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

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

together with boundary conditions, which can take the following form

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

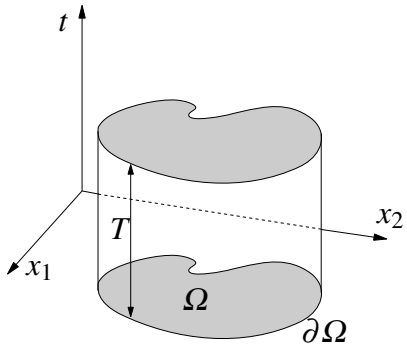

Fig. 5.1. The cylinder $Q_T = \Omega \times (0,T)$, $\Omega \subset \mathbb{R}^2$

A. Quarteroni: *Numerical Models for Differential Problems*, 2nd Ed.
MS&A – Modeling, Simulation & Applications 8
DOI 10.1007/978-88-470-5522-3_5, © Springer-Verlag Italia 2014

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 metal bar of length d occupying the interval $[0,d]$, whose thermal conductivity is ν and whose endpoints are kept at a constant temperature of zero degrees. The function u_0 describes the initial temperature, while f represents the heat generated (per unit length) by the bar. For this reason, (5.4) is called *heat equation*. For a particular case, see Example 1.5 of Chapter 1.

5.1 Weak formulation and its approximation

In order to solve problem (5.1)–(5.3) numerically, 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 $u(t) \in V$ such that

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

where $u(0) = u_0$, $a(\cdot,\cdot)$ 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

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

yielding 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 Sect. 2.7. For the proof, see [QV94, Sect. 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 $u_h(t) \in V_h$ such that

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

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, since 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 coefficients $\{u_j(t)\}$ 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 difference quotient and replaces the other terms with 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 $\Delta t = t^{k+1} - t^k$, $k = 0, 1, \ldots$, denotes the discretization step (here assumed to be constant), while the superscript 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 accurate to order one with respect to Δt;
- for $\theta = 1$ we have the *backward Euler* (or *implicit* Euler) method

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

  also of 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 of second order in Δt. (More precisely, $\theta = 1/2$ is the only value for which we obtain a second-order method.)

Let us consider the two extremal 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 M diagonal, we actually decouple the system. This operation is performed by the so-called *lumping* of the mass matrix (see Sect. 12.5). However, this scheme is not unconditionally stable (see Sect. 5.4) and in the case where V_h is a subspace of finite elements we have the following stability condition (see Sect.~5.4)

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

so Δt icannot be chosen irrespective of h.

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

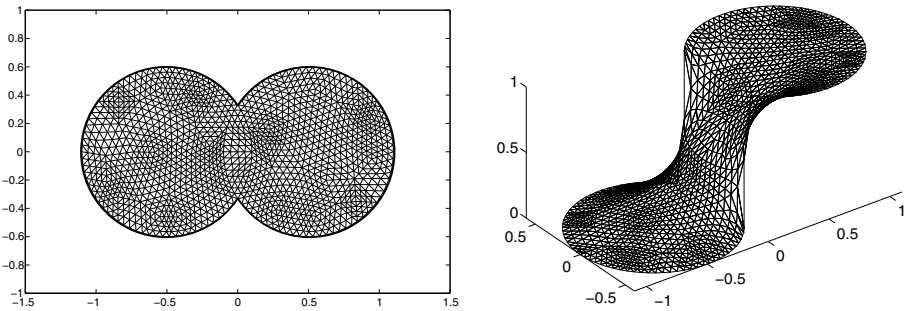

Fig. 5.2. Solution of the heat equation for the problem of Example 5.1

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 set $v = u(t)$ (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

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

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 *Young's inequality*

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

which descends from the elementary inequality

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

Using first Poincaré' inequality (2.13) and Young's inequality, we obtain

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

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

$$\|u(t)\|_{L^2(\Omega)}^2 + \alpha \int_0^t \|\nabla u(s)\|_{L^2(\Omega)}^2 ds \leq \|u_0\|_{L^2(\Omega)}^2 + \frac{C_\Omega^2}{\alpha} \int_0^t \|f(s)\|_{L^2(\Omega)}^2 ds. \tag{5.14}$$

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

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

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

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

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

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

This is a further a priori estimate.

