynomial of order $N/2$,

$$\tilde{u}_0(x) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \alpha_k e^{ikx} \tag{12.37}$$

which interpolates it at the nodes. Note that $\tilde{u}_0$ is a periodic function with period 2π. We will have, thanks to (12.36),

$$u_j^0 = u_0(x_j) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \alpha_k e^{ikjh}, \quad u_j^n = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \alpha_k \gamma_k^n e^{ikjh}.$$

We note that

$$\|\mathbf{u}^n\|_{\Delta,2}^2 = h \sum_{j=0}^{N-1} \sum_{k,m=-\frac{N}{2}}^{\frac{N}{2}-1} \alpha_k \overline{\alpha}_m (\gamma_k \overline{\gamma}_m)^n e^{i(k-m)jh}.$$

As

$$h \sum_{j=0}^{N-1} e^{i(k-m)jh} = 2\pi \delta_{km}, \quad -\frac{N}{2} \le k, m \le \frac{N}{2} - 1,$$

(see e.g. [QSS07, Lemma 10.2]) we find

$$\|\mathbf{u}^n\|_{\Delta,2}^2 = 2\pi \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\alpha_k|^2 |\gamma_k|^{2n}.$$

Thanks to the assumption made on $|\gamma_k|$ we have

$$\|\mathbf{u}^n\|_{\Delta,2}^2 \le (1+\beta\Delta t)^{\frac{2n}{m}} 2\pi \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\alpha_k|^2 = (1+\beta\Delta t)^{\frac{2n}{m}} \|\mathbf{u}^0\|_{\Delta,2}^2 \ \ \forall n \ge 0.$$

As $1 + \beta\Delta t \le e^{\beta\Delta t}$, we deduce that

$$\|\mathbf{u}^n\|_{\Delta,2} \le e^{\frac{\beta\Delta t n}{m}} \|\mathbf{u}^0\|_{\Delta,2} = e^{\frac{\beta T}{m}} \|\mathbf{u}^0\|_{\Delta,2} \quad \forall n \ s.t. \quad n\Delta t \le T.$$

This proves the theorem. ◇

Remark 12.4 Should strong stability be required, the condition $|\gamma_k| \leq 1$ indicated in Theorem 12.3 is also necessary. •

In the case of the upwind scheme (12.20), as

$$|\gamma_k|^2 = [1 - |a|\lambda(1 - \cos kh)]^2 + a^2\lambda^2 \sin^2 kh, \quad k \in \mathbb{Z},$$

we obtain

$$|\gamma_k| \leq 1 \text{ if } \Delta t \leq \frac{h}{|a|}, \quad k \in \mathbb{Z}, \tag{12.38}$$

that is we find that the CFL condition guarantees the strong stability in the $||\cdot||_{\Delta,2}$ norm.

Proceeding in a similar way, we can verify that (12.38) also holds for the Lax-Friedrichs scheme.

The centered backward Euler scheme BE/C instead is unconditionally strongly stable in the $\|\cdot\|_{\Delta,2}$ norm, as $|\gamma_k| \leq 1$ for each k and for each possible choice of Δt and h, as we previously obtained in Theorem 12.2 by following a different procedure.

In the case of the centered forward Euler method FE/C we have

$$|\gamma_k|^2 = 1 + \frac{a^2\Delta t^2}{h^2}\sin^2(kh) \leq 1 + \frac{a^2\Delta t^2}{h^2}, \quad k \in \mathbb{Z}.$$

If $\beta > 0$ is a constant such that

$$\Delta t \leq \beta\frac{h^2}{a^2} \tag{12.39}$$

then $|\gamma_k| \leq (1 + \beta\Delta t)^{1/2}$. Hence, applying Theorem 12.3 (with $m = 2$) we deduce that the FE/C scheme is stable, albeit with a more restrictive CFL condition, as previously obtained following a different path in Theorem 12.2.

We can find a strong stability condition for the centered forward Euler method applied to the transport-reaction equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + a_0 u = 0, \tag{12.40}$$

with $a_0 > 0$. In this case we have for each $k \in \mathbb{Z}$

$$|\gamma_k|^2 = 1 - 2a_0\Delta t + \Delta t^2 a_0^2 + \lambda^2 \sin^2(kh) \leq 1 - 2a_0\Delta t + \Delta t^2 a_0^2 + \left(\frac{a\Delta t}{h}\right)^2$$

and thus the scheme is strongly stable in the $||.||_{\Delta,2}$ norm under the condition

$$\Delta t < \frac{2a_0}{a_0^2 + h^{-2}a^2}. \tag{12.41}$$

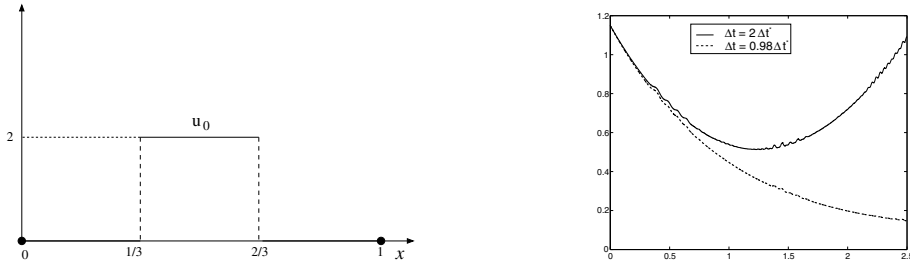


Fig. 12.4. The figure on the right displays the behavior of $\|\mathbf{u}^n\|_{\Delta,2}$, where $\mathbf{u}^n$ is the solution of equation (12.40) (with $a = a_0 = 1$) obtained using the FE/C method, for two values of Δt, one smaller and one greater than the critical value Δt^*. On the left, the initial datum used

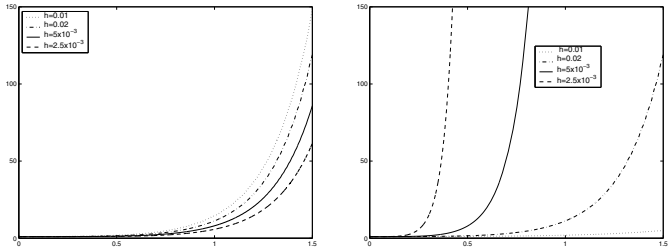


Fig. 12.5. Behavior of $\|\mathbf{u}^n\|_{\Delta,2}$ where $\mathbf{u}^n$ is the solution obtained using the FE/C mehod, in the $a_0 = 0$ case and for different values of h. On the left, the case where Δt satisfies the stability condition (12.39). On the right, the results obtained by maintaining the CFL number constant and equal to 0.1, violating the condition (12.39)

Example 12.1 In order to numerically verify the stability condition (12.41), we have considered equation (12.40) in the $(0, 1)$ interval with periodic boundary conditions. We have chosen $a = a_0 = 1$ and the initial datum u_0 equal to 2 in the $(1/3, 2/3)$ interval and 0 elsewhere. As the initial datum is a square wave, its Fourier development has all its α_k coefficients not null. On the right of Fig. 12.4, we report $\|\mathbf{u}^n\|_{\Delta,2}$ in the time interval $(0, 2.5)$ for two values of Δt, one larger and one smaller than the critical value $\Delta t^* = 2/(1 + h^{-2})$, provided by (12.41). Note that for $\Delta t < \Delta t^*$ the norm is decreasing, while, in the opposite case, after an initial decrease it grows exponentially. Fig. 12.5 shows the result for $a_0 = 0$ obtained with FE/C using the same initial datum. In the figure on the left, we display the behavior of $\|\mathbf{u}^n\|_{\Delta,2}$ for different values of h and using $\Delta t = 10h^2$, that is varying the time step based on the restriction provided by inequality (12.39) and taking $\beta = 10$. Note how the norm of the solution remains bounded for decreasing values of h. At the right-hand side of the same figure, we illustrate the result obtained for the same values of h taking as condition $\Delta t = 0.1h$, which corresponds to a constant CFL number equal to 0.1. In this case, the discrete norm of the numerical solution blows up as h decreases, as expected. ■

12.4.4 Dissipation and dispersion

Besides allowing to enquire about the stability of a numerical scheme, the analysis of amplification coefficients is also useful to study its dissipation and dispersion properties.
To understand what this is about, let us consider the exact solution of the problem (12.1); for such solution, we have the following relation

$$u(x, t^n) = u_0(x - an\Delta t), \quad n \geq 0, \quad x \in \mathbb{R},$$

as $t^n = n\Delta t$. In particular, using (12.35) we obtain

$$u(x_j, t^n) = \sum_{k=-\infty}^{\infty} \alpha_k e^{ikjh}(g_k)^n \quad \text{with} \quad g_k = e^{-iak\Delta t}. \tag{12.42}$$

Comparing (12.42) with (12.36) we can note that the amplification coefficient γ_k (generated by the specific numerical scheme) is the correspondent of g_k.

We observe that $|g_k| = 1$ for each $k \in \mathbb{Z}$, while it must be $|\gamma_k| \leq 1$ in order to guarantee the strong stability of the scheme. Thus, γ_k is a *dissipative* coefficient. The smaller $|\gamma_k|$ is, the larger will be the reduction of the amplitude α_k and, consequently, the larger will be the dissipation of the numerical scheme.
The ratio $\epsilon_a(k) = \frac{|\gamma_k|}{|g_k|}$ is called *amplification error* (or *dissipation error*) of the k-th harmonic associated to the numerical scheme (and in our case it coincides with the amplification coefficient).
Having set

$$\phi_k = kh,$$

as $k\Delta t = \lambda\phi_k$ we obtain

$$g_k = e^{-ia\lambda\phi_k}. \tag{12.43}$$

The real number ϕ_k, here expressed in radiants, is called *phase angle* of the k-th harmonic. We rewrite γ_k in the following way

$$\gamma_k = |\gamma_k| e^{-i\omega\Delta t} = |\gamma_k| e^{-i\frac{\omega}{k}\lambda\phi_k},$$

and comparing such relation to (12.43), we can deduce that the ratio $\frac{\omega}{k}$ represents the *propagation rate* of the numerical scheme, relatively to the k-th harmonic. The ratio

$$\epsilon_d(k) = \frac{\omega}{ka} = \frac{\omega h}{\phi_k a}$$

between such rate and the propagation rate a of the exact solution is called *dispersion error* ϵ_d relative to the k-th harmonic.

The amplification (or dissipation) error and the dispersion error for the numerical schemes analyzed up to now vary as a function of the phase angle ϕ_k and of the CFL number $a\lambda$, as reported in Fig. 12.6. For symmetry reasons we have considered the

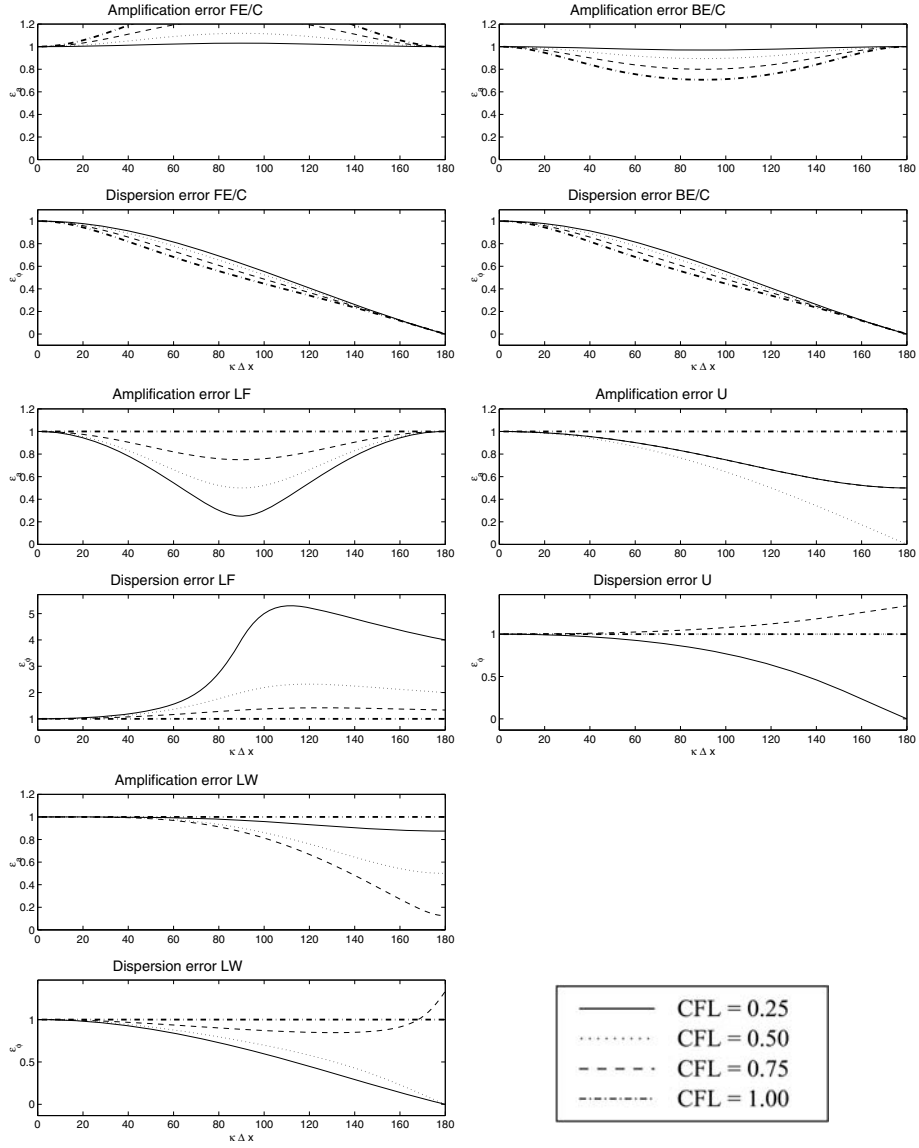


Fig. 12.6. Amplification and dispersion errors for different numerical schemes as a function of the phase angle $\phi_k = kh$ and for different values of the CFL number

interval $0 \leq \phi_k \leq \pi$ and we have used degrees instead of radiants in the abscissa to indicate ϕ_k. Note how the forward/centered Euler scheme denotes a curve of the amplification factor with values above one for all the CFL schemes we have considered, in accordance with the fact that such scheme is never strongly stable.

Example 12.2 In Fig. 12.7 we compare the obtained numerical results by solving equation (12.1) with $a = 1$ and initial datum u_0 composed by a packet of two sinusoidal waves of equal length l centered at the origin ($x = 0$). In the figures on the left $l = 20h$, while in the right ones we have $l = 8h$. As $k = \dfrac{2\pi}{l}$, we have $\phi_k = \dfrac{2\pi}{l}h$ and therefore the values of the phase angle of the wave packet are $\phi_k = \pi/20$ at the left and $\phi_k = \pi/8$ at the right. The numerical solution has been computed for the value 0.75 of the CFL number, using the different (stable) schemes illustrated previously. We can note how the dissipative effect is very strong at high frequencies ($\phi_k = \pi/4$) and in particular for the first-order upwind, backward/centered Euler and Lax-Friedrichs methods.

In order to appreciate the dispersion effects, the solution for $\phi_k = \pi/4$ after 8 time steps is reported in Fig. 12.8. We can note how the Lax-Wendroff method is the least dissipative. Moreover, by attentively observing the position of the numerical wave crests with respect to those of the numerical solution, we can verify that the Lax-Friedrichs method features a positive dispersion error. Indeed, the numerical wave results to anticipate the exact one. The upwind method is also weakly dispersive for a CFL number equal to 0.75, while the dispersion of the Lax-Friedrichs and backward Euler methods is evident (even after only 8 time steps!). ■

12.5 Equivalent equations

To each numerical scheme, we can associate a family of differential equations, called equivalent equations.

12.5.1 The upwind scheme case

Let us first focus on the upwind scheme. Suppose there exists a regular function $v(x,t)$ satisfying the difference equation (12.20) at each point $(x,t) \in \mathbb{R} \times \mathbb{R}^+$ (and not only at the grid nodes (x_j, t^n)!). We can then write (in the case where $a > 0$)

$$\frac{v(x,t+\Delta t) - v(x,t)}{\Delta t} + a\frac{v(x,t) - v(x-h,t)}{h} = 0. \tag{12.44}$$

Using the Taylor developments with respect to x and t relative to the point (x,t) and supposing that v is of class C^4 with respect to x and t, we can write

$$\frac{v(x,t+\Delta t) - v(x,t)}{\Delta t} = v_t + \frac{\Delta t}{2}v_{tt} + \frac{\Delta t^2}{6}v_{ttt} + \mathcal{O}(\Delta t^3),$$

$$a\frac{v(x,t) - v(x-h,t)}{h} = av_x - \frac{ah}{2}v_{xx} + \frac{ah^2}{6}v_{xxx} + \mathcal{O}(h^3),$$

where the right-hand side derivatives are all evaluated at point (x,t). Thanks to (12.44) we deduce that, at each point (x,t), the v function satisfies the relation

$$v_t + av_x = R^U + \mathcal{O}(\Delta t^3 + h^3), \tag{12.45}$$

with

$$R^U = \frac{1}{2}(ah\, v_{xx} - \Delta t\, v_{tt}) - \frac{1}{6}(ah^2 v_{xxx} + \Delta t^2 v_{ttt}).$$