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

$$\begin{aligned}
&\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2 ds \\
&\quad \le \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u(s)\|_{L^2(\Omega)} ds \\
&\quad \le \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \cdot (\|u_0\|^2_{L^2(\Omega)} + \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau) ds \\
&\text{(using (5.15))} \\
&\quad = \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u_0\|_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau \\
&\quad = (\|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\| ds)^2. \qquad (5.16)
\end{aligned}$$

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

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

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):

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

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):

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

and

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

5.3 Convergence analysis of the semi-discrete problem

Let us consider 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

$$\begin{aligned}\alpha\|(u-u_h)(t)\|^2_{\mathrm{H}^1(\Omega)} &\le a((u-u_h)(t),(u-u_h)(t))\\ &= a((u-u_h)(t),(u-v_h)(t))\\ &\quad + a((u-u_h)(t),(v_h-u_h)(t)) \qquad \forall v_h : v_h(t)\in V_h,\ \ \forall t>0.\end{aligned}$$

For the sake of clarity, we suppress the dependence from t. 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 $(v,w) = \int_\Omega vw$ is the scalar product of $L^2(\Omega)$. Then

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

We analyze the two right-hand side terms separately:

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

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

- 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

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

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

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

Integrating by parts and using Young's inequality, we obtain

$$\begin{aligned}
&\int_0^t \left(\frac{\partial}{\partial t}(u-u_h)(s), (u-v_h)(s) \right) ds = -\int_0^t \left((u-u_h)(s), \frac{\partial}{\partial t}((u-v_h)(s)) \right) ds \\
&\qquad + ((u-u_h)(t), (u-v_h)(t)) - ((u-u_h)(0), (u-v_h)(0)) \\
&\leq \int_0^t \|(u-u_h)(s)\|_{\mathrm{L}^2(\Omega)} \, \| \frac{\partial((u-v_h)(s))}{\partial t} \|_{\mathrm{L}^2(\Omega)} \, ds + \frac{1}{4} \|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} \\
&\qquad + \|(u-v_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2} \|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2} \|u(0)-v_h(0)\|^2_{\mathrm{L}^2(\Omega)}.
\end{aligned}$$

From (5.23) we thus obtain

$$\begin{aligned}
&\frac{1}{2} \|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \alpha \int_0^t \|(u-u_h)(s)\|^2_{\mathrm{H}^1(\Omega)} \, ds \\
&\qquad \leq 2\|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)} + \frac{M^2}{\alpha} \int_0^t \|(u-v_h)(s)\|^2_{\mathrm{H}^1(\Omega)} \, ds \\
&\qquad\quad + 2 \int_0^t \|(u-u_h)(s)\|_{\mathrm{L}^2(\Omega)} \| \frac{\partial((u-v_h)(s))}{\partial t} \|_{\mathrm{L}^2(\Omega)} \, ds \\
&\qquad\quad + 2\|(u-v_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + \|u(0)-v_h(0)\|^2_{\mathrm{L}^2(\Omega)}.
\end{aligned} \tag{5.24}$$

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, $v_h(t) = \Pi_h^r u(t)$, the interpolant of $u(t)$ in V_h (see (4.20)). Thanks to (4.69) we have, assuming that u is sufficiently regular,

$$h\|u(t) - \Pi_h^r u(t)\|_{\mathrm{H}^1(\Omega)} + \|u(t) - \Pi_h^r u(t)\|_{\mathrm{L}^2(\Omega)} \leq C_2 h^{r+1} |u(t)|_{\mathrm{H}^{r+1}(\Omega)}.$$

Let us consider and bound from above some of the summands of the right-hand side of inequality (5.24):

$$\begin{aligned}
E_1 &= 2\|(u-u_h)(0)\|^2_{\mathrm{L}^2(\Omega)} \leq C_1 h^{2r} |u_0|^2_{\mathrm{H}^r(\Omega)}. \\
E_2 &= \frac{M^2}{\alpha} \int_0^t \|u(s) - v_h(s)\|^2_{\mathrm{H}^1(\Omega)} \, ds \leq C_2 h^{2r} \int_0^t |u(s)|^2_{\mathrm{H}^{r+1}(\Omega)} \, ds, \\
E_3 &= 2\|u(t) - v_h(t)\|^2_{\mathrm{L}^2(\Omega)} \leq C_3 h^{2r} |u(t)|^2_{\mathrm{H}^r(\Omega)}.
\end{aligned}$$

Finally

$$E_4(s) = \| \frac{\partial(u(s) - v_h(s))}{\partial t} \|_{\mathrm{L}^2(\Omega)} \leq C_4 h^r \left| \frac{\partial u(s)}{\partial t} \right|_{H^r(\Omega)}.$$

Consequently,

$$E_1 + E_2 + E_3 + E_4 \leq C h^{2r} N(u),$$

where $N(u)$ is a suitable function depending on u and on $\frac{\partial u}{\partial t}$, and C is a suitable positive constant. In this way, from (5.24) we obtain the inequality

$$\|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha \int_0^t \|(u-u_h)(s)\|^2_{\mathrm{H}^1(\Omega)}\, ds \\ \leq Ch^{2r}N(u) + 2C_4 h^r \int_0^t \left| \frac{\partial u(s)}{\partial t} \right|_{H^r(\Omega)} \|(u-u_h)(s)\|_{\mathrm{L}^2(\Omega)} ds.$$

Finally, applying the Gronwall lemma (Lemma 2.2 *ii*)), we obtain the a priori error estimate for all $t > 0$

$$\left\{ \|(u-u_h)(t)\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha \int_0^t \|u-u_h\|^2_{\mathrm{H}^1(\Omega)} \right\}^{1/2} \\ \leq \bar{C}h^r \left(\sqrt{N(u)} + \int_0^t \left| \frac{\partial u(s)}{\partial t} \right|_{H^r(\Omega)} ds \right) \forall t > 0 \quad (5.25)$$

for a suitable positive constant $\bar{C}$.
An alternative proof that does not make use of Gronwall' lemma goes as follows. If we subtract (5.6) from (5.5) and set $E_h = u - u_h$, we obtain that (the dependence of E_h on t is understood)

$$\left(\frac{\partial E_h}{\partial t}, v_h \right) + a(E_h, v_h) = 0 \; \forall v_h \in V_h, \forall t > 0.$$

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, Sect. 3.5]) that there exists a constant $C > 0$ such that, $\forall w \in V \cap H^{r+1}(\Omega)$,

$$\|\Pi^r_{1,h} w - w\|_{\mathrm{H}^1(\Omega)} + h^{-1} \|\Pi^r_{1,h} w - w\|_{\mathrm{L}^2(\Omega)} \leq Ch^p |w|_{\mathrm{H}^{p+1}(\Omega)}, 0 \leq p \leq r. \quad (5.27)$$

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

$$(\frac{\partial e_h}{\partial t}, v_h) + a(e_h, v_h) = -(\frac{\partial \sigma_h}{\partial t}, v_h) - a(\sigma_h, v_h) \; \forall v_h \in V_h, \forall t > 0.$$

If we take at every $t > 0$, $v_h = e_h(t)$, and proceed as done in Sect. 5.2 to deduce the a priori estimates on the semi-discrete solution u_h, we obtain

$$\begin{aligned} \tfrac{1}{2} \tfrac{d}{dt} \|e_h(t)\|^2_{L^2(\Omega)} \; & + \alpha \|\nabla e_h(t)\|^2_{L^2(\Omega)} \\ & \leq |a(\sigma_h(t), e_h(t))| + |(\tfrac{\partial}{\partial t}\sigma_h(t), e_h(t))|. \end{aligned} \quad (5.29)$$