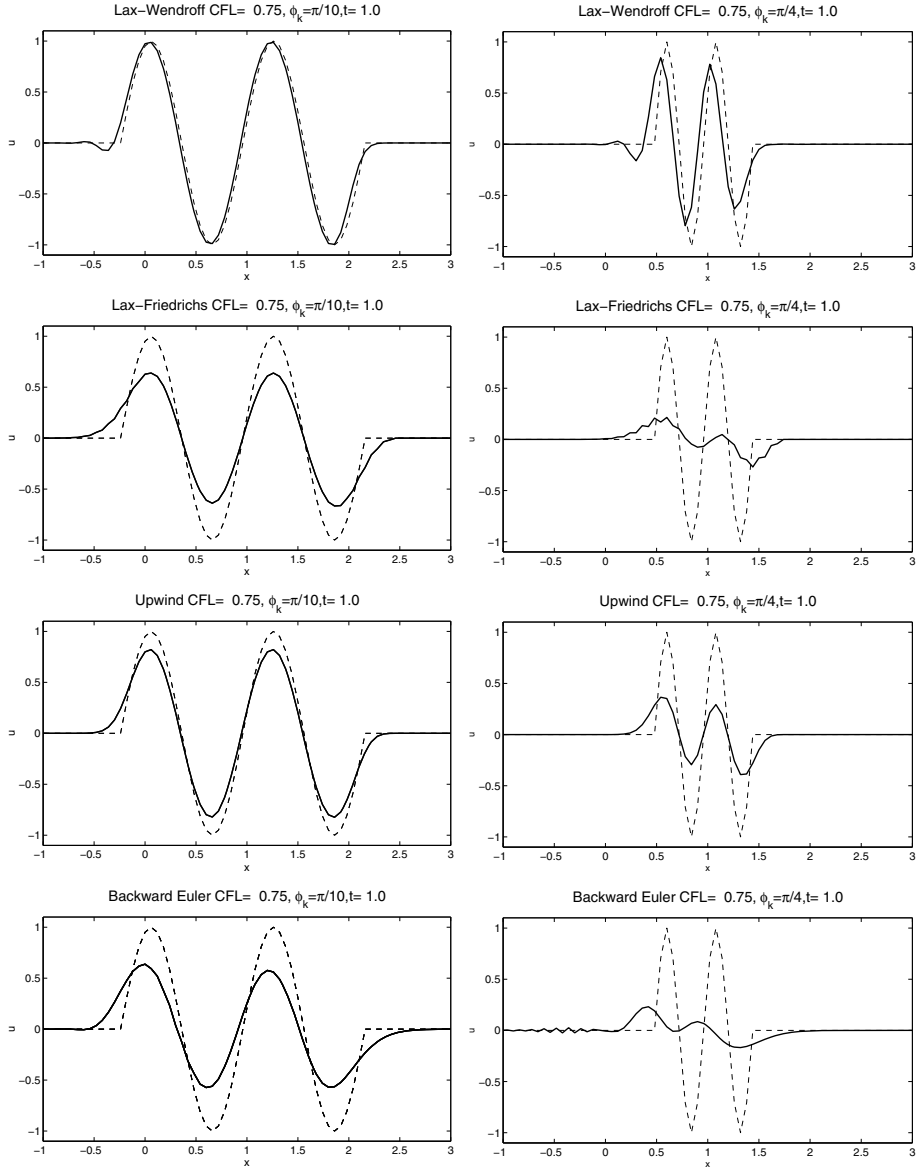


Fig. 12.7. Numerical solution of the convective transport equation of a sinusoidal wave packet with different wavelengths ($l = 20h$ at the left, $l = 8h$ at the right) obtained with different numerical schemes. The numerical solution for $t = 1$ is displayed in solid line, while the exact solution at the same time instant is displayed in etched line

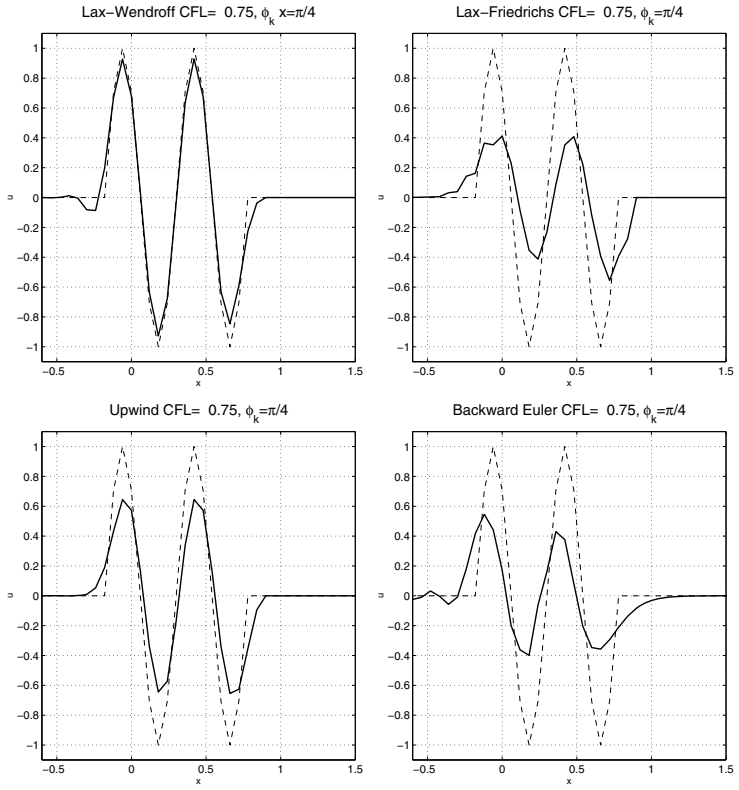


Fig. 12.8. Numerical solution of the convective transport of a packet of sinusoidal waves. The solid line represents the solution after 8 time steps. The etched line represents the corresponding exact solution at the same time level

Formally differentiating such equation with respect to t, we find

$$v_{tt} + av_{xt} = R_t^U + \mathcal{O}(\Delta t^3 + h^3).$$

Instead, differentiating it with respect to x, we have

$$v_{xt} + av_{xx} = R_x^U + \mathcal{O}(\Delta t^3 + h^3). \tag{12.46}$$

Hence,

$$v_{tt} = a^2 v_{xx} + R_t^U - aR_x^U + \mathcal{O}(\Delta t^3 + h^3) \tag{12.47}$$

which allows to obtain from (12.45)

$$v_t + av_x = \mu v_{xx} - \frac{1}{6}(ah^2 v_{xxx} + \Delta t^2 v_{ttt}) - \frac{\Delta t}{2}(R_t^U - aR_x^U) + \mathcal{O}(\Delta t^3 + h^3), \tag{12.48}$$

having set

$$\mu = \frac{1}{2}ah(1 - (a\lambda)) \tag{12.49}$$

and, as usual, $\lambda = \Delta t/h$. Now, formally differentiating (12.47) with respect to t, hence (12.46) with respect to x, we find

$$\begin{aligned} v_{ttt} &= a^2 v_{xxt} + R^U_{tt} - aR^U_{xt} + \mathcal{O}(\Delta t^3 + h^3) \\ &= -a^3 v_{xxx} + a^2 R^U_{xx} + R^U_{tt} - aR^U_{xt} + \mathcal{O}(\Delta t^3 + h^3). \end{aligned} \tag{12.50}$$

Moreover, we have that

$$\begin{aligned} R^U_t &= \frac{1}{2} ah\, v_{xxt} - \frac{\Delta t}{2} v_{ttt} - \frac{ah^2}{6} v_{xxxt} - \frac{\Delta t^2}{6} v_{tttt}, \\ R^U_x &= \frac{1}{2} ah\, v_{xxx} - \frac{\Delta t}{2} v_{ttx} - \frac{ah^2}{6} v_{xxxx} - \frac{\Delta t^2}{6} v_{tttx}. \end{aligned} \tag{12.51}$$

Using the relations (12.50) and (12.51) in (12.48) we obtain

$$\begin{aligned} v_t + av_x &= \mu v_{xx} - \frac{ah^2}{6}\Big(1 - \frac{a^2\Delta t^2}{h^2} - \frac{3a\Delta t}{2h}\Big) v_{xxx} \\ &+ \underbrace{\frac{\Delta t}{4}\Big(\Delta t\, v_{ttt} - ah\, v_{xxt} - a\Delta t\, v_{ttx}\Big)}_{(\mathrm{A})} \\ &+ \frac{\Delta t}{12}(\Delta t^2 v_{tttt} - a\Delta t^2 v_{tttx} + ah^2 v_{xxxt} - a^2 h^2 v_{xxxx}) \\ &- \frac{a^2\Delta t^2}{6} R^U_{xx} - \frac{\Delta t^2}{6} R^U_{tt} + a\frac{\Delta t^2}{6} R^U_{xt} + \mathcal{O}(\Delta t^3 + h^3). \end{aligned} \tag{12.52}$$

Let us now focus on the third derivatives of v contained in the term (A). Thanks to (12.50), (12.46) and (12.47), respectively, we find:

$$\begin{aligned} v_{ttt} &= -a^3 v_{xxx} + r_1, \\ v_{xxt} &= -a\, v_{xxx} + r_2, \\ v_{ttx} &= a^2 v_{xxx} + r_3, \end{aligned}$$

where r_1, r_2 and r_3 are terms containing derivatives of v of order no less than four, as well as terms of order $\mathcal{O}(\Delta t^3 + h^3)$. (Note that it follows from the definition of R^U that its derivatives of order two are expressed through derivatives of v of order no less than four.) Regrouping the coefficients that multiply v_{xxx}, we therefore deduce from (12.52) that

$$v_t + av_x = \mu v_{xx} + \nu v_{xxx} + R_4(v) + \mathcal{O}(\Delta t^3 + h^3), \tag{12.53}$$

having set

$$\nu = -\frac{ah^2}{6}(1 - 3a\lambda + 2(a\lambda)^2), \tag{12.54}$$

and having indicated with $R_4(v)$ the set of terms containing the derivatives of v of order no less than four.
We can conclude that the v function satisfies, respectively, the equations:

$$v_t + a v_x = 0 \tag{12.55}$$

if we neglect the terms containing derivatives of order above the first;

$$v_t + a v_x = \mu v_{xx} \tag{12.56}$$

if we neglect the terms containing derivatives of order above the second;

$$v_t + a v_x = \mu v_{xx} + \nu v_{xxx} \tag{12.57}$$

if we neglect the derivatives of order above the third. The coefficients μ and ν have been defined in (12.49) and (12.54). Equations (12.55), (12.56) and (12.57) are called *equivalent equations* (at the first, second resp. third order) relative to the upwind scheme.

12.5.2 The Lax-Friedrichs and Lax-Wendroff case

Proceeding in a similar way, we can derive the equivalent equations of any numerical scheme. For instance, in the case of the Lax-Friedrichs scheme, having denoted by v a hypothetic function that verifies the equation (12.16) at each point (x, t), having observed that

$$\frac{1}{2}\big(v(x+h,t) + v(x-h,t)\big) = v + \frac{h^2}{2} v_{xx} + \mathcal{O}(h^4),$$

$$\frac{1}{2}\big(v(x+h,t) - v(x-h,t)\big) = h\, v_x + \frac{h^3}{6} v_{xxx} + \mathcal{O}(h^4),$$

we obtain

$$v_t + a v_x = R^{LF} + \mathcal{O}\Big(\frac{h^4}{\Delta t} + \Delta t^3\Big), \tag{12.58}$$

having set

$$R^{LF} = \frac{h^2}{2\Delta t}(v_{xx} - \lambda^2 v_{tt}) - \frac{a\, h^2}{6}(v_{xxx} + \frac{\lambda^2}{a} v_{ttt}).$$

Proceeding as we did previously, tedious computation allows us to deduce from (12.58) the equivalent equations (12.55)-(12.57), in this case having however

$$\mu = \frac{h^2}{2\Delta t}(1 - (a\lambda)^2), \quad \nu = \frac{a h^2}{3}(1 - (a\lambda)^2).$$

In the case of the *Lax-Wendroff* scheme, the equivalent equations are characterized by the following parameters

$$\mu = 0, \qquad \nu = \frac{a h^2}{6}((a\lambda)^2 - 1).$$

12.5.3 On the meaning of coefficients in equivalent equations

In general, in the equivalent equations, the term μv_{xx} represents a dissipation, while νv_{xxx} represents a dispersion. We can provide a heuristic proof of this by examining the solution to the problem

$$\begin{cases} v_t + a v_x = \mu v_{xx} + \nu v_{xxx}, & x \in \mathbb{R},\ t > 0, \\ v(x,0) = e^{ikx}, & k \in \mathbb{Z}. \end{cases} \tag{12.59}$$

By applying the Fourier transform we find, if $\mu = \nu = 0$,

$$v(x,t) = e^{ik(x-at)},$$

while for μ and ν arbitrary real numbers (with $\mu > 0$) we have

$$v(x,t) = e^{-\mu k^2 t} e^{ik[x-(a+\nu k^2)t]}.$$

Comparing these two relations, we see that for growing μ, the modulus of the solution gets smaller. Such effect gets more remarkable as the frequency k increases (a phenomenon we have already registered in the previous section, albeit with partly different arguments).

The term μv_{xx} in (12.59) therefore has a dissipative effect on the solution. In turn, ν modifies the propagation rate of the solution, increasing it in the $\nu > 0$ case, and decreasing it if $\nu < 0$. Also in this case, the effect is more notable at high frequencies. Hence, the third derivative term νv_{xxx} introduces a dispersive effect.

In general, in the equivalent equation, even order spatial derivatives represent diffusive terms, while odd order derivatives represent dispersive terms. For first-order schemes (such as the upwind scheme) the dispersive effect is often barely visible, as it is disguised by the dissipative one. Taking Δt and h of the same order, from (12.56) and (12.57) we evince that $\nu \ll \mu$ for $h \to 0$, as $\nu = O(h^2)$ and $\mu = O(h)$. In particular, if the CFL number is $\frac{1}{2}$, the third-order equivalent equation of the upwind method features a null dispersion, in accordance with the numerical results seen in the previous section.

Conversely, the dispersive effect is evident for the Lax-Friedrichs scheme, as well as for the Lax-Wendroff scheme which, being of the second order, does not feature a dissipative term of type μv_{xx}. However, being stable, the latter cannot avoid being dissipative. Indeed, the fourth-order equivalent equation for the Lax-Wendroff scheme is

$$v_t + a v_x = \frac{ah^2}{6}[(a\lambda)^2 - 1] v_{xxx} - \frac{ah^3}{6} a\lambda [1 - (a\lambda)^2] v_{xxxx},$$

where the last term is dissipative if $|a\lambda| < 1$, as it can easily be verified by applying the Fourier transform. We then recover, also for the Lax-Wendroff scheme, the CFL condition.

12.5.4 Equivalent equations and error analysis

The technique applied to obtain the equivalent equation denotes a strong analogy with the so-called *backward analysis* that we encounter during the numerical solution of

linear systems, where the computed (not exact) solution is interpreted as the exact solution of a perturbed linear system (see [QSS07, Chap. 3]). As a matter of fact, the perturbed system plays a similar role to that of the equivalent equation.

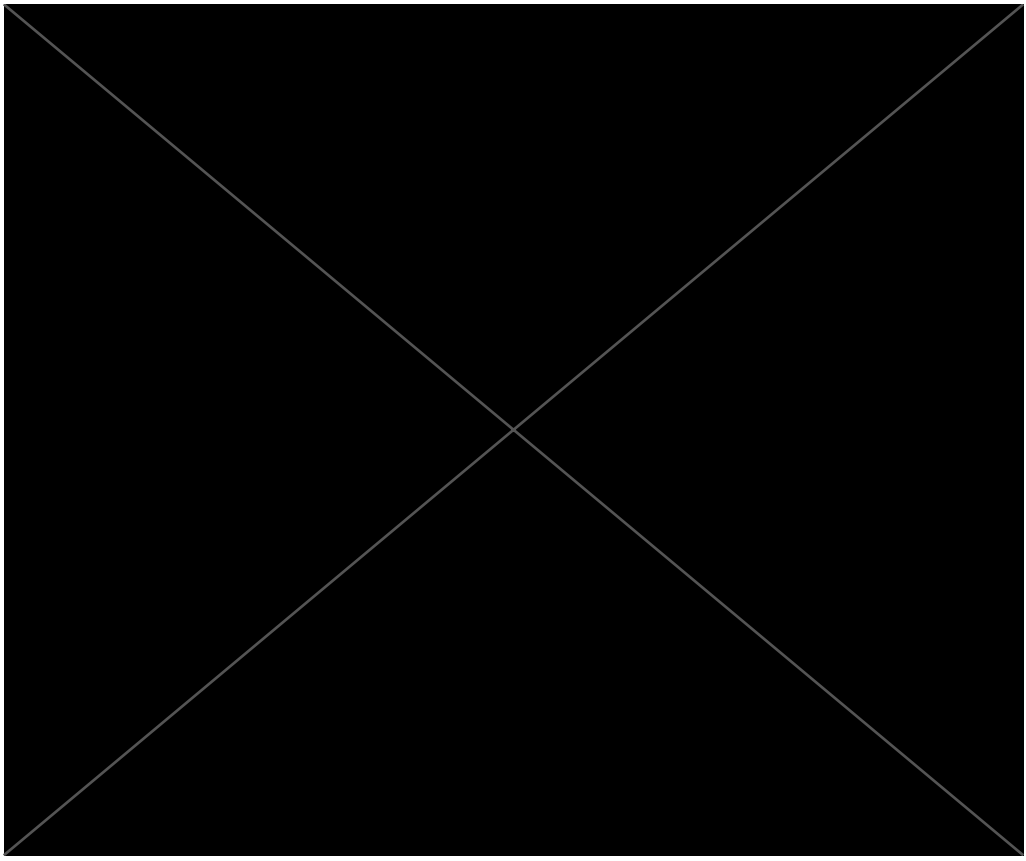
Moreover, we observe that an error analysis of a numerical scheme can be deduced from the knowledge of the equivalent equation associated to it. Indeed, by generically denoting by $r = \mu v_{xx} + \nu v_{xxx}$ the right-hand side of the equivalent equation, by comparison with (12.1) we obtain the error equation

$$e_t + ae_x = r,$$

where $e = v - u$. Multiplying such equation by e and integrating in space and time (between 0 and t) we obtain

$$\|e(t)\|_{\mathrm{L}^2(\mathbb{R})} \leq C(t) \left(\|e(0)\|_{\mathrm{L}^2(\mathbb{R})} + \sqrt{\int_0^t \|r(s)\|^2_{\mathrm{L}^2(\mathbb{R})} ds} \right), \quad t > 0$$

having used the a priori estimate (12.4). We can assume $e(0) = 0$ and therefore observe that $\|e(t)\|_{\mathrm{L}^2(\mathbb{R})}$ tends to zero (for h and Δt tending to zero) at order 1 for the upwind or Lax-Friedrichs schemes, and at order 2 for the Lax-Wendroff scheme (having supposed v to be sufficiently regular).



13

Finite elements and spectral methods for hyperbolic equations

In this chapter, we will illustrate how to apply Galerkin methods, and in particular the finite element method and the spectral one, to the spatial and/or temporal discretization of scalar hyperbolic equations. We will treat both the continuous as well as discontinuous finite element cases.

Let us consider the transport problem (12.3) and let us set for simplicity $(\alpha, \beta) = (0, 1)$, $\varphi = 0$. Moreover, let us suppose that a is a positive constant and a_0 a non-negative constant.

To start with, we proceed with a spatial discretization based on continuous finite elements. We therefore attempt a semidiscretization of the following form:

$\forall t > 0$, find $u_h = u_h(t) \in V_h$ s.t.

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a\left(\frac{\partial u_h}{\partial x}, v_h\right) + a_0\,(u_h, v_h) = (f, v_h)\ \forall v_h \in V_h, \tag{13.1}$$

u_h^0 being the approximation of the initial datum. We have set

$$V_h = \{v_h \in X_h^r :\ v_h(0) = 0\}, \quad r \geq 1.$$

The space X_h^r is defined as in (4.14), provided that we replace (a, b) with $(0, 1)$.

13.1 Temporal discretization

For the temporal discretization of problem (13.1) we will use finite difference schemes such as those introduced in Chap. 12.

As usual, we will denote by u_h^n, $n \geq 0$, the approximation of u_h at time $t^n = n\Delta t$.