Using the continuity of the bilinear form $a(\cdot,\cdot)$ (M being the continuity constant) and Young's inequality (5.12), 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's inequality it follows that

$$\begin{aligned}|(\frac{\partial}{\partial t}\sigma_h(t), e_h(t))| &\leq \|\frac{\partial}{\partial t}\sigma_h(t)\|_{L^2(\Omega)} C_\Omega \|\nabla e_h(t)\|_{L^2(\Omega)} \\ &\leq \frac{\alpha}{4}\|\nabla e_h(t)\|^2_{L^2(\Omega)} + \frac{C^2_\Omega}{\alpha}\|\frac{\partial}{\partial t}\sigma_h(t)\|^2_{L^2(\Omega)}.\end{aligned}$$

Using these bounds in (5.29) we obtain, after integrating with respect to t:

$$\begin{aligned}&\|e_h(t)\|^2_{L^2(\Omega)} + \alpha\int_0^t \|\nabla e_h(t)\|^2_{L^2(\Omega)} ds \\ &\leq \|e_h(0)\|^2_{L^2(\Omega)} + \frac{2M^2}{\alpha}\int_0^t \|\nabla \sigma_h(s)\|^2_{L^2(\Omega)} ds + \frac{2C^2_\Omega}{\alpha}\int_0^t \|\frac{\partial}{\partial t}\sigma_h(s)\|^2_{L^2(\Omega)} ds, \quad t > 0.\end{aligned}$$

At this point we can use (5.27) to bound the errors on the right-hand side:

$$\begin{aligned}&\|\nabla \sigma_h(t)\|_{L^2(\Omega)} \leq Ch^r |u(t)|_{H^{r+1}(\Omega)}, \\ &\left\|\frac{\partial}{\partial t}\sigma_h(t)\right\|_{L^2(\Omega)} = \left\|\left(\frac{\partial u}{\partial t} - \Pi^r_{1,h}\frac{\partial u}{\partial t}\right)(t)\right\|_{L^2(\Omega)} \leq Ch^r \left|\frac{\partial u(t)}{\partial t}\right|_{H(\Omega)}.\end{aligned}$$

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 such that

$$\begin{aligned}&\{\|u(t) - u_h(t)\|^2_{L^2(\Omega)} + \alpha\int_0^t \|\nabla u(s) - \nabla u_h(s)\|^2_{L^2(\Omega)} ds\}^{1/2} \\ &\leq Ch^r\{|u_0|^2_{H^r(\Omega)} + \int_0^t |u(s)|^2_{H^{r+1}(\Omega)} ds + \int_0^t |\frac{\partial u(s)}{\partial t}|^2_{H^{r+1}(\Omega)} ds\}^{1/2}.\end{aligned}$$

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

$$\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h\right) = \theta F^{k+1}(v_h) + (1-\theta)F^k(v_h) \qquad \forall v_h \in V_h, \quad (5.30)$$

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 ($\theta = 1$) 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 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

$$\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \|u_h^{k+1}\|_V^2 \leq \|u_h^k\|_{\mathrm{L}^2(\Omega)}^2. \quad (5.31)$$

By summing over k from 0 to $n-1$ we deduce that

$$\|u_h^n\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \sum_{k=0}^{n-1} \|u_h^{k+1}\|_V^2 \leq \|u_{0h}\|_{\mathrm{L}^2(\Omega)}^2.$$

When $f \neq 0$, using the discrete Gronwall lemma (see Sect. 2.7) it can be proved in a similar way that

$$\|u_h^n\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \sum_{k=1}^{n} \|u_h^k\|_V^2 \leq C(t^n)\left(\|u_{0h}\|_{\mathrm{L}^2(\Omega)}^2 + \sum_{k=1}^{n} \Delta t \|f^k\|_{\mathrm{L}^2(\Omega)}^2\right). \quad (5.32)$$

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 positive, real eigenvalues forming 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 N_h in number (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 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 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 work, 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 for 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 impose 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 largest 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

$$\lambda_h^{N_h} = \frac{a(w_{N_h}, w_{N_h})}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq \frac{M\|w_{N_h}\|^2_V}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq M(1+C^2h^{-2}).$$

The constant $C > 0$ which appears in the latter step derives from the following *inverse inequality*

$$\exists C > 0 \ : \ \|\nabla v_h\|_{\mathrm{L}^2(\Omega)} \leq Ch^{-1}\|v_h\|_{\mathrm{L}^2(\Omega)} \qquad \forall v_h \in V_h,$$

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], Sect. 11.3.1, for the proof in the general case.

Let $\Pi_{1,h}^r$ be the orthogonal projector operator introduced in (5.26). Then

$$\|u(t^k) - u_h^k\|_{\mathrm{L}^2(\Omega)} \leq \|u(t^k) - \Pi_{1,h}^r u(t^k)\|_{\mathrm{L}^2(\Omega)} + \|\Pi_{1,h}^r u(t^k) - u_h^k\|_{\mathrm{L}^2(\Omega)}. \tag{5.38}$$

The first term can be estimated by referring to (5.27). To analyze the second term, where $\varepsilon_h^k = u_h^k - \Pi_{1,h}^r u(t^k)$, 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, $\forall v_h \in V_h$,

$$(\delta^{k+1}, v_h) = (f^{k+1}, v_h) - \frac{1}{\Delta t}(\Pi_{1,h}^r(u(t^{k+1}) - u(t^k)), v_h) - a(u(t^{k+1}), v_h) \tag{5.40}$$

and having exploited on the last summand the orthogonality (5.26) of the operator $\Pi_{1,h}^r$. The sequence $\{\varepsilon_h^k,\, k = 0, 1 \ldots\}$ satisfies problem (5.39), which is similar 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 $n \geq 1$,

$$\|\varepsilon_h^n\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \sum_{k=1}^{n} \|\varepsilon_h^k\|_V^2 \leq C(t^n) \left(\|\varepsilon_h^0\|_{\mathrm{L}^2(\Omega)}^2 + \sum_{k=1}^{n} \Delta t \|\delta^k\|_{\mathrm{L}^2(\Omega)}^2 \right). \tag{5.41}$$

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 suitable regularity requirements on the function u with respect to the temporal variable. By now using the fundamental theorem of calculus and exploiting the commutativity between the projection operator $\Pi^r_{1,h}$ and the temporal derivative, we obtain

$$\left(I - \Pi^r_{1,h}\right)\left(u(t^{k+1}) - u(t^k)\right) = \int_{t^k}^{t^{k+1}} \left(I - \Pi^r_{1,h}\right)\left(\frac{\partial u}{\partial t}\right)(s)\, ds. \tag{5.45}$$

By choosing $v_h = \delta^{k+1}$ in (5.43), thanks to (5.44) and (5.45), we can deduce the following upper bound

$$\begin{aligned}
&\|\delta^{k+1}\|_{\mathrm{L}^2(\Omega)} \\
&\leq \left\| \frac{1}{\Delta t} \int_{t^k}^{t^{k+1}} (s - t^k) \frac{\partial^2 u}{\partial t^2}(s)\, ds \right\|_{\mathrm{L}^2(\Omega)} + \left\| \frac{1}{\Delta t} \int_{t^k}^{t^{k+1}} \left(I - \Pi^r_{1,h}\right)\left(\frac{\partial u}{\partial t}\right)(s)\, ds \right\|_{\mathrm{L}^2(\Omega)} \\
&\leq \int_{t^k}^{t^{k+1}} \left\| \frac{\partial^2 u}{\partial t^2}(s) \right\|_{\mathrm{L}^2(\Omega)} ds + \frac{1}{\Delta t} \int_{t^k}^{t^{k+1}} \left\| \left(I - \Pi^r_{1,h}\right)\left(\frac{\partial u}{\partial t}\right)(s) \right\|_{\mathrm{L}^2(\Omega)} ds.
\end{aligned} \tag{5.46}$$