13.1.1 The forward and backward Euler schemes

In case we use the forward Euler scheme, the discrete problem becomes:

A. Quarteroni: Numerical Models for Differential Problems.

$\forall n \geq 0$, find $u_h^{n+1} \in V_h$ such that

$$\frac{1}{\Delta t}\left(u_h^{n+1} - u_h^n, v_h\right) + a\left(\frac{\partial u_h^n}{\partial x}, v_h\right) + a_0\left(u_h^n, v_h\right) = \left(f^n, v_h\right) \quad \forall v_h \in V_h, \quad (13.2)$$

where $(u, v) = \int_0^1 u(x)v(x)dx$ denotes as usual the scalar product of $\mathrm{L}^2(0, 1)$.

In the case of the backward Euler method, instead of (13.2) we will have

$$\frac{1}{\Delta t}\left(u_h^{n+1} - u_h^n, v_h\right) + a\left(\frac{\partial u_h^{n+1}}{\partial x}, v_h\right) + a_0\left(u_h^{n+1}, v_h\right) = \left(f^{n+1}, v_h\right) \forall v_h \in V_h. \quad (13.3)$$

Theorem 13.1 *The backward Euler scheme is strongly stable with no restriction on Δt. Instead, the forward Euler method, is strongly stable only for $a_0 > 0$, provided we suppose that*

$$\Delta t \leq \frac{2a_0}{(aCh^{-1} + a_0)^2} \quad (13.4)$$

for a given constant $C = C(r)$.

Proof. Choosing $v_h = u_h^n$ in (13.2), we obtain (in the $f = 0$ case)

$$\left(u_h^{n+1} - u_h^n, u_h^n\right) + \Delta t a\left(\frac{\partial u_h^n}{\partial x}, u_h^n\right) + \Delta t a_0 \|u_h^n\|^2_{\mathrm{L}^2(0,1)} = 0.$$

For the first term, we use the identity

$$(v - w, w) = \frac{1}{2}\left(\|v\|^2_{\mathrm{L}^2(0,1)} - \|w\|^2_{\mathrm{L}^2(0,1)} - \|v - w\|^2_{\mathrm{L}^2(0,1)}\right) \forall v, w \in \mathrm{L}^2(0, 1) \quad (13.5)$$

which generalizes (12.34). For the second addendum, integrating by parts and using the boundary conditions, we find

$$\left(\frac{\partial u_h^n}{\partial x}, u_h^n\right) = \frac{1}{2}(u_h^n(1))^2.$$

Thus, we obtain

$$\begin{aligned} \|u_h^{n+1}\|^2_{\mathrm{L}^2(0,1)} + a\Delta t(u_h^n(1))^2 + 2a_0\Delta t\|u_h^n\|^2_{\mathrm{L}^2(0,1)} \\ = \|u_h^n\|^2_{\mathrm{L}^2(0,1)} + \|u_h^{n+1} - u_h^n\|^2_{\mathrm{L}^2(0,1)}. \end{aligned} \quad (13.6)$$

We now seek an estimate for the term $\|u_h^{n+1} - u_h^n\|^2_{\mathrm{L}^2(0,1)}$. To this end, setting in (13.2) $v_h = u_h^{n+1} - u_h^n$, we obtain

$$\begin{aligned} \|u_h^{n+1} - u_h^n\|^2_{\mathrm{L}^2(0,1)} &\leq \Delta t a\left|\left(\tfrac{\partial u_h^n}{\partial x}, u_h^{n+1} - u_h^n\right)\right| + \Delta t a_0\left|\left(u_h^n, u_h^{n+1} - u_h^n\right)\right| \\ &\leq \Delta t\left[a\|\tfrac{\partial u_h^n}{\partial x}\|_{\mathrm{L}^2(0,1)} + a_0\|u_h^n\|_{\mathrm{L}^2(0,1)}\right]\|u_h^{n+1} - u_h^n\|_{\mathrm{L}^2(0,1)}. \end{aligned}$$

By now using the inverse inequality (11.39) (referred to the interval $(0,1)$), we obtain

$$\|u_h^{n+1} - u_h^n\|_{\mathrm{L}^2(0,1)} \leq \Delta t \left(aC_I h^{-1} + a_0\right) \|u_h^n\|_{\mathrm{L}^2(0,1)}.$$

Finally, (13.6) becomes

$$\begin{aligned} &\|u_h^{n+1}\|^2_{\mathrm{L}^2(0,1)} + a\Delta t (u_h^n(1))^2 \\ &\quad + \Delta t \left[2a_0 - \Delta t (aC_I h^{-1} + a_0)^2\right] \|u_h^n\|^2_{\mathrm{L}^2(0,1)} \leq \|u_h^n\|^2_{\mathrm{L}^2(0,1)}. \end{aligned} \tag{13.7}$$

If (13.4) is satisfied, then $\|u_h^{n+1}\|_{\mathrm{L}^2(0,1)} \leq \|u_h^n\|_{\mathrm{L}^2(0,1)}$ and we therefore have strong stability in $\mathrm{L}^2(0,1)$ norm.

In the case where $a_0 = 0$ the obtained stability condition is meaningless. However, if we suppose that

$$\Delta t \leq \frac{Kh^2}{a^2 C_I^2},$$

for a given constant $K > 0$, then we can apply the discrete Gronwall lemma (see Lemma 2.3) to (13.7) and we find that the method is stable with a stability constant which in this case depends on the final time T. Precisely,

$$\|u_h^n\|_{\mathrm{L}^2(0,1)} \leq \exp(Kt^n) \|u_h^0\|_{\mathrm{L}^2(0,1)} \leq \exp(KT) \|u_h^0\|_{\mathrm{L}^2(0,1)}.$$

In the case of the backward Euler method (13.3), we choose instead $v_h = u_h^{n+1}$. By then using the relation

$$(v - w, v) = \frac{1}{2}\Big(\|v\|^2_{\mathrm{L}^2(0,1)} - \|w\|^2_{\mathrm{L}^2(0,1)} + \|v - w\|^2_{\mathrm{L}^2(0,1)}\Big) \; \forall v, w \in \mathrm{L}^2(0,1) \tag{13.8}$$

which generalizes (12.32), we find,

$$(1 + 2a_0\Delta t)\|u_h^{n+1}\|^2_{\mathrm{L}^2(0,1)} + a\Delta t (u_h^{n+1}(1))^2 \leq \|u_h^n\|^2_{\mathrm{L}^2(0,1)}. \tag{13.9}$$

Hence, we have strong stability in $\mathrm{L}^2(0,1)$, unconditioned (that is for each Δt) and for each $a_0 \geq 0$. ◇

13.1.2 The upwind, Lax-Friedrichs and Lax-Wendroff schemes

The generalization to the finite elements case of the Lax-Friedrichs (LF), Lax-Wendroff (LW) and upwind (U) finite difference schemes can be made in different ways.

We start by observing that (12.16), (12.18), and (12.20) can be rewritten in the following common form

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2h} - \mu\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + a_0 u_j^n = 0. \tag{13.10}$$

(Note however that $a_0 = 0$ in (12.16), (12.18) and (12.20).) The second term is the discretization via centered finite differences of the convective term $au_x(t^n)$, while the

third one is a numerical diffusion term and corresponds to the discretization via finite differences of $-\mu u_{xx}(t^n)$. The numerical viscosity coefficient μ is given by

$$\mu = \begin{cases} h^2/2\Delta t & \text{(LF)}, \\ a^2\Delta t/2 & \text{(LW)}, \\ ah/2 & \text{(U)}. \end{cases} \tag{13.11}$$

Equation (13.10) suggests the following finite element version for the approximation of problem (12.3): $\forall n \geq 0$, find $u_h^{n+1} \in V_h$ such that

$$\begin{aligned} &\frac{1}{\Delta t}\left(u_h^{n+1} - u_h^n, v_h\right) + a\left(\frac{\partial u_h^n}{\partial x}, v_h\right) + a_0\left(u_h^n, v_h\right) \\ &+\mu\left(\frac{\partial u_h^n}{\partial x}, \frac{\partial v_h}{\partial x}\right) - \mu\gamma\frac{\partial u_h^n}{\partial x}(1)v_h(1) = (f^n, v_h) \quad \forall v_h \in V_h, \end{aligned} \tag{13.12}$$

where $\gamma = 1, 0$ depending on whether or not we want to take the boundary contribution into account in the integration by parts of the numerical viscosity term.

For the stability analysis, in the case $\gamma = 0$, $a_0 = 0$, $a > 0$, let us set $v_h = u_h^{n+1} - u_h^n$, in order to obtain, thanks to inequality (4.52)

$$\|u_h^{n+1} - u_h^n\|_{\mathrm{L}^2(0,1)} \leq \Delta t(a + \mu C_I h^{-1})\|\frac{\partial u_h^n}{\partial x}\|_{\mathrm{L}^2(0,1)}.$$

Having now set $v_h = u_h^n$, thanks to (13.5) we obtain

$$\begin{aligned} \|u_h^{n+1}\|^2_{\mathrm{L}^2(0,1)} &- \|u_h^n\|^2_{\mathrm{L}^2(0,1)} + a\Delta t(u_h^n(1))^2 + 2\Delta t\mu\|\frac{\partial u_h^n}{\partial x}\|^2_{\mathrm{L}^2(0,1)} \\ &= \|u_h^{n+1} - u_h^n\|^2_{\mathrm{L}^2(0,1)} \leq \Delta t^2(a + \mu C_I h^{-1})^2\|\frac{\partial u_h^n}{\partial x}\|^2_{\mathrm{L}^2(0,1)}. \end{aligned}$$

A sufficient condition for strong stability (i.e. to obtain an estimate such as (12.27), with respect to the $\|\cdot\|_{L^2(0,1)}$ norm) is therefore

$$\Delta t \leq \frac{2\mu}{(a + \mu C_I h^{-1})^2}.$$

Thanks to (13.11), in the case of the upwind method this is equivalent to

$$\Delta t \leq \frac{h}{a}\left(\frac{1}{1 + C_I/2}\right)^2.$$

In the case of linear finite elements, $C_I \simeq 2\sqrt{3}$, therefore we deduce that

$$\frac{a\Delta t}{h} \lesssim \left(\frac{1}{1 + \sqrt{3}}\right)^2.$$

The stability analysis we have just developed is based on the *energy method* and, in this case, leads to sub-optimal results. A better indicator can be obtained by resorting to the von Neumann analysis, as done in Sec. 12.4.3. To this end we observe that, in the case of linear finite elements with constant spacing h, (13.12) with $f = 0$ can be rewritten in the following way for each internal node x_j:

$$\begin{aligned} &\frac{1}{6}(u_{j+1}^{n+1} + 4u_j^{n+1} + u_{j-1}^{n+1}) + \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{a_0}{6}(u_{j+1}^n + 4u_j^n + u_{j-1}^n) \\ &-\mu \Delta t \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = \frac{1}{6}(u_{j+1}^n + 4u_j^n + u_{j-1}^n). \end{aligned} \tag{13.13}$$

By comparing such relation to (13.10), we can note that the difference only resides in the term arising from the temporal derivative and from the term of order zero, and has to be attributed to the presence of the mass matrix in the case of finite elements. On the other hand, we have previously seen in Sec. 11.5 that we can apply the mass-lumping technique to approximate the mass matrix using a diagonal matrix. By proceeding this way, the scheme (13.13) can effectively be reduced to (13.10) (see Exercise 1).

Remark 13.1 Note that the provided relations (13.13) refer to the internal nodes. The approach used to handle boundary conditions with the finite element method generally yields different relations than those obtained via the finite difference method. •

These observations allow us to extend all the schemes seen in Sec. 12.3.1 to analogous schemes, generated by discretizations in space with continuous linear finite elements. To this end, it will be sufficient to replace the term $u_j^{n+1} - u_j^n$ with

$$\frac{1}{6}[(u_{j-1}^{n+1} - u_{j-1}^n) + 4(u_j^{n+1} - u_j^n) + (u_{j+1}^{n+1} - u_{j+1}^n)].$$

Thus, the general scheme (12.13) is replaced by

$$\frac{1}{6}(u_{j-1}^{n+1} + 4u_j^{n+1} + u_{j-1}^{n+1}) = \frac{1}{6}(u_{j-1}^n + 4u_j^n + u_{j-1}^n) - \lambda(H_{j+1/2}^{n*} - H_{j-1/2}^{n*}), \tag{13.14}$$

where

$$H_{j+1/2}^{n*} = \begin{cases} H_{j+1/2}^n & \text{for explicit time-advancing schemes,} \\ H_{j+1/2}^{n+1} & \text{for implicit time-advancing schemes.} \end{cases}$$

Note that, even if we adopted a numerical flux corresponding to an explicit time-advancing scheme, the resulting scheme would no longer yield to a diagonal system (indeed, it becomes a tridiagonal one) because of the mass matrix terms. It could therefore appear that the use of an explicit time-advancing scheme for finite elements is inconvenient with respect to a similar full finite difference scheme. However, such a scheme has interesting features. In particular, let us consider its amplification and dispersion coefficients, using the von Neumann analysis illustrated in Sec. 12.4.3. To this end, let us suppose that the differential equation be defined on all of $\mathbb{R}$, or,

alternatively, let us consider a bounded interval, however imposing periodic boundary conditions. In either case, we can assume that relation (13.14) holds for all values of the index j. A simple computation leads us to writing the following relation between the amplification coefficient γ_k of a finite difference scheme (see Table 12.1) and the amplification coefficient γ_k^{FEM} of the corresponding finite element scheme

$$\gamma_k^{\text{FEM}} = \frac{3\gamma_k - 1 + \cos(\phi_k)}{2 + \cos(\phi_k)}, \tag{13.15}$$

where we denote again with ϕ_k the phase angle relative to the k-th harmonic (see Sec. 12.4.3).

We can thus compute the amplification and dispersion errors, which are reported in Fig. 13.1. Comparing them with the analogous errors relating to the corresponding finite difference scheme (reported in Fig. 12.6) we can make the following remarks. The forward Euler scheme is still unconditionally unstable (in the sense of strong stability). The upwind scheme (FEM) is strongly stable if the CFL number is less than $\frac{1}{3}$ (hence, a less restrictive result than the one found using the energy method), while the Lax-Friedrichs (FEM) method *never satisfies* the condition $\gamma_k^{FEM} \leq 1$ (in this case, in accordance with the result that we would find using the energy method).
More generally, we can say that in the case of schemes with an explicit temporal treatment the "finite element" version requires more restrictive stability conditions than the corresponding finite difference one. In particular, for the Lax-Wendroff finite element scheme, that we will denote with LW (FEM), the CFL number must now be less than $\frac{1}{\sqrt{3}}$, instead of 1 as in the finite differences case. However, the LW (FEM) scheme (for the CFL values for which it is stable), results to be slightly less diffusive and dispersive than the equivalent finite difference scheme for a wide range of values of the phase angle $\phi_k = kh$. The implicit Euler scheme remains unconditionally stable also in the FEM version (coherently with what we obtained using the energy method in Sec. 13.1.1).

Example 13.1 The previous conclusions have been experimentally verified as follows. We have repeated the case of Fig. 12.7 (right), where we have now considered a CFL value of 0.5. The numerical solutions obtained via the classical Lax-Wendroff method (LW) and via LW (FEM) for $t = 2$ are reported in Fig. 13.2. We can note how the LW (FEM) scheme provides a solution that is more accurate and especially more in phase with the exact solution. This result is confirmed by the value of the $\|\cdot\|_{\Delta,2}$ norm of the error in the two cases. Indeed, by calling u the exact solution and u_{LW} resp. $u_{LW(FEM)}$ the one obtained using the two numerical schemes, $\|u_{LW} - u\|_{\Delta,2} = 0.78, \quad \|u_{LW(FEM)} - u\|_{\Delta,2} = 0.49.$
Further tests conducted by using non periodic boundary conditions confirm the stability properties previously derived. ■

13.2 Taylor-Galerkin schemes

We now illustrate a class of finite element schemes named "Taylor-Galerkin" schemes. These are derived in a similar way to the Lax-Wendroff scheme, and we will indeed see that the LW (FEM) version is part of their class.

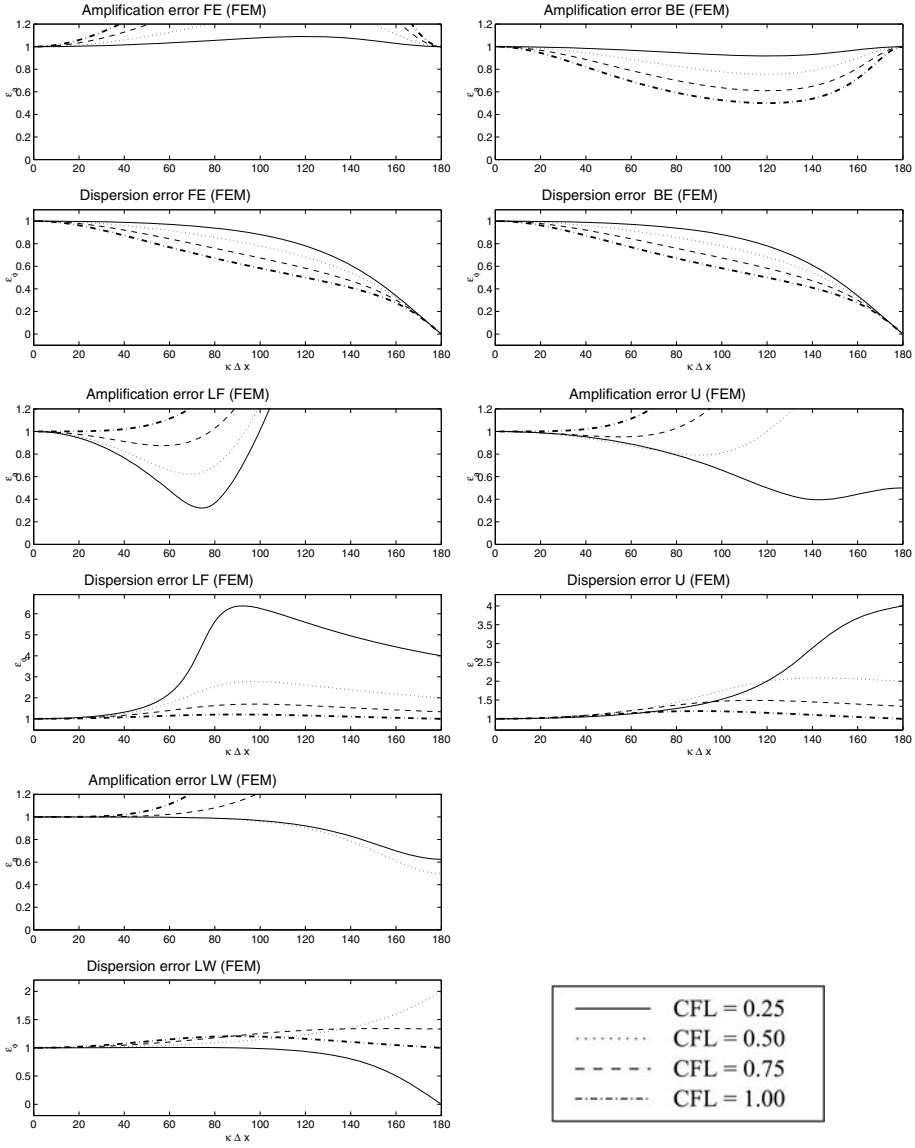


Fig. 13.1. Amplification and dispersion errors for several finite element schemes obtained from the general scheme (13.14)

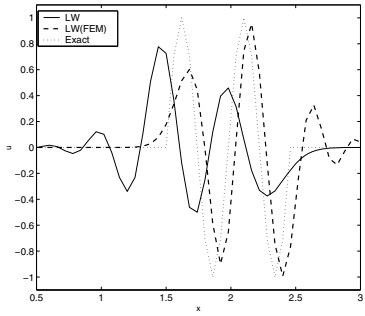


Fig. 13.2. Comparison between the solution obtained via the Lax-Wendroff finite difference scheme (LW) and its finite element version (LW (FEM)) ($\phi_k = \pi/4$, $t = 2$)

For simplicity, we will refer to the pure transport problem (12.1). The Taylor-Galerkin method consists in combining the Taylor formula truncated to the first order

$$u(x, t^{n+1}) = u(x, t^n) + \Delta t \frac{\partial u}{\partial t}(x, t^n) + \int_{t^n}^{t^{n+1}} (s - t^n) \frac{\partial^2 u}{\partial t^2}(x, s)\, ds \qquad (13.16)$$

with equation (12.1), thanks to which we obtain

$$\frac{\partial u}{\partial t} = -a \frac{\partial u}{\partial x},$$

and, by formal derivation,

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}(-a \frac{\partial u}{\partial x}) = -a \frac{\partial}{\partial x} \frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}.$$

From (13.16) we then obtain

$$u(x, t^{n+1}) = u(x, t^n) - a\Delta t \frac{\partial u}{\partial x}(x, t^n) + a^2 \int_{t^n}^{t^{n+1}} (s - t^n) \frac{\partial^2 u}{\partial x^2}(x, s)\, ds. \quad (13.17)$$

We approximate the integral in the following way

$$\int_{t^n}^{t^{n+1}} (s - t^n) \frac{\partial^2 u}{\partial x^2}(x, s)\, ds \approx \frac{\Delta t^2}{2} \left[\theta \frac{\partial^2 u}{\partial x^2}(x, t^n) + (1 - \theta) \frac{\partial^2 u}{\partial x^2}(x, t^{n+1}) \right], \quad (13.18)$$

obtained by evaluating the first factor in $s = t^n + \frac{\Delta t}{2}$ and the second one through a linear combination using $\theta \in [0, 1]$ as a parameter of its values in $s = t^n$ and $s = t^{n+1}$. We denote by $u^n(x)$ the approximating function $u(x, t^n)$.

Let us consider two remarkable situations. If $\theta = 1$, the resulting semi-discretized scheme is explicit in time and is written as

$$u^{n+1} = u^n - a\Delta t \frac{\partial u^n}{\partial x} + \frac{a^2 \Delta t^2}{2} \frac{\partial^2 u^n}{\partial x^2}.$$

If we now discretize in space by finite differences or finite elements, we re-encounter the previously examined LW and LW (FEM) schemes.
Instead, if we take $\theta = \frac{2}{3}$, the approximation error in (13.18) becomes $O(\Delta t^4)$ (supposing that u has the required regularity). De facto, such choice corresponds to approximating $\frac{\partial^2 u}{\partial x^2}$ between t^n and t^{n+1} with its linear interpolant. The resulting semi-discretized scheme is written

$$\left[1 - \frac{a^2\Delta t^2}{6}\frac{\partial^2}{\partial x^2}\right] u^{n+1} = u^n - a\Delta t\frac{\partial u^n}{\partial x} + \frac{a^2\Delta t^2}{3}\frac{\partial^2 u^n}{\partial x^2}, \tag{13.19}$$

and the truncation error of the semi-discretized scheme in time (13.19) is $\mathcal{O}(\Delta t^3)$.

At this point, a discretization in space using the finite element method leads to the following scheme, called Taylor-Galerkin (TG):
for $n = 0, 1, \ldots$ find $u_h^{n+1} \in V_h$ such that

$$\begin{aligned} A(u_h^{n+1}, v_h) &= (u_h^n, v_h) - a\Delta t\left(\frac{\partial u_h^n}{\partial x}, v_h\right) - \frac{a^2\Delta t^2}{3}\left(\frac{\partial u_h^n}{\partial x}, \frac{\partial v_h}{\partial x}\right) \\ &\quad + \gamma\frac{a^2\Delta t^2}{3}\frac{\partial u_h^n}{\partial x}(1)v_h(1) \quad \forall v_h \in V_h, \end{aligned} \tag{13.20}$$

where

$$A(u_h^{n+1}, v_h) = \left(u_h^{n+1}, v_h\right) + \frac{a^2\Delta t^2}{6}\left(\frac{\partial u_h^{n+1}}{\partial x}, \frac{\partial v_h}{\partial x}\right) - \gamma\frac{a^2\Delta t^2}{6}\frac{\partial u_h^{n+1}}{\partial x}(1)v_h(1),$$

and $\gamma = 1, 0$ depending on whether or not we want to take into account the boundary contribution in the integration by parts of the second derivative term.

The latter yields a linear system whose matrix is

$$\mathrm{A} = \mathrm{M} + \frac{a^2(\Delta t)^2}{6}\mathrm{K},$$

M being the mass matrix and K being the stiffness matrix, potentially taking the boundary contribution as well into account (if $\gamma = 1$).

In the case of linear finite elements, the von Neumann analysis leads to the following amplification factor for the scheme (13.20)

$$\gamma_k = \frac{2 + \cos(kh) - 2a^2\lambda^2(1 - \cos(kh)) + 3ia\lambda\sin(kh)}{2 + \cos(kh) + a^2\lambda^2(1 - \cos(kh))}. \tag{13.21}$$

It can be proven that the scheme is strongly stable in the $\|\cdot\|_{\Delta,2}$ norm under the CFL condition $\frac{a\Delta t}{h} \leq 1$. Thus, it has a *less restrictive* stability condition than the Lax-Wendroff (FEM) scheme.
Fig. 13.3 shows the behavior of the amplification and dispersion error for the scheme (13.20), as a function of the phase angle, analogously to what we have seen for other schemes in Sec. 12.4.4.

In the case of linear finite elements the truncation error of the TG scheme results to be $\mathcal{O}(\Delta t^3) + \mathcal{O}(h^2) + \mathcal{O}(h^2\Delta t)$.

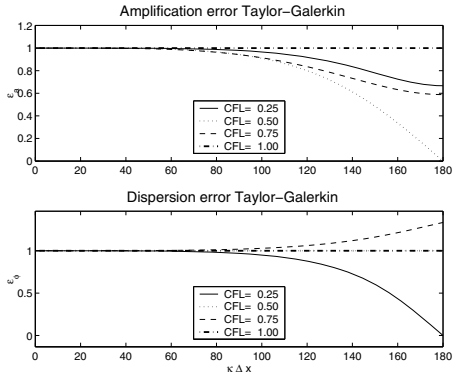


Fig. 13.3. Amplification (top) and dispersion (bottom) error of the Taylor-Galerkin scheme (13.20), as a function of the phase angle $\phi_k = kh$ and for different values of the CFL number

Example 13.2 To compare the accuracy of the schemes presented in the last two sections, we have considered the problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial x} = 0, & x \in (0, 0.5), t > 0, \\ u(x, 0) = u_0(x), & x \in (0, 0.5), \end{cases}$$

with periodic boundary conditions, $u(0,t) = u(0.5,t)$, for $t > 0$. The initial datum is $u_0(x) = 2\cos(4\pi x) + \sin(20\pi x)$, and is illustrated in Fig. 13.4 (left). The latter superposes two harmonics, one with low frequency one and one with high frequency.

We have considered the Taylor-Galerkin, Lax-Wendroff (FEM), (finite difference) Lax-Wendroff and upwind schemes. In Fig. 13.4 (right), we show the error in discrete norm $\|u - u_h\|_{\Delta,2}$ obtained at time $t = 1$ for different values of Δt and at a fixed CFL number

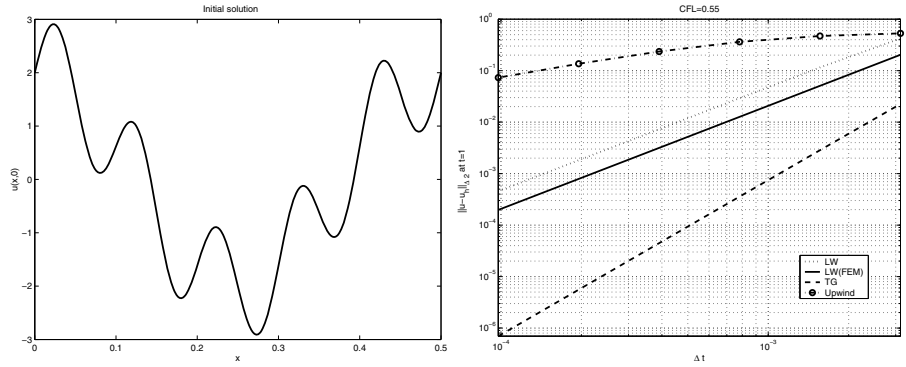


Fig. 13.4. Initial condition u_0 for the simulation of example 13.2 (left) and error $\|u - u_h\|_{\Delta,2}$ at $t = 1$ for varying Δt at fixed CFL for different numerical schemes (right)

of 0.55. We can note a better convergence of the Taylor-Galerkin scheme, while the two versions of the Lax-Wendroff scheme show the same order of convergence, but with a smaller error for the finite element version with respect to the finite difference one. The upwind scheme is less accurate: it features a larger absolute error and a lower convergence rate. Moreover, it can be verified that for a fixed CFL, the error of the upwind scheme is $\mathcal{O}(\Delta t)$, that of both variants of the Lax-Wendroff scheme is $\mathcal{O}(\Delta t^2)$, while the error of the Taylor-Galerkin scheme is $\mathcal{O}(\Delta t^3)$. ■

We report in Fig. 13.5 and 13.6 the numerical approximations and corresponding errors in the maximum norm for the transport problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x} = 0, & x \in (0, 2\pi),\ t > 0 \\ u(x,0) = \sin\big(\pi\ \cos(x)\big), & x \in (0, 2\pi) \end{cases}$$

and periodic boundary conditions. Such approximations are obtained using finite differences of order 2 and 4 (ufd2, ufd4), compact finite differences of order 4 and 6 (ucp4, ucp6), and by the Galerkin spectral method using Fourier basis (ugal). For the

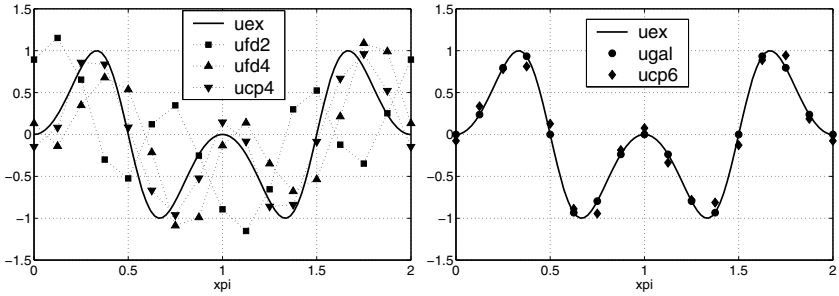


Fig. 13.5. Approximation of the solution of a wave propagation problem using finite difference methods (of order 2, 4), compact finite difference methods (of order 4 and 6) and with the Fourier Galerkin spectral method (from [CHQZ06])

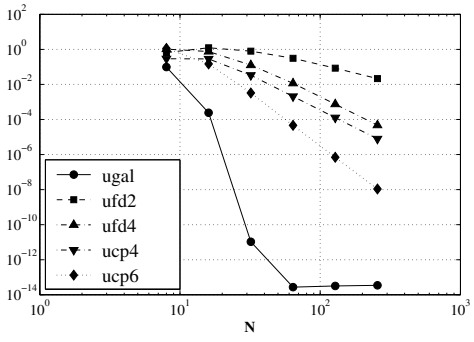


Fig. 13.6. Behavior of the error in the maximum norm for the different numerical methods reported in Fig. 13.5 (from [CHQZ06])

sake of comparison, we also report the exact solution $u(x,t) = \sin\big(\pi\,\cos(x+t)\big)$ (uex).

13.3 The multi-dimensional case

Let us now move to the multi-dimensional case and let us consider the following first-order, linear and scalar hyperbolic transport-reaction problem in the domain $\Omega \subset \mathbb{R}^d$, with $d = 2, 3$

$$\begin{cases} \dfrac{\partial u}{\partial t} + \mathbf{a}\cdot\nabla u + a_0 u = f, & \mathbf{x}\in\Omega,\ t>0 \\ u = \varphi, & \mathbf{x}\in\partial\Omega^{in},\ t>0, \\ u_{|t=0} = u_0, & \mathbf{x}\in\Omega, \end{cases} \tag{13.22}$$

where $\mathbf{a} = \mathbf{a}(\mathbf{x})$, $a_0 = a_0(\mathbf{x},t)$ (optionally null), $f = f(\mathbf{x},t)$, $\varphi = \varphi(\mathbf{x},t)$ and $u_0 = u_0(\mathbf{x})$ are given functions. The inflow boundary $\partial\Omega^{in}$ is defined by

$$\partial\Omega^{in} = \{\mathbf{x}\in\partial\Omega :\ \mathbf{a}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) < 0\}, \tag{13.23}$$

$\mathbf{n}$ being the outward unit normal vector to $\partial\Omega$.

For simplicity, we have supposed that $\mathbf{a}$ does not depend on t; this way, the inflow boundary $\partial\Omega^{in}$ does not change with time.

13.3.1 Semi-discretization: strong and weak treatment of the boundary conditions

To obtain a semi-discrete approximation of problem (13.22), similar to that used in the one-dimensional case (13.1), we define the spaces

$$V_h = X_h^r, \qquad V_h^{in} = \{v_h\in V_h :\ v_{h|\partial\Omega^{in}} = 0\},$$

where r is an integer ≥ 1 and X_h^r has been introduced in (4.38).

We denote by $u_{0,h}$ and φ_h two suitable finite element approximations of u_0 and φ, respectively, and we consider the problem: for each $t>0$ find $u_h(t)\in V_h$ such that

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h\,d\Omega + \int_\Omega \mathbf{a}\cdot\nabla u_h(t) v_h\,d\Omega + \int_\Omega a_0(t) u_h(t) v_h\,d\Omega \\ \qquad = \displaystyle\int_\Omega f(t) v_h\,d\Omega \qquad \forall\, v_h\in V_h^{in}, \\ u_h(t) = \varphi_h(t) \quad \text{on } \partial\Omega^{in}, \end{cases} \tag{13.24}$$

with $u_h(0) = u_{0,h}\in V_h$.

To obtain a stability estimate, we assume for simplicity that φ, and therefore φ_h, is identically null. In this case $u_h(t) \in V_h^{in}$, and taking, for every t, $v_h = u_h(t)$, we get the following inequality

$$\begin{aligned} \|u_h(t)\|^2_{\mathrm{L}^2(\Omega)} \quad &+ \int_0^t \mu_0 \|u_h(\tau)\|^2_{\mathrm{L}^2(\Omega)}\, d\tau + \int_0^t \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\; u_h^2(\tau)\, d\gamma\, d\tau \\ &\leq \|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \frac{1}{\mu_0}\|f(\tau)\|^2_{\mathrm{L}^2(\Omega)} d\tau\,. \end{aligned} \tag{13.25}$$

We have assumed that there exists a positive constant μ_0 s.t., for all $t > 0$ and for each $\mathbf{x}$ in Ω

$$0 < \mu_0 \leq \mu(\mathbf{x}, t) = a_0(\mathbf{x}, t) - \frac{1}{2}\mathrm{div}\mathbf{a}(\mathbf{x}). \tag{13.26}$$

In the case where such hypothesis is not verified (for instance if $\mathbf{a}$ is a constant field and $a_0 = 0$), then by using the Gronwall Lemma 2.2 we obtain

$$\begin{aligned} &\|u_h(t)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\, u_h^2(\tau) d\gamma\, d\tau \\ &\leq \left(\|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|f(\tau)\|^2_{\mathrm{L}^2(\Omega)}\, d\tau \right) \exp \int_0^t [1 + 2\mu^*(\tau)]\, d\tau, \end{aligned} \tag{13.27}$$

where we have set $\mu^*(t) = \max_{\mathbf{x}\in\bar{\Omega}} |\mu(\mathbf{x}, t)|$.

Supposing for simplicity that $f = 0$, if $u_0 \in H^{r+1}(\Omega)$ we have the following convergence result

$$\begin{aligned} \max_{t\in[0,T]} \|u(t) - u_h(t)\|_{\mathrm{L}^2(\Omega)} \quad &+ \left(\int_0^T \int_{\partial\Omega} |\mathbf{a}\cdot\mathbf{n}|\, |u(t) - u_h(t)|^2\, d\gamma\, dt \right)^{1/2} \\ &\leq C h^r \|u_0\|_{\mathrm{H}^{r+1}(\Omega)}. \end{aligned}$$

For the proofs, we refer to [QV94, Chap. 14], [Joh87] and to the references cited thereby.

In problem (13.24) the boundary condition has been imposed in a *strong* (or essential) way. An alternative option is the *weak* (or natural) treatment that derives from the integration by parts of the transport term in the first equation in (13.24), where we now consider $v_h \in V_h$ (i.e. we no longer require that the test function be null on the inflow boundary). We obtain

$$\begin{aligned} &\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h\, d\Omega - \int_\Omega \mathrm{div}(\mathbf{a}\, v_h)) u_h(t)\, d\Omega \\ &+ \int_\Omega a_0 u_h(t) v_h\, d\Omega + \int_{\partial\Omega} \mathbf{a}\cdot\mathbf{n}\, u_h(t) v_h\, d\gamma = \int_\Omega f(t) v_h\, d\Omega. \end{aligned}$$

The boundary condition is imposed by replacing u_h with φ_h on the inflow boundary part, obtaining

$$\begin{aligned} &\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega - \int_\Omega \mathrm{div}(\mathbf{a}v_h) u_h(t) \, d\Omega \\ + &\int_\Omega a_0 u_h(t) v_h \, d\Omega + \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\, u_h(t) v_h \, d\gamma \\ = &\int_\Omega f(t) v_h \, d\Omega - \int_{\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\, \varphi_h(t) v_h \, d\gamma \quad \forall v_h \in V_h \,. \end{aligned} \tag{13.28}$$

Clearly, the solution u_h found in this way only satisfies the boundary condition in an approximate way.