By reverting to the stability estimate (5.41) and exploiting (5.42) and the estimate (5.46) with suitably scaled indices, we have

$$\begin{aligned}
\|\varepsilon_h^n\|^2_{\mathrm{L}^2(\Omega)} \quad &\leq \quad C(t^n)\left(h^{2r} |u_0|^2_{\mathrm{H}^r(\Omega)} + \sum_{k=1}^{n} \Delta t \left[\left(\int_{t^{k-1}}^{t^k} \left\| \frac{\partial^2 u}{\partial t^2}(s) \right\|_{\mathrm{L}^2(\Omega)} ds \right)^2 \right.\right. \\
&+ \quad \left.\left. \frac{1}{\Delta t^2} \left(\int_{t^{k-1}}^{t^k} \left\| \left(I - \Pi^r_{1,h}\right)\left(\frac{\partial u}{\partial t}\right)(s) \right\|_{\mathrm{L}^2(\Omega)} ds \right)^2 \right] \right),
\end{aligned}$$

Then, using the Cauchy-Schwarz inequality and estimate (5.27) for the projection operator $\Pi^r_{1,h}$, we obtain

$$\begin{aligned}
\|\varepsilon_h^n\|^2_{\mathrm{L}^2(\Omega)} \leq C(t^n) \Bigg(& h^{2r} |u_0|^2_{\mathrm{H}^r(\Omega)} + \sum_{k=1}^{n} \Delta t \left[\Delta t \int_{t^{k-1}}^{t^k} \left\| \frac{\partial^2 u}{\partial t^2}(s) \right\|^2_{\mathrm{L}^2(\Omega)} ds \right. \\
& \left. + \frac{1}{\Delta t^2} \left(\int_{t^{k-1}}^{t^k} h^r \left| \frac{\partial u}{\partial t}(s) \right|_{\mathrm{H}^r(\Omega)} ds \right)^2 \right] \Bigg)
\end{aligned}$$

$$\leq C(t^n)\left(h^{2r}|u_0|^2_{\mathrm{H}^r(\Omega)} + \Delta t^2 \sum_{k=1}^{n} \int_{t^{k-1}}^{t^k} \left\| \frac{\partial^2 u}{\partial t^2}(s) \right\|^2_{\mathrm{L}^2(\Omega)} ds \right.$$
$$\left. + \frac{1}{\Delta t} h^{2r} \sum_{k=1}^{n} \Delta t \int_{t^{k-1}}^{t^k} \left| \frac{\partial u}{\partial t}(s) \right|^2_{\mathrm{H}^r(\Omega)} ds \right). \tag{5.47}$$

The result now follows using (5.38) and estimate (5.27). ◇

More stability and convergence estimates can be found in [Tho84].

5.6 Exercises

1. Verify that the mass matrix M introduced in (5.7) is positive definite.
2. Consider the problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial}{\partial x}\left(\alpha \dfrac{\partial u}{\partial x}\right) - \beta u = 0 & \text{in } Q_T = (0,1)\times(0,\infty), \\ u = u_0 & \text{for } x \in (0,1),\ t = 0, \\ u = \eta & \text{for } x = 0,\ t > 0, \\ \alpha \dfrac{\partial u}{\partial x} + \gamma u = 0 & \text{for } x = 1,\ t > 0, \end{cases}$$

   where $\alpha = \alpha(x)$, $u_0 = u_0(x)$ are given functions and β, γ, $\eta \in \mathbb{R}$ (with positive β).
   a) Prove existence and uniqueness of the weak solution for varying γ, providing suitable limitations on the coefficients and suitable regularity hypotheses on the functions α and u_0.
   b) Introduce the spatial semi-discretization of the problem using the Galerkin-finite element method, and carry out its stability and convergence analysis.
   c) In the case where $\gamma = 0$, approximate the same problem with the explicit Euler method in time and carry out its stability analysis.
3. Consider the following problem: find $u(x,t)$, $0 \leq x \leq 1$, $t \geq 0$, such that

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, & 0 < x < 1,\ t > 0, \\ v + \alpha(x)\dfrac{\partial u}{\partial x} - \gamma(x)u = 0, & 0 < x < 1,\ t > 0, \\ v(1,t) = \beta(t),\ u(0,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & 0 < x < 1, \end{cases}$$

   where α, γ, β, u_0 are given functions.

a) Introduce an approximation based on finite elements of degree two in x and the implicit Euler method in time and prove its stability.
b) How will the error behave as a function of the parameters h and Δt?
c) Suggest a way to provide an approximation for v starting from the one for u as well as its approximation error.