A further option consists in counter-integrating (13.28) by parts, thus getting to the following formulation: for each $t > 0$, find $u_h(t) \in V_h$ such that

$$\begin{aligned} &\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + \int_\Omega \mathbf{a}\cdot\nabla u_h(t) v_h \, d\Omega + \int_{\partial\Omega^{in}} v_h(\varphi_h(t) - u_h(t))\mathbf{a}\cdot\mathbf{n}\, d\gamma \\ &= \int_\Omega f(t) v_h \, d\Omega \quad \forall v_h \in V_h \,. \end{aligned} \tag{13.29}$$

We note that the (13.28) and (13.29) formulations are equivalent: the only difference is the way boundary terms are highlighted. In particular, the boundary integral in formulation (13.29) can be interpreted as a penalization term with which we evaluate how different u_h is from the data φ_h on the inflow boundary. Assuming that hypothesis (13.26) is still true, having chosen $v_h = u_h(t)$ in (13.29), integrating the convective term by parts and using the Cauchy-Schwarz and Young inequalities we get the following stability estimate

$$\begin{aligned} &\|u_h(t)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \mu_0 \|u_h(\tau)\|^2_{\mathrm{L}^2(\Omega)} \, d\tau + \int_0^t \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\, u_h^2(\tau) \, d\gamma \, d\tau \\ &\le \|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \int_{\partial\Omega^{in}} |\mathbf{a}\cdot\mathbf{n}| \varphi_h^2(\tau) d\gamma \, d\tau + \int_0^t \frac{1}{\mu_0} \|f(\tau)\|^2_{\mathrm{L}^2(\Omega)} d\tau. \end{aligned} \tag{13.30}$$

In absence of hypothesis (13.26), inequality (13.30) would change in an analogous way to what we have previously seen, provided we use the Gronwall Lemma 2.2 as done to derive (13.27).

Remark 13.2 In the case where the boundary condition for problem (13.22) takes the form $\mathbf{a}\cdot\mathbf{n}\, u = \psi$, we could again impose it weakly by adding a penalization term, that

in such case would take the form

$$\int_{\partial\Omega^{in}} (\psi_h(t) - \mathbf{a}\cdot\mathbf{n}\, u_h(t)) v_h \, d\gamma,$$

ψ_h being a suitable finite element approximation of the datum ψ. •

Alternatively to the strong and weak imposition of the boundary conditions, i.e. to formulations (13.24) and (13.29), we could adopt a Petrov-Galerkin approach by imposing in a strong way the condition $u_h(t) = \varphi_h(t)$ on the inflow boundary $\partial\Omega^{in}$ and requiring $v_h = 0$ on the outflow boundary $\partial\Omega^{out}$, yielding the following discrete formulation. Set $V_h^{out} = \{v_h \in V_h : \ v_{h|\partial\Omega^{out}} = 0\}$ then, for each $t > 0$, find $u_h(t) \in V_h = X_h^r$ such that

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + \int_\Omega (\mathbf{a}\cdot\nabla u_h(t))\, v_h \, d\Omega + \int_\Omega a_0(t)\, u_h(t)\, v_h \, d\Omega \\ \qquad\qquad = \displaystyle\int_\Omega f(t)\, v_h \, d\Omega \quad \forall v_h \in V_h^{out}, \\ u_h(t) = \varphi_h(t) \quad \text{on } \partial\Omega^{in}. \end{cases}$$

We recall that for a Petrov-Galerkin formulation, the well-posedness analysis cannot be based on the Lax-Milgram lemma any longer.

Instead, if the inflow condition were inposed in a weak way, we would have the following formulation:
for each $t > 0$, find $u_h(t) \in V_h = X_h^r$ such that, for each $v_h \in V_h^{out}$,

$$\begin{aligned} &\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega - \int_\Omega \mathrm{div}(\mathbf{a}\, v_h)\, u_h(t) \, d\Omega + \int_\Omega a_0(t)\, u_h(t)\, v_h \, d\Omega \\ &= \int_\Omega f(t)\, v_h \, d\Omega - \int_{\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}\, \varphi_h(t)\, v_h \, d\gamma. \end{aligned}$$

For further details, the reader can refer to [QV94, Chap. 14].

13.3.2 Temporal discretization

For an illustrative purpose, let us limit ourselves to considering the Galerkin semi-discrete problem (13.24). Using the backward Euler scheme for the temporal discretization, we get to the following fully-discrete problem:

$\forall n \geq 0$ find $u_h^n \in V_h$ s.t.

$$\begin{cases} \frac{1}{\Delta t}\displaystyle\int_\Omega (u_h^{n+1} - u_h^n) v_h \, d\Omega + \int_\Omega \mathbf{a} \cdot \nabla u_h^{n+1} v_h \, d\Omega + \int_\Omega a_0^{n+1} u_h^{n+1} v_h \, d\Omega \\ = \displaystyle\int_\Omega f^{n+1} v_h \, d\Omega \quad \forall v_h \in V_h^{in}, \\ u_h^{n+1} = \varphi_h^{n+1} \text{ on } \partial\Omega^{in}, \end{cases}$$

with $u_h^0 = u_{0,h} \in V_h$ being a suitable approximation in V_h of the initial datum u_0.

Let us limit ourselves to the homogeneous case, where $f = 0$ and $\varphi_h = 0$ (in this case $u_h^n \in V_h^{in}$ for every $n \geq 0$). Having set $v_h = u_h^{n+1}$ and using identities (13.8) and (13.26), we obtain, for each $n \geq 0$

$$\frac{1}{2\Delta t}\left(\|u_h^{n+1}\|^2_{\mathrm{L}^2(\Omega)} - \|u_h^n\|^2_{\mathrm{L}^2(\Omega)}\right) + \tfrac{1}{2}\int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}(u_h^{n+1})^2 \, d\gamma + \mu_0 \|u_h^{n+1}\|^2_{\mathrm{L}^2(\Omega)} \leq 0.$$

For each $m \geq 1$, summing over n from 0 to $m-1$ we obtain

$$\|u_h^m\|^2_{\mathrm{L}^2(\Omega)} + 2\Delta t\left(\mu_0 \sum_{n=0}^m \|u_h^n\|^2_{\mathrm{L}^2(\Omega)} + \tfrac{1}{2}\sum_{n=0}^m \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}(u_h^n)^2 \, d\gamma\right) \leq \|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)}.$$

In particular, as $\mathbf{a}\cdot\mathbf{n} \geq 0$ on $\partial\Omega \setminus \partial\Omega^{in}$, we conclude that

$$\|u_h^m\|_{\mathrm{L}^2(\Omega)} \leq \|u_{0,h}\|_{\mathrm{L}^2(\Omega)} \quad \forall m \geq 0.$$

As expected, this method is strongly stable, with no condition on Δt and h. We now consider the discretization in time using the forward Euler method

$$\begin{cases} \frac{1}{\Delta t}\displaystyle\int_\Omega (u_h^{n+1} - u_h^n) v_h \, d\Omega + \int_\Omega \mathbf{a} \cdot \nabla u_h^n v_h \, d\Omega + \int_\Omega a_0^n u_h^n v_h \, d\Omega \\ = \displaystyle\int_\Omega f^n v_h \, d\Omega \quad \forall v_h \in V_h, \\ u_h^{n+1} = \varphi_h^{n+1} \text{ on } \partial\Omega^{in}. \end{cases} \tag{13.31}$$

We suppose again that $f = 0$, $\varphi = 0$ and that the condition (13.26) is verified. Moreover, we suppose that $\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)} < \infty$ and that, for each $t > 0$, $\|a_0\|_{\mathrm{L}^\infty(\Omega)} < \infty$. Setting $v_h = u_h^n$, exploiting identity (13.5) and integrating the convective term by parts, we obtain

$$\frac{1}{2\Delta t}\left(\|u_h^{n+1}\|^2_{\mathrm{L}^2(\Omega)} - \|u_h^n\|^2_{\mathrm{L}^2(\Omega)} - \|u_h^{n+1} - u_h^n\|^2_{\mathrm{L}^2(\Omega)}\right) + \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}(u_h^n)^2 \, d\gamma + (-\frac{1}{2}\operatorname{div}(\mathbf{a}) + a_0^n, (u_h^n)^2) = 0,$$

and then, after a few steps,

$$\begin{aligned} \|u_h^{n+1}\|^2_{\mathrm{L}^2(\Omega)} &+ 2\Delta t \int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}(u_h^n)^2\, d\gamma + 2\Delta t\mu_0\|u_h^n\|^2_{\mathrm{L}^2(\Omega)} \\ &\leq \|u_h^n\|^2_{\mathrm{L}^2(\Omega)} + \|u_h^{n+1}-u_h^n\|^2_{\mathrm{L}^2(\Omega)}. \end{aligned} \tag{13.32}$$

It is now necessary to control the term $\|u_h^{n+1}-u_h^n\|^2_{\mathrm{L}^2(\Omega)}$. To this end, we set $v_h = u_h^{n+1}-u_h^n$ in (13.31) and obtain

$$\|u_h^{n+1}-u_h^n\|^2_{\mathrm{L}^2(\Omega)} = -\Delta t(\mathbf{a}\nabla u_h^n, u_h^{n+1}-u_h^n) - \Delta t(a_0^n u_h^n, u_h^{n+1}-u_h^n)$$

$$\leq \Delta t\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)}|(\nabla u_h^n, u_h^{n+1}-u_h^n)| + \Delta t\|a_0^n\|_{\mathrm{L}^\infty(\Omega)}|(u_h^n, u_h^{n+1}-u_h^n)|$$

$$\leq \Delta t\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)}\|\nabla u_h^n\|_{\mathrm{L}^2(\Omega)}\|u_h^{n+1}-u_h^n\|_{\mathrm{L}^2(\Omega)} +$$

$$\Delta t\|a_0^n\|_{\mathrm{L}^\infty(\Omega)}\|u_h^n\|_{\mathrm{L}^2(\Omega)}\|u_h^{n+1}-u_h^n\|_{\mathrm{L}^2(\Omega)}.$$

Using the inverse inequality (4.52), we obtain

$$\begin{aligned} \|u_h^{n+1}-u_h^n\|^2_{\mathrm{L}^2(\Omega)} &\leq \Delta t(C_I h^{-1}\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)} + \\ &\quad \|a_0^n\|_{\mathrm{L}^\infty(\Omega)})\|u_h^n\|_{\mathrm{L}^2(\Omega)}\|u_h^{n+1}-u_h^n\|_{\mathrm{L}^2(\Omega)}, \end{aligned}$$

and then

$$\|u_h^{n+1}-u_h^n\|_{\mathrm{L}^2(\Omega)} \leq \Delta t\left(C_I h^{-1}\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)} + \|a_0^n\|_{\mathrm{L}^\infty(\Omega)}\right)\|u_h^n\|_{\mathrm{L}^2(\Omega)}.$$

Using such result to find an upper bound to the term in (13.32), we have

$$\|u_h^{n+1}\|^2_{\mathrm{L}^2(\Omega)} + 2\Delta t\int_{\partial\Omega\setminus\partial\Omega^{in}} \mathbf{a}\cdot\mathbf{n}(u_h^n)^2\, d\Omega +$$

$$\Delta t\left[2\mu_0 - \Delta t\,\left(C_I h^{-1}\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)} + \|a_0^n\|_{\mathrm{L}^\infty(\Omega)}\right)^2\right]\|u_h^n\|^2_{\mathrm{L}^2(\Omega)}$$

$$\leq \|u_h^n\|^2_{\mathrm{L}^2(\Omega)}.$$

The integral on $\partial\Omega\setminus\partial\Omega^{in}$ is positive because of the hypotheses on the boundary conditions; hence, if

$$\Delta t \leq \frac{2\mu_0}{\left(C_I h^{-1}\|\mathbf{a}\|_{\mathrm{L}^\infty(\Omega)} + \|a_0^n\|_{\mathrm{L}^\infty(\Omega)}\right)^2} \tag{13.33}$$

we have $\|u_h^{n+1}\|_{\mathrm{L}^2(\Omega)} \leq \|u_h^n\|_{\mathrm{L}^2(\Omega)}$, that is the scheme is strongly stable. Note that the stability condition (13.33) is of parabolic type, similar to the one found in (12.31) for the case of finite difference discretizations.

Remark 13.3 In the case where $\mathbf{a}$ is constant and $a_0 = 0$ we have that $\mu_0 = 0$ and the stability condition (13.33) can never be satisfied by a positive Δt. Thus, the result in (13.33) does not contradict the one we have previously found for the forward Euler scheme. •

13.4 Discontinuous finite elements

An alternative approach to the one adopted so far is based on the use of *discontinuous* finite elements. The resulting method is called the discontinuous Galerkin method (DG in short). This choice is motivated by the fact that, as we have previously observed, the solutions of (even linear) hyperbolic problems can be discontinuous.

For a given mesh $\mathcal{T}_h$ of Ω, the space of discontinuous finite elements is

$$W_h = Y_h^r = \{v_h \in \mathrm{L}^2(\Omega) \,|\, v_{h|K} \in \mathbb{P}_r, \ \forall\, K \in \mathcal{T}_h\}, \tag{13.34}$$

that is the space of piecewise polynomial functions of degree less than or equal to r, with $r \geq 0$, which are not necessarily continuous across the finite element interfaces.

13.4.1 The one-dimensional case

In the case of the one-dimensional problem (12.3), the DG finite element method takes the following form: $\forall t > 0$, find a function $u_h = u_h(t) \in W_h$ such that

$$\begin{aligned}
&\int_\alpha^\beta \frac{\partial u_h(t)}{\partial t} v_h \, dx \\
&+ \sum_{i=0}^{m-1} \Big[\int_{x_i}^{x_{i+1}} \left(a \frac{\partial u_h(t)}{\partial x} + a_0 u_h(t) \right) v_h \, dx \\
&+ a(x_i)(u_h^+(t) - U_h^-(t))(x_i) v_h^+(x_i) \Big] \\
&= \int_\alpha^\beta f(t) v_h \, dx \quad \forall v_h \in W_h,
\end{aligned} \tag{13.35}$$

where we have supposed that $a(x)$ is a continuous function. We have set, for each $t > 0$,

$$U_h^-(t)(x_i) = \begin{cases} u_h^-(t)(x_i), & i = 1, \ldots, m-1, \\ \varphi_h(t)(x_0), \end{cases} \tag{13.36}$$

where $\{x_i, \ i = 0, \cdots, m\}$ are the nodes, $x_0 = \alpha$, $x_m = \beta$, h is the maximal distance between two consecutive nodes, $v_h^+(x_i)$ denotes the right limit of v_h in x_i, $v_h^-(x_i)$ the left one. For simplicity of notation, the dependence of u_h and f on t will often be understood when this does not yield to ambiguities.

We now derive a stability estimate for the solution u_h of (13.35), supposing, for simplicity, that the forcing term f be identically null. Having then chosen $v_h = u_h$ in

(13.35), we have (setting $\Omega = (\alpha, \beta)$)

$$\frac{1}{2}\frac{d}{dt}\|u_h\|^2_{\mathrm{L}^2(\Omega)} \quad + \quad \sum_{i=0}^{m-1}\Big[\int_{x_i}^{x_{i+1}}\Big(\frac{a}{2}\frac{\partial}{\partial x}(u_h)^2 + a_0 u_h^2\Big)\,dx + a(x_i)(u_h^+ - U_h^-)(x_i)u_h^+(x_i)\Big] = 0.$$

Now, integrating the convective term by parts, we have

$$\frac{1}{2}\frac{d}{dt}\|u_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{i=0}^{m-1}\int_{x_i}^{x_{i+1}}\Big(a_0 - \frac{\partial}{\partial x}\Big(\frac{a}{2}\Big)\Big)u_h^2\,dx + \sum_{i=0}^{m-1}\Big[\frac{a}{2}(x_{i+1})\,(u_h^-(x_{i+1}))^2 + \frac{a}{2}(x_i)\,(u_h^+(x_i))^2 - a(x_i)\,U_h^-(x_i)\,u_h^+(x_i)\Big] = 0. \tag{13.37}$$

Isolating the contribution associated to node x_0 and exploiting definition (13.36), we can rewrite the second sum in the previous equation as

$$\begin{aligned}
&\sum_{i=0}^{m-1}\Big[\frac{a}{2}(x_{i+1})\,(u_h^-(x_{i+1}))^2 + \frac{a}{2}(x_i)\,(u_h^+(x_i))^2 - a(x_i)\,U_h^-(x_i)\,u_h^+(x_i)\Big] \\
= \;& \frac{a}{2}(x_0)\,(u_h^+(x_0))^2 - a(x_0)\,\varphi_h(x_0)\,u_h^+(x_0) + \frac{a}{2}(x_m)\,(u_h^-(x_m))^2 + \\
&\sum_{i=1}^{m-1}\Big[\frac{a}{2}(x_i)\Big((u_h^-(x_i))^2 + (u_h^+(x_i))^2\Big) - a(x_i)\,u_h^-(x_i)\,u_h^+(x_i)\Big] \\
= \;& \frac{a}{2}(x_0)\,(u_h^+(x_0))^2 - a(x_0)\,\varphi_h(x_0)\,u_h^+(\alpha) + \\
&\frac{a}{2}(x_m)\,(u_h^-(x_m))^2 + \sum_{i=1}^{m-1}\frac{a}{2}(x_i)\,\big[u_h(x_i)\big]^2,
\end{aligned} \tag{13.38}$$

having denoted by $\big[u_h(x_i)\big] = u_h^+(x_i) - u_h^-(x_i)$ the jump of function u_h at node x_i. We now suppose, analogously to the multi-dimensional case (see (13.26)), that

$$\exists \gamma \geq 0 \;\text{ s.t. }\; a_0 - \frac{\partial}{\partial x}\Big(\frac{a}{2}\Big) \geq \gamma. \tag{13.39}$$

Returning to (13.37) and using the relation (13.38) and the Cauchy-Schwarz and Young inequalities, we have

$$\frac{1}{2}\frac{d}{dt}\|u_h\|^2_{\mathrm{L}^2(\Omega)} + \gamma\|u_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{i=1}^{m-1}\frac{a}{2}(x_i)\,\big[u_h(x_i)\big]^2 + \frac{a}{2}(x_0)\,(u_h^+(x_0))^2 + \frac{a}{2}(x_m)\,(u_h^-(x_m))^2 = a(x_0)\,\varphi_h(x_0)\,u_h^+(x_0) \leq \frac{a}{2}(x_0)\,\varphi_h^2(x_0) + \frac{a}{2}(x_0)\,(u_h^+(x_0))^2,$$

that is, integrating with respect to time as well, $\forall t > 0$,

$$\|u_h(t)\|^2_{\mathrm{L}^2(\Omega)} + 2\,\gamma \int_0^t \|u_h(t)\|^2_{\mathrm{L}^2(\Omega)}\, dt + \sum_{i=1}^{m-1} a(x_i) \int_0^t \big[u_h(x_i,t)\big]^2\, dt + a(x_m)\,(u_h^-(x_m))^2 \le \|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + a(x_0) \int_0^t \varphi_h^2(x_0,t)\, dt. \tag{13.40}$$

Such estimate represents the desired stability result.
Note that, in case the forcing term is no longer null, we can replicate the previous analysis by suitably using the Gronwall Lemma 2.2 to handle the contribution of f. This would lead to an estimate similar to (13.40), however this time the right-hand side of the inequality would become

$$e^t\Big(\|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + a(x_0) \int_0^t \varphi_h^2(x_0,t)\, dt + \int_0^t (f(\tau))^2\, d\tau\Big). \tag{13.41}$$

In the case where the γ constant in inequality (13.39) is strictly positive, we could avoid using the Gronwall lemma, getting an estimate such as (13.40) where in the first term $2\,\gamma$ is replaced by γ, while the second term takes the form (13.41), however without the exponential e^t.

Because of the discontinuity of test functions, (13.35) can be rewritten in an equivalent way as follows, $\forall i = 0, \ldots, m-1$,

$$\begin{aligned} &\int_{x_i}^{x_{i+1}} \left(\frac{\partial u_h}{\partial t} + a\frac{\partial u_h}{\partial x} + a_0 u_h\right) v_h dx + a(u_h^+ - U_h^-)(x_i) v_h^+(x_i) \\ &\qquad = \int_\alpha^\beta f v_h dx \quad \forall v_h \in \mathbb{P}_r(I_i), \end{aligned} \tag{13.42}$$

with $I_i = [x_i, x_{i+1}]$. In other terms, the approximation via discontinuous finite elements yields to element-wise "independent" relations, the only term connecting an element and its neighbors is the jump term $(u_h^+ - U_h^-)$ that can also be interpreted as the attribution of the boundary datum on the inflow boundary of the element under exam.

We then have a set of small problems to be solved in each element, precisely $r+1$ equations for each interval $[x_i, x_{i+1}]$. Let us write them in compact form as

$$M_h \dot{\mathbf{u}}_h(t) + L_h \mathbf{u}_h(t) = \mathbf{f}_h(t) \qquad \forall t > 0, \quad \mathbf{u}_h(0) = \mathbf{u}_{0,h}, \tag{13.43}$$

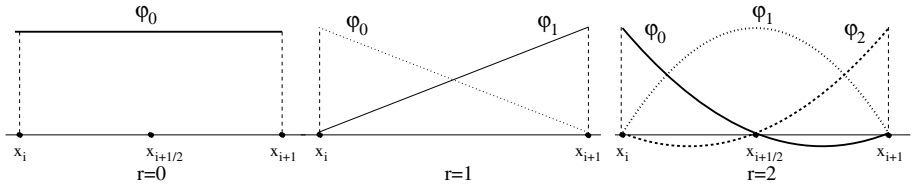


Fig. 13.7. The Lagrange bases for $r = 0$, $r = 1$ and $r = 2$

M_h being the mass matrix, L_h the matrix associated to the bilinear form and to the jump relation, $\mathbf{f}_h$ the known term:

$$(M_h)_{pq} = \int_{x_i}^{x_{i+1}} \varphi_p \varphi_q \, dx, \quad (L_h)_{pq} = \int_{x_i}^{x_{i+1}} (a\varphi_{q,x} + a_0 \varphi_q)\, \varphi_p \, dx + (a\varphi_q \varphi_p)(x_i),$$

$$(\mathbf{f}_h)_p = \int_{x_i}^{x_{i+1}} f\varphi_p \, dx + aU_h^-(x_i)\varphi_p(x_i), \quad p, q = 0, \ldots, r\,.$$

We have denoted by $\{\varphi_q,\ q = 0, \ldots, r\}$ a basis for $\mathbb{P}_r([x_i, x_{i+1}])$ and by $\mathbf{u}_h(t)$ the coefficients of $u_h(x,t)|_{[x_i,x_{i+1}]}$ in the development with respect to the basis $\{\varphi_q\}$. If we take the Lagrange basis we will have, for instance, the functions reported in Fig. 13.7 (for the $r = 0$, $r = 1$ and $r = 2$ cases) and the values of $\{\mathbf{u}_h(t)\}$ are the ones taken by $u_h(t)$ at nodes ($x_{i+1/2}$ for $r = 0$, x_i and x_{i+1} for $r = 1$, x_i, $x_{i+1/2}$ and x_{i+1} for $r = 2$). Note that all the previous functions are identically null outside the interval $[x_i, x_{i+1}]$. Also, in the case of discontinuous finite elements it is perfectly acceptable to use polynomials of degree $r = 0$, in which case the transport term $a\dfrac{\partial u_h}{\partial x}$ will provide a null contribution on each element.

Aiming at diagonalizing the mass matrix, it can be interesting to use as a basis for $\mathbb{P}_r([x_i, x_{i+1}])$ the Legendre polynomials $\varphi_q(x) = L_q(2(x - x_i)/h_i)$, $h_i = x_i - x_i$ and $\{L_q,\ q = 0, 1, \ldots\}$ being the orthogonal Legendre polynomials defined over the interval $[-1, 1]$, that we have introduced in Sec. 10.2.2. Indeed, in such a way we obtain $(M_h)_{pq} = \dfrac{h_i}{2p+1}\delta_{pq},\ p, q = 0, \ldots r$. Obviously, in this case the unknown values $\{\mathbf{u}_h(t)\}$ will no longer be interpretable as nodal values of $u_h(t)$, but rather as the Legendre coefficients of the expansion of $u_h(t)$ with respect to the new basis.

The diagonalization of the mass matrix turns out to be particularly interesting when we use explicit time advancing schemes (such as e.g. second- and third-order Runge-Kutta schemes, introduced in Chap. 14). In this case, indeed, we will have a fully explicit problem to solve on each small interval.

For illustrative purposes, we present below some numerical results obtained for problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial x} = 0, & x \in (-5, 5),\ t > 0, \\ u(-5, t) = 0, & t > 0, \end{cases} \tag{13.44}$$

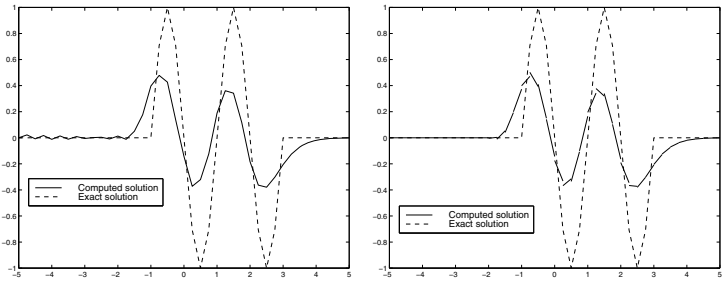


Fig. 13.8. Solution at time $t = 1$ of problem (13.44) with $\phi_k = \pi/2$, $h = 0.25$, obtained using continuous (left) and discontinuous (right) linear finite elements and backward Euler time discretization

using the initial condition

$$u(x,0) = \begin{cases} \sin(\pi x), & x \in (-2,2), \\ 0 & \text{otherwise.} \end{cases} \tag{13.45}$$

The problem has been discretized using linear finite elements in space, both continuous and discontinuous. For the temporal discretization, we have used the backward Euler scheme in both cases. We have chosen $h = 0.25$ and a time step $\Delta t = h$; for such value of h the phase number associated to the sinusoidal wave is $\phi_k = \pi/2$.

In Fig. 13.8 we report the numerical solution at time $t = 1$ together with the corresponding exact solution. We can note the strong numerical diffusion of the scheme that, however, denotes small oscillations in the posterior part in the case of continuous elements. Furthermore, we can observe that the numerical solution obtained using discontinuous finite elements, although being discontinuous, it no longer features an oscillatory behavior in the posterior part.

Let us now consider the following problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial x} = 0, & x \in (0,1),\ t > 0, \\ u(0,t) = 1, & t > 0, \\ u(x,0) = 0, & x \in [0,1], \end{cases} \tag{13.46}$$

which represents the transport of a discontinuity entering the domain. We have considered continuous linear finite elements, with both strong and weak treatment of the boundary conditions, as well as discontinuous linear finite elements. This time as well, we have used the backward Euler method for the temporal discretization. The grid-size is $h = 0.025$ and the time step is $\Delta t = h$.

The results at time $t = 0.5$ are represented in Fig. 13.9. We can note how the Dirichlet datum is well represented also by schemes with weak boundary treatment. To this end, for the case of continuous finite elements with weak boundary treatment, we have computed the behavior of $|u_h(0) - u(0)|$ for $t = 0.1$ for several values of h, Δt being constant. We can note a linear convergence to zero with respect to h.

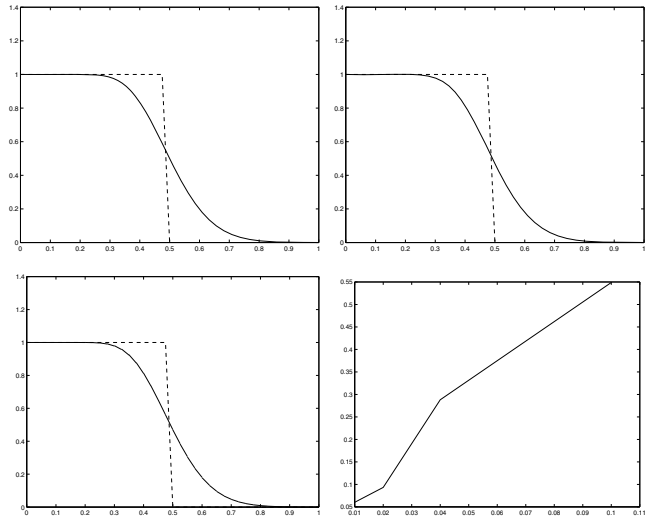


Fig. 13.9. Solution to problem (13.46) for $t = 0.5$ with $h = 0.025$ obtained using continuous linear finite elements and strong (top left) and weak (top right) treatment of the boundary Dirichlet condition, while discontinuous elements in space have been used in the bottom left case. Finally, we show in the bottom left the behavior of $|u_h(0) - u(0)|$ as a function of h for $t = 0.1$, in weak treatment of the Dirichlet condition

13.4.2 The multi-dimensional case

Let us now consider the case of the multi-dimensional case (13.22). Let W_h be the space of discontinuous piecewise polynomials of degree r on each element $K \in \mathcal{T}_h$, introduced in (13.34). The discontinuous Galerkin (DG) finite element semi-discretization of problem (13.22) becomes: for each $t > 0$ find $u_h(t) \in W_h$ such that

$$\begin{aligned} &\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h d\Omega + \sum_{K\in\mathcal{T}_h} \left[a_K(u_h(t), v_h) - \int_{\partial K^{in}} \mathbf{a}\cdot\mathbf{n}_K \, [u_h(t)] v_h^+ d\gamma \right] \\ &= \int_\Omega f(t) v_h d\Omega \qquad \forall v_h \in W_h, \end{aligned} \tag{13.47}$$

with $u_h(0) = u_{0,h}$, where $\mathbf{n}_K$ denotes the outward normal unit vector on ∂K, and

$$\partial K^{in} = \{\mathbf{x} \in \partial K : \ \mathbf{a}(\mathbf{x}) \cdot \mathbf{n}_K(\mathbf{x}) < 0\}.$$

The bilinear form a_K is defined in the following way

$$a_K(u, v) = \int_K \left(\mathbf{a}\cdot\nabla u \, v + a_0 u \, v\right) d\mathbf{x},$$

while

$$[u_h(\mathbf{x})] = \begin{cases} u_h^+(\mathbf{x}) - u_h^-(\mathbf{x}), & \mathbf{x} \notin \partial\Omega^{in}, \\ u_h^+(\mathbf{x}) - \varphi_h(\mathbf{x}), & \mathbf{x} \in \partial\Omega^{in}, \end{cases}$$

$\partial\Omega^{in}$ being the inflow boundary (13.23) and with

$$u_h^{\pm}(\mathbf{x}) = \lim_{s \to 0^{\pm}} u_h(\mathbf{x} + s\mathbf{a}), \quad \mathbf{x} \in \partial K.$$

For each $t > 0$, the stability estimate obtained for problem (13.47) is (thanks to the hypothesis (13.26))

$$\begin{aligned} \|u_h(t)\|^2_{\mathrm{L}^2(\Omega)} \quad &+ \quad \int_0^t \left(\mu_0 \|u_h(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \sum_{K \in \mathcal{T}_h} \int_{\partial K^{in}} |\mathbf{a} \cdot \mathbf{n}_K| \, [u_h(\tau)]^2 \right) d\tau \\ &\leq \quad C \left[\|u_{0,h}\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \left(\|f(\tau)\|^2_{\mathrm{L}^2(\Omega)} + |\varphi_h|^2_{\mathbf{a},\partial\Omega^{in}} \right) d\tau \right], \end{aligned}$$

having introduced, for each subset Γ of $\partial\Omega$ of positive measure, the seminorm

$$|v|_{\mathbf{a},\Gamma} = \left(\int_\Gamma |\mathbf{a} \cdot \mathbf{n}| v^2 \, d\gamma \right)^{1/2}.$$

Supposing for simplicity that $f = 0$, $\varphi = 0$, and that $u_0 \in H^{r+1}(\Omega)$, we can prove the following a priori error estimate

$$\begin{aligned} \max_{t \in [0,T]} \|u(t) - u_h(t)\|_{\mathrm{L}^2(\Omega)} \quad &+ \left(\int_0^T \sum_{K \in \mathcal{T}_h} \int_{\partial K^{in}} |\mathbf{a} \cdot \mathbf{n}_K| \, [u(t) - u_h(t)]^2 \, dt \right)^{\frac{1}{2}} \\ &\leq C h^{r+1/2} \|u_0\|_{\mathrm{H}^{r+1}(\Omega)}. \end{aligned} \tag{13.48}$$

For the proofs, we refer to [QV94, Chap. 14], [Joh87], and to the references cited thereby.

Other formulations are possible, based on different forms of stabilization. Let us consider a diffusion and reaction problem such as (13.22) but written in conservative form

$$\frac{\partial u}{\partial t} + \mathrm{div}(\mathbf{a}u) + a_0 u = f, \quad \mathbf{x} \in \Omega, \ t > 0. \tag{13.49}$$

Having now set

$$a_K(u_h, v_h) = \int_K \big(- u_h(\mathbf{a} \cdot \nabla v_h) + a_0 u_h \, v_h \big) \, d\mathbf{x},$$

we consider the following approximation: for each $t > 0$, find $u_h(t) \in W_h$ such that, $\forall v_h \in W_h$,

$$\begin{aligned}
&\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + \sum_{K \in \mathcal{T}_h} a_K(u_h(t), v_h) + \sum_{e \not\subset \partial\Omega^{in}} \int_e \{\mathbf{a}\, u_h(t)\} [\![v_h]\!] \, d\gamma \\
&+ \sum_{e \not\subset \partial\Omega} \int_e c_e(\gamma) \, [\![u_h(t)]\!] \, [\![v_h]\!] \, d\gamma \\
&= \int_\Omega f(t) v_h \, d\Omega - \sum_{e \subset \partial\Omega^{in}} \int_e (\mathbf{a} \cdot \mathbf{n}) \, \varphi(t) \, v_h \, d\gamma \, .
\end{aligned} \tag{13.50}$$

The notations are the following: we denote by e any side of the grid $\mathcal{T}_h$ shared by two triangles, say K_1 and K_2. For each scalar function ψ, piecewise regular on the mesh, with $\psi^i = \psi\big|_{K_i}$, we have defined its jump on e as follows:

$$[\![\psi]\!] = \psi^1 \mathbf{n}_1 + \psi^2 \mathbf{n}_2,$$

$\mathbf{n}_i$ being the outward normal unit vector to element K_i. Instead, if $\boldsymbol{\sigma}$ is a vector function, then its average on e is defined as

$$\{\boldsymbol{\sigma}\} = \frac{1}{2} (\boldsymbol{\sigma}^1 + \boldsymbol{\sigma}^2).$$

Note that the jump $[\![\psi]\!]$ through e of a scalar function ψ is a vector parallel to the normal to e.
These definitions do not depend on the ordering of the elements.
If e is a side belonging to the boundary $\partial\Omega$, then

$$[\![\psi]\!] = \psi \, \mathbf{n}, \qquad \{\boldsymbol{\sigma}\} = \boldsymbol{\sigma}.$$

Concerning $c_e(\gamma)$, this is a non-negative function which will typically be chosen to be constant on each side. Choosing, for instance, $c_e = |\mathbf{a} \cdot \mathbf{n}|/2$ on each internal side, $c_e = -\mathbf{a} \cdot \mathbf{n}/2$ on $\partial\Omega^{in}$, $c_e = \mathbf{a} \cdot \mathbf{n}/2$ on $\partial\Omega^{out}$, the formulation in (13.50) is reduced to the standard upwind formulation

$$\begin{aligned}
&\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + \sum_{K \in \mathcal{T}_h} a_K(u_h(t), v_h) + \sum_{e \not\subset \partial\Omega^{in}} \int_e \{\mathbf{a}\, u_h(t)\}_{up} [\![v_h]\!] \, d\gamma \\
&= \int_\Omega f(t) v_h \, d\Omega - \sum_{e \subset \partial\Omega^{in}} \int_e (\mathbf{a} \cdot \mathbf{n}) \, \varphi(t) \, v_h \, d\gamma \qquad \forall v_h \in W_h.
\end{aligned} \tag{13.51}$$

Here $\{\mathbf{a}\, u_h\}_{up}$ denotes the upwind value of $\mathbf{a}\, u_h$, that is coincides with $\mathbf{a}\, u_h^1$ if $\mathbf{a} \cdot \mathbf{n}_1 > 0$, with $\mathbf{a}\, u_h^2$ if $\mathbf{a} \cdot \mathbf{n}_1 < 0$, and finally with $\mathbf{a}\{u_h\}$ if $\mathbf{a} \cdot \mathbf{n}_1 = 0$. Finally, if $\mathbf{a}$ is constant (or divergence free), then $\mathrm{div}(\mathbf{a} u_h) = \mathbf{a} \cdot \nabla u_h$ and (13.51) coincides with (13.47). The formulation (13.50) is called discontinuous Galerkin method with *jump stabilization*. The latter is stable if $c_e \geq \theta_0 \, |\mathbf{a} \cdot \mathbf{n}_e|$ (for a suitable $\theta_0 > 0$) for each internal side e,

and also convergent with optimal order. Indeed, in the case of the stationary problem it can be proven that

$$\|u - u_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{e \in \mathcal{T}_h} \|\sqrt{c_e}\,[\![u - u_h]\!]\|^2_{\mathrm{L}^2(e)} \leq C\, h^{2r+1}\, \|u\|^2_{\mathrm{H}^{r+1}(\Omega)}.$$

For the proof and for other formulations with jump stabilization, including the case of advection-diffusion equations, we refer the reader to [BMS04].