4. Consider the following (diffusion-transport-reaction) initial-boundary value problem: find $u : (0,1) \times (0,T) \to \mathbb{R}$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial}{\partial x}\left(\alpha \dfrac{\partial u}{\partial x}\right) + \dfrac{\partial}{\partial x}(\beta u) + \gamma u = 0, & 0 < x < 1,\ 0 < t < T, \\ u = 0 & \text{for } x = 0,\ 0 < t < T, \\ \alpha \dfrac{\partial u}{\partial x} + \delta u = 0 & \text{for } x = 1,\ 0 < t < T, \\ u(x,0) = u_0(x), & 0 < x < 1,\ t = 0, \end{cases}$$

where $\alpha = \alpha(x)$, $\beta = \beta(x)$, $\gamma = \gamma(x)$, $\delta = \delta(x)$, $u_0 = u_0(x)$, $x \in [0,1]$ are given functions.

a) Write its weak formulation.
b) In addition to the hypotheses:

$$a.\ \ \exists \beta_0,\ \alpha_0,\ \alpha_1 > 0 :\ \forall x \in (0,1)\ \alpha_1 \geq \alpha(x) \geq \alpha_0,\ \beta(x) \leq \beta_0,$$

$$b.\ \ \tfrac{1}{2}\beta'(x) + \gamma(x) \geq 0 \quad \forall x \in (0,1),$$

provide further possible hypotheses on the data so that the problem is well-posed. Moreover, give an a priori estimate of the solution. Treat the same problem with non-homogeneous Dirichlet data $u = g$ for $x = 0$ and $0 < t < T$.
c) Consider a semi-discretization based on the linear finite elements method and prove its stability.
d) Finally, provide a full discretization where the temporal derivative is approximated using the implicit Euler scheme and prove its stability.

5. Consider the following fourth-order initial-boundary value problem:
find $u : \Omega \times (0,T) \to \mathbb{R}$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} - \operatorname{div}(\mu \nabla u) + \Delta^2 u + \sigma u = 0 & \text{in } \Omega \times (0,T), \\ u(\mathbf{x},0) = u_0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = u = 0 & \text{on } \Sigma_T = \partial\Omega \times (0,T), \end{cases}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded open domain with "regular" boundary $\partial\Omega$, $\Delta^2 = \Delta\Delta$ is the bi-harmonic operator, $\mu(\mathbf{x})$, $\sigma(\mathbf{x})$ and $u_0(\mathbf{x})$ are known functions defined in

Ω. It is known that

$$\sqrt{\int_\Omega |\Delta u|^2 \mathrm{d}\Omega} \simeq \|u\|_{\mathrm{H}^2(\Omega)} \qquad \forall\, u \in \mathrm{H}_0^2(\Omega)\,,$$

that is the two norms are equivalent, where

$$\mathrm{H}_0^2(\Omega) = \{u \in \mathrm{H}^2(\Omega) :\ u = \partial u/\partial n = 0 \text{ on } \partial\Omega\}. \tag{5.48}$$

a) Write its weak formulation and verify that the solution exists and is unique, formulating suitable regularity hypotheses on the data.
b) Consider a semi-discretization based on triangular finite elements and provide the minimum degree that such elements must have in order to solve the given problem adequately. (We note that, if $\mathcal{T}_h$ is a triangulation of Ω and $v_{h|K}$ is a polynomial for each $K \in \mathcal{T}_h$, then $v_h \in \mathrm{H}^2(\Omega)$ if and only if $v_h \in C^1(\overline{\Omega})$, that is v_h and its first derivatives are continuous across the interfaces of the elements of $\mathcal{T}_h$.)