13.5 Approximation using spectral methods

In this section, we will briefly discuss the approximation of hyperbolic problems with spectral methods. For simplicity, we will limit our discussion to one-dimensional problems. We will first treat the G-NI approximation in a single interval, then the SEM approximation corresponding to a decomposition in sub-intervals where we use discontinuous polynomials when we move from an interval to its neighbors. This provides a generalization of discontinuous finite elements, in the case where we consider polynomials of "high" degree on each element, and the integrals on each element are approximated using the GLL integration formula (10.18).

13.5.1 The G-NI method in a single interval

Let us consider the first-order hyperbolic transport-reaction problem (12.3) and let us suppose that $(\alpha, \beta) = (-1, 1)$. Then we approximate in space by a spectral collocation method, with strong imposition of the boundary conditions. Having denoted by $\{x_0 = -1, x_1, \ldots, x_N = 1\}$ the GLL nodes introduced in Sec. 10.2.3, the semi-discretized problem is:
for each $t > 0$, find $u_N(t) \in \mathbb{Q}_N$ (the space of polynomials (10.1)) such that

$$\begin{cases} \Big(\dfrac{\partial u_N}{\partial t} + a\,\dfrac{\partial u_N}{\partial x} + a_0\,u_N\Big)(x_j,t) = f(x_j,t), \quad j = 1,\ldots,N, \\ u_N(-1,t) = \varphi(t), \\ u_N(x_j,0) = u_0(x_j), \qquad\qquad j = 0,\ldots,N. \end{cases} \tag{13.52}$$

Suitably using the discrete GLL scalar product defined in (10.25), the G-NI approximation of problem (13.52) becomes: for each $t > 0$, find $u_N(t) \in \mathbb{Q}_N$ such that

$$\begin{cases} \Big(\dfrac{\partial u_N(t)}{\partial t}, v_N\Big)_N + \Big(a\,\dfrac{\partial u_N(t)}{\partial x}, v_N\Big)_N + \big(a_0\,u_N(t), v_N\big)_N = \big(f(t), v_N\big)_N \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall v_N \in \mathbb{Q}_N^-, \\ u_N(-1,t) = \varphi(t), \\ u_N(x,0) = u_{0,N}, \end{cases} \tag{13.53}$$

where $u_{0,N} \in \mathbb{Q}_N$ is a suitable approximation of u_0, and having set $\mathbb{Q}_N^- = \{v_N \in \mathbb{Q}_N : v_N(-1) = 0\}$. At the inflow, the solution u_N then satisfies the imposed condition at each time $t > 0$, while the test functions vanish.
In fact, the solutions of problems (13.52) and (13.53) coincide if $u_{0,N}$ in (13.53) is chosen as the interpolated $\Pi_N^{GLL} u_0$. To prove this, it is sufficient to choose in (13.53) v_N coinciding with the characteristic polynomial ψ_j (defined in (10.12), (10.13)) associated to the GLL node x_j, for each $j = 1, \ldots, N$.

Let us now derive a stability estimate for the formulation (13.53) in the norm (10.53) induced from the discrete scalar product (10.25). For simplicity, we choose a homogeneous inflow datum, that is $\varphi(t) = 0$, for each t, and a and a_0 constant. Having chosen, for each $t > 0$, $v_N = u_N(t)$, we obtain

$$\frac{1}{2}\frac{\partial}{\partial t}\|u_N(t)\|_N^2 + \frac{a}{2}\int_{-1}^{1}\frac{\partial u_N^2(t)}{\partial x}\,dx + a_0\|u_N(t)\|_N^2 = \big(f(t), u_N(t)\big)_N.$$

Suitably rewriting the convective term, integrating with respect to time and using the Young inequality, we have

$$\begin{aligned}
\|u_N(t)\|_N^2 &+ a\int_0^t \big(u_N(1,\tau)\big)^2\,d\tau + 2\,a_0\int_0^t \|u_N(\tau)\|_N^2\,d\tau \\
&= \|u_{0,N}\|_N^2 + 2\int_0^t \big(f(\tau), u_N(\tau)\big)_N\,d\tau \\
&\le \|u_{0,N}\|_N^2 + a_0\int_0^t \|u_N(\tau)\|_N^2\,d\tau + \frac{1}{a_0}\int_0^t \|f(\tau)\|_N^2\,d\tau,
\end{aligned}$$

that is

$$\begin{aligned}
\|u_N(t)\|_N^2 &+ a\int_0^t \big(u_N(1,\tau)\big)^2\,d\tau + a_0\int_0^t \|u_N(\tau)\|_N^2\,d\tau \\
&\le \|u_{0,N}\|_N^2 + \frac{1}{a_0}\int_0^t \|f(\tau)\|_N^2\,d\tau.
\end{aligned} \tag{13.54}$$

The norm of the initial data can be bounded as follows

$$\|u_{0,N}\|_N^2 \le \|u_{0,N}\|_{\mathrm{L}^\infty(-1,1)}^2 \left(\sum_{i=0}^{N}\alpha_i\right) = 2\,\|u_{0,N}\|_{\mathrm{L}^\infty(-1,1)}^2,$$

and a similar bound holds for $\|f(\tau)\|_N^2$ provided that f be a continuous function. Hence, reverting to (13.54) and using inequality (10.54) to bound the norms of the

left-hand side, we deduce

$$\begin{aligned}\|u_N(t)\|^2_{\mathrm{L}^2(-1,1)} \;+\; & a\int_0^t \big(u_N(1,\tau)\big)^2\,d\tau + a_0\int_0^t \|u_N(\tau)\|^2_{\mathrm{L}^2(-1,1)}\,d\tau \\ \leq\; & 2\,\|u_{0,N}\|^2_{\mathrm{L}^\infty(-1,1)} + \frac{2}{a_0}\int_0^t \|f(\tau)\|^2_{\mathrm{L}^2(-1,1)}\,d\tau.\end{aligned}$$

The reinterpretation of the G-NI method as a collocation method is less immediate in the case where the convective term a is not constant and we start from a conservative formulation of the differential equation in (13.52), that is the second term on the left-hand side is replaced by $\partial(au)/\partial x$. In such case, we can show again that the G-NI approximation is equivalent to the collocation approximation where the convective term is replaced by $\partial\big(\Pi_N^{GLL}(au_N)\big)/\partial x$, i.e. by the interpolation derivative (10.40).

Also in the case of a G-NI approximation, we can resort to a weak imposition of the boundary conditions. Such approach is more flexible than the one considered above and more suitable for the generalization to multi-dimensional problems or systems of equations. As we have seen in the previous section, the starting point for a weak imposition of boundary conditions is a suitable integration by parts of the transport terms. Referring to the one-dimensional problem (13.52), we have (if a is constant)

$$\begin{aligned}\int_{-1}^{1} a\,\frac{\partial u(t)}{\partial x}\,v\,dx \;&= -\int_{-1}^{1} a\,u(t)\frac{\partial v}{\partial x}\,dx + \big[a\,u(t)\,v\big]_{-1}^{1} \\ &= -\int_{-1}^{1} a\,u(t)\frac{\partial v}{\partial x}\,dx + a\,u(1,t)\,v(1) - a\,\varphi(t)\,v(-1).\end{aligned}$$

Thanks to the above identity, we can immediately formulate the G-NI approximation to problem (13.52) with a weak treatment of boundary conditions:
for each $t>0$, find $u_N(t)\in\mathbb{Q}_N$ such that

$$\begin{aligned}&\Big(\frac{\partial u_N(t)}{\partial t},v_N\Big)_N - \Big(a\,u_N(t),\frac{\partial v_N}{\partial x}\Big)_N + \big(a_0\,u_N(t),v_N\big)_N \\ &+ a\,u_N(1,t)\,v_N(1) = \big(f(t),v_N\big)_N + a\,\varphi(t)\,v_N(-1)\;\;\forall v_N\in\mathbb{Q}_N,\end{aligned} \tag{13.55}$$

with $u_N(x,0)=u_{0,N}(x)$. We note that both the solution u_N and the test function v_N are free at the boundary.
An equivalent formulation to (13.55) is obtained by suitably counter-integrating the convective term by parts:
for each $t>0$, find $u_N(t)\in\mathbb{Q}_N$ such that

$$\begin{aligned}&\Big(\frac{\partial u_N(t)}{\partial t},v_N\Big)_N + \Big(a\,\frac{\partial u_N(t)}{\partial x},v_N\Big)_N + \big(a_0\,u_N(t),v_N\big)_N \\ &+ a\,\big(u_N(-1,t)\,-\varphi(t)\big)v_N(-1) = \big(f,v_N\big)_N\;\;\forall v_N\in\mathbb{Q}_N.\end{aligned} \tag{13.56}$$

It is now possible to reinterpret such weak formulation as a suitable collocation method. To this end, it is sufficient to choose in (13.56) the test function v_N coinciding with the characteristic polynomials (10.12), (10.13) associated to the GLL nodes. Considering first the internal and outflow nodes, and choosing therefore $v_N = \psi_i$, with $i = 1, \ldots, N$, we have

$$\Big(\frac{\partial u_N}{\partial t} + a\,\frac{\partial u_N}{\partial x} + a_0\,u_N\Big)(x_i,t) = f(x_i,t), \tag{13.57}$$

having previously simplified the weight α_i common to all the terms of the equation. On the other hand, by choosing $v_N = \psi_0$ we obtain the following relation at the inflow node

$$\begin{aligned}\Big(\frac{\partial u_N}{\partial t} &+ a\,\frac{\partial u_N}{\partial x} + a_0\,u_N\Big)(-1,t)\\ &+ \frac{1}{\alpha_0}\,a\,\big(u_N(-1,t) - \varphi(t)\big) = f(-1,t),\end{aligned} \tag{13.58}$$

$\alpha_0 = 2/(N^2+N)$ being the GLL weight associated to node $x_0 = -1$. From equations (13.57) and (13.58) it then follows that a reformulation in terms of collocation is possible at all the GLL nodes except for the inflow node, for which we find the relation

$$a\,\big(u_N(-1,t) - \varphi(t)\big) = \alpha_0\,\Big(f - \frac{\partial u_N}{\partial t} - a\,\frac{\partial u_N}{\partial x} - a_0\,u_N\Big)(-1,t). \tag{13.59}$$

The latter can be interpreted as the fulfillment of the boundary condition of the differential problem (13.52) up to the residue associated to the u_N approximation. Such condition is therefore satisfied exactly only at the limit, for $N \longrightarrow \infty$ (i.e. in a natural way).

In accordance with what we previously noted, the formulation (13.56) would be complicated for instance in case of a non-constant convective field a. Indeed,

$$-\Big(a\,u_N(t), \frac{\partial v_N}{\partial x}\Big)_N = \Big(a\,\frac{\partial u_N(t)}{\partial x}, v_N\Big)_N - a\,u_N(1,t)\,v_N(1) + a\,\varphi(t)\,v_N(-1),$$

would not be true as, in this case, the product $a\,u_N(t)\,\dfrac{\partial v_N}{\partial x}$ no longer identifies a polynomial of degree $2N-1$, so the exactness of the numerical integration formula would not hold in this case. It is therefore necessary to apply the interpolation operator Π_N^{GLL}, introduced in Sec. 10.2.3, before counter-integrating by parts, yielding

$$\begin{aligned} - &\Big(a\,u_N(t), \frac{\partial v_N}{\partial x}\Big)_N = -\Big(\Pi_N^{GLL}\big(a\,u_N(t)\big), \frac{\partial v_N}{\partial x}\Big)_N\\ = &-\Big(\Pi_N^{GLL}\big(a\,u_N(t)\big), \frac{\partial v_N}{\partial x}\Big)\\ = &\Big(\frac{\partial \Pi_N^{GLL}\big(a\,u_N(t)\big)}{\partial x}, v_N\Big) - \big[\big(a\,u_N(t)\big)\,v_N\big]_{-1}^{1}.\end{aligned}$$

In this case, the formulation (13.56) then becomes:
for each $t > 0$, find $u_N(t) \in \mathbb{Q}_N$ such that

$$\Big(\frac{\partial u_N(t)}{\partial t}, v_N\Big)_N + \Big(\frac{\partial \Pi_N^{GLL}\big(a\, u_N(t)\big)}{\partial x}, v_N\Big) + \big(a_0\, u_N(t), v_N\big)_N + a(t)\,\big(u_N(-1,t) - \varphi(t)\big) v_N(-1) = \big(f(t), v_N\big)_N \quad \forall v_N \in \mathbb{Q}_N, \tag{13.60}$$

with $u_N(x,0) = u_{0,N}(x)$. Also the collocation reinterpretation of formulation (13.56), represented by relations (13.57) and (13.59), will need to be modified with the introduction of the interpolation operator Π_N^{GLL} (that is by replacing the exact derivative with the interpolation derivative). Precisely, we obtain

$$\Big(\frac{\partial u_N}{\partial t} + \frac{\partial \Pi_N^{GLL}\big(a\, u_N\big)}{\partial x} + a_0\, u_N\Big)(x_i, t) = f(x_i, t),$$

for $i = 1, \ldots, N$, and

$$a(-1)\,\big(u_N(-1,t) - \varphi(t)\big) = \alpha_0 \Big(f - \frac{\partial u_N}{\partial t} - \frac{\partial \Pi_N^{GLL}\big(a\, u_N\big)}{\partial x} - a_0\, u_N\Big)(-1, t),$$

at the inflow node $x = -1$.

13.5.2 The DG-SEM-NI method

As anticipated, we will introduce in this section an approximation based on a partition in sub-intervals, in each of which the G-NI method is used. Moreover, the solution will be discontinuous between an interval and its neighbors. This explains the DG (*discontinuous Galerkin*), SEM (*spectral element method*), NI (*numerical integration*) acronym.

Let us reconsider problem (13.52) on the generic interval (α, β). On the latter, we introduce a partition in M subintervals $\Omega_m = (\overline{x}_{m-1}, \overline{x}_m)$ with $m = 1, \ldots, M$. Let

$$W_{N,M} = \{v \in \mathrm{L}^2(\alpha, \beta) \,:\, v\big|_{\Omega_m} \in \mathbb{Q}_N,\ \forall m = 1, \ldots, M\}$$

be the space of piecewise polynomials of degree $N(\geq 1)$ on each sub-interval. We observe that the continuity is not necessarily guaranteed in correspondence of the points $\{\overline{x}_i\}$. Thus, we can formulate the following approximation of problem (13.52):
for each $t > 0$, find $u_{N,M}(t) \in W_{N,M}$ such that

$$\sum_{m=1}^{M} \Big[\Big(\frac{\partial u_{N,M}}{\partial t}, v_{N,M}\Big)_{N,\Omega_m} + \Big(a\,\frac{\partial u_{N,M}}{\partial x}, v_{N,M}\Big)_{N,\Omega_m} + \big(a_0\, u_{N,M}, v_{N,M}\big)_{N,\Omega_m} + a(\overline{x}_{m-1})\,\big(u^+_{N,M} - U^-_{N,M}\big)(\overline{x}_{m-1})\, v^+_{N,M}(\overline{x}_{m-1})\Big] = \sum_{m=1}^{M} \big(f, v_{N,M}\big)_{N,\Omega_m} \tag{13.61}$$

for all $v_{N,M} \in W_{N,M}$, with

$$U^-_{N,M}(\overline{x}_i) = \begin{cases} u^-_{N,M}(\overline{x}_i), & i = 1, \ldots, M-1, \\ \varphi(\overline{x}_0), & \text{for } i = 0, \end{cases} \tag{13.62}$$

and where $(\cdot,\cdot)_{N,\Omega_m}$ denotes the approximation via the GLL formula (10.25) of the scalar product L^2 restrained to the element Ω_m. To simplify the notations we have omitted to explicitly indicate the dependence on t of $u_{N,M}$ and f. Given the discontinuous nature of the test functions, we can reformulate equation (13.61) on each of the M sub-intervals, by choosing the test function $v_{N,M}$ so that $v_{N,M}\big|_{[\alpha,\beta]\setminus\Omega_m} = 0$. Proceeding this way, we obtain

$$\begin{aligned} &\Big(\frac{\partial u_{N,M}}{\partial t}, v_{N,M}\Big)_{N,\Omega_m} + \Big(a\,\frac{\partial u_{N,M}}{\partial x}, v_{N,M}\Big)_{N,\Omega_m} + \big(a_0\,u_{N,M}, v_{N,M}\big)_{N,\Omega_m} \\ &\qquad + a(\overline{x}_{m-1})\,\big(u^+_{N,M} - U^-_{N,M}\big)(\overline{x}_{m-1})\,v^+_{N,M}(\overline{x}_{m-1}) = \big(f, v_{N,M}\big)_{N,\Omega_m}, \end{aligned}$$

for each $m = 1, \ldots, M$. We note that, for $m = 1$, the term

$$a(\overline{x}_0)\,\big(u^+_{N,M} - \varphi\big)(\overline{x}_0)\,v^+_{N,M}(\overline{x}_0)$$

can be regarded as the imposition in weak form of the inflow boundary condition. On the other hand for $m = 2, \ldots, M$, the term

$$a(\overline{x}_{m-1})\,\big(u^+_{N,M} - U^-_{N,M}\big)(\overline{x}_{m-1})\,v^+_{N,M}(\overline{x}_{m-1}),$$

can be interpreted as a penalization term that provides a weak imposition of the continuity of the solution $u_{N,M}$ at the extrema $\overline{x}_i$, $i = 1, \ldots, M-1$.

We now want to interpret the formulation (13.61) as a suitable collocation method. To this end, we introduce on each sub-interval Ω_m, the $N+1$ GLL nodes $x^{(m)}_j$, with $j = 0, \ldots, N$, and we denote by $\alpha^{(m)}_j$ the corresponding weights (see (10.71)). We now identify the test function $v_{N,M}$ in (13.61) as the characteristic Lagrangian polynomial $\psi^{(m)}_j \in \mathbb{P}^N(\Omega_m)$ associated to node $x^{(m)}_j$ and extended by zero outside the domain Ω_m. Given the presence of the jump term, we will have a non-univocal rewriting for equation (13.61). We start by considering the characteristic polynomials associated to the nodes $x^{(m)}_j$, with $j = 1, \ldots, N-1$, and $m = 1, \ldots, M$. In this case we will have no contribution of the penalization term, yielding

$$\Big[\frac{\partial u_{N,M}}{\partial t} + a\,\frac{\partial u_{N,M}}{\partial x} + a_0\,u_{N,M}\Big](x^{(m)}_j) = f(x^{(m)}_j).$$

For this choice of nodes we thus find exactly the collocation of the differential problem (13.52).

Instead, in the case where function $\psi^{(m)}_j$ is associated to a node of the partition $\{\overline{x}_i\}$, that is $j = 0$, with $m = 1, \ldots, M$ we have

$$\begin{aligned} &\alpha^{(m)}_0\,\Big[\frac{\partial u_{N,M}}{\partial t} + a\,\frac{\partial u_{N,M}}{\partial x} + a_0\,u_{N,M}\Big](x^{(m)}_0) \\ &\qquad + a(x^{(m)}_0)\,\big(u^+_{N,M} - U^-_{N,M}\big)(x^{(m)}_0) = \alpha^{(m)}_0\,f(x^{(m)}_0), \end{aligned} \tag{13.63}$$

recalling that $U^-_{N,M}(x_0^{(1)}) = \varphi(\overline{x}_0)$. We have implicitly adopted the convention that the sub-interval Ω_m should not include $\overline{x}_m$, as the discontinuous nature of the adopted method would take us to processing twice each node $\overline{x}_i$, with $i = 1, \ldots, M-1$. Equation (13.63) can be rewritten as

$$\Big[\frac{\partial u_{N,M}}{\partial t} + a\,\frac{\partial u_{N,M}}{\partial x} + a_0\,u_{N,M} - f\Big](x_0^{(m)}) = -\frac{a(x_0^{(m)})}{\alpha_0^{(m)}}\,\big(u^+_{N,M} - U^-_{N,M}\big)(x_0^{(m)}).$$

We observe that while the left-hand side represents the residue of the equation at node $x_0^{(m)}$, the right-hand side one is, up to a multiplicative factor, the residue of the weak imposition of the continuity of $u_{N,M}$ in $x_0^{(m)}$.

13.6 Numerical treatment of boundary conditions for hyperbolic systems

We have seen different strategies to impose the inflow boundary conditions for the scalar transport equation. When considering hyperbolic systems, the numerical treatment of boundary conditions requires more attention. We will illustrate this issue on a linear system with constant coefficients in one dimension,

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + A\,\dfrac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & -1 < x < 1,\ \ t > 0, \\ \mathbf{u}(x,0) = \mathbf{u}_0(x), & -1 < x < 1, \end{cases} \tag{13.64}$$

completed with suitable boundary conditions. Following [CHQZ07], we choose the case of a system made of two hyperbolic equations, identifying in (13.64) $\mathbf{u}$ with the vector $(u, v)^T$ and A with the matrix

$$A = \begin{bmatrix} -1/2 & -1 \\ -1 & -1/2 \end{bmatrix},$$

whose eigenvalues are $-3/2$ and $1/2$. We make the choice

$$u(x,0) = \sin(2x) + \cos(2x), \quad v(x,0) = \sin(2x) - \cos(2x)$$

for the initial conditions and

$$\begin{aligned} u(-1,t) &= \sin(-2+3t) + \cos(-2-t) = \varphi(t), \\ v(1,t) &= \sin(2+3t) + \cos(2-t) = \psi(t) \end{aligned} \tag{13.65}$$

for the boundary conditions.
Let us now consider the (right) eigenvector matrix

$$W = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix},$$

whose inverse is

$$W^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Exploiting the diagonalization

$$\Lambda = W^{-1}AW = \begin{bmatrix} -3/2 & 0 \\ 0 & 1/2 \end{bmatrix},$$

we can rewrite the differential equation in (13.64) in terms of the characteristic variables

$$\mathbf{z} = W^{-1}\mathbf{u} = \begin{bmatrix} u+v \\ u-v \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \tag{13.66}$$

as

$$\frac{\partial \mathbf{z}}{\partial t} + \Lambda \frac{\partial \mathbf{z}}{\partial x} = \mathbf{0}. \tag{13.67}$$

The characteristic variable z_1 propagates toward the left at rate $3/2$, while z_2 propagates toward the right at rate $1/2$.

This suggests to assign a condition for z_1 at $x = 1$ and one for z_2 at $x = -1$. The boundary values of z_1 and z_2 can be generated by using the boundary conditions for u and v as follows. From relation (13.66), we have

$$\mathbf{u} = W\mathbf{z} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1/2\,(z_1+z_2) \\ 1/2\,(z_1-z_2) \end{bmatrix},$$

that is, exploiting the boundary values (13.65) assigned for u and v,

$$\frac{1}{2}(z_1+z_2)(-1,t) = \varphi(t), \quad \frac{1}{2}(z_1-z_2)(1,t) = \psi(t). \tag{13.68}$$

The conclusion is that, in spite of the diagonal structure of system (13.67), the characteristic variables are in fact coupled by the boundary conditions (13.68).

We are therefore confronted with the problem of how to handle, from a numerical viewpoint, the boundary conditions for systems like (13.64). Indeed, difficulties can arise even from the discretization of the corresponding scalar problem (for a constant > 0)

$$\begin{cases} \dfrac{\partial z}{\partial t} + a\dfrac{\partial z}{\partial x} = 0, & -1 < x < 1, \ \ t > 0, \\ z(-1,t) = \phi(t), & t > 0, \\ z(x,0) = z_0(x), & -1 < x < 1, \end{cases} \tag{13.69}$$

if we do not use an appropriate discretization scheme. We will illustrate the procedure for a spectral approximation method. As a matter of fact, a correct treatment of the boundary conditions for high order methods is even more vital than for a finite element or finite difference method, as with spectral methods boundary errors would be propagated inwards with an infinite rate.

Having introduced the partition $x_0 = -1 < x_1 < \ldots < x_{N-1} < x_N = 1$ of the interval $[-1, 1]$, if we decide to use, e.g., a finite difference scheme, we encounter problems essentially in getting the value of z at the outflow node x_N, unless we use the first order upwind scheme. As a matter of fact, higher order FD schemes such as the centered finite difference scheme would not be able to provide us with such an approximation unless we introduce additional nodes outside the definition interval $(-1, 1)$.
In contrast, a spectral discretization does not involve any boundary problem. For instance, the collocation scheme corresponding to problem (13.69) can be written as follows:
$\forall n \geq 0$, find $z_N^n \in \mathbb{Q}_N$ such that

$$\begin{cases} \dfrac{z_N^{n+1}(x_i) - z_N^n(x_i)}{\Delta t} + a\, \dfrac{\partial z_N^n}{\partial x}(x_i) = 0, \qquad i = 1, \ldots, N, \\ z_N^{n+1}(x_0) = \phi(t^{n+1}). \end{cases}$$

One equation results to be associated to each node, be it an internal or a boundary one, and the outflow node is treated as any other internal node. However, when moving to system (13.64), two unknowns and two equations are associated at each internal node x_i, with $i = 1, \ldots, N-1$, while at the boundary nodes x_0 and x_N we still have two unknowns but a single equation. Thus, we will need additional conditions for these points: in general, at the extremum $x = -1$ we will need as many conditions as the positive eigenvalues, while for $x = 1$ we will need to provide as many additional conditions as the negative eigenvalues.
Let us look for a solution to this problem by drawing inspiration from the spectral Galerkin method. Let us suppose we apply a collocation method to system (13.64); then, we want to find $\mathbf{u}_N = (u_{N,1}, u_{N,2})^T \in (\mathbb{Q}_N)^2$ such that

$$\frac{\partial \mathbf{u}_N}{\partial t}(x_i) + A\, \frac{\partial \mathbf{u}_N}{\partial x}(x_i) = \mathbf{0}, \quad i = 1, \ldots, N-1, \tag{13.70}$$

and with

$$u_{N,1}(x_0, t) = \varphi(t), \quad u_{N,2}(x_N, t) = \psi(t). \tag{13.71}$$

The simplest idea to obtain the two missing equations for $u_{N,1}$ and $u_{N,2}$ at x_N resp. x_0, is to exploit the vector equation (13.70) together with the known vectors $\varphi(t)$ and $\psi(t)$ in (13.71). The solution computed this way results however to be strongly unstable.
To seek an alternative approach, the idea is to add to the $2(N-1)$ collocation relations (13.70) and to the "physical" boundary conditions (13.71), the equations of the outgoing characteristics at points x_0 and x_N. More in detail, the characteristic outgoing from the domain at point $x_0 = -1$ is the one associated to the negative eigenvalue of matrix A, and has equation

$$\frac{\partial z_1}{\partial t}(x_0) - \frac{3}{2}\, \frac{\partial z_1}{\partial x}(x_0) = 0, \tag{13.72}$$

while the one associated with the point $x_N = 1$ is highlighted by the positive eigenvalue $1/2$ and is given by

$$\frac{\partial z_2}{\partial t}(x_N) + \frac{1}{2}\frac{\partial z_2}{\partial x}(x_N) = 0. \tag{13.73}$$

The choice of the outgoing characteristic is motivated by the fact that the latter carries information from the inside of the domain to the corresponding outflow point, where it makes sense to impose the differential equation.
Equations (13.72) and (13.73) allow us to have a closed system of $2N + 2$ equations in the $2N + 2$ unknowns $u_{N,1}(x_i, t) = u_N(x_i, t)$, $u_{N,2}(x_i, t) = v_N(x_i, t)$, with $i = 0, \ldots, N$.
For completeness, we can rewrite the characteristic equations (13.72) and (13.73) in terms of the unknowns u_N and v_N, as

$$\frac{\partial(u_N + v_N)}{\partial t}(x_0) - \frac{3}{2}\frac{\partial(u_N + v_N)}{\partial x}(x_0) = 0,$$

and

$$\frac{\partial(u_N - v_N)}{\partial t}(x_N) + \frac{1}{2}\frac{\partial(u_N - v_N)}{\partial x}(x_N) = 0,$$

respectively, or in matrix terms as

$$\begin{aligned} \begin{bmatrix} W_{11}^{-1} & W_{12}^{-1} \end{bmatrix} \left[\frac{\partial \mathbf{u}_N}{\partial t}(x_0) + A\frac{\partial \mathbf{u}_N}{\partial x}(x_0) \right] = 0, \\ \begin{bmatrix} W_{21}^{-1} & W_{22}^{-1} \end{bmatrix} \left[\frac{\partial \mathbf{u}_N}{\partial t}(x_N) + A\frac{\partial \mathbf{u}_N}{\partial x}(x_N) \right] = 0. \end{aligned} \tag{13.74}$$

Such additional equations are called *compatibility* equations: they represent a linear combination of the differential equations of the problem at the boundary points with coefficients given by the components of matrix W^{-1}.

Remark 13.4 Due to their global nature, spectral methods (either collocation, Galerkin, or G-NI) propagate immediately and on the whole domain every possible numerical perturbation introduced at the boundary. As such, spectral methods represent a good testbed for investigating the suitability of numerical strategies for the boundary treatment of hyperbolic systems. •

13.6.1 Weak treatment of boundary conditions

We now want to generalize the approach based on compatibility equations moving from pointwise relations, such as (13.74), to integral relations, that are more suitable for numerical approximations such as, e.g., finite elements or G-NI.
Let us again consider the constant coefficient system (13.64) and the notations used in Sec. 13.6. Let A be a real, symmetric and non-singular matrix of order d, Λ the diagonal matrix whose diagonal entries are the real eigenvalues of A, and W the square matrix whose columns are the (right) eigenvectors of A. Let us suppose that W is

orthogonal, which guarantees that $\Lambda = W^T A W$. The characteristic variables, defined as $\mathbf{z} = W^T \mathbf{u}$, satisfy the diagonal system (13.67). We introduce the splitting $\Lambda = \text{diag}(\Lambda^+, \Lambda^-)$ of the eigenvalue matrix, respectively grouping the positive eigenvalues (Λ^+) and the negative ones (Λ^-). Both such sub-matrices result to be diagonal, Λ^+ positive definite of order p, Λ^- negative definite of order $n = d - p$.
Analogously, we can rewrite $\mathbf{z}$ as $\mathbf{z} = (\mathbf{z}^+, \mathbf{z}^-)^T$, having denoted by $\mathbf{z}^+$ ($\mathbf{z}^-$, respectively) the characteristic variables that are constant along the characteristic lines with positive (respectively, negative) slope, that is that move towards the right (respectively, left) on the (x, t) reference frame. In correspondence of the right extremum $x = 1$, $\mathbf{z}^+$ is associated to the outgoing characteristic variables while $\mathbf{z}^-$ is associated to the incoming ones. Clearly, the roles are switched at the left boundary point $x = -1$.
A simple case occurs when, as boundary conditions, we assign the values of the incoming characteristics at both the domain extrema, that is p conditions at $x = -1$ and n conditions at $x = 1$. In this case, (13.67) represents a full-fledged decoupled system. Much more frequently however, the values of suitable linear combinations of the physical variables are assigned at both boundary points. Re-reading them in terms of the $\mathbf{z}$ variables, these yield linear combinations of the characteristic variables. None of the outgoing characteristics will in principle be determined by these combinations as the resulting values will generally be incompatible with the ones propagated inwards from the hyperbolic system. In contrast, the boundary conditions should allow to determine the incoming characteristic variables as a function of the outgoing ones and of the problem data.
For the sake of illustration, let us consider the following boundary conditions

$$B_L \mathbf{u}(-1, t) = \mathbf{g}_L(t), \quad B_R \mathbf{u}(1, t) = \mathbf{g}_R(t), \quad t > 0, \tag{13.75}$$

where $\mathbf{g}_L$ and $\mathbf{g}_R$ are assigned vectors and B_L, B_R are suitable matrices. At the left extremum, $x = -1$ we have p incoming characteristics, then B_L will have dimension $p \times d$. Setting $C_L = B_L W$ and using the splitting $\mathbf{z} = (\mathbf{z}^+, \mathbf{z}^-)^T$ introduced for $\mathbf{z}$ and the corresponding splitting $W = (W^+, W^-)^T$ for the eigenvector matrix, we have

$$C_L \mathbf{z}(-1, t) = C_L^+ \mathbf{z}^+(-1, t) + C_L^- \mathbf{z}^-(-1, t) = \mathbf{g}_L(t),$$

where $C_L^+ = B_L W^+$ is a $p \times p$ matrix, while $C_L^- = B_L W^-$ has dimension $p \times n$. We formulate the requirement that matrix C_L^+ be non-singular. Then the incoming characteristic at the $x = -1$ extremum can be obtained by

$$\mathbf{z}^+(-1, t) = S_L \mathbf{z}^-(-1, t) + \mathbf{z}_L(t), \tag{13.76}$$

$S_L = -(C_L^+)^{-1} C_L^-$ being a $p \times n$ matrix and $\mathbf{z}_L(t) = (C_L^+)^{-1} \mathbf{g}_L(t)$. In a similar way, we can assign at the right extremum $x = 1$ the incoming characteristic variable as

$$\mathbf{z}^-(1, t) = S_R \mathbf{z}^+(1, t) + \mathbf{z}_R(t), \tag{13.77}$$

S_R being a $n \times p$ matrix.
Matrices S_L and S_R are called *reflection matrices*.
The hyperbolic system (13.64) will thus be completed by the boundary conditions

(13.75) or, equivalently, by conditions (13.76)-(13.77).

Let us see which advantages can be brought by such a choice for boundary conditions. We start from the weak formulation of problem (13.64), integrating by parts the term containing the space derivative

$$\int_{-1}^{1} \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial t}\, dx - \int_{-1}^{1} \Big(\frac{\partial \mathbf{v}}{\partial x}\Big)^T A\mathbf{u}\, dx + \big[\mathbf{v}^T A\mathbf{u}\big]_{-1}^{1} = 0,$$

for each $t > 0$, $\mathbf{v}$ being an arbitrary, differentiable test function. We want to rewrite the boundary term $\big[\mathbf{v}^T A\mathbf{u}\big]_{-1}^{1}$ by exploiting the boundary equations (13.76) - (13.77). Introducing the characteristic variable $W^T\mathbf{v} = \mathbf{y} = (\mathbf{y}^+, \mathbf{y}^-)^T$ associated to the test function $\mathbf{v}$, we will have

$$\mathbf{v}^T A\mathbf{u} = \mathbf{y}^T \Lambda \mathbf{z} = (\mathbf{y}^+)^T \Lambda^+ \mathbf{z}^+ + (\mathbf{y}^-)^T \Lambda^- \mathbf{z}^-.$$

Using the relations (13.76)-(13.77), it then follows that

$$\begin{aligned}
&\int_{-1}^{1} \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial t}\, dx - \int_{-1}^{1} \Big(\frac{\partial \mathbf{v}}{\partial x}\Big)^T A\mathbf{u}\, dx \\
- \quad & (\mathbf{y}^+)^T(-1,t)\Lambda^+ S_L \mathbf{z}^-(-1,t) - (\mathbf{y}^-)^T(-1,t)\Lambda^- \mathbf{z}^-(-1,t) \\
+ \quad & (\mathbf{y}^+)^T(1,t)\Lambda^+ \mathbf{z}^+(1,t) + (\mathbf{y}^-)^T(1,t)\Lambda^- S_R\, \mathbf{z}^+(1,t) \\
= \quad & (\mathbf{y}^+)^T(-1,t)\Lambda^+ \mathbf{z}_L(t) - (\mathbf{y}^-)^T(1,t)\Lambda^- \mathbf{z}_R(t).
\end{aligned} \tag{13.78}$$

We observe that the boundary conditions (13.76)-(13.77) are naturally incorporated in the right-hand side of the system. Moreover, integrating again by parts, it is possible to obtain an equivalent formulation to (13.78) where the boundary conditions are imposed in a weak way

$$\begin{aligned}
&\int_{-1}^{1} \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial t}\, dx + \int_{-1}^{1} \mathbf{v}^T A \frac{\partial \mathbf{u}}{\partial x}\, dx \\
+ \quad & (\mathbf{y}^+)^T(-1,t)\Lambda^+ \big(\mathbf{z}^+(-1,t) - S_L \mathbf{z}^-(-1,t)\big) \\
- \quad & (\mathbf{y}^-)^T(1,t)\Lambda^- \big(\mathbf{z}^-(1,t) - S_R \mathbf{z}^+(1,t)\big) \\
= \quad & (\mathbf{y}^+)^T(-1,t)\Lambda^+ \mathbf{z}_L(t) - (\mathbf{y}^-)^T(1,t)\Lambda^- \mathbf{z}_R(t).
\end{aligned} \tag{13.79}$$

Finally, we recall that the following assumption, called *dissipation hypothesis*, is usually made on the reflection matrices S_L and S_R

$$\|S_L\|\, \|S_R\| < 1. \tag{13.80}$$

The matrix norm in (13.80) must be understood as the euclidean norm of a rectangular matrix, that is the square root of the maximum eigenvalue of $S_L^T S_L$ and that of $S_R^T S_R$, respectively.
This assumption is sufficient to guarantee the stability of the previous scheme in the L^2 norm. Formulation (13.78) (or (13.79)) is suitable for Galerkin approximations such as the Galerkin-finite elements, the spectral Galerkin method, the spectral method with Gaussian numerical integration in a single domain (G-NI), the spectral element version, in both cases of continuous (SEM-NI) or discontinuous (DG-SEM-NI) spectral elements